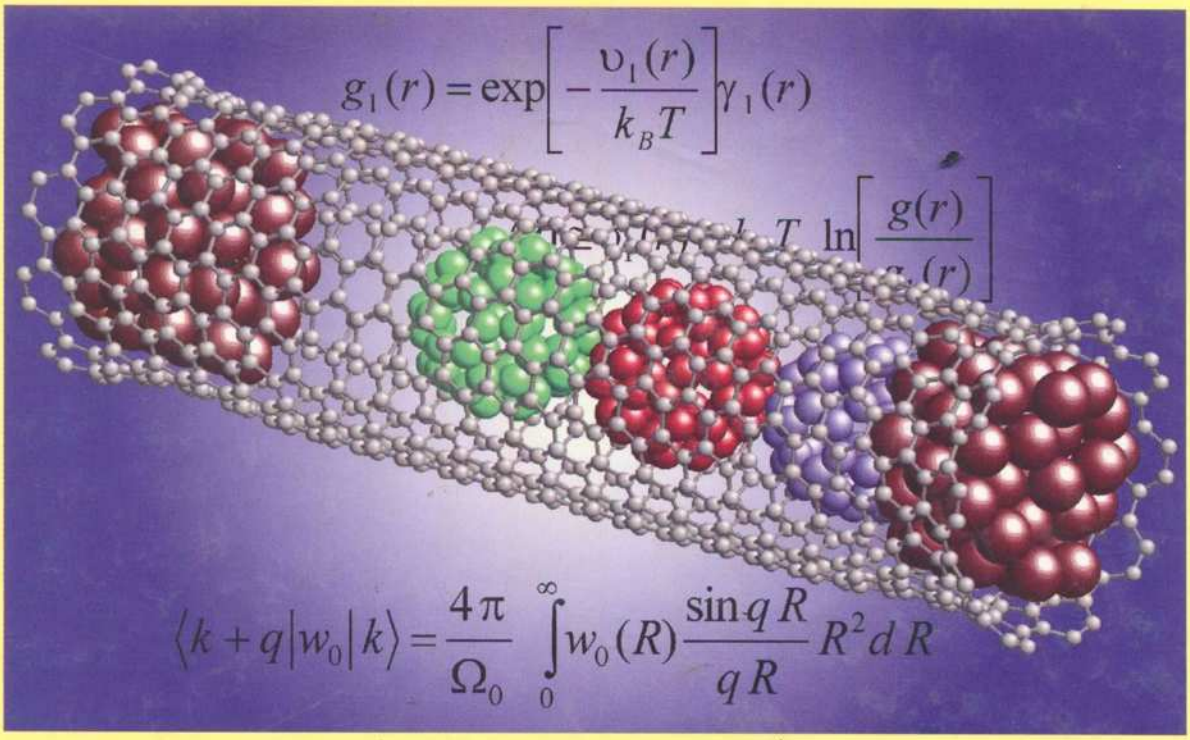
Handbook of THEORETICAL and COMPUTATIONAL NANOTECHNOLOGY

6

Bioinformatics, Nanomedicine, and Drug Design



Edited by

Volfram Schommers

Foreword by

Pierre-Gilles de Gennes, Nobel Prize Laureate

ASP
AMERICAN SCIENTIFIC PUBLISHERS

Handbook of

THEORETICAL and COMPUTATIONAL NANOTECHNOLOGY

Titles in Nanotechnology Book Series

Founding Editor

Dr. Hari Singh Nalwa

1. **Encyclopedia of Nanoscience and Nanotechnology, 10-Volume Set**
   Edited by Hari Singh Nalwa
2. **Handbook of Theoretical and Computational Nanotechnology, 10-Volume Set**
   Edited by Michael Rieth and Wolfram Schommers
3. **Bottom-up Nanofabrication: Supramolecules, Self-Assemblies, and Organized Films, 6-Volume Set**
   Edited by Katsuhiko Ariga and Hari Singh Nalwa
4. **Handbook of Semiconductor Nanostructures and Nanodevices, 5-Volume Set**
   Edited by A. A. Balandin and K. L. Wang
5. **Handbook of Organic-Inorganic Hybrid Materials and Nanocomposites, 2-Volume Set**
   Edited by Hari Singh Nalwa
6. **Handbook of Nanostructured Biomaterials and Their Applications in Nanobiotechnology, 2-Volume Set**
   Edited by Hari Singh Nalwa
7. **Handbook of Electrochemical Nanotechnology, 2-Volume Set**
   Edited by Yuehe Lin and Hari Singh Nalwa
8. **Polymeric Nanostructures and Their Applications, 2-Volume Set**
   Edited by Hari Singh Nalwa
9. **Soft Nanomaterials, 2-Volume Set**
   Edited by Hari Singh Nalwa
10. **Functional Nanomaterials**
    Edited by Kurt E. Geckeler and Edward Rosenberg
11. **Synthesis, Functionalization and Surface Treatment of Nanoparticles**
    Edited by I. M. Baraton
12. **Quantum Dots and Nanowires**
    Edited by S. Bandyopadhyay and Hari Singh Nalwa
13. **Nanoclusters and Nanocrystals**
    Edited by Hari Singh Nalwa
14. **Molecular Nanoelectronics**
    Edited by Mark A. Reed and T. Lee
15. **Magnetic Nanostructures**
    Edited by Hari Singh Nalwa
16. **Nanoparticles for Pharmaceutical Applications**
    Edited by J. Domb, Y. Tabata, M. N. V. Ravi Kumar, and S. Farber
17. **Cancer Nanotechnology**
    Edited by Hari Singh Nalwa and Thomas Webster
18. **Biochips Nanotechnology**
    Edited by Nongyue He and Hari Singh Nalwa
19. **Nanotoxicology**
    Edited by Yuliang Zhao and Hari Singh Nalwa
20. **Polymer Nanocomposites and Their Applications**
    Written by Suprakas Sinha Ray and Mosto Bousmina
21. **Nanoscale Science and Engineering Education**
    Edited by Aldrin E. Sweeney and Sudipta Seal
22. **Hard Nanomaterials**
    Edited by Hari Singh Nalwa

Additional Volumes in Preparation

Visit: www.aspbs.com

Handbook of
THEORETICAL and COMPUTATIONAL NANOTECHNOLOGY

Volume 6

Bioinformatics, Nanomedicine, and Drug Design

Edited by

Michael Rieth and **Wolfram Schommers**

Forschungszentrum Karlsruhe, Karlsruhe, Germany

AMERICAN SCIENTIFIC PUBLISHERS

25650 North Lewis Way

Stevenson Ranch, California 91381-1439, USA

AMERICAN SCIENTIFIC PUBLISHERS
25650 North Lewis Way, Stevenson Ranch, California 91381-1439, USA
Tel.: (661) 254-0807
Fax: (661) 254-1207
E-mail: order@aspbs.com
WEB: www.aspbs.com

Handbook of Theoretical and Computational Nanotechnology
edited by Michael Rieth and Wolfram Schommers.

The image on the cover of this handbook was provided by Professor Jeong Won Kang, Chung-Ang University, Seoul, Korea. See Jeong Won Kang, Won Young Choi, and Ho Jung Hwang, *Journal of Computational and Theoretical Nanoscience*, Vol. 1(2), pp. 199–203 (2004). Copyright © 2004. American Scientific Publishers.

This book is printed on acid-free paper. ♾

Library of Congress Control Number: 2003111818

International Standard Book Number: 1-58883-042-X (Set)
International Standard Book Number: 1-58883-043-8 (Volume 1)
International Standard Book Number: 1-58883-044-6 (Volume 2)
International Standard Book Number: 1-58883-045-4 (Volume 3)
International Standard Book Number: 1-58883-046-2 (Volume 4)
International Standard Book Number: 1-58883-047-0 (Volume 5)
International Standard Book Number: 1-58883-048-9 (Volume 6)
International Standard Book Number: 1-58883-049-7 (Volume 7)
International Standard Book Number: 1-58883-050-0 (Volume 8)
International Standard Book Number: 1-58883-051-9 (Volume 9)
International Standard Book Number: 1-58883-052-7 (Volume 10)

Dr. Michael Rieth
Forschungszentrum Karlsruhe
Institute of Materials Research I
D-76021 Karlsruhe, GERMANY

Prof. Dr. Wolfram Schommers
Forschungszentrum Karlsruhe
Institute for Scientific Computing
D-76021 Karlsruhe, GERMANY

PRINTED IN THE UNITED STATES OF AMERICA
10 9 8 7 6 5 4 3 2 1

Foreword

Nanoscience is fashionable. All administrations in the Western world have stressed their interest in nanoobjects and nanotechnologies. As usual, this type of large scientific movement has its pluses and minuses. Many scientists join the crowd without necessarily changing anything in their actual work. Most chemists, for instance, build new molecules that may be called nanoobjects; but again, as usual, the movement does generate significant new content.

Let us, for instance, follow the role of nanostructures in *chemistry*. On one side, nature has provided us with beautiful, robust objects such as fullerenes and carbon tubes, which have some admirable properties. The current challenge is to obtain them in large amounts and at a reasonable price. Here is the real problem.

A completely different sector is obtained from *chemical nanomachines*, for which a molecular unit of nanometric size moves with respect to another one through a change in redox potential or pH. Some of these machines have been built. At the moment, I feel rather skeptical about them because they are extremely costly, extremely fragile (sensitive to poisons), and not easy to protect with a suitable coating—or by a local "antipoison" center. But, here again, there is a challenge.

Let us now turn to *biology*. Here we find an immense group of working nanomachines, enzymes, ionic channels, sensor proteins, adhesion molecules, and so on. They are extremely impressive, but of course they represent progressive construction by trial and error over more than a billion years. Should we try to mimic these machines or, rather, use them for technological purposes, *as they are*, for instance, to grow plants or create proteins at an industrial level according to the techniques of molecular genetics? This is a major question.

A third, open side is *quantum physics* and the (remote) possibility of quantum computers. In my youth, I had hopes for digital storage via quantized flux quanta: The corresponding technology, based on Josephson functions, was patiently built by IBM, but they ultimately dropped out. This shows the hardship of nanotechnologies even when they are handled by a large, competent group. But the cause is not lost, and it may well be that our children use some unexpected form of quantum computers.

Thus, we are facing real challenges, not just the vague recommendations of some anonymous boards. And, we need the tools. We need to know the behavior of materials at the nanolevel, the clever tricks of physical chemistry required to produce nanoparticles or nanopores, the special properties of small cooperative systems (nanomagnets, nanosuperconductors, nanoferroelectrics, etc.), the ability for assembling functional units, and so on.

The aim of the present handbook is to help us with the tools by suitable modelizations. It is written by leading experts, starting from general theoretical principles and progressing to detailed recipes.

In the second half of the 18th century, all the knowledge (fundamental and practical) of the Western world was condensed into an outstanding encyclopedia constructed energetically by Denis Diderot just after the industrial revolution started. Here, at a more modest level, we can hope for something similar. Soon after the first wave, including this handbook, a certain form of nanoindustry may be born.

The discussions started in this handbook will continue in a journal (*Journal of Computational and Theoretical Nanoscience*) launched by the present editors. I wish them the best.

Professor Pierre-Gilles de Gennes

Nobel Prize Laureate, Physics
Collège de France
Paris, France

Preface

This is the first handbook that deals with theoretical and computational developments in nanotechnology. The 10-volume compendium is an unprecedented single reference source that provides ideal introduction and overview of the most recent advances and emerging new aspects of nanotechnology spanning from science and engineering to neurogenetics. Many works in the field of theoretical and computational nanotechnology have been published to date, but no book or handbook has focused on all aspects in this field that deal with nanomachines, electronics, devices, quantum computing, nanostructured materials, nanorobotics, medicine, biology, biotechnology, and more.

There is no doubt that nanoscience will be the dominant direction for technology in this new century, and this science will influence our lives to an extent impossible in years past: Specific manipulations of matter at its ultimate level will open completely new perspectives on all scientific and technological disciplines. To be able to produce optimal nanosystems with tailor-made properties, it is necessary to analyze and construct such systems in advance by adequate theoretical and computational methods. The handbook gives a complete overview of the essential methods, models, and basic pictures.

But, as is well known, there are also threats connected with nanotechnology, specifically with respect to biological systems: Self-assembly can be an uncontrolled process, and the final state of a developing system is in general uncertain in such cases. To avoid undesirable developments, the theoretical (computational) analysis of such processes is not only desirable but also absolutely necessary. Thus, the computational and theoretical methods of nanoscience are essential for the prediction of new and custom nanosystems and can help keep nanoscience under control. There is basically no alternative. Therefore, one possible answer to the question, "Why a book on theoretical and computational nanotechnology?" is *to give nanotechnology a direction!*

In the design of macroscopic and microscopic systems, engineering is essentially based on *intuitive concepts*, which are tailored to observations in everyday life. Classical mechanics is also based on these macroscopic observations, and its notions have been chosen with respect to our intuitive demands for *visualizability*. However, when we approach the nanolevel, the tools used for the design of macroscopic and microscopic systems become more and more useless. At the nanolevel, *quantum phenomena* are dominant, and the main features in connection with quantum effects are not accessible to our intuitive concepts, which are merely useful at the macroscopic level; the framework of quantum theory is in striking conflict with our intuitive demands for visualizability, and we are forced to use abstract physical laws expressed by mathematical equations. In other words, effects at the nanolevel are (almost) not accessible to our usual engineering concepts. Therefore, here we rely on the abstract mathematical relations of theoretical physics. In nanotechnology functional systems, machines and the like cannot be adequately designed without the use of these abstract theoretical laws and the application of suitable computational methods. Therefore, in nanotechnology, theoretical and computational methods are centrally important: This makes the present handbook an indispensable compendium.

Nanometer-scale units are by definition very small atomic structures and functional systems; it is the smallest level at which functional matter can exist. We already learned to manipulate matter at this ultimate level: Atoms can be moved experimentally in a controlled manner from one position to another. This is astonishing because one nanometer only corresponds to one millionth of a millimeter. For example, an electrical nanogenerator could be designed consisting of various parts that included a very fast revolving rotator. One million of these generators could be arranged side by side on a length of two centimeters; it is remarkable that not only *static* nanostructures could in principle be produced and significantly manipulated but also artificial *dynamical* nanosystems. But, the downscaling of functional structures from the macroscopic to the nanometer scale is only one of the essential

points in connection with nanotechnology. In addition—and maybe much more important—nanosystems provide unique properties in comparison to those we observe at the macroscopic level. For example, a metal nanocluster shows a melting temperature that strongly deviates from that of a macroscopic piece of metal; its melting point is significantly lower. A decrease down to a fraction of only 20% is typical, depending, however, on the material and particle number.

A professional treatment of the various problems in nanoscience and nanotechnology makes the application and development of theoretical and computational methods in this field absolutely necessary. In other words, the discipline of theoretical and computational nanotechnology has to be considered as a key topic to be able to treat nanotechnology adequately and to reach optimal solutions for certain tasks. It is therefore desirable to get a timely overview about the specific topics presently relevant in this field. In this respect, the handbook gives a complete overview of the specific topics so far established in nanotechnology. Each chapter gives a certain overview of actual activities of the envisaged topic and in most cases an adequate description of the basics, so advanced students also can benefit from the handbook. It was our strategy to provide consistent and complete representations so the reader would be able to study each chapter without consulting other works. This of course leads to certain overlaps, which was also part of our strategy to enable an approach to the same topic from various points of view.

The handbook reflects the spectrum of questions and facts that are and could be relevant in the field of nanotechnology. Not only formal developments and methods are outlined, but also descriptions of a broad variety of applications in particular are typical for the handbook. All relevant topics have been taken into account, from functional structures—like an electrical nanogenerator—or quantum computing to questions that deal directly with basic physics. Almost all fields related to theoretical and computational nanotechnology could be covered, including *multiscale modeling*, which is important for the transition from microscale to nanoscale and vice versa.

All theoretical and computational methods used in connection with the various topics in nanoscience are directly based on the *same* theoretical physical laws. At the nanolevel, all properties of our world emerge at the level of the *basic* theoretical laws. In traditional technologies, engineers do not work at the ultimate level. They use more or less phenomenological descriptions that generally cannot be deduced from the basic physical theoretical laws. We have as many phenomenological descriptions as there are technological disciplines, and each is tailor-made to a specific topic. An exchange of concepts is either not possible or rather difficult. In contrast, at the ultimate nanolevel the world is based on only one theory for all disciplines, and this is expressed by basic theoretical physics. This situation opens the possibility for interconnections between the various topics in nanotechnology to bring about new effects and chances for further applications. In other words, nanotechnology and nanoscience can be considered interdisciplinary. Clearly, the handbook reflects the interdisciplinary character of this new science and technology.

The *Handbook of Theoretical and Computational Nanotechnology* includes 138 chapters written by hundreds of the world's leading scientists. Topics cover mainly the following areas:

(i) Computational biology: DNA, enzymes, proteins, biomechanisms, neurogenetic information processing, and nanomedicine
(ii) Computational chemistry: quantum chemistry, molecular design, chemical reactions, drugs, and design
(iii) Computational methods and simulation techniques from *ab initio* to multiscale modeling
(iv) Materials behavior at the nanolevel, such as mechanics, defects, diffusion, and dynamics
(v) Nanoscale processes: membranes, pores, diffusion, growth, friction, wear, catalysis
(vi) Nanostructured materials: metals, composites, polymers, liquid crystals, photonic crystals, colloids, and nanotubes
(vii) Nanostructures: fullerenes, nanotubes, clusters, layers, quantum dots, thin films, surfaces, and interfaces
(viii) Nanoengineering and nanodesign: nanomachines, nano-CAD, nanodevices, and logic circuits

(ix) Nanoelectronics: molecular electronics, nanodevices, electronic states, and nanowires
(x) Nanomagnetism: magnetic properties of nanostructures and nanostructured materials
(xi) Nanooptics: optical response theory, quantum dots, luminescence, and photonic crystals
(xii) Quantum computers: theoretical aspects, devices, and computational methods for simulating quantum computers and algorithms

The handbook provides broad information on all basic and applied aspects of theoretical and computational nanotechnology by considering more than two decades of pioneering research. It is the only scientific work of its kind since the beginning of nanotechnology, bringing together core knowledge and the very latest advances. The handbook is written for audiences of various levels while providing the latest up-to-date information to active scientists and experts in the field. This handbook is an indispensable source for research professionals and developers seeking the most up-to-date information on theoretical and computational nanotechnology among a wide range of disciplines, from science and engineering to medicine.

This handbook was written by leading experts, and we are highly grateful to all contributing authors for their tremendous efforts in writing these outstanding state-of-the-art chapters that altogether form a unified whole. K. Eric Drexler (designer of nanomachines, founder of the Foresight Institute, coiner of the term *nanotechnology*) gives an excellent introductory chapter about possible trends of future nanotechnology. We especially express our sincere gratitude to Dr. Drexler for his instructive and basic representation.

We cordially extend our special thanks to Professor Pierre-Gilles de Gennes for his valuable and insightful Foreword.

The editors are particularly thankful to Dr. Hari Singh Nalwa, President and CEO of American Scientific Publishers, for his continuous support of the project and the enthusiastic cooperation in connection with all questions concerning the development of the handbook. Furthermore, we are grateful to the entire team at Bytheway Publishing and especially to Kate Brown for copyediting.

Dr. Michael Rieth
Prof. Dr. Wolfram Schommers
Karlsruhe, Germany

Contents

CHAPTER 3. Biomolecular Machines

Artur Baumgaertner

CHAPTER 4. Knots, Bubbles, Unwinding, and Breathing: Probing the Topology of DNA and Other Biomolecules

Ralf Metzler, Andreas Hanke

CHAPTER 5. Receptor Flexibility in Ligand Docking

Claudio N. Cavasotto, Andrew J. W. Orry, Ruben A. Abagyan

CHAPTER 6. Enzyme Reactivity Studied by Computer Simulations

Giorgio Colombo, Massimiliano Meli, Giacomo Carrea

CHAPTER 7. Simulating Enzyme-Catalyzed Reactions

Anna Bowman, Adrian Mulholland

CHAPTER 8. Modeling of Biologically Motivated Soft Matter Systems

Ilpo Vattulainen, Mikko Karttunen

CHAPTER 9. Theoretical and Computational Treatments of DNA and RNA Molecules

Haijun Zhou, Yang Zhang, Zhong-Can Ou-Yang

CHAPTER 10. Self-Consistent P3M Simulation of Ion Channels

Campbell Millar, Asen Asenov

CHAPTER 11. Nanomagnetics in Biotechnology

Ching Jen Chen, Yousef Haik, Jhunu Chatterjee

CHAPTER 12. Computational and Theoretical Approaches to Unraveling the Permeation Dynamics in Biological Nanotubes

Shin-Ho Chung, D. Peter Tieleman

CHAPTER 13. Progress in Nanomedicine and Medical Nanorobotics

Robert A. Freitas, Jr.

CHAPTER 14. Molecular Engineering in Nanotechnology: Engineered Drug Delivery

István Majoros, Thommey Thomas, James R. Baker, Jr.

CHAPTER 15. Design of Protein and Enzyme Mimetics

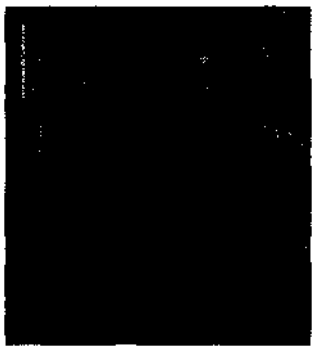
Garland R. Marshall, Dennis P. Riley

CHAPTER 16. Quantum Information Processing in Nanostructures

Alexandra Olaya-Castro, Neil F. Johnson

About the Editors

Dr. Michael Rieth has been a research scientist at the Institute of Materials Research I (IMF-I) in the Forschungszentrum Karlsruhe, Germany, since 2002. He has been head of the consulting company AIFT, Karlsruhe, since 1987. He worked as a researcher at the Institute of Materials Research II (IMF-II), Forschungszentrum Karlsruhe, from 1995 to 1999 and at the Engineering Science Department of the University of Patras (Greece) from 1999 to 2000. He was product manager at AMA Systems, Pforzheim, Germany, from 2000 to 2001. He received his master of science (German Dipl. Ing.) degree in electrical engineering from the University of Karlsruhe in 1991 and his doctoral degree in physics from the University of Patras (Greece) in 2001. Dr. Rieth published 23 research articles in refereed journals, 2 book chapters, and four patents. He is the author of *Nano-Engineering in Science and Technology* (World Scientific, Singapore, 2003) and was the editor-in-chief of the *Journal of Computational and Theoretical Nanoscience* (2004–2005). His main scientific interests are in atomistic modeling of metallic nanosystems and materials development for advanced fusion reactor applications.



Prof. Dr. Wolfram Schommers is a theoretical physicist and is presently at the Research Center of Karlsruhe in Germany. He is also professor of theoretical physics, professor of physics and materials sciences, and distinguished professor in Europe, China, and the United States. He began his studies of theoretical physics at the Technical University of Munich and continued his course work at the University of Münster, receiving a diploma in physics. After a brief intermezzo in the industry, Professor Schommers joined the Research Center of Karlsruhe. He received his doctoral degree (Dr. rer. nat.) in theoretical physics from the University of Karlsruhe.

Professor Schommers concentrates his scientific activities on computational and theoretical physics. His main fields of interest include foundations of physics, liquids, solids, and gases; superionic conductors; surface science; and nanophysics as the basis for the investigation of properties of nanometer-scale atomic devices, junctions, quantum dots, and nanomachines. He has published the results of his research and thoughts in various scientific journals (214 articles and book chapters).

Some topics concerning liquids, solids, and gases concern interaction potentials, single-particle motion, diffusion, generalized phonon density of states, collective motion, and liquid-solid phase transition. Selected articles include "The Effect of van der Waals-Type Interactions in Metals: A Pseudopotential Model" (*Zeitschrift für Physik B* 121, 1976); "Liquid Argon: The Influence of Three-Body Interactions on Atomic Correlations" (*Physical Review A* 16, 327, 1977); "Theoretical Investigation of the Liquid Solid Transition. A Study for Gallium" (*Solid State Communications* 21, 65, 1977); "Pair Potentials in Disordered Many-Particle Systems: A Study for Liquid Gallium" (*Physical Review A* 28, 3599, 1983); "Many-Body Polarization and Overlap Effects in the Dynamic Structure Factor of Dense Krypton Gas" (with P. A. Egelstaff, J. J. Salacuse, and J. Ram; *Physical Review A* 34, 1516, 1986); and "Comment on 'Pair Interaction from Structural Data for Dense Classical Liquids'" (*Physical Review Letters* 58, 427, 1987).

Topics in connection with superionic conductors involve structure and dynamics, correlated motions, and collective behavior. Selected articles include "Correlations in the Motions of Particles in AgI: A Molecular-Dynamics Study" (*Physical Review Letters* 38, 1536, 1977);

"Current-Current Correlations in AgI" (*Physical Review B* 16, 327, 1977); "Structure and Dynamics of Superionic Conductors" (*Physical Review B* 21, 847, 1979); "Triplet Correlations in Solid Electrolytes" (*Solid State Ionics* 1, 473, 1980).

Topics concerning surface physics touch on temperature effects, structure, dynamics, and interaction. Selected works are as follows: "Structural and Dynamical Behaviour of Noble-Gas Surfaces" (*Physical Review A* 32, 6845, 1985); "Statistical Mechanics of the Liquid Surface and the Effect of Premelting," in *Structure and Dynamics of Surfaces II* (Springer-Verlag, Heidelberg, 1987); "The Effect of Non-Linear Interactions at the Surface of Solids" (*Surface Science* 269/270, 180, 1992); and "Steps, Point Defects and Thermal Expansion at the Au(100) Surface" (with H. Zimmermann, M. Nold, U. Romahn, and P. von Blanckenhagen; *Surface Science* 287/288, 76, 1993).

Some details regarding the work of Professor Schommers on nanophysics include study of nanoclusters, nanostructures, and nanomachines; temperature effects; and electronic states. Selected works include "Phonons and Structure in Nano-Clusters: A Molecular Dynamics Study for Al" (*Nanostructured Materials* 9, 621, 1997); "Excited Nano-Clusters" (*Applied Physics A* 68, 187, 1999); "Thermal Stability and Specific Properties of Nanosystems" (with S. Baskoutas and M. Rieth; *Modern Physics Letters B* 14, 621, 2000); "Computational Atomic Nanodesign," in *Encyclopedia of Nanoscience and Nanotechnology* (with M. Rieth; American Scientific Publishers, Stevenson Ranch, CA, 2004); "Computational Engineering of Metallic Nanostructures and Nanomachines" (with M. Rieth; *Journal of Nanoscience and Nanotechnology* 2, 679, 2002); and "Electron in an Interaction Potential of General Shape" (with M. Rieth; *Journal of Computational and Theoretical Nanoscience* 2, 362, 2005).

Concerning the foundations of physics, Professor Schommers has discussed new aspects in connection with reality, and his basic ideas can be summarized as follows: Information about reality outside flows via sense organs into the body of the observer, and the brain forms a picture of reality. On the basis of many facts, Schommers concluded that the symbols in this picture of reality should have in general no similarity with the objects in the outside world; that is, the reality outside is transformed. On the one hand, we have the reality; on the other hand, we have a picture of reality. The reality is projected on space and time, and we obtain a picture of reality; the structures in the pictures are different from those in the reality outside. This conception is discussed mathematically by Professor Schommers in connection with quantum phenomena leading to new aspects in connection with relevant basic topics. Like both Whitehead and Bergson, Schommers argues for the primacy of processes and shows that space and time are closely tied to real processes. Selected work in this regard are "Inertial Frames of Reference: Mass Coupling to Space and Time" (*International Journal of Theoretical Physics* 20, 411, 1981); "Raum-Zeit. Quantentheorie und Bilder von der Wirklichkeit" (*Philosophia Naturalis* 23, 238, 1986); "Being and Becoming at the Microscopic Level" (*International Journal of Modern Physics B* 3, 1, 1989); "Space-Time and Quantum Phenomena," in *Quantum Theory and Pictures of Reality* (Springer-Verlag, Heidelberg, 1989); and "Truth and Knowledge," in *What Is Life?* (World Scientific, Singapore, 2002).

Professor Schommers is author and editor of the following books: *Fundamentals of Nanometer Structuring; Structure and Dynamics of Surfaces I and II; Quantum Theory and Pictures of Reality; The Visible and the Invisible; Das Sichtbare und das Unsichtbare; Elemente des Lebens; What is Life?; Formen des Kosmos; Space and Time, Matter and Mind; Symbols, Pictures and Quantum Reality.*

Professor Schommers is the editor-in-chief of the *Journal of Computational and Theoretical Nanoscience*. He is also an editorial board member of various scientific journals, and he is principal editor-in-charge of the book series *Foundations of Natural Science and Technology*. He is an invited member of the Humboldt Academy, an invited member of the Academic Board of the Humboldt Society, and an invited member of the Advisory Board of Medical Ethics of the 21st Century. Professor Schommers is also deputy governor of the American Biographical Institute (inauguration 2000).

Professor Schommers has been honored by various awards, medals, and appointments. He has been cited in *Who's Who in the World, Who's Who in Science and Technology, Living Science, The Europe 500, The Barons 500, 2000 Outstanding Intellectuals of the 21st Century, Leading Intellectuals of the World, 500 Leaders of Influence, and International Register of Profiles* (no. 123 of 200), and elsewhere.

List of Contributors

Number in parentheses indicates the page on which the author's contribution begins.

Ruben A. Abagyan (217)
Molsoft LLC, La Jolla, California, USA and The Scripps Research Institute, La Jolla, California, USA

Asen Asenov (489)
Device Modelling Group, Department of Electronics and Electrical Engineering, University of Glasgow, Glasgow, Scotland

James R. Baker, Jr. (673)
Center for Biologic Nanotechnology, University of Michigan, Ann Arbor, Michigan, USA

Artur Baumgaertner (73)
Forschungszentrum, Institut für Festkorperforschung, Jülich, Germany

Lubica Benuskova (779)
Knowledge Engineering and Discovery Research Institute, School of Information Technology, Auckland University of Technology, Auckland, New Zealand

Anna Bowman (305)
Computational Chemistry Centre, School of Chemistry, University of Bristol, Bristol, United Kingdom

Giacomo Carrea (259)
Istituto di Chimica del Riconoscimento Molecolare, Milano, Italy

Claudio N. Cavasotto (217)
Molsoft LLC, La Jolla, California, USA

Jhunu Chatterjee (545)
FAMU-FSU College of Engineering, Tallahassee, Florida, USA

Ching Jen Chen (545)
FAMU-FSU College of Engineering, Tallahassee, Florida, USA

Shin-Ho Chung (567)
Department of Theoretical Physics, Australian National University, Canberra, Australia

Giorgio Colombo (259)
Istituto di Chimica del Riconoscimento Molecolare, Milano, Italy

Dimiter S. Dimitrov (1)
Laboratory of Experimental and Computational Biology, CCR, NCI-Frederick, National Institutes of Health, Frederick, Maryland, USA

Robert A. Freitas, Jr. (619)
Institute for Molecular Manufacturing, Palo Alto, California, USA

Masami Hagiya (817)
Department of Computer Science, Graduate School of Information Science and Technology, University of Tokyo, Japan

Yousef Haik (545)
FAMU-FSU College of Engineering, Tallahassee, Florida, USA

Andreas Hanke (163)
Department of Physics and Astronomy, University of Texas at Brownsville, Brownsville, Texas, and NanoTech Institute, University of Texas at Dallas, Richardson, Texas, USA

Neil F. Johnson (757)
Centre for Quantum Computation, Clarendon Laboratory, Department of Physics, University of Oxford, Oxford, United Kingdom

Mikko Karttunen (361)
Biophysics and Statistical Mechanics Group, Laboratory of Computational Engineering, Helsinki University of Technology, Helsinki, Finland

Nikola Kasabov (1, 779)
Knowledge Engineering and Discovery Research Institute, School of Computer and Information Sciences, Auckland University of Technology, Auckland, New Zealand

Michiel van Lun (47)
Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden

István Majoros (673)
Center for Biologic Nanotechnology, University of Michigan, Ann Arbor, Michigan, USA

Garland R. Marshall (719)
Center for Computational Biology, Departments of Biochemistry and Molecular Biophysics and of Biomedical Engineering, Washington University, St. Louis, Missouri, USA

Massimiliano Meli (259)
Istituto di Chimica del Riconoscimento Molecolare, Milano, Italy

Ralf Metzler (163)
NORDITA–Nordic Institute for Theoretical Physics, DK-2100 Copenhagen, Denmark

Campbell Millar (489)
Device Modelling Group, Department of Electronics and Electrical Engineering, University of Glasgow, Glasgow, Scotland

Adrian Mulholland (305)
Computational Chemistry Centre, School of Chemistry, University of Bristol, Bristol, United Kingdom

Alexandra Olaya-Castro (757)
Centre for Quantum Computation, Clarendon Laboratory, Department of Physics, University of Oxford, Oxford, United Kingdom

Andrew J. W. Orry (217)
The Scripps Research Institute, La Jolla, California, USA

Zhong-Can Ou-Yang (419)
Institute of Theoretical Physics, The Chinese Academy of Sciences, Beijing, China

Dennis P. Riley (719)
Kereos, Inc., St. Louis, Missouri, USA

Igor A. Sidorov (1)
Laboratory of Experimental and Computational Biology, CCR, NCI-Frederick, National Institutes of Health, Frederick, Maryland, USA

David van der Spoel (47)
Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden

Thommey Thomas (673)
Center for Biologic Nanotechnology, University of Michigan, Ann Arbor, Michigan, USA

D. Peter Tieleman (567)
Department of Biological Sciences, University of Calgary, Canada

Ilpo Vattulainen (361)
Laboratory of Physics and Helsinki Institute of Physics, Helsinki University of Technology, Helsinki, Finland

Yang Zhang (419)
Center of Excellence in Bioinformatics, University at Buffalo, Buffalo, New York, USA

Haijun Zhou (419)
Max-Planck-Institute of Colloids and Interfaces, Potsdam, Germany

Handbook of Theoretical and Computational Nanotechnology

Edited by

Michael Rieth and Wolfram Schommers

Volume 1. BASIC CONCEPTS, NANOMACHINES, AND MEDICAL NANODEVICES

Volume 2. ATOMISTIC SIMULATIONS—ALGORITHMS AND METHODS

Volume 3. QUANTUM AND MOLECULAR COMPUTING, QUANTUM SIMULATIONS

Volume 4. NANOMECHANICS AND MULTISCALE MODELING

Volume 5. TRANSPORT PHENOMENA AND NANOSCALE PROCESSES

Volume 6. BIOINFORMATICS, NANOMEDICINE, AND DRUG DESIGN

Volume 7. MAGNETIC NANOSTRUCTURES AND NANO-OPTICS

Volume 8. FUNCTIONAL NANOMATERIALS, NANOPARTICLES, AND POLYMER DESIGN

Volume 9. NANOCOMPOSITES, NANO-ASSEMBLIES, AND NANOSURFACES

Volume 10. NANODEVICE MODELING AND NANOELECTRONICS

CHAPTER 1

Computational Biology

Dimiter S. Dimitrov, Igor A. Sidorov
Laboratory of Experimental and Computational Biology, CCR, NCI-Frederick, National Institutes of Health, Frederick, Maryland, USA

Nikola Kasabov
Knowledge Engineering and Discovery Research Institute, School of Computer and Information Sciences, Auckland University of Technology, Auckland, New Zealand

CONTENTS

ISBN: 1-58883-048-9

Handbook of Theoretical and Computational Nanotechnology
Edited by Michael Rieth and Wolfram Schommers
Volume 6: Pages (1–46)

1. INTRODUCTION

A major goal of computational biology is to discover knowledge or to enhance knowledge discovery for biological systems through computation. Computational biology, a term coined from analogy to the role of computing in the physical sciences, is now coming into its own as a major element of contemporary biological and biomedical research [1]. Information and computational sciences provide essential tools for the next-generation biological science efforts, from focusing the direction of experimental studies to providing knowledge and insight that cannot otherwise be obtained. Going beyond the revolution in biology reflected in the successes of the genome project and driven by the power of molecular biology techniques, computational approaches will provide an underpinning for the integration of broad disciplines for development of a quantitative systems approach to understanding the mechanisms determining the life of the cell and organism. Another aspect of the integration of computation and biology is that biological systems can be viewed as special computing devices. This view emerges from considerations of how information is stored in and retrieved from the genes. Genes can only specify the properties of the proteins they code for, and any integrative properties of the system must be "computed" by their interactions. This provides a framework for analysis by simulation and sets practical bounds on what can be achieved by reductionist models [2].

Recent advances in many areas of biology, especially in genomics, are heavily rooted in engineering technology, from the capillary electrophoresis units used in large DNA sequencing projects to the photolithography and robotics technology used in chip manufacture, to the confocal imaging systems used to read those chips, to the beam and detector technology driving high-throughput mass spectroscopy. Further advances in materials science and nanotechnology promise to improve the sensitivity and cost of these technologies greatly in the near future [3]. Current research makes it possible to look at biological phenomena on a scale not previously possible: all genes in a genome, all transcripts in a cell, and all metabolic processes in a tissue.

These modern approaches produce massive quantities of data. GenBank, for example, now accommodates more than 10^{10} nucleotides of nucleic acid sequence data and continues to more than double in size every year. New technologies for assaying gene expression patterns, protein structure, protein–protein interactions, and so forth will provide even more data. How to handle these data, make sense of them, and render them accessible to biologists working on a wide variety of problems is the challenge facing computational biology and bioinformatics seeking to integrate computer science with applications derived from molecular biology.

One core aspect of research in computational biology focuses on database development: how to integrate and optimally query data from genomic DNA sequence, spatial and temporal patterns of mRNA expression, protein structure, immunological reactivity, clinical outcomes, publication records, and other sources. A second focus involves pattern-recognition algorithms for such areas as nucleic acid or protein sequence assembly, sequence alignment for similarity comparisons or phylogeny reconstruction, motif recognition in linear sequences or higher-order structure, and common patterns of gene expression. Both database integration and pattern recognition depend absolutely on accessing data from diverse sources and on being able to integrate, transform, and reproduce these data in new formats.

Computational biology is a fundamentally collaborative discipline, owing its very existence to the availability of rich and extensive data sets for analysis, integration, and manipulation.

Data accessibility and usability are therefore critical, raising concerns about data release policies—what constitutes primary data, who owns this resource, when and how data should be released, and what restrictions may be placed on further use.

2. COMPUTATIONAL GENOMICS AND PROTEOMICS

2.1. DNA Sequencing

DNA (deoxyribonucleic acid) is a nucleic acid polymer consisting of individual units termed nucleotides. Each nucleotide consists of one of four distinct nucleosides (deoxypentose sugar plus one of four bases (adenine [A], guanine [G], cytosine [C], and thymine [T])) and a phosphate group. Thymine is replaced by uracil (U) in RNA (ribonucleic acid). With respect to similarity in structure, nucleosides are divided in two classes: pyrimidines and purines. Nucleosides A, T, G, and C are capable of being linked together to form a long chain. The bases along the polymer can interact with complementary bases in the other strand: Adenine is capable of forming hydrogen bonds with thymine (A:T), and cytosine can pair with guanine (C:G). Thus, the DNA consists of two antiparallel strands and can be written, for instance, as

```
... A T C G C G T T A G C T A G C T ...
    | | | | | | | | | | | | | | | |
... T A G C G C A A T C G A T C G A ...
```

The main steps of DNA sequencing are the following: purified fragments of DNA are denatured to a single chain, and then one strand is hybridized to an oligonucleotide primer with small amounts of one of four chain-terminating nucleotides. After synthesis, the mixture, consisting of DNA fragments ending with one of the nucleotides, is electrophoresed to separate fragments by size. After this, one can calculate the probable order of the bands and predict the sequence. For the sequencing of larger molecules of DNA, the molecules are first randomly sheared, the fragments are sequenced then, and finally the sequence of the large molecule is assembled from the overlaps found.

After sequencing, the information about the sequence can be submitted to one of the data banks, including GenBank at the National Center of Biotechnology Information, National Library of Medicine, Washington, DC (http://www.ncbi.nlm.nih.gov/Entrez); the European Molecular Biology Laboratory (EMBL) Outstation at Hixton, England (http://www.ebi.ac.uk/embl/index.html); and the DNA DataBank of Japan (DDBJ) at Mishima, Japan (http://www.ddbj.nig.ac.jp/). A more extensive list of the data banks can be found in DBCAT (Public Catalog of Databases) located at http://www.infobiogen.fr/services/dbcat/. Different data banks may have different formats for storing the data, but most of them have the same features: sequence name and identification code, source organism, keywords to look up this entry, dates of entry and modification, and so forth.

2.2. DNA and RNA Structures

A nucleoside is one of the four DNA bases attached covalently to the sugar. The sugar in deoxynucleosides is 2′-deoxyribose, and ribose in ribonucleosides. The four different nucleosides of DNA are deoxyadenosine (dA), deoxyguanosine (dG), deoxycytosine (dC), and deoxythymidine (dT). A nucleotide is a nucleoside with one or more phosphate groups covalently attached to the 3′- or 5′-hydroxyl group or groups. The DNA backbone is a polymer with an alternating sugar–phosphate sequence. DNA is a normally double-stranded macromolecule with two polynucleotide chains (the double-helical nature of DNA was discovered in 1953 [4]). These chains are noncovalently held together by weak intermolecular forces and form a DNA molecule. Two DNA strands form a helical spiral, winding around a helix axis in a right-handed spiral with two polynucleotide chains running in opposite directions. The sugar–phosphate backbones wind around the helix axis. The bases of the individual nucleotides are on the inside of the helix. For DNA duplexes, the right-handed double helix

has 10 pairs per complete turn. Within the DNA double helix, the adenine:thymine base pair has two hydrogen bonds, compared to three in the guanine:cytosine pair. The two base pairs are required to be identical in dimensions by the Watson–Crick model. High-resolution X-ray crystallographic analysis of the ribodinucleoside monophosphate duplexes (G:C and A:U) showed that the distances between the glycosidic carbon atoms in the base pairs are close (10.67 and 10.48 Å, respectively).

RNA molecules are polynucleotides containing ribose sugars connected by phosphodiester linkages. Although RNA is generally single-stranded, double-stranded RNA molecules can be formed where uracil participates in a U:A pair. Single-stranded RNA have a tendency to fold back on themselves to form double-stranded structures like stacked double-helixes for the regions with paired bases and different loops (bulge, hairpin, internal, and multibranch) for unpaired bases. These elements form the RNA secondary structure.

Prediction of RNA secondary structure requires intensive computational resources. Usually, RNA resultant structure corresponds to the local minima of the free energy, and overall free energy of the molecular folded is the sum of the energies of the stacked base pairs and loops. However, the molecular environment and folding pathway, which can have a significant effect on this structure, should be accounted.

In Structurelab [5], dynamic programming algorithm and genetic algorithm were used for determination of the folding of the RNA molecules. This system allows researchers to pursue interactively and methodically a multiperspective analysis of RNA structure (multiple and individual). It uses various software modules and hardware complexes [6]. Secondary structure representation of RNA molecular structures is based on LISP's nested list notations, for instance [N(H)(H)(BH)(H)(H)(H)(BBBIH)], where the symbols are H, hairpin loop; B, bulge loop; I, internal loop; and M, multibranch loop.

Other packages that can be used for secondary structure prediction and presentation are *mfold*, http://www.bioinfo.rpi.edu/~zukerm/rna/, http://bioweb.pasteur.fr/seqanal/interfaces/mfold-simple.html, http://biotools.idtdna.com/mfold/, http://www.bioinfo.rpi.edu/applications/mfold/old/rna/form1.cgi; RNAfold, http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi; ESSA, http://www.inra.fr/bia/T/essa/Doc/essa_home.html; and CARD, http://rrna.uia.ac.be/card.html.

2.3. Protein Sequencing

Proteins are linear polymers of amino acids linked by a peptide bond. The primary structure of the protein (protein sequence) determines the ultimate three-dimensional structure of the protein. Amino acids are small molecules consisting of an amino group (NH_2), a carboxyl group (COOH), a hydrogen atom attached to the central carbon (α), and a side chain (or R group) attached to the central carbon. There are 20 standard amino acids, which can be grouped into classes based on the chemical properties conferred by their side chains. Amino acids can be charged (+/−), hydrophobic/hydrophilic, polar/nonpolar, and capable of H-bonding—allowing for weak interactions. Amino acids can form peptide bonds with each other through reaction of the carboxyl and amino groups.

2.4. Protein Structure

Protein structure is typically considered on several levels.

The primary structure is the sequence of amino acids that make up a protein. This sequence determines the ultimate three-dimensional structure of the protein. The secondary structure involved local folding of peptide, creating distinctive structures shared by many proteins including alpha (α) helices and beta (β) pleated sheets. These structures were predicted theoretically before the experimental determination of protein structure, and they are the only regular secondary structural elements present in proteins (there are also irregular structural elements: loop and coil). Helix is created by a curving of the polypeptide backbone, and sheet is formed by hydrogen bonds between adjacent polypeptide chains rather than within a single chain. There are two configurations for both elements: rightward/leftward for helix and parallel/antiparallel for sheet.

The tertiary structure is a global, three-dimensional structure of the polypeptide chain. At this level of structure, the side chains play a major role in creating the final structure. Protein folding is the process of forming a final three-dimensional tertiary structure. It is interesting to note that random polypeptide sequences almost never fold into an ordered structure, so protein sequences were selected by the evolution to achieve reproducible stable structure [7].

Finally, the quaternary structure is the way multiple subunits of a protein interact. Many proteins are formed from more than one polypeptide chain (i.e., they exist as a noncovalent association of two or more identical or different polypeptides folded independently). The quaternary structure describes the way in which the different subunits are packed together to form the overall structure of the protein. For example, the human hemoglobin molecule is made of four subunits. Other examples of the combination of nonidentical subunits are immunoglobulins and bovine hemoglobin.

One task that has been explored in the literature is predicting the secondary structure from the primary one. Segments of a protein can have different shapes in their secondary structure, which is defined by many factors—one of them being the amino-acid sequence itself.

Protein secondary structure prediction can be performed by different packages, among them

1. NNPREDICT (http://www.empharm.ucsf.edu/~nomi/nnpredict.html) predicts using a two-layer, feed-forward neural network and a separate program—a modification of the Parallel Distributed Programming suite (see for details; see Refs. [8, 9]).
2. PHDsec (http://cubic.bioc.columbia.edu/predictprotein) predicts secondary structure from multiple sequence alignments. Secondary structure is predicted by a system of neural networks rating for the three states, helix, strand, and loop, at an expected average accuracy greater than 72% [10–12].
3. PROFsec (http://cubic.bioc.columbia.edu/predictprotein) is an improved version of PHDsec, a profile-based neural network prediction of protein secondary structure.
4. JPRED (http://www.compbio.dundee.ac.uk/~www-jpred/) is a consensus method for protein secondary structure prediction [13].

Qian and Sejnowski [14] investigated the use of multilayer perceptrons (MLPs) for the task of predicting the secondary structure based on available labeled data. In Ref. [15], an Evolving Fuzzy Neural Network (EFuNN) is trained on the data from Ref. [14] to predict the shape of an arbitrary new protein segment. A window of 13 aminoacids was used; there were 273 inputs and three outputs, and 18,000 examples for training were used. The block diagram of the EFuNN model is given in Fig. 1 (from Ref. [15]).

Prediction of the three-dimensional structure of proteins by homology modeling (described in detail in the next section) is based on the similarity of primary sequences of the protein being analyzed to those of a protein of experimentally determined structure (this method

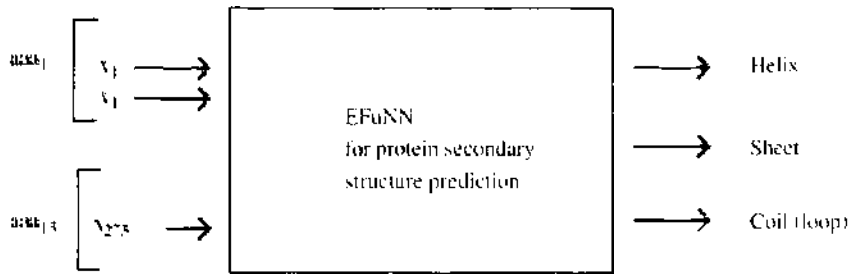


Figure 1. An artificial neural network model (in this case it is an evolving fuzzy neural network [15]) for the prediction of the protein secondary structure. Reprinted with permission from [15], N. Kasabov, "Evolving Connectionist Systems —Methods and Applications in Bioinformatics, Brain Study and Intelligent Machines," Springer, New York, 2002. © 2002, Springer.

assumes that significant identity between the two sequences exists). Algorithms for homology modeling can be found in the following servers:

1. SWISS-MODEL (http://www.expasy.ch/swissmod/SWISS-MODEL.html), an Automated Protein Modeling Server running at the GlaxoWellcome Experimental Research in Geneva, Switzerland (see, for details, http://www.expasy.ch/swissmod/SWISS-MODEL.html and [16, 17]).
2. CPHmodels (http://www.cbs.dtu.dk/services/CPHmodels/), Centre for Biological Sequence Analysis, The Technical University of Denmark, Denmark: Methods and databases developed to predict protein structures include Sowhat, a neural network–based method to predict contacts between C-alpha atoms from the amino acid sequence, and RedHom, a tool to find a subset with low sequence similarity in a database (see Ref. [18]).
3. 3D-JIGSAW Comparative Modelling Server (http://www.bmm.icnet.uk/~3djigsaw/); BioMolecular Modeling Group, Imperial Cancer Research Fund, London: an automated system to build three-dimensional models for proteins based on homologs of known structure ([19–21]).
4. SDSC1-SDSC Structure Homology Modeling Server (http://cl.sdsc.edu/hm.html) and Databases and Tools for 3-D Protein Structure Comparison and Alignment (http://cl.sdsc.edu/ce.html); San Diego Supercomputing Center, San Diego, CA [22, 23].

2.5. Searching for Motifs in Sequences

To find a motif (or pattern or consensus) in the sequence, one first needs to define it. Let us consider a sequence as a vector of symbols

$$X = (x_1, x_2, \ldots, x_L)$$

where L is sequence length and all symbols x_i belong to a finite set of symbols (or alphabet)

$$x_i \in A = \{a_1, \ldots, a_k\} \quad i = 1, \ldots, L$$

For DNA sequences, the alphabet is simply set of four letters

$$A = \{A, T, G, C\}$$

Searching for a functional motif can be considered like comparing two sequences: one is the target sequence (S), and the other is the motif (M)

$$S = (s_1, s_2, \ldots, s_{L_S})$$
$$M = (m_1, m_2, \ldots, m_{L_M})$$

having length L_S and L_M, respectively.

Let now assume that function $w_{ij} = w(a_i, a_j)$ expresses the weight of a combination of two symbols in comparing sequences. The simplest equation for this function is

$$w(a_i, a_j) = \begin{cases} 1 & a_i = a_j \\ 0 & a_i \neq a_j \end{cases}$$

In this case, position j of the motif in target sequence can be found as a value giving maximum for the following score

$$Q_j = \sum_{i=1}^{L_M} w(m_i, s_{i+j-1}) \quad j = 1, \ldots, L_S - L_M + 1$$

Motif can be also represented as a probability (frequency or weight) $p_i(a_j)$ to find the jth symbol a_j from alphabet A in position i of the motif. Using the values of matrix $\|p_i(a_j)\|$ (position weight matrix, PWM), the score value for each position can be calculated as

$$Q_j = \sum_{i=1}^{L_M} p_i(s_{i-j+1}) \quad j = 1, \ldots, L_S - L_M + 1$$

If for the ith position in the motif $p_i(m_i) = 1$ and equals 0 for all other symbols, then this approach is equal to the simplest case of the weight function. Graphically, both methods can be represented as shown in the Fig. 2. The matrix shown in this picture is a dot plot (i.e., a visual representation of the similarity between two sequences). Each axis represents one of the two sequences to be compared, and each cell represents weight or probability function values. To calculate Q_j for position j in a target sequence, one should summarize the values on the diagonal, starting from the cell $(1, j)$. For a case in which weight function has a simplest representation (1 for the equal symbols and 0 for different ones) and the values of this function are represented as empty and filled cells for 0 and 1, respectively, then the whole diagonal will represent sequences sharing similarity. For sequences that share only patches of similarity, diagonal stretches will be shown.

The method of dot matrix analysis was first described in Ref. [24]. It can be also useful for finding inverted repeats and self-complimentary repeats. The use of an enhanced dot plot for nucleic and protein sequences was described in Ref. [25]. An additional description of the method can be found in Ref. [26].

The package for detection of patterns and structural motif in nucleotide sequences (PatSearch) is described in Ref. [27]. It allows scanning for specific combinations of oligonucleotide consensus sequences with defined order, orientation, and spacing, and it also allows mismatches and mispairing below a user-fixed threshold (available at http://bighost.area.ba.cnr.it/BIG/PatSearch). The possible pattern units for this package are as follows: string, palindrome, hairpin loop, position weight matrix, repeat. It also uses logical patterns such as "either/or" and length constraints for specific combination of pattern units.

The TRANSFAC database (available at http://www.gene-regulation.com) on eukaryotic transcriptional regulation comprises data on transcriptional factors, their target genes and regulatory binding sites [28], and tools for a matrix-based search of transcription-factors' binding sites (MATCH). The algorithm of MATCH uses two values to score hints: the matrix similarity score and the core similarity score, which is close to the MatInspector algorithm [29].

2.6. Sequence Alignment

There are two main type of sequence alignment: pairwise (comparing two sequences) and multiple sequence (comparing more than two sequences). Multiple-sequence alignment is the procedure of comparing sequences by searching for the similarity in the subsets that are

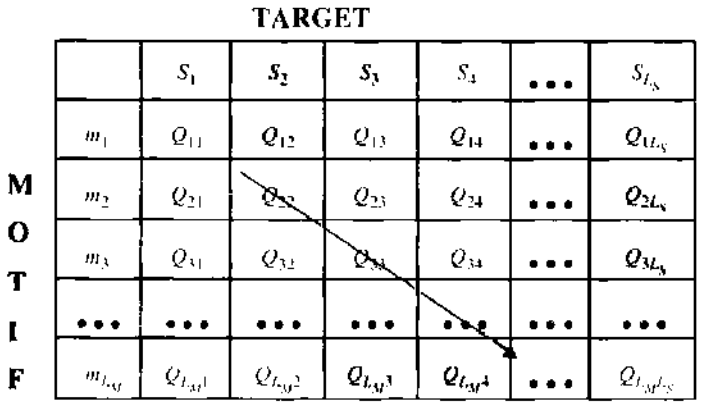


Figure 2. Searching for the motif in target sequence. Arrow shows the direction of calculating the score function $Q_j = \sum_{i=1}^{L_M} Q_{i,\,j+i-1}$, where Q_{ij} equals $w(m_i, s_j)$ or $p_i(s_j)$.

in the same order in the sequences. Each subset can consist of one or more characters of the sequence and the gap or gaps between them.

Comparison of two or more sequences has a lot of biological rationales. Among them is the idea that gene sequences may have derived from common ancestral sequences, and thus the changes in sequence (mutation, insertion, and deletion) can show us the evolutionary course of the particular molecule. Another reason for multiple-sequence alignment is indicating the regions of common origin that may in turn coincide with regions of similar structure or similar function. Results of alignment can be used as a starting point for solving various tasks (predicting *de novo* the secondary structure of proteins and other knowledge-based structure predictions, resolving phylogenetic issues, and interpreting data from the human genome).

Let us have n sequences $S_1, S_2, \ldots, S_n$, and have each sequence be represented as a vector

$$S_i = (s_1^i, s_2^i, \ldots, s_{L_i}^i) \quad i = 1, \ldots, n$$

where N_i is the length of the ith sequence, and

$$s_j^i \in A = \{a_1, \ldots, a_K\} \quad \text{for all } i \text{ and } j$$

Let us now assume that each sequence can be represented with gaps (insertions/deletion), so

$$S_i = (g_0^i, s_1^i, g_1^i, s_2^i, g_2^i, \ldots, g_{L_i-1}^i, s_{L_i}^i, g_{L_i}^i)$$

where g_j^i, $j = 0, \ldots, L_i$ are the gaps inserted in the ith sequence at jth place. Alignment of n sequences can be represented as a matrix $R = \|r_{ij}\|$, with the following properties: $r_{ij} \in A \cup \{\text{gap}\}$, so the gap is included in the alphabet; each row matrix represents the ith sequence with gaps $r_i = S_i$; and each column cannot consist only of gaps.

Score function for multiple alignments depends on the weight function [$w_{ij} = w(a_i, a_j)$, scoring matrix] and the so-called gap penalty function. The latter describes the decrease in score for gaps of given length and consists of the constant term-describing penalty for opening the gap (a) and of the penalty for each element in gap (b). The usual formula for the penalty for the gap having length (so-called affine gap penalty) is

$$Q(g) = a + bg$$

and one of its extensions

$$Q(g) = \begin{cases} a + b(g - q) & g > q \\ a & g \leq q \end{cases}$$

where q means that gap penalty for each element will be added only when gap size is greater than q.

It is obvious that gap-penalty functions have to be appropriate to the weight function to obtain a reasonable alignment. If the gap-penalty function is high enough with respect to the scoring matrix values, final alignments will never have gaps. However, too-small values of the gap-penalty function will lead to the alignment having gaps everywhere.

Two alignments can be compared using the same score function. The typical matrix of alignment is represented in Fig. 3. It should be noted that several different alignments can provide approximately the same alignment score.

A key element in evaluating the quality of a sequence alignment is the score matrix (or substitution score matrix) $w_{ij} = w(a_i, a_j)$, which assigns a score for aligning any possible pair of sequence elements. The theory of amino acid substitution matrices is described in Ref. [30] and is applied to DNA sequence comparison in Ref. [31].

Percent accepted mutation matrices (PAM, or Dayhoff amino acid substitution matrices) list the probability of change from one amino acid to another in homologous proteins during evolution. In deriving the PAM matrices, each change in the current amino acid at a particular site is assumed to be independent of previous mutational events at that site [32], so amino acid substitutions can be viewed as a Markov model. To calculate the values of

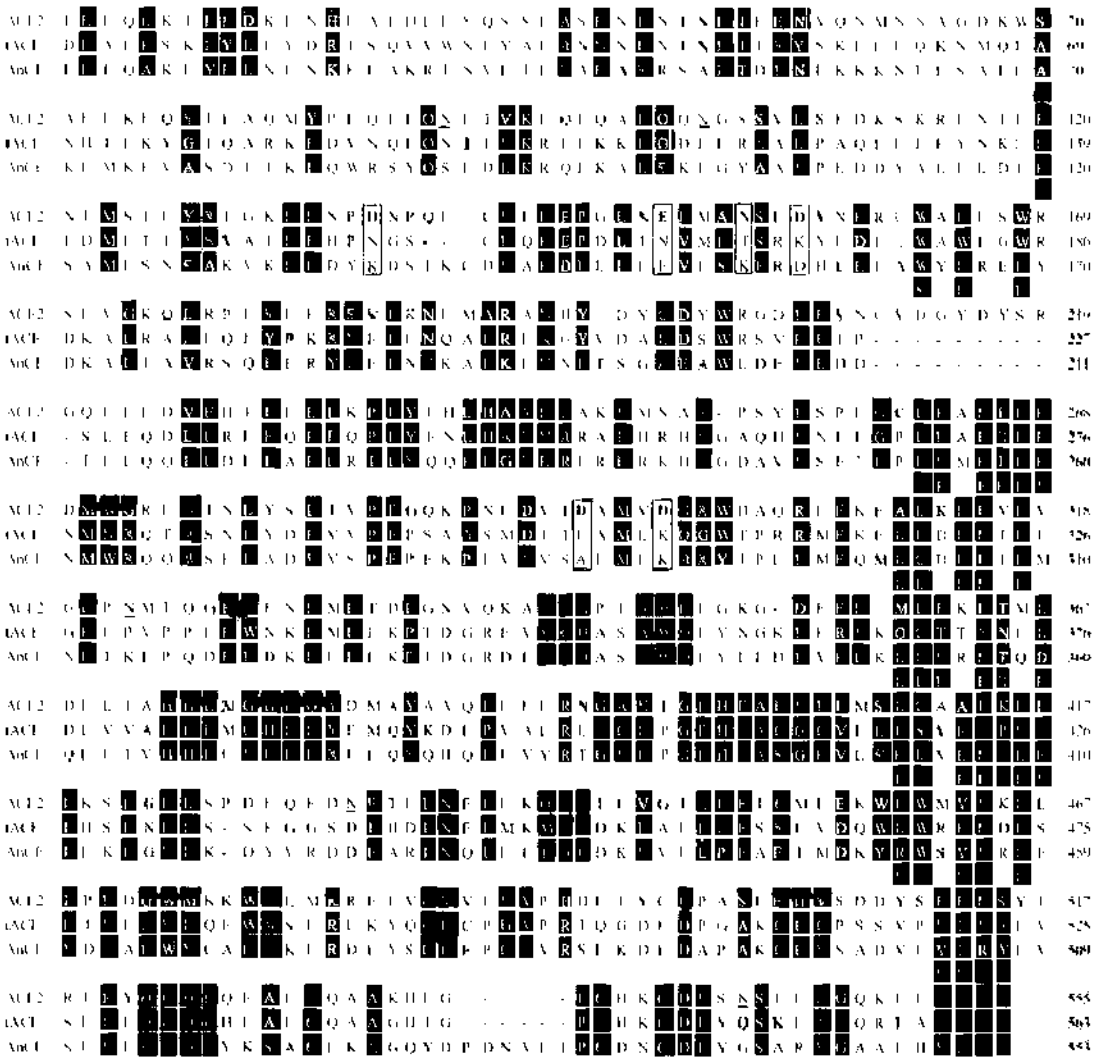


Figure 3. Multiple sequence alignment of angiotensin-converting enzyme 2 (ACE2), testis-specific ACE2 (tACE), and the *Drosophila* homolog of ACE2 (AnCE) using the program CLUSTALW. The sequence numbering is the same as in the crystal structures. The *N*-glycosylation sites are underlined; colored in blue, the putative binding residues in ACE2 are in boldface letters and boxed along with the corresponding aligned residues in tACE and AnCE. The identical and similar residues are shown in black and gray backgrounds, respectively. Conserved residues are in green, and the critical binding residues are in red.

Dayhoff matrices, amino acid substitutions that occur in groups of evolving proteins were estimated using 1572 changes in 71 groups of protein sequences with at least 85% similarity.

There are several approaches related to PAM. They can be based on exhaustive matching of the entire protein sequence database [33] or rapid generation of mutation data matrices from protein sequences [34], as well as accounting for the different patterns of mutation at low and high sequence divergence [35].

The BLOSSOM substitution matrix [36] is used for scoring protein sequence alignment and is based ondifferent type of sequence analysis and a much larger data set then the Dayhoff matrices. A 400 × 400 dipeptide substitution matrix that reports empirical probabilities for the interconversion of all pairs of dipeptides in proteins undergoing divergent evolution was presented in Ref. [37]. In Ref. [38], a mutation data matrix was calculated for membrane spanning segments.

Detailed descriptions of the methods and algorithms for alignment of pair of sequences and multiple sequence alignment can be found in Ref. [39].

In the late 1980s, fast algorithms for comparison DNA and protein sequences were developed that are capable of searching sequence databases, evaluating similarity scores, and identifying periodic structures based on local sequence similarity FASTP, FASTA, and

LFASTA [40, 41]. (FASTA stands for FAST-All, reflecting the fact that it can be used for a fast protien comparison or a fast nucleotide comparison: FASTP can search protein or DNA sequence databases and compare a protein sequence to a DNA sequence database; LFASTA can display all the regions of local similarity between two sequences.) These algorithms achieve much of their speed and selectivity by using a look-up table to locate all identities between two sequences.

A new method, FastM, for the development of simple models of transcriptionally regulatory units was developed in Ref. [42]. It combines a search algorithm for individual transcriptional factor binding sites (MatInspector, see Ref. [29]) with a distance correlation function. The models are composed from various individual elements (hairpins, direct repeats, short multiple repeats, and terminal repeats; see Ref. [43] for details). FastM now is part of the GenomatixSuite, and the whole package (including ElDorado, a portal to explore various genomes; Gene2Promoter, for retrieval and analysis of promoters; BiblioSphere, for literature network mining; GEMS Launcher, Genomatix' genome exploring and modeling software package; MatInspector, the de facto standard for transcription-factor binding-site search; and PromoterInspector, highly specific prediction of mammalian promoter regions) is available at http://www.genomatix.de/cgi-bin/fastm2/fastm.pl.

A basic local alignment search tool (BLAST) for rapid sequence comparison was developed in 1990 [44]. It directly approximates alignments that optimize a measure of local similarity, the maximal segment pair (MSP) score. The basic algorithm is simple and robust, and it was applied in straightforward DNA and protein sequence database searches, motif searches, gene identification searches, and the analysis of multiple regions of similarity in long DNA sequences. A new generation of this algorithm for searching databases (gapped BLAST and PSI-BLAST) is described in Ref. [45].

2.7. Proteomics

Understanding the structures, interactions, and functions of all of a cell's or organism's proteins has been given a disciplinary title of its own: proteomics. The ultimate goal of proteomics is to characterize the information flow through protein networks.

The word proteome, coined in 1994 as a linguistic equivalent to the concept of genome, indicates proteins expressed by a genome. This term was coined by Marc Wilkins and colleagues and appeared for the first time in 1995 [46]. The term "proteome" is used to describe the complete set of proteins that is expressed, and modified following expression, by the entire genome in the lifetime of a cell. It can be used also as a description of proteins expressed by a cell at any stage. The generation of messenger RNA expression profiles is referred to as transcriptomics, as they are based around the process of transcription, and the mRNAs transcribed from a cell's genome is the transcriptome.

The field of proteomics has rapidly expanded and includes diverse technologies:

- Two-dimensional gel electrophoresis and mass spectrometry-based methods for protein profiling: These methods are currently the major experimental technologies for large-scale and high-throughput analysis of proteomes focusing on the proteins' identification and on their qualitative and quantitative comparison.
- Protein microarrays: Proteome profiling is a very powerful tool in clinical medicine for the identification of diagnostic markers. Clinical applications of proteomics can also provide information on drug targets, the mechanism of drug action, and drug-mediated toxicity.
- Yeast two-hybrid system [47]: Yeast two-hybrid assays are the main technology for large-scale interaction network construction. They can detect interactions between two known proteins or polypeptides. Two strategies (the matrix approach and the library-screening approach) have been tested to find the most efficient way to explore interactions within the proteome [48].
- Protein–protein interaction pathways and cell-signaling networks: This functional proteomics approach involves the generation of large-scale protein–protein interaction networks.

- High-throughput protein structural studies using mass spectrometry, nuclear magnetic resonance, and X-ray crystallography: The majority of drug target molecules are proteins, and for a large number of proteins, information on three-dimensional structures is required for drug discovery. Although high-throughput X-ray and nuclear magnetic resonance methods hold much promise for protein three-dimensional structure determination in the future, at present the only efficient way to determine the structure is to implement large-scale homology modeling programs to accompany structural proteomics initiatives.
- High-throughput computational methods for protein three-dimensional structure and function determination.

The two key steps in classical proteomics are the separation of proteins derived from cells or tissues and their subsequent identification. One of the best methods of separation is two-dimensional gel electrophoresis. In this method, spots of a prepared mixture of proteins extracted from cells or tissues are applied to a polyacrylamide gel. The proteins, which can number in the tens of thousands, are separated along the gel in one direction according to two different properties: their molecular charge, by applying an electric field (isoelectric focusing), and their molecular mass, using SDS–PAGE (polyacrylamide gel) electrophoresis. A typical gel can reliably separate 2000 protein spots in this way, and the "best" ones can separate up to 11,000 protein spots. Proteins separated on the gel are stained using Coomassie blue dye, silver stains, or fluorescent dyes or by radiolabeling, and the proteins are then quantified using spectroscopic or radiographic techniques. Although two-dimensional gel electrophoresis gives the highest resolution of all available methods, there are difficulties in its using hydrophobic proteins, such as the cell-membrane-spanning receptor proteins (the most attractive drug targets), which usually do not dissolve in solvents used for isoelectric focusing (as well as proteins with very high relative molecular mass); low-abundance proteins, which cannot be recognized on the background of high-abundance "housekeeping" proteins (unfortunately, there is no amplification method for proteins, like there is a polymerase chain reaction method for genes); and proteins having very high molecular charges, or very low molecular mass, which will not separate on gels.

After separation, identification of the proteins begins with their digestion into fragments by specific proteases, and then the fragments are analyzed by mass spectroscopy in a process called peptide mass fingerprinting. In this approach, proteins are identified by comparing the mass of the peptide fragments with data predicted by genetic or protein sequence information.

Any experiment that involves a limited number of proteins can avoid the step with the two-dimensional gel separation by using other methods, such as high-performance liquid chromatography, gel filtration, or one-dimensional gel chromatography or microarray technology.

Protein microarray assays allow the identification and quantification of a large number of proteins from a small amount of a sample. They can be used for the analysis of interactions between proteins with other proteins, peptides, low–molecular weight compounds, oligosaccharides, or DNA.

The protein microarray is powerful tool for diagnostic and therapeutic purposes as well as for basic research. Array-based methods to study proteins allow high-throughput determination of protein functions in parallel [49]. It was demonstrated that a high-density antibody microarray could be applied for the global analysis of expression profiles of proteins, and different types of protein and peptide microarrays have been reported to be useful for immunoassays and for analyzes of enzymatic activity [50–52].

The major advantages of protein array technologies are based on the following features of the approach [53]: having a highly parallel and small solid-phase assay system, having a highly sensitive system, being useful for very high throughput approaches, having low consumptions of reagent samples, and having potentially attractive manufacturing costs.

The core technologies for protein microarrays currently practiced are surface chemistry for immobilization of proteins or capture agents, capture molecules that are immobilized onto a solid support and used for capturing target proteins or molecules, and systems to

detect protein–protein interactions based on fluorescence, chemiluminescence, mass spectrometry (MS), and electrochemical or surface plasmon resonance (SPR). Capture on microarray can be specific (affibodies, antibodies, aptamers, and antibody sandwich formation) or unspecific based on electrostatic, van der Waals–hydrophobic, or metal–chelate interactions. Specific interaction microarrays have been described for receptor–ligand protein–protein, protein–DNA, and enzyme–substrate interactions (see Ref. [54] for review).

Several studies illustrate the application of functional proteomics for the identification of regulated targets in specific pathways [55, 56].

Proteome profiling of microorganisms can also generates valuable knowledge that can be used for the development of metabolic and cellular engineering strategies. This approach includes the following steps [57]:

1. Obtain the proteome profiles of the microorganism under different conditions of interest.
2. Analyze the proteome profiling results based on biological, biochemical, and biotechnological information.
3. Develop a rational strategy for the engineering of the microorganism.
4. Compare and analyze the results (phenotypes) obtained by employing the microorganism before and after engineering.
5. Repeat the above steps until the results are satisfactory.

Protein microarrays have been used for the screening of molecular markers and pathway targets in patient-matched human tissue during disease progression. In contrast to protein arrays, in which immobilized capture molecules are directed against certain target proteins (e.g., an antibody), reverse-phase protein microarrays immobilize the whole repertoire of sample proteins that represent the state of individual tissue cell populations undergoing disease transitions.

Detailed review of the informatic tools for proteome profiling can be found in Refs. [58–60]. Development of software for two-dimensional gel image analysis began about 35 years ago [61–63] with further improvement that were made in the late 1980s [64–68]. There are many commercially available packages now, including DeCyder 2D Analysis, ImageMaster 2D Elite, http://www1.amershambiosciences.com/; Delta2D, http://www.decodon.com; GELLAB II+, www.scanalytics.com and http://www.lecb.ncifcrf.gov/gellab/index.html; GeneData Impressionist system, http://www.genedata.com; ImagepIQ, http://www.proteomesystems.com; Melanie 3, http://www.genebio.com; ProteinMine, http://www.scimagix.com; and TotalLab, http://www.nonlinear.com/products/totallab. Some of the two-dimensional gel image analysis packages can interact with automatic robotic systems.

Another approach used for identification of proteins is mass spectroscopy, and there are three different methods for identification based on mass-spectrometric data:

1. Peptide mass fingerprint or peptide mass map analysis: comparisons of peptide molecular weights determined by mass spectrometry with the theoretical masses of peptides produced *in silico* by digestion of sequences in a target database [69–73]. Database search tools include Mascot, http://www.matrixscience.com; Mowse, http://www.hgmp.mrc.ac.uk; PeptideSearch, http://www.narrador.embl-heidelberg.de; and Peptident, http://www.expasy.ch.
2. Peptide sequence or peptide sequence tag query: peptide tandem mass spectrometry (MS/MS) data are combined with amino acid sequence or composition data to identify the protein from a protein or nucleotide sequence database [74, 75]. Database search tools include TagIdent, www.expasy.ch, and MS-Seq, http://prospector.ucsf.edu.
3. MS/MS ion search analysis: uninterpreted MS/MS data from a series of peptides in a complete LC-MS/MS run are matched with protein sequences in a protein- or nucleotide-sequence database to identify proteins without any manual sequence interpretation. The input data is usually a list of fragment ion mass and intensity values [73, 74]. Database search tools include PepFrag, http://prowl.rockefeller.edu, and Sequest, http://fields.scripps.edu.

The isotope-coded affinity tag (ICAT) approach and tandem MS were applied [76, 77] to quantitative protein profiling. Individual tandem MS spectra were searched against a human

sequence database, and a variety of recently developed, publicly available software applications were used to sort, filter, analyze, and compare the results. In particular, robust statistical modeling algorithms were used to assign measures of confidence to both peptide sequences and the proteins from which they were likely derived, which were identified via the database searches. It was shown that these statistical tools allow the estimation of the accuracy of peptide and protein identifications made. Data flow for automated database searching and statistical data analysis was represented in this work as follows: first, acquired MS/MS spectra are submitted to SEQUEST [74] for searching protein sequence databases to identify peptides and protein sequence matches for each recorded MS/MS spectrum; second, the search results are then submitted to PeptideProphet, a statistical data modeling algorithm [78], and this algorithm generates its own discriminant score for the peptide sequence assigned to each MS/MS spectrum, based on the weighting of a number of parameters for the peptide, including the various SEQUEST scores, the mass differential between the observed and calculated mass for the sequence in question, and so forth. third, combined SEQUEST/PeptideProphet output was displayed via the interface INTERACT, a Web-based application that allows the user to view the data as well as sort or filter it according to a range of user-definable parameters [79] (among other things, INTERACT can list all MS/MS scan-file locations with their assigned peptide sequences according to SEQUEST and their corresponding SEQUEST and PeptideProphet score values); finally, ProteinProphet [80] takes the INTERACT data file and derives a list of protein identifications and their corresponding scores from the observed peptide data.

It is not uncommon to find that mass spectrometric data cannot be correlated with any sequence in the database searched. This can happen for the following reasons [58]:

1. The sequence is absent in the database searched: the peptides may be derived from novel proteins or from variants (allelic or strain- or species-derived variants) of known proteins, or there may be errors in the reported sequence. *De novo* sequence analysis of peptides from the MS/MS spectra may be used to determine related proteins using homology-based database search methods (for instance, CIDentify [81], a homology-based sequence database search program).
2. The presence of unexpected co- or posttranslational modifications or chemical modifications (as artifacts of sample handling): FindMod (http://www.expasy.ch/tools/findmod) can be applied for high-throughput determination of protein posttranslational modification from peptide mass fingerprint data [82].
3. The peptide is produced by an unexpected or nonspecific cleavage.
4. The quality of the spectrum is poor or the spectrum originates from a nonpeptide contaminant. Purification and concentration of peptide samples before mass spectrometric analysis can be applied to improve the quality of the spectrum.

3. COMPUTATIONAL STRUCTURAL BIOLOGY

Technical advances have expanded the applicability of existing methods in structural biology and provide a basis for the discovery of general structural principles that underlie all cellular processes [83]. A major goal of computational structural biology is to help reach this goal and to predict structure and the structural basis of the function of biologically related molecules. Of all major classes of biomolecules including proteins, DNA, RNA, carbohydrates, and small molecules with biological activity, protein structures have been mostly studied computationally because of their importance and the variety of structures known. In the rest of this chapter, we will review the current state of three major areas in the structural biology of proteins: prediction of three-dimensional structures from sequence including homology modeling (for a recent review, see Ref. [84]) and fold recognition (recent review [85]); docking of proteins with known structures; and simulation of protein dynamics, with an emphasis on perhaps the most accurate methodology—homology modeling of protein three-dimensional structure.

Comparative (homology) modeling remains the only computational biology method at present that can provide models with a root-mean-square (rms) error lower than 2 Å. Computational methods for protein structure prediction based on related proteins of known

structures were developed more than three decades ago [86]. Later, Greer outlined a basic protocol that is still followed today [87, 88]. Most homology modeling methods consist of four sequential steps [89]. The first step is to identify the proteins with known three-dimensional structures that are related to the target sequence. The second step is to align them with the target sequence and to pick those known structures that will be used as templates. Any corrections in the alignment are made at this stage. The third step is to build the model for the target sequence given its alignment with the template structures. In the fourth step, the model is evaluated using a variety of criteria. If necessary, the alignment and model building are repeated until a satisfactory model is obtained. The main difference between the different comparative modeling methods is how the three-dimensional model is calculated from a given alignment (step 3 above). Because of the importance of step 3, sometimes it is divided into four stages [84]: backbone generation, loop modeling, side-chain modeling, and model optimization.

The level of sequence identity that is critical for the success of the homology modeling can be evaluated by sequence alignment programs such as BLAST, FASTA, and CLUSTALW. The latter program is used for multiple alignment, and the results contain significant amounts of additional information about the structural context that can be used to improve the alignment. It is especially useful, for example, to position deletions or insertions in places in which the sequences are widely divergent.

A widely used method is building a model by rigid-body assembly. The method constructs the model from a few core regions and from loops and side chains, which are obtained from dissected related structures. This assembly involves fitting the rigid bodies on the framework, which is defined as the average of the Cα atoms in the conserved regions of the fold. Another method is based on the approximate positions of conserved atoms from the templates to calculate the coordinates of other atoms. The third group of methods, modeling by satisfaction of spatial restraints, uses either distance geometry or optimization techniques to satisfy spatial restraints obtained from the alignment of the target sequence with homologous templates of known structure. In general, generation of the backbone is not difficult, especially when two or more template structures are available so complementary regions that are accurately determined are used.

The most significant problems in homology modeling remain the prediction of loop structures and side-chain modeling. The two main approaches for loop modeling are based on the use of already determined structures of loops and true *ab initio* prediction by using energy minimization or molecular dynamics techniques. These methods typically work relatively well for relatively short loops containing not more than six to eight residues. Most successful approaches for side-chain modeling are based on the use of libraries of common rotamers extracted from high-resolution X-ray structures. The prediction accuracy is typically much higher for residues in hydrophobic cores than for surface residues. This is mostly because of the flexibility of the side chains at the surface that can adopt multiple conformations.

After building the model, it can be optimized by several methods including iterative approaches. Such approaches are based on the idea of sequential prediction of loops and side-chains orientation, as well as energy minimization. The success of such an approach critically depends on the accuracy of the function describing the energy of the whole molecule. At present, the accuracy is not sufficient for accurate prediction, and typically only few hundred steps of energy minimization are used to avoid accumulation of small errors that can lead to structures completely different than the real one. Another straightforward approach is a molecular dynamics simulation of the model. Again, a major problem is the lack of sufficient accuracy of the force fields used for simulation; more accurate force fields are needed.

The best comparative techniques are able to produce models with good stereochemistry and overall structural accuracy when the modeling alignment is correct. The errors in comparative models depend on the level of sequence similarity between template and target sequences and on the errors in the template structure. They can be divided into five categories: side-chain errors, distortions and rigid-body changes in regions that are aligned correctly (e.g., loops, helices), distortions and rigid-body changes in insertions (e.g., loops), distortions in incorrectly aligned regions (loops and longer segments with low sequence identity to the templates), and incorrect folding resulting from an incorrect choice of a

template. The consequence of these errors is that the comparative method can result in models with a main-chain rms error as low as 1 Å for 90% of the main-chain residues if a sequence is at least 40% identical to one or more of the templates. In this range of sequence similarity, the alignment is mostly straightforward to construct, there are not many gaps, and structural differences between the proteins are usually limited to loops and side chains. When sequence identity is between 30% and 40%, the structural differences become larger and the gaps in the alignment are more frequent and longer. As a result, the main-chain rms error rises to ~1.5 Å for about 80% of residues. The rest of the residues are modeled with large errors because the methods generally cannot model structural distortions and rigid-body shifts and cannot recover from misalignments. Insertions longer than about eight residues usually cannot be modeled accurately at this times whereas shorter loops frequently can be modeled successfully.

An example of homology modeling with implications for elucidations of mechanisms of virus entry is described in more detail in the following. Enveloped viruses enter cells by binding their envelope glycoproteins to cell-surface receptors, followed by conformational changes leading to membrane fusion and delivery of the genome in the cytoplasm [90]. By using homology modeling, we have recently analyzed the three-dimensional structure of the angiotensin-converting enzyme 2 (ACE2) [91] that was recently identified as a functional receptor for the SARS virus [92], and its binding site on the SARS-CoVS glycoprotein was localized between amino acid residues 303 and 537 [93]. ACE2 is a homolog of the metalloprotease angiotensin-converting enzyme ACE [94, 95] and was found to be an essential regulator of heart function [96]. ACE exists in two isoforms: somatic ACE, which has two homologous domains each containing an active catalytic site, and testis-specific ACE (tACE), which corresponds to the C domain of somatic ACE and has only one active site. ACE2 has a high level of similarity (sequence identities 43% and 35% and similarities 61% and 55%, respectively) to tACE and the *Drosophila* homolog of ACE (AnCE). Recently, the crystal structures of tACE [97] and the *Drosophila* ACE homologue AnCE [98] have been determined at resolutions of 2.0 and 2.4 Å respectively.

These crystal structures were used as templates to build an accurate (rms deviation [rmsd] less than 0.5 Å) three-dimensional model of ACE2 by comparative (homology) modeling. Based on the ACE2 model, an analysis of the receptor-binding domain (RBD) of the SARS-CoV S glycoprotein, and similarity with other interactions of viral envelope glycoproteins (Env) with receptors [99], we proposed a possible mechanism of the ACE2 function as a receptor for the SARS virus. The analysis of the ACE2 model could also help in the design of experiments to further elucidate the structure and the dual function of ACE2.

The sequences of ACE2, tACE and AnCE, were aligned using the multiple sequence alignment program CLUSTALW [100]. The comparative modeling procedure COMPOSER [101, 102] implemented in SYBYL6.9 (Tripos Inc., St. Louis, MO) was used to build a three-dimensional model of the ACE2 structure. We used the tACE and AnCE structures to find out topologically equivalent residues based on structural alignment, and the structurally conserved regions (SCRs) were modeled. The structurally variable regions (loops) were modeled by using loops either from the corresponding location of the homologous protein or from the general protein database. The three-dimensional model of ACE2 was then subjected to energy minimization by using standard Tripos force fields and was finally validated with the PROCHECK program [103]. The coordinates were deposited to the protein data bank (PDB) (code: 1RIX).

The sequences of both ACE2 and the S RBD (spike protein of SARS receptor-binding domain) were scanned against the PROSITE [104] motifs to locate potential glycosylation sites. Six *N*-glycosylation sites with high probability of occurrences on ACE2 were predicted by PROSITE. Fully surface-exposed asparagine (*N*) residues were found at five of these sites, which were modeled by attaching *N*-acetylglucosamine moieties. Three *N*-glycosylation sites were found in the S RBD fragment and were modeled similarly. The areas of solvent accessibility (ASAs) were calculated with probe radius 1.4 Å by using the Lee and Richards's algorithm [105]. Electrostatistic potentials were calculated by using the program GRASP [106] with the following parameters: a protein dielectric constant of 2.0, a solvent dielectric constant of 80, an ion exclusion radius of 2.0 Å, a probe radius of 1.4 Å, and an ionic

strength of 0.14 M. The calculated potentials were displayed at the solvent-accessible surface. The visualization of solvent accessibility, superpositioning of molecules, and calculation of surface hydrophobicity were performed by using InsightII. The hydrophobicity of the surface residues was calculated according to the Kyte–Doolittle method [107] with a window size of 5 amino acids and hydrophobic and hydrophilic levels of 0.7 and −2.4, respectively.

To begin to understand the interactions between the SARS-CoV S glycoprotein and its recently identified receptor ACE2, we attempted to develop an accurate model of the ACE2 three-dimensional structure. We found two proteins, tACE and AnCE, with available high-resolution crystal structures and ACE2 sequence identities of 43% and 35% (sequence similarities are 61% and 55%), respectively; the sequence alignment of ACE2 with tACE and AnCE2 is shown in Fig. 3. Therefore, homology modeling was used to build an accurate three-dimensional model of ACE2.

The architecture of the ACE2 model is very similar to the crystal structure of tACE (Fig. 4A). The superposition of the ACE2 model structure with the template structures of tACE and AnCE (Fig. 4B) shows very small deviation (rmsd less than 0.5 Å). A major feature of the ACE2 structure (and the template structures) is a deep channel on the top of the molecule that contains the catalytic site (Fig. 5A). A comprehensive analysis of the structure and function of the catalytic site was very recently reported after our model was completed [108]; here we will not discuss the enzymatic function of ACE2 but rather use the enzymatic site location for reference purposes. The channel is surrounded by ridges containing loops, helices and a portion of a β sheet. The long loop between N210 and Q221 that is missing in tACE and AnCE (Fig. 3) is on the ACE2 surface (Fig. 4B); note that the orientation of ACE2 in Fig. 4A is different than in Fig. 4B show this loop. Potential *N*-glycosylation sites were identified at six positions, 53, 90, 103, 322, 432, and 546, but only two of them (53 and 90) were aligned with the tACE structure (Fig. 3). They shared the pattern NXTX (except 103) and were modeled with a *N*-acetylglucosamine moiety (Fig. 5B). The direction of the main chain is illustrated in Fig. 5C.

Interactions of viral attachment proteins with protein receptor molecules are mostly determined by complementarity in surface charge distribution, hydrophobic interactions, and geometry; typically, carbohydrates are excluded from the binding sites [99]. In an attempt to provide a working hypothesis for possible regions involved in the interaction of the S glycoprotein with its receptor, we analyzed the ACE2 surface potential, solvent accessibility,

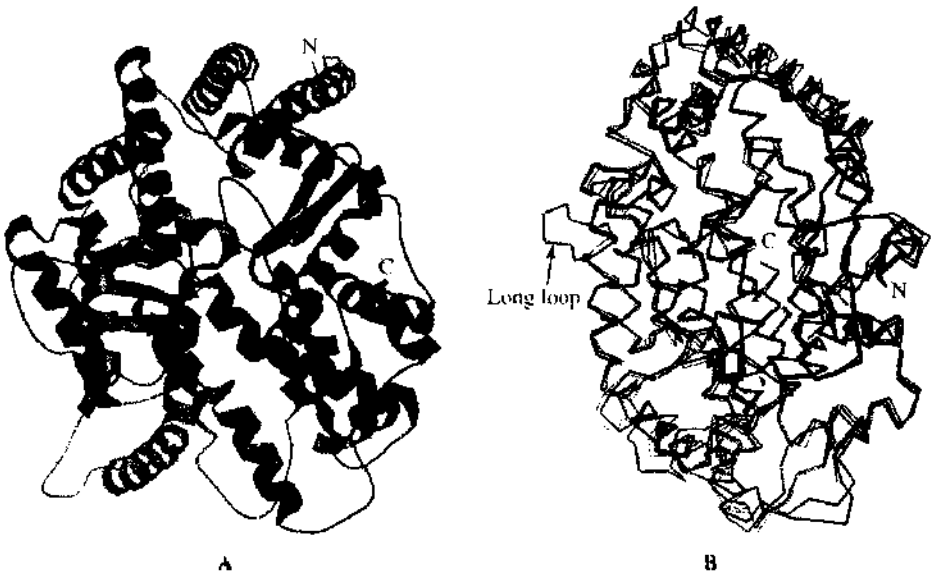


Figure 4. Model of the angiotensin-converting enzyme 2 (ACE2) structure. (A) A ribbon representation of the ACE2 model. The N and C termini are indicated. (B) Superposition of the ACE2 model structure with the crystal structures of testis-specific ACE (tACE) and the Drosophila homolog of ACE (AnCE) based on the Cα-atoms of ACE2, tACE, and AnCE (ACE2, dark gray; tACE, light gray; and AnCE, black). The long loop inserted between N210 and Q221 that is unique for ACE2 is indicated.

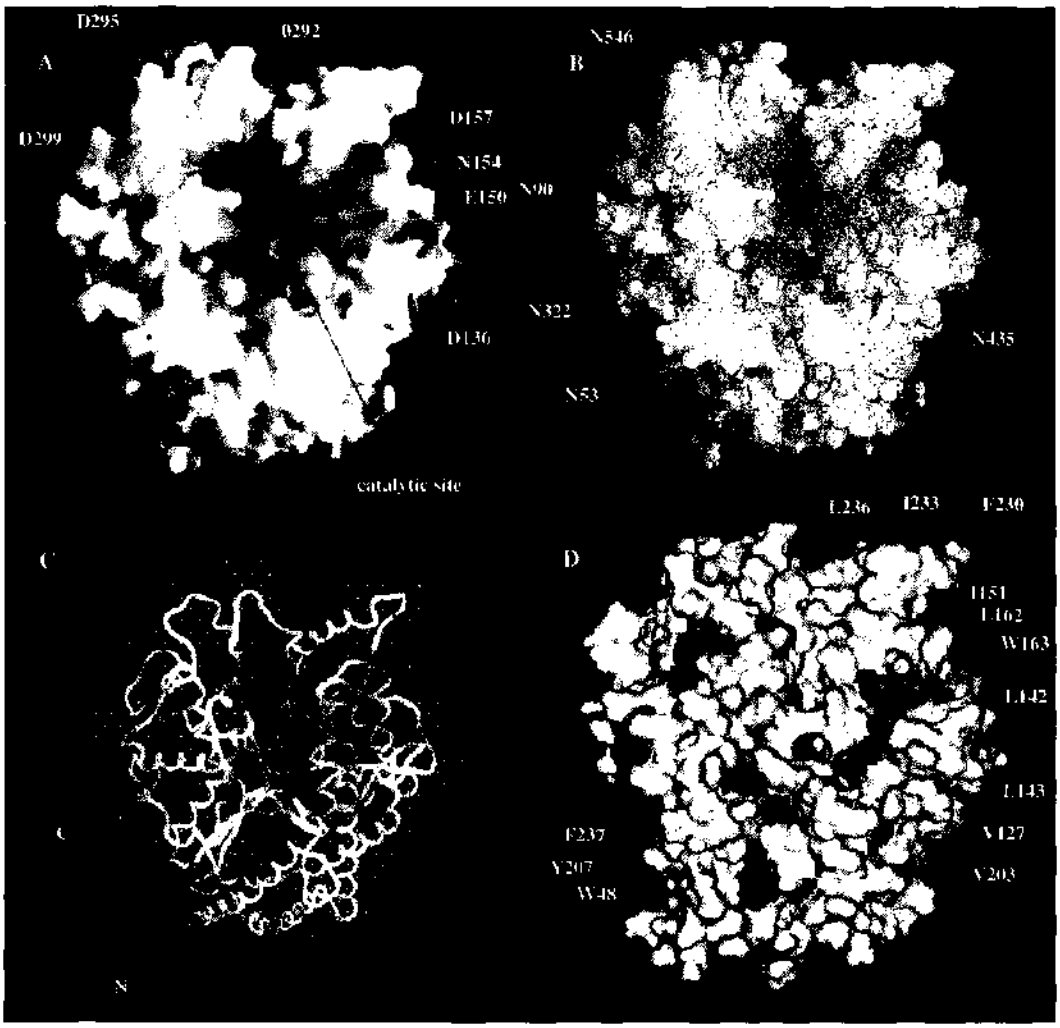


Figure 5. Analysis of the angiotensin-converting enzyme 2 model structure. (A) SARS S protein. receptor-binding domain model (fragment of amino acids 300- 500 in S1 subunit). Ribbon diagram illustrating the receptor-binding domain of SARS S protein model (aa 300–499).

hydrophobicity, and carbohydrate distribution. The surface of the deep channel at the top of ACE2 and the surrounding ridges is highly negatively charged (Fig. 5A). These ridges contain residues D136, E150, N154, D157, D292, D295, and D299, some of which have large ASA values (e.g., D136, N154, and D157 have values of 109, 108, and 80 Å^2, respectively; Fig. 6). Comparison of these residues with the corresponding residues from ACE that do not support fusion mediated by the S glycoprotein [92], and mouse ACE2 that binds to S but with somewhat lower affinity than human ACE2 (M. Farzan, personal communication) (Fig. 7), support the possibility that some of these residues contribute to specific binding. The hydrophobicity analysis revealed distinct hydrophobic patches in close proximity to the negatively charged ridges (Fig. 5D). There are at least three hydrophobic regions, comprising different residues including Phe, Trp, and Tyr, that could contribute to binding, in addition to the charged binding surface. All carbohydrate sites are topologically separate from the electronegative surface at the top of the molecule (Fig. 5B).

The sequence similarity of the S glycoprotein from the SARS virus with S glycoproteins from other coronoviruses or other proteins whose structures are available in the PDB is about 20% or lower. The sequence similarity of the attachment glycoprotein (S1) from the SARS-CoV to other coronanvirus S1 glycoproteins or other proteins with known three-dimensional structures is even lower. Such low sequence similarity does not allow accurate homology modeling. Because of the absence of significant sequence similarity, we built a model by threading (data not shown) a fragment (amino acid residues 300–537) containing

LYS 26 155.75	PRO 154 111.27	SER 291 57.13	SER 429 89.34
THR 27 106.68	GLN 155 [illegible]	PRO 295 [illegible]	GLY 432 48.12
ASP 30 90.32	GLU 141 [illegible]	ILE 259 147.19	PRO 42[illegible] 120 83
LYS 31 149.51	PRO 146 62.39	SER 260 98.10	ASP 433 102 68
HIS 34 103.46	**GLU 150 82.31**	GLU 2[illegible] 43.[illegible]	ILE 428 152 59
GLU 35 67.72	**ASN 154 107.65**	GLN 287 160.31	GLN 429 62.11
GLN 42 88.04	LEU 156 [illegible]	PRO 289 123.13	GLU 430 117 90
ASN 49 72.35	**ASP 157 79.60**	ASN 290 94.[illegible]	ASP 431 138 66
GLU 56 152.15	ASN 159 97.65	**ASP 292 83.20**	GLU 435 [illegible]4.17
GLU 57 139.26	GLU 160 83.50	THR 294 52.[illegible]	THR 434 58.11
GLN 60 128.47	GLU 171 87.60	**ASP 295 54.90**	ILE 436 80.10
ASN 61 79.64	GLN 173 91.65	VAL 297 87.15	LEU 437 71.32
ASN 64 90.94	PRO 174 99.64	**ASP 299 95.50**	GLY 466 [illegible]1.7
ASP 67 53.92	VAL 185 70.32	GLN 300 96.[illegible]	PRO 469 71.13
LYS 68 109.99	ARG 192 [illegible]31.40	ALA 301 90.[illegible]	LYS 470 122 67
ALA 71 60.43	HIS 195 [illegible]68.69	ASP 304 76.71	ASP 471 119 93
GLU 75 103.10	GLU 197 [illegible]14.87	GLN 306 97.61	GLN 472 -9.35
THR 78 87.46	ASP 200 71.42	ARG 306 96.24	ASP 473 132.50
LEU 79 73.79	GLU 208 [illegible]46.30	LYS 309 81.89	GLU 479 41.38
GLN 81 110.23	VAL 209 76.52	LYS 313 139.52	ARG 482 104.76
MET 82 127.25	ASN 210 167.[illegible]	GLY 309 [illegible]1.7[illegible]	GLU 483 117.49
PRO 84 74.73	TYR 211 [illegible]43.[illegible]	ASN 322 [illegible]1.6[illegible]	PRO 492 [illegible]3.80
GLN 86 139.19	SER 216 84.84	THR 324 74.83	ASP 494 84.35
GLU 87 128.66	ARG 219 [illegible]67.[illegible]	HIS 325 120.91	GLU 495 [illegible]35.47
GLN 90 [illegible]49.[illegible]	GLN 221 [illegible]9.65	GLU 329 [illegible]24.13	THR 496 45.33
ASN 98 69.13	GLU 224 [illegible]2.60	ASN 330 66.41	TYR 510 83.32
LEU 91 131.77	ASP 225 [illegible]5.61	THR 334 [illegible]5.[illegible]	GLN 524 [illegible]01.90
THR 92 69.98	HIS 228 88.04	ASP 335 76.82	GLU 527 [illegible]5.03
LEU 95 83.29	GLU 231 [illegible]1.05	GLY 337 84.09	ALA 528 47.83
GLN 98 93.66	GLU 232 [illegible]1.[illegible]4[illegible]	ASN 338 [illegible]16.[illegible]	GLN 531 [illegible]38.22
SER 105 [illegible]5.43	LYS [illegible]34 [illegible]1.[illegible]	VAL 339 [illegible]60.62	LYS 534 [illegible]57.11
SER 109 72.67	PRO 235 74.61	LYS 355 [illegible]09.[illegible]	GLU 536 [illegible]37.88
GLU 110 129.44	ASP 236 [illegible]4.[illegible]	ASP 367 76.[illegible]	GLY 537 [illegible]1.61
ASP 111 97.02	GLN 238 84.20	ALA 385 [illegible]8.[illegible]	PRO 538 [illegible]7.51
LYS 114 116.29	HIS 239 [illegible]27.[illegible]9	ALA 387 [illegible]6.[illegible]1	LEU 539 [illegible]5.65
ARG 115 111.65	ALA 242 [illegible]6.[illegible]6	GLN 388 [illegible]7.[illegible]6	LYS 541 [illegible]3.91
THR 118 54.89	ALA 248 [illegible]5.[illegible]1	ASN 394 96.46	ASN 546 96.10
THR 129 [illegible]3.46	ASN 250 [illegible]17.69	ALA 396 81.36	THR 548 84.73
PRO 135 145.59	PRO 253 91.48	PRO 413 49.77	GLU 549 96.53
ASP 136 108.67	SER 254 [illegible]0.[illegible]	LYS 416 [illegible]49.[illegible]3	GLN 552 114.05
ASN 13[illegible] 58.31	TYR 255 141.94	LYS 419 93.[illegible]6	LYS 553 118.29

Figure 6. Solvent-accessible surface areas (right column, in Angstroms squared) for angiotensin-converting enzyme 2 amino acid residues that are significantly exposed to solvent at the surface of the molecule. The cutoff for significant surface exposure here is assumed to be 45% ratio value, defined as the ratio of side-chain surface area to a "random coil" value per residue in the tripeptide Gly–X–Gly. The middle column represents the amino acid residue number.

the S RBD that we have recently identified [93]. The electrostatic analysis of the model revealed mostly positive charges on the surface and, in particular, an electronegative loop containing residues K439, R441, R444, H445, and K447. The hydrophobic analysis indicated several patches of hydrophobic residues around the positively charged loop region. One must note that although the size of the fragment is relatively small, the S RBD modeling has limitations in the absence of a template structure or structures with high sequence identity. Thus, the RBD model could significantly deviate or even be completely different from the real structure. In this aspect, modeling of the much larger S1 and S2 units is even less reliable. For example, a recent model [109] of S1 and S2 proposed putative-receptor (thought to be CD13) binding regions located in S2 instead of S1, where RBDs of coronaviruses should be. This is why we used our RBD model mostly as an illustration of possible complementary charged surfaces, hydrophobic patches, and β sheets and complemented it with an analysis of the secondary structure of the RBD fragment that also revealed the predominance of β

```
                155  167  171  174  300 303  307
Human ACE       ASN-ASN-THR-LSY-ASP-GLU-LYS
                     .    .    .    |   :    .
Human ACE2      ASP-GLU-ASN-ASP-ASP-ASP-ASP
                     .    |    .    |   |    :
Mouse ACE2      LYS-GLU-THR-ASP-ASP-ASP-ASN
                136  130  154  157  292 295  299
```

Figure 7. Conservation of amino acid residues in human angiotensin-converting enzyme (ACE), human ACE2, and mouse ACE2 that could contribute to interactions with the S glycoprotein. Identities are marked by a pipe (|), highly conservative replacements by a colon (:), and replacements with lower scores by a dot (.). The numbers denote the amino acid residue positions in the sequence. Note that the similarity of these ACE2 residues with the corresponding residues of mouse ACE2 is much higher than with the respective human ACE residues.

sheets (data not shown). In progress are our experiments for the SARS-CoV S glycoprotein RBD crystallization and determination of its three-dimensional structure.

Typically, virus receptors contain ridges that bind to cavities or to structures containing loops, cavities, and channels in the proteins mediating entry [99]. The model structure of ACE2 indicates that some or most of the ridges surrounding the cavity at the top of the molecule (Fig. 8) could serve as a likely binding region for the S glycoprotein for the following reasons. First, the top of the molecule is far away of the membrane and is likely to be easier to reach than membrane proximal regions. Second, protruding structures are likely to be used for binding by viral proteins; they will also ensure geometric complementarity with concave surfaces, as the S RBD domain could be based on our illustrative model (Fig. 8). Third, the negative charges of the ridges complement the positive charges of the RBD. Fourth, the hydrophobic patches around the charges could contribute to binding. Finally, the lack of carbohydrates at the top of the molecule could ensure high-affinity binding. Experiments currently in progress will determine the specific amino acid residues and their relative contribution to the interaction of ACE2 with the S glycoprotein. The ACE2 model developed here, and this proposition of binding regions, could help in the design and analysis of the experimental data and of the virus-binding function of ACE2.

4. COMPUTATIONAL CELL BIOLOGY

4.1. Introduction

As noted on the official Web site of the First International Symposium on Computational Cell Biology [http://caboy.uchc.edu/conference/], "The formulation of hypotheses based on complex experimental data is often impossible without the construction of computational models. Computational cell biology is an emerging discipline that responds to the need for computational methods to analyze and organize the abundance of experimental data on the structure and function of the cell."

Historically [110], mathematical biology has had limited success, turning in time into a somewhat abstract discipline. Several examples of early success in mathematical modeling in biology was demonstrated in following works: the Lotka–Volterra predator–prey model in ecology [111, 112], Hodgkin–Huxley's model of nerve conduction [113], Manfred Eigen's

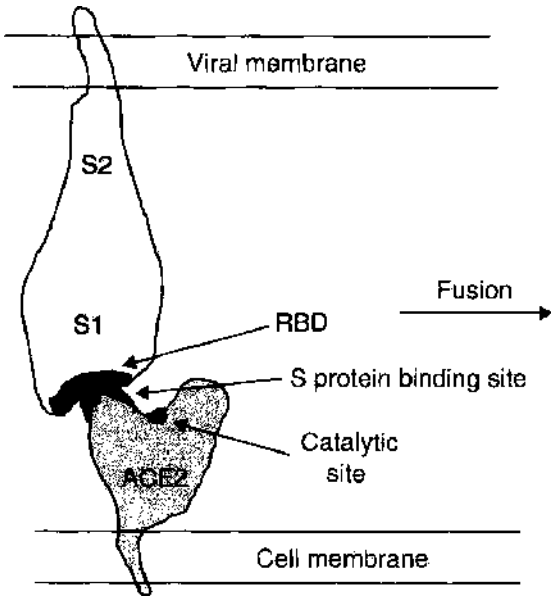


Figure 8. Schematic representation of the interaction between angiotensin-converting enzyme 2 (ACE2) and the SARS-CoV S glycoprotein receptor-binding domain (RBD) leading to binding and fusion. The RBD is depicted as a surface containing a cavity that binds a ridge close to the deep channel containing the catalytic site.

theory of molecular evolution [114], and the Gierer–Meinhardt theory of biological pattern formation [115]. Extensions of these models were considered later by many biologists as a mathematical refinement with limited practical utility (with some exceptions for the mathematical models in neurobiology and cardiology). Theoretical and experimental biology have grown largely apart, and the reason is that biology can rarely present simple experimental system for the evaluation of theoretical models. Biological systems are complex and open, and various factors can change the behavior of the system, so even simple computational modeling requires continuous interaction between model building and experimental verification. The interaction between theoretical models and experimental design can be represented schematically, as shown in Fig. 9 (modified from more detailed diagrams [110, 116]).

The complexity of biological objects can be defined as a "large number of functionally diverse, and frequently multifunctional, sets of elements which interact selectively and nonlinearly to produce coherent rather then complex behavior" [117]. Biological events occurring at various levels (such as organism, tissue, cell, and molecular), and the complexity of the system being modeled, lead to the need for integration of the different models. For instance, for modeling transduction of the activation signal into a cell, one may need to include gene regulatory network, models of proteins pathways, models of membrane, diffusion of molecules and ions into cell, and so forth. Some of these models may be available for investigators, but they are most likely different in format, programming language, and computing platforms, so one may need to develop the tools for unification of the models and of communication between them. At present, there are two ongoing projects for introducing standards in the model communication: System Biology Markup Language (http://www.cds.caltech.edu/erato/sbml/docs) and CellML (http://www.cellml.org)

The next challenge related to the complex nature of biological systems is the interpretation of the results of modeling, model prediction, and finding how the mode should be changed with respect to new experimental data. For a model consisting of hundreds of equations and parameters (even for smaller numbers), the problems such as fitting to the set of experimental data, investigating and optimizing the model behavior, and correcting the model design with respect to the experimental data are not trivial. One of the possible ways to solve this problem is to identify the semiautonomous functional units in the model with known parameters and system behavior. So building a complex model or estimating model parameters can be accomplished in a stepwise or module-based manner. As an example, the modeling of activation of MAP kinase through receptor tyrosine kinases and GTP-binding protein RAS can include the modeling of three modules: RAS activation through signaling from the epidermal growth factor (EGF) receptor, recruitment of adaptor proteins Shc and Grb2, and activation of the exchange factor Sos; cascade of proteins kinases from Raf1 to MAP; and modulation of Raf1 by an inhibitory phosphorilation event by protein kinase A (PKA). All three modules have been modeled [118, 119] and can be combined in one system. Another example of using the model-based approach is represented in [120], where a simple model of T-lymphocyte proliferation was combined with the model of viral hepatitis B, and then the final, rather simple, model, consisting of only 17 equations and about 70 parameters, was fitted to the data describing the so-called "generalized picture" of hepatitis B [121].

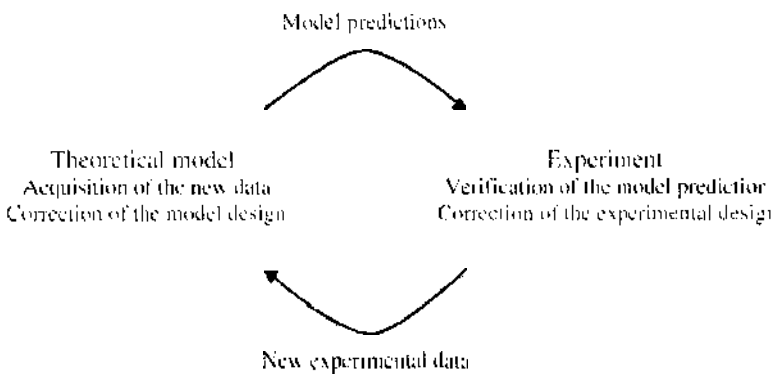


Figure 9. Interaction between theoretical model and experiment.

In this work, three functional units corresponding to cell proliferation of helper and effector T-lymphocytes and production and consumption of soluble factors interleukin 1 and interleukin 2 were used, and the model was fitted to the data in three steps. By analogy to how biologists generally deal with the complexity of biological systems, this modeling approach is based on a hierarchical view of the biological object, that is composed of functional units with specific inputs and outputs. These units can communicate with other modules or units and can themselves be composed of functional units.

Additional problems that arise in computational biology are the optimization of the computational tools, improving the efficiency of existing algorithms, and creating new efficient packages. Complex computational problems require better algorithm development and powerful computers. It should be noted that it is not entirely a mathematical or computer problem, and in some cases, reconsideration of the underlying assumptions of the model and reformulation of these assumptions (that will lead to the changing in model equations) may provide significant improvement of the algorithm.

Another consistently raising (in the direction from experimental data toward theoretical model) issue, mentioned in Ref. [122] is the need for cell biologists to produce more quantitative results about biological processes. Some of the data are often difficult to obtain, and they have been insufficiently valued by biomedical researchers.

4.2. Computational Modeling for Cell Biology

There are five main phases of information processing and problem solving in most of the bioinformatics systems (as in a detailed description of the interaction between theoretical model and experimental design described above).

1. Data collection: Collecting biological samples and processing them, and primary processing of the data (quantification of the experimental results, normalization, statistical analysis of the data).
2. Analysis and extraction of the model features: Defining which features (parameters, variables, modules, etc.) are more relevant and which, therefore, should be used when creating a model for a particular problem (e.g., classification, prediction, decision making), making an assumptions model, and choosing the tools and algorithms for modeling.
3. Modeling the problem: Defining inputs, outputs, and type of model (e.g., probabilistic, rule-based, connectionist), training the model, and statistical verification.
4. Knowledge discovery *in silico*: Making calculation experiments, fitting model to the data, and gaining new knowledge through the analysis of the modeling results and the model itself.
5. Verifying the discovered knowledge *in vitro* and *in vivo*: Making biological experiments in both laboratory and in real life to confirm the discovered knowledge or predicted model behavior. Planning new experiments, changing experimental design, and collecting a new data set (go to the first phase above).

It is not uncommon in bioinformatics to find that models are characterized by small data sets (100 or fewer samples); static data sets, in which data do not change in time once the set was used to create a model; and no need for online adaptation and training on new data. For these tasks, the traditional statistical and artificial intelligence (AI) techniques are well suited. The traditional, offline modeling methods assume that data is static and that no new data are going to be added to the model. Before a model is created, data are analyzed and relevant features are selected again in an offline mode. The offline mode usually requires many iterations of data propagation for estimating the model parameters. Such methods for data analysis and feature extraction use principle component analysis (PCA), correlation analysis, offline clustering techniques (such as K-means, fuzzy C-means, etc.), self-organizing maps (SOMs), and many more techniques. Many modeling techniques are applicable for

these tasks; for example, statistical techniques (regression analysis, support vector machines), AI techniques (decision trees, hidden Markov models, finite automata), and neural-network techniques (MLP, learning vector quantization, fuzzy neural networks).

Some of the modeling techniques allow for extracting knowledge (e.g., rules from the models) that can be used for explanation or knowledge discovery. Such models are the decision trees and the knowledge-based neural networks (KBNNs) [123].

Some of the tasks for data analysis and modeling in bioinformatics are characterized by large data sets, which are updated regularly, a need for online learning and adaptation and online new model creation from input data streams changing with time.

Knowledge adaptation based on a continuous stream of new data.

When creating models of complex processes in molecular biology, the following issues must be considered: how to model complex interactions between genes and proteins—between the human genome and the environment; modeling both stability and repetitiveness, because genes are relatively stable carriers of information; and dealing with uncertainty (when modeling gene expressions, there are many sources of uncertainty; among them are alternative splicing [generation of different mRNA isoforms from a single transcript] and mutation in genes caused by ionizing radiation, chemical contamination, replication errors, viruses that insert genes into host cells, etc). Mutated genes are expressed differently and may cause the production of different proteins. For large data sets and for continuously incoming data streams that require the model and the system to rapidly adapt to new data, it is more appropriate to use online, knowledge-based techniques and Evolving Connectionist System (ECOS) in particular, as it will be demonstrated below.

4.3. Microarray Gene Expression Data Analysis and Disease Profiling

One of the contemporary directions while searching for efficient drugs for many terminal illnesses, such as cancer or HIV, is the creation of gene profiles of these diseases and the subsequent finding of targets for treatment through gene-expression regulation. A gene profile is a pattern of expression of a number of genes that is typical for all, or for some, of the known samples of a particular disease. A disease profile would look like the following.

IF	(gene A is highly expressed)	AND
	...	
	(gene B is low expressed)	AND
	(gene C is very highly expressed)	
THEN	most probably this is cancer of type N	

Having such profiles for a particular disease makes it possible to set early diagnostic tests, so that a sample can be taken from a patient, the data related to the sample can be processed, and a profile obtained. This profile can be matched against existing gene profiles, and based on similarity, it can be predicted with certain probability whether the patient is in an early phase of a disease or whether he or she is at risk of developing the disease in the future with certain probability.

A methodology that consists of training an evolving system and extracting rules, which are presented as disease profiles, is illustrated schematically in Fig. 10. Each profile is a rule extracted from a trained ECOS, which on the figure is shown using colors: The higher the level of a gene expression, the brighter the color. Five profiles are visualized in Fig. 10. The first three represent a group of samples of class 1 (disease), the second two represent class 2 (normal cases). Each column in the condition part of the rules (profiles) represents the expression of one gene out of the 100 relevant genes used in this example.

4.3.1. Gene-Expression Data: A Biological Perspective

Microarray equipment is used widely at present to evaluate the level of gene expression [124]. Each point (pixel, cell) in a microarray represents the level of expression of a single gene. Five principal steps in the microarray technology are shown in Fig. 11. They are tissue collection, RNA extraction, microarray gene-expression calculation, scanning and image processing, and data analysis.

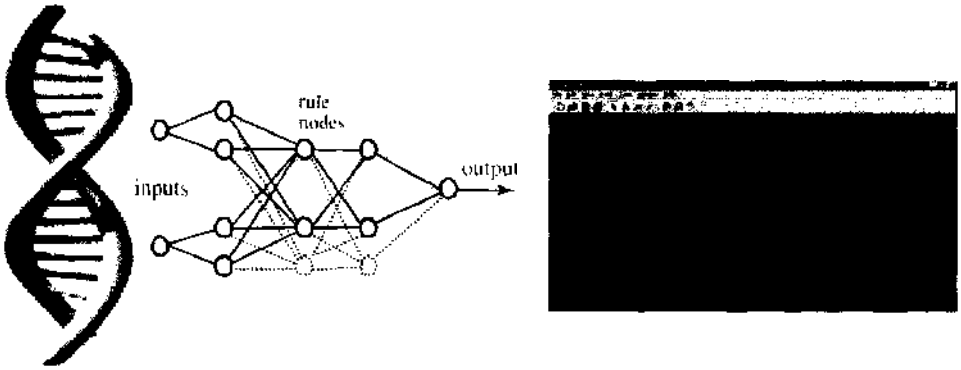


Figure 10. The pathway from DNA to disease profiling: DNA is used to obtain RNA microarray gene expression data; this data is used to train a model (a knowledge-based neural network in this case); profiles (rules) are extracted from the trained model—each rule, presented as one row, represents a gene expression pattern (the "IF" part of the rule) and the class (cancer or normal) associated with this pattern (the "THEN" part of the rule). Reprinted with permission from [15], N. Kasabov, "Evolving Connectionist Systems—Methods and Applications in Bioinformatics, Brain Study and Intelligent Machines," Springer, New York, 2002. © 2002, Springer.

The recent advent of cDNA microarray and genechip technologies means that it is now possible to simultaneously interrogate thousands of genes. The potential applications of this technology are numerous and include identifying markers for classification, diagnosis, disease outcome prediction, therapeutic responsiveness, and target identification. Microarray analysis might not identify unique markers (e.g., a single gene) of clinical utility for a disease because of the heterogeneity of the disease, but a prediction of the biological state of disease is likely to be more sensitive when identifying clusters of gene expression (profiles) [125].

For example, gene-expression clustering has been used to distinguish normal colon samples from tumors from within a 6500-gene set, although clustering according to clinical parameters was not undertaken [126]. Although distinction between normal and tumor tissue can be easily made using microscopy, this analysis represented one of the early attempts to classify biological samples through gene-expression clustering. The above data set is used in this section to extract profiles of colon cancer and normal tissue, using an EFuNN [127]. Another example of profiling is determining the distinction between two subtypes of leukemia; namely, acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) [128].

There are several traditional approaches to analyzing gene regulatory networks: logical or binary, chemical kinetic, stochastic kinetic, and so forth. A detailed discussion of these methods can be found in Refs. [118, 129–135].

Neural networks have already been used to create classification systems based on gene expression data. In Ref. [136], multilayer perceptron neural networks were used to achieve a classification of 93% of Ewings sarcomas, 96% of rhabdomyosarcomas, and 100% of neuroblastomas. From within a set of 6567 genes, 96 genes were used as variables in the classification system. Whether these results would be different using different classification methods needs further exploration.

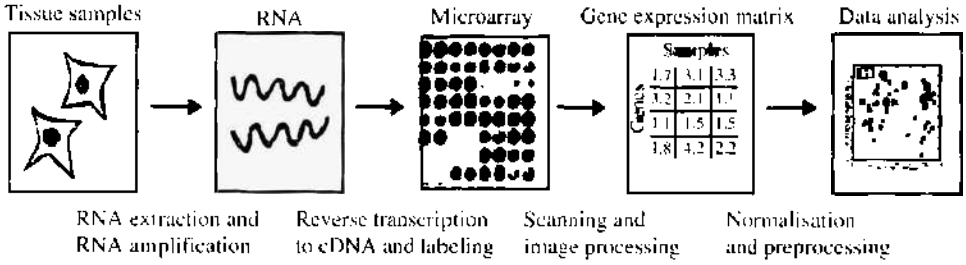


Figure 11. The pathway of the microarray technology: from DNA data to data analysis results.

A methodology for profiling of gene expression data from microarrays is described in Ref. [15]. It consists of the following phases:

1. Microarray data preprocessing: This phase aims at eliminating the low expressed genes, or genes that are not expressed sufficiently across the classes (e.g., controlled vs. tumor samples, or metastatic vs. nonmetastatic tumors, etc). Log transformation of the data can be applied to reduce the range of gene expression data. An example of how this transformation "squeezes" the gene expression values plotted in the two-dimensional principal components is given in Fig. 12. There are only two samples used (two cell lines) and only 150 genes, out of the 4000 on the microarray, that distinguish these samples.
2. Selecting a set of significant differentially expressed genes across the classes: Usually, the t-test is applied at this stage, with an appropriate threshold [137]
3. Finding subsets of both underexpressed and overexpressed genes from the selected ones in the previous step: Statistical analysis of these subsets is performed.
4. Clustering of the gene sets from step 3 that would reveal preliminary profiles of jointly overexpressed or underexpressed genes across the classes. An example of the hierarchical clustering of 12 microarray vectors (samples), each containing the expression of 50 genes after steps 1 to 3 were applied on the initial 4000-gene-expresson data from the microarrays, is given in Fig. 13 [(a) samples in two-dimensional Sanmon's projection space of the 50-dimensional gene expression space; (b) the similarity between the samples (columns), based on the 50 selected genes, and the similarity between the genes (rows) based on their expression in the 12 samples].
5. Building a classification model and extracting rules that define the profiles for each class: The rules would represent the fine grades of the common expression level of groups of genes. Through using thresholds, smaller or larger groups of genes can be selected from the profile.
6. Further training of the model on new data and updating the profiles: With the arrival of new labeled data (samples), the model needs to be updated (e.g., traned on additional data) and possibly have modified rules (profiles) extracted.

Two data sets are used here to illustrate the above methodology, which explores evolving systems for microarray data analysis.

4.3.2. Case Study: Gene Profiling of Two Classes of Leukemia

A data set of 72 classification examples for leukemia is used that consists of two classes and a large input space—the expression values of 6817 genes monitored by Affymetrix arrays [128]. The two types of leukemia are AML and ALL. The latter type can be subdivided further into

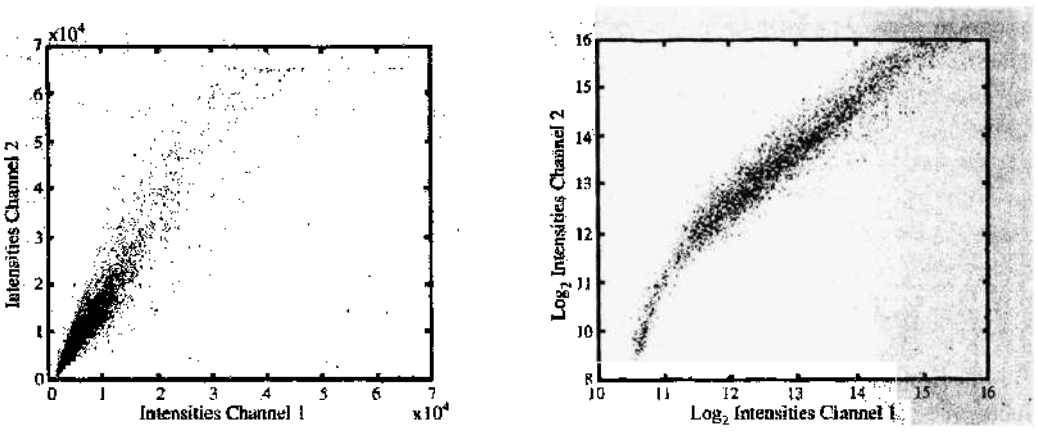


Figure 12. An illustration of the importance of the preprocessing phase in microarray gene expression data analysis on two channel expression measurements taken from the same tissue. Each spot is the expression of a gene. It can be seen from the figures on the left and on the right, respectively, that using log transformation of the values makes the values from the two channels more similar to each other (closer to the desirable diagonal perfect match line).

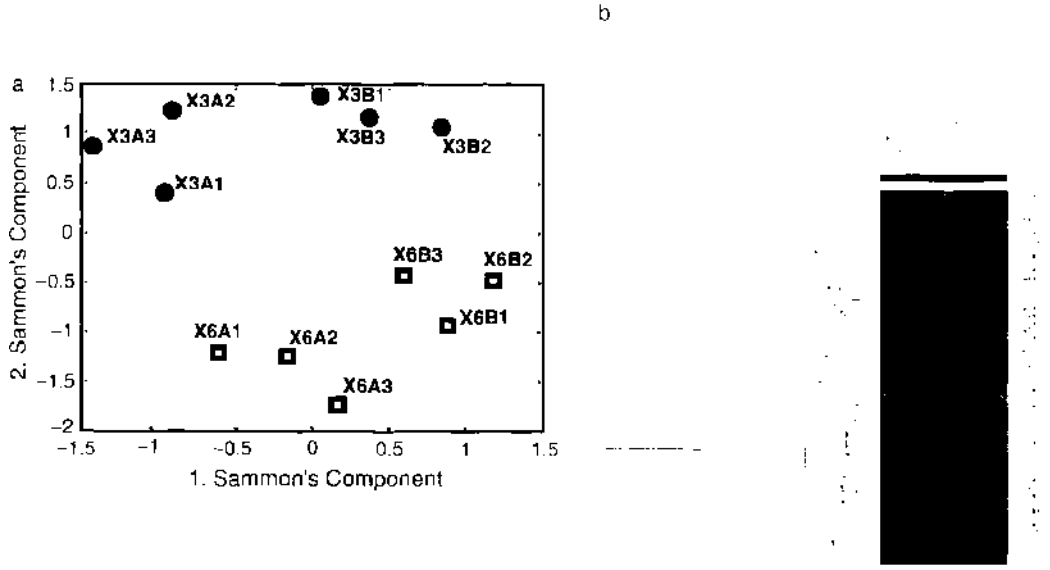


Figure 13. (a) Sammon's projection of 50 gene-expression variables of 12 gene-expression vectors (samples, taken from 12 tissue samples). (b) Hierarchical clustering of the data. The rows are labeled by the gene names, and the columns represent different samples. The lines link similar items (similarity is measured as correlation) in a hierarchical fashion.

T-cell and B-cell lineage classes. In Ref. [128], the data set was split into 38 cases (27 ALL and 11 AML) for training and 34 cases (20 ALL and 14 AML) for validation of a classifier system. The test set shows a higher heterogeneity with regard to tissue and age of patients, making any classification more difficult. So, the tasks are to, find a set of genes distinguishing ALL and AML, construct a classifier based on these data, and find a gene profile of each of the classes.

After having applied points 1 and 2, 100 genes were selected. A preliminary analysis on the separability of the two classes can be done through plotting the 72 samples in the two-dimensional PCA space. PCA consists of a linear transformation from the original set of variables (100 genes) to a new (smaller, two-dimensional) set of orthogonal variables (principal components) so that the variance of the data is maximal and ordered according to the principal components (see Fig. 14a).

The extracted rules for each class make up a profile of this class. One way of visually representing of these profiles is illustrated in Fig. 14b, where rules were extracted from a trained EFuNN with 100 genes.

To choose the model for gene-profiling and classification tasks, the gene-profiling task requires that the model meets the following requirements: it can be continuously trained on

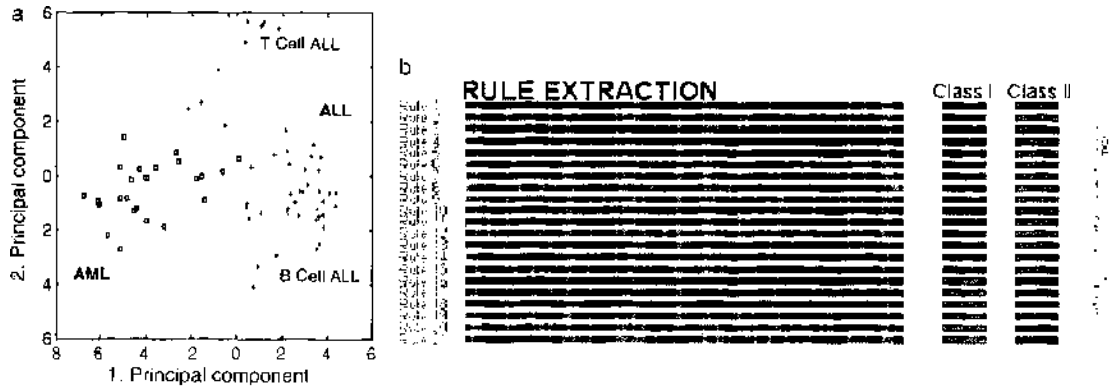


Figure 14. (a) The first two principal components of the leukemia 100 genes selected after a t-test is applied. (b) Some of the rules extracted from an Evolving Fuzzy Neural Network trained on the leukemia data and visualized as profile patterns (see Ref. [15]). AML, acute myeloid leukemia; ALL, acute lymphoblastic leukemia.

new data, the knowledge-based model is extracted (knowledge in the form of profiles), and the model gives an evaluation for the validity of the profiles.

4.4. Clustering the Time-Course Gene-Expression Data

The expression of genes in cell changes with time. Measuring the expression rate of each gene over time gives the gene a temporal profile of its expression level. Genes can be grouped together according to their similarity of temporal expression profiles.

This is illustrated here using case study data. To demonstrate the method, we used yeast gene-expression data that are available in public databases. We analyzed the gene expression during the mitotic cell cycle of different synchronized cultures as reported by [138] and by [139]. The data sets consisted of expression profiles for over 6100 open reading frames (ORFs).

In this study we did not reduce the original data set by applying a filter in form of a minimum variance. This leads to a higher number of clusters of weakly regulated genes, however, it diminished the possibility of missing co-regulated genes during the clustering.

To find upstream regulatory sequences we used Hughes' compiled set of upstream regions for the ORFs in yeast (Church lab, http://atlas.med.harvard.edu/).

One of the reasons for cluster analysis of time-course gene-expression data is to infer the function of novel genes by grouping them with genes of well-known functionality. This is based on the observation that genes that show similar activity patterns over time (coexpressed genes) are often functionally related and controlled by the same mechanisms of regulation (coregulated genes). The gene clusters generated by cluster analysis often relate to certain functions (e.g., DNA replication or protein synthesis). If a novel gene of unknown function falls into such a cluster, it is likely that this gene serves the same function as the other members of this cluster. This "guilt-by-association" method makes it possible to assign functions to a large number of novel genes by finding groups of coexpressed genes across a microarray experiment [140].

Different clustering algorithms have been applied to the analysis of time-course gene-expression data; k-means, SOM, and hierarchical clustering, to name just a few [140]. They all assign genes to clusters based on the similarity of their activity patterns. Genes with similar activity patterns should be grouped together, whereas genes with different activation patterns should be placed in distinct clusters. The cluster methods used so far have been restricted to a one-to-one mapping: one gene belongs to exactly one cluster. Although this principle seems reasonable in many fields of cluster analysis, it might be too limited for the study of microarray time-course gene-expression data. Genes can participate in different genetic networks and are frequently coordinated by a variety of regulatory mechanisms. For the analysis of microarray data, we may therefore expect that single genes can belong to several clusters. Several researchers have noted that genes were frequently highly correlated with multiple classes and that the definition of clear borders between gene-expression clusters often seemed arbitrary [141]. This is a motivation to use fuzzy clustering to assign single objects to several clusters.

A second reason for applying fuzzy clustering is the large noise component in microarray data resulting from biological and experimental factors. The activity of genes can show large variations under minor changes of the experimental conditions. Numerous steps in the experimental procedure contribute to additional noise and bias. A usual procedure to reduce the noise in microarray data is setting a threshold for a minimum variance of the abundance of a gene. Genes below this threshold are excluded from further analysis. However, the exact value of the threshold remains arbitrary because of the lack of an established error model and the use of filtering as preprocessing.

Because we usually have little information about the data structure in advance, a crucial step in the cluster analysis is selection of the number of clusters. Finding the "correct" number of clusters addresses the issue of cluster validity. This has turned out to be a rather difficult problem, as it depends on the definition of a cluster. Without prior information, a common method is the comparison of partitions resulting from different numbers of clusters. For assessing the validity of the partitions, several cluster validity functionals have been introduced [142]. These functionals should reach an optimum if the correct number of clusters

is chosen. When using evolving clustering techniques, the number of clusters does not need to be defined *a priori*.

In Ref. [15], an ESOM is evolved from the yeast gene temporal profiles used as input vectors. The number of clusters did not need to be specified in advance (Fig. 15).

It can be seen from Fig. 15 that clusters 72 and 70 are represented on the ESOM as neighboring nodes. The ESOM on the figure is plotted as a two-dimensional PCA projection. Cluster 72 has 43 members (genes that have similar temporal profiles), cluster 70 has 61 members, and cluster 5 has only three genes as cluster members. New cluster vectors can be created in an online mode if the distance between existing clusters and the new data vectors are above a chosen threshold.

5. COMPUTATIONAL SYSTEMS BIOLOGY

5.1. Introduction

The aim of computational systems biology is to understand complex biological objects in their entirety (i.e., at system level). It involves the integration of different approaches and tools: computer modeling, large-scale data analysis, and biological experimentation. One of the major challenges of the systems biology is the identification of the logic and dynamics of gene-regulatory and biochemical networks.

In Ref. [143], general systems theory was applied to biology, psychology, economics, and social science. In the view of this work, old-fashioned science tried to explain observable phenomena by reducing them to elementary units that can be investigated independently. However, contemporary science empowers the importance of wholeness, which can be defined as impossibility of understanding the systems of various orders by investigation of their respective parts in isolation. This remains an effective definition of systems biology as practiced today, with the integration and application of mathematics, engineering, physics, and computer science to understanding a range of complex biological regulatory systems.

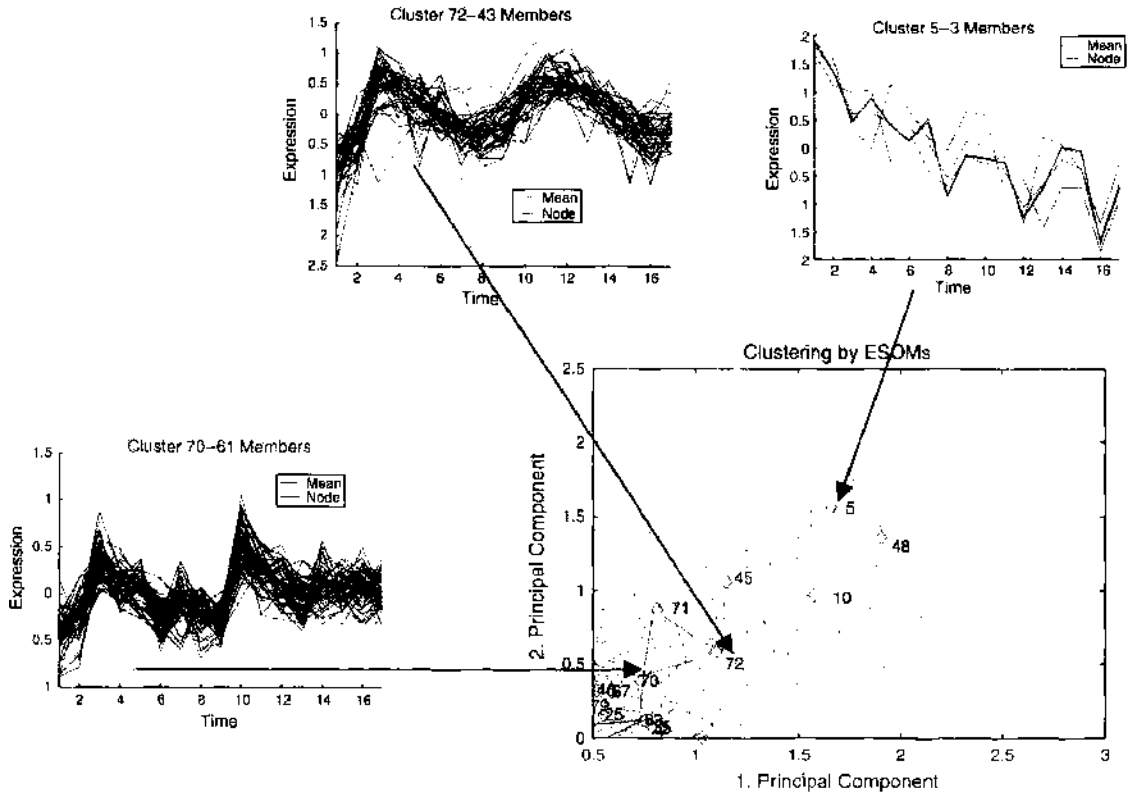


Figure 15. Using a clustering model, an evolving self-organized maps (ESOM) to cluster time-course yeast gene expression data (see Ref. [15]). Genes with similar time-course profiles are gathered in a same cluster.

Computational cell biology is the most immediate beneficiary of the flood o the data that has emerged from the genomics and proteomics that are the main sources of enormously large data sets providing information about genes and proteins interaction. The main goal of systems biology is to integrate all these data in the whole, providing explanation and prediction of behavior of the system under study. The most feasible applicaton of systems biology is to create a detailed model of cell regulation to provide system-level insights into mechanism-based drug discoveries [144–146]. The main subjects of computational systems biology are the following [147]: the structure of the systems, such as genes, metabolism, and signal transduction networks and physical structures; the dynamics of such sysems; methods to control the systems; and methods to design and modify the systems for desied properties.

5.2. System-Level Understanding

System–level understanding is a recurrent theme in biology and has a long hisory [148–150]. The term "system–level understanding" was described in Ref. [147] as the shift of focus in understanding a system's structure and dynamics as a whole, rather than as the particular objects and their interactions. The property of a whole biological system, like a cell, cannot be understood by drawing a diagram of interconnection of the genes and proteins—this can give us only a static picture of the dynamics. System-level understanding of a biological system, however, can be derived from insight into four key properties [117]:

1. System structures: These include the gene regulatory network and biochemical pathways. They can also include the mechanisms of modulation for the phyical properties of intracellular and multicellular structures by interaction.
2. System dynamics: System behavior over time under various conditionscan be understood by identifying essential mechanisms underlying specific behavios and through various approaches depending on the systems nature: metabolic analysis finding a basis of elementary flux modes that describe the dominant reaction pathways within the network), sensitivity analysis (the study of how the variation in the output of a model can be apportioned, qualitatively or quantitatively, to different source of variation), dynamic analysis methods such as phase portrait (geometry of the trajectories of the system in state space), and bifurcation analysis (bifurcation analysis traces time-varying changes in the state of the system in a multidimensional space, where each dimension represents a particular system parameter—concentration of the biochemical factor involved, rate of reactions/interactions, etc). As parameters vary, changes may occur in the qualitative structure of the solutions for certain parameter values.These changes are called bifurcations, and the parameter values are called bifurcationvalues.
3. The control method: Mechanisms that systematically control the state o the cell can be modulated to change system behavior and optimize potential therapeutic effect targets of the treatment.
4. The design method: Strategies to modify and construct biological systems having desired properties can be devised based on definite design principles and simulations, instead of blind trial-and-error.

As mentioned above, in reality, analysis of system dynamics and understanding the system structure are overlapping processes. In some cases, analysis of the syster dynamics can give useful predictions in system structure (new interactions, additional members of the system). Different methods can be used to study the dynamical properties of the system: analysis of steady states allows us to find the systems states when there are no dynamical changes in system components; stability and sensitivity analyses provide insights into how systems behavior changes when stimuli and rate constants are modified to reflect dynamic behavior; bifurcation analysis, in which a dynamic simulator is coupled with analysis tools, can provide a detailed illustration of dynamic behavior [151, 152]; and flux balance analysis [153] can be used to predict the different metabolic patterns, as was done, for instance, in [154] for predicting the switching in of the metabolic pathways in *Escherichia coli* under different nutritional conditions, based on knowledge of only the metabolic network structure. The choice of the analytical methods depends on the availability of the data that can be incorporated into the model, and on the nature of the model.

5.3. Properties of the Complex System

Robustness is a central issue in all complex systems, and it is essential for understanding the biological object functioning at the system level. Robust behavior in biochemical networks was reported a long time ago in Refs. [155, 156], as well as in more recent papers [157–160].

Robustness can be defined as the preservation of particular characteristics despite uncertainty in components or the environment [161]. Robust systems exhibit the following phenomenological properties [117]: adaptation, which denotes the ability to cope with environmental changes; parameter insensitivity, which indicates a system's relative insensitivity (to a certain extent) to specific kinetic parameters; and graceful degradation, which reflects the characteristic slow degradation of a system's functions after damage, rather than catastrophic failure.

These properties correspond to the following properties attained to robust systems in engineering: a form of system control such as negative feedback and feed-forward control; redundancy, whereby multiple components with equivalent functions are introduced for backup; and structural stability, where intrinsic mechanisms are built to promote stability and modularity, and where subsystems are physically or functionally insulated so that failure in one module does not spread to other parts and lead to system-wide catastrophe.

Conducting system-level analysis requires a comprehensive set of experimental data. Comprehensiveness in measurements requires consideration of three aspects [117]: first, factor comprehensiveness, which reflects the number of variables that can be measured at once; second, time-line comprehensiveness, which represents the time frame within which measurements are made; and third, item comprehensiveness, which refers to the simultaneous measurement of multiple items (instance, concentrations, localization, etc).

Some systems may have a property of "spiraling complexity," which means that each module in the system that provides some advantage can lead also to the fragility of the system. To overcome a new threat, it is necessary to build a new module/subsystem that can also lead to the new failure/fragility. This evolution leads to an excess of complexity for the system, which in its turn leads to the robustness of the system.

The main feature of the concepts in evolutionary biology, the converged evolution, is that it leads to nearly optimal systems with similar gross characteristics, so simple arguments based on optimal design can explain functional relations between variables across many scales [162, 163]. Three other key elements (discovered with the use of computational modeling and experimentation) of the organizational principles used by cells are noted in Refs. [116, 132, 164].

1. Ultrasensivity: A response that is more sensitive to ligand concentration as compared to standard responses defined by the Michaelis–Menten equations [165–167];
2. Multistability: An existence of two or more stable states for the regulating network [168, 169];
3. Rhythmic behavior, functioning as a systemic oscillator: In Ref. [170], the gene regulatory network with this property was described: three transcriptional repressors were used to build an oscillating network in *E. coli*.

More general principles that seem to be necessary for the operation of a living system (and peculiar to the complex biological systems) were presented in Ref. [171]:

- Program: plan describing ingredients and interactions between them as living system persist through time.
- Improvisation: the ability to change the program with respect to changes in environment.
- Compartmentalization: division of the whole organisms into smaller compartments to centralize and specialize certain functions.
- Energy: living organism is an open system metabolizing energy.
- Regeneration: resynthesis of the constituents of the system.
- Adaptability: fast response that allows survival in quickly changing environments.
- Seclusion: the ability to allow thousands of reactions to occur with high efficiency in the tiny volume of living cells.

5.4. Representation of Gene-Regulatory and Biochemical Networks

Theoretically, each system can be described as a set of modules and protocols. Modules are components or subsystems of larger system that may have the following features [161]: possess identifiable interfaces to other modules, can be modified and evolved somewhat independently, facilitate simplified or abstract modeling, maintain some identity when isolated or rearranged, derive additional identity from the rest of the system.

The protocol is the set of rules describing and managing the relationship between modules and subsystems and processes in the system. It allow interfaces between modules and permit system functions. The set of protocols for a particular system can include such relationships between modules and components as activation, inhibition, feedback and feed-forward controls, and so forth.

Developing software for building networks and simulating experiments with the use of standardized technology and common infrastructure is important for systems biology. To solve the problem of software interoperability, two related packages were developed [172]: ERATO Systems Biology Workbench (SBW; a modular, broker-based message-passing framework for simplified interconnection between applications), and the Systems Biology Markup Language (SBML: open, extensible markup language (XML) based format for representing biochemical reaction networks). Initially, the SBWs focus was to provide interoperability for the following existing simulation tools: BioSpice [173], DBSolve [174, 175], e-Cell [176, 177], Gepasi [178, 179], Jarnac [180, 181], StochSim [182, 183], and Virtual Cell [184, 185].

SMBL is an XML-based language. XML [186] and originally it was designed to meet the challenges of large-scale electronic publishing (http://www.w3.org/XML/) and it is a dialect of the Standard Generalized Markup Language. The two draft versions of the SMBL were developed and released by the Caltech ERATO team in 2000, and the base-level definition of SBML was delivered in March 2001. Model definition in SMBL consists of the following components [187]:

- Unit definition: A name for a unit used in the expression of quantities in a model. Units may be supplied in a number of contexts in an SBML model, and it is convenient to have a facility both for setting default units and for allowing combinations of units to be given abbreviated names.
- Compartment: A container of finite volume for substances, in which reactions take place. Compartments do not necessarily have to correspond to actual structures inside or outside of a cell.
- Specie: A substance or entity that takes part in a reaction. Some example species are ions such as Ca^{2+} and molecules such as glucose or adenosine triphosphate. The primary qualities associated with specie are its initial amount and the compartment in which it is located.
- Parameter: A quantity that has a symbolic name; this name can be used in formulas in place of the value. Parameters can be global to a model or local to a single reaction.
- Reaction: A statement describing some transformation, transport or binding process that can change the amount of one or more species. For example, a reaction may describe how certain entities (reactants) are transformed into certain other entities (products). Reactions have associated rate laws describing how quickly they take place. Reactions are defined using lists of reactant species and products, their stoichiometric coefficients, and kinetic rate laws.
- Rule: A mathematical expression that is added to the differential equations constructed from the set of reactions and that can be used to set parameter values, establish constraints between quantities, and so forth.

Models of different complexity can be written in SMBL and then read by software packages and translated to an internal format. After that, one can model the dynamics, study the model behavior, and represent the results in plots.

CellML language is close to SMBL and is an open and XML-based standard [188]. CellML is being developed by the Bioengineering Institute, University of Auckland (http://www.cellml.org). It was designed for storing and exchanging biological simulation

models. Two other projects are closely affiliated to CellML: AnatML, for exchanging information at the organ level—it can be used to store geometric information and documentation that was generated during a skeleton digitization project; and FieldML, to provide a description of spatially and temporally varying field information using finite elements—it is appropriate for storing geometry information inside AnatML, for spatial distribution of parameters inside compartments in CellML, or for spatial distribution of cellular model parameters across an entire organ.

Together, these XML-based technologies provide a complete vocabulary for describing "virtual" biological systems from the cellular to the organism level.

Various attempts were made to standardize the graphical representation of the biochemical and gene networks [189–193]. In Ref. [193], the following requirements for graphical notation system were formulated:

- Expressiveness: The ability to describe every possible relationship between objects.
- Semantically unambiguous: Different semantics should be assigned to different symbols.
- Visually unambiguous: Symbol should be clearly identified and not be mistaken for other symbols.
- Extension capability: The notation system should be easily extended.
- Mathematical translation: Availability to be directly applied for numerical analysis.
- Software support: Support of notations by software for drawing, viewing, editing, and translation into mathematical formalism.

To support the graphical notation system proposed in this work, a new process diagram editor (Cell Designer) for gene-regulatory and biochemical network was developed [194].

A research program aimed at creating a framework, experimental infrastructure, and computational environment for understanding, experimenting with, manipulating, and modifying a diverse set of fundamental biological processes at multiple scales and spatiotemporal modes is described in Ref. [195]. From a biological viewpoint, the basic issues of these projects are understanding common and shared structural motifs among biological processes, modeling biological noise resulting from interactions among a small number of key molecules or loss of synchrony, explaining the robustness of these systems in spite of such noise, and cataloging multistatic behavior and adaptation exhibited by many biological processes.

5.5. Artificial Life

5.5.1. E-Cell

Several other projects that aimed at computer modeling of the cell should be noted. The first one is the E-Cell Project—an international research project aiming to model and reconstruct biological phenomena *in silico* and developing necessary theoretical supports, technologies, and software platforms to allow precise whole-cell simulation (www.e-cell.org). This project started in 1996 and led first to the design and development of the first working version of the E-Cell simulation environment in 1996. Then, the self-sustaining cell model was constructed by abstracting the gene set of *Mycoplasma genitalium*—the smallest known genome whose complete 580 kb genome sequence was determined in 1995. Next, an attempt was made to model real cells and to develop a more sophisticated simulation environment for biological simulations, and new modeling projects for modeling a human erythrocyte, mitochondrion, *E. coli chemotaxis*, and gene expression/replication were run. A list of publications and a Windows version of the software (E-Cell, version 2) can be found in Web site listed earlier.

5.5.2. Virtual Cell

Virtual Cell (National Resource for Cell Analysis and Modeling, http://www.nrcam.uchc.edu/index.html) is another project that is aimed at providing a remote-user modeling and simulation environment using Java's Remote Method Invocation (RMI). The Virtual Cell provides a formal framework for modeling biochemical, electrophysiological, and

transport phenomena and considers localization in cell of the molecules that take part in these reactions [196]. This localization can take the form of a three-dimensiona arbitrarily shaped cell; molecular species might be heterogeneously distributed in the cell. The geometry of the cell, including the locations and shapes of subcellular organelles, can be imported directly from microscope images. Such a model considers the diffusion of the molecules within the geometry. Users can create biological models of various types and run simulations on a remote server. A general-purpose solver is used to translate the initial biological description into a set of concise mathematical problems. The generated results can be reviewed in the software or exported in a variety of popular formats.

5.5.3. GEPASI

GEPASI (http://www.gepasi.org) is a software package intended for modeling biochemical systems [197, 198]. With the help of this package, one can simulate the kinetics of systems of biochemical reactions as well as fit models to data, optimize functions of the model, and perform metabolic control analysis and linear stability analysis. GEPASI simplifies the task of model building with its user-friendly interface and helps to translate the language of chemistry (reactions) to that of mathematics (matrices and differential equations). This package uses a set of sophisticated numerical algorithms that ensure that the results obtained are fast and accurate [199]. GEPASI is intended primarily for research purpose, but it also can be used for education.

5.5.4. In Silico Cell

In Silico Cell architecture supports the hierarchical modeling of biological system, and the creation of more complex models from simpler ones. In Silico Cell allows researchers to interface with the technology in a fashion that is most intuitive to their particular scientific background. This process is enabled by the use of CellML, an XML-based markup language for describing biological processes at the cellular and subcellular levels.

5.6. Computational System Biology: Modeling Issues

Tomita stated in his paper [200, p. 210] that, "the cell is never conquered until its total behavior is understood, and the total behavior of the cell is never understood until it is modeled and simulated."

Modeling living cells *in silico* (in a computer) has many implications, one of which is testing new drugs through simulation rather than on patients. According to Ref. [201], human trials fail for 70–75% of the drugs that enter them.

Computer modeling of the processes in living cells is an extremely difficult task. There are several reasons for that, including that the processes in a cell are dynamic and depend on many variables, some of which are related to a changing environment; and that the processes of DNA transcription and protein translation are not fully understood.

A starting point for dynamic modeling of a cell would be dynamic modeling of a single gene-regulation process. In Ref. [202] the following methods for single-gene-regulation modeling are discussed, taking into account different aspects of the processes (chemical reactions, physical chemistry, kinetic changes of states, and thermodynamics): Boolean models, based on Boolean logic (true/false logic); differential equation models; stochastic models; hybrid Boolean/differential equation models; hybrid differential equations/stochastic models; neural network models; and hybrid connectionist-statistical models.

Some of these methods are briefly described below.

5.6.1. Boolean Models

Consider a set of N objects at time t_k $\{x_1^k, x_2^k, \ldots, x_N^k\}$, and each object can be in only two different states: on/off, 1/0, False/True, and so forth. For simplicity, let us assume

$$x_i^k \in \{0, 1\} \quad i = 1, \ldots, N$$

The state of the system at a given moment of time can be described as the states of all objects in this set. The state of a given object at the next time step t_{k+1} can be determined by a Boolean logic function (returning only two values, 0 or 1) whose input is the current state of the system.

$$x_i^{k+1} = B_i(x_1^k, x_2^k, \ldots, x_N^k)$$

Boolean function $B = \{B_1, B_2, \ldots, B_N\}$ can be represented as a truth table that consists of all possible system states (2^N) and corresponding states calculated using Boolean function. Because this function represents relations between all systems' states, it can be easily represented as diagram.

As an example, let us consider a system of two genes, A and B, which can be expressed (1) or not expressed (0), and a particular regulatory network that can be described as the following truth table.

All Possible System States	Next State
00	01
01	10
10	11
11	11

The graphical representation of the network is in Fig. 16.

One can see that state 11 in this diagram is stable (i.e., it leads to no change in system states). The Boolean function that corresponds to the truth table and diagram is

$$A^{k+1} = A^k | B^k$$
$$B^{k+1} = A^k | \neg B^k$$

where: | is logical OR and ¬ is logical NOT.

5.6.2. Kinetic Logic Models

This type of model is the extension of the Boolean one: each gene has finite number L of discrete values of states

$$x_i^k \in \{X_1, X_2, \ldots, X_L\} \quad i = 1, \ldots, N$$

So, for each gene, Boolean function should return one of the L possible values. In addition, genes may have different rates of changing their states. This type of relations is described by the more sophisticated function.

For the example shown above, let us assume that genes A and B can be not expressed, 0; expressed at low level, 1; or expressed, 2. Therefore, the total number of possible system states is nine, and one of the possible representations of the system is shown in Fig. 17.

For some models of this type, objects may have a different number of discrete values of states ($L = L_i$); moreover, there may be more than one possible next state for the

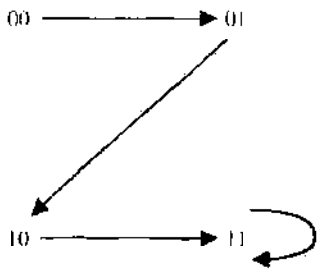


Figure 16. Diagram representation of the Boolean network for the set of two genes.

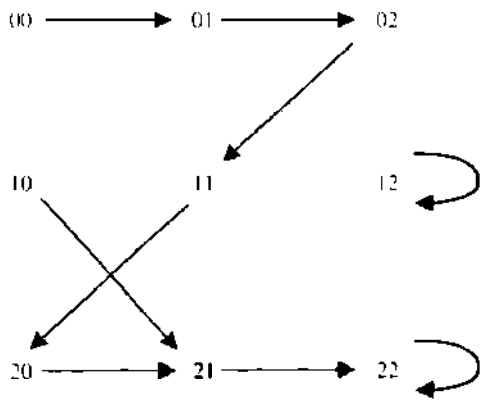


Figure 17. Diagram representation of the kinetic logic model for the set of two genes and three states for gene expression. System has two stable states ("12" and "22").

object in the system, and objects may change their states asynchronously. All these features make kinetic logic models more complex but also more tied to the biological system being modeled.

5.6.3. Ordinary Differential Equation Models

In their turn, differential equation models are the extension of the kinetic logic model. They are used usually to deal with continuous values characterizing the system state. Time is also continuous for these models, and changing of object states with time can be written as

$$\frac{dx_i}{dt} = F_i(t, x) \quad i = 1, \ldots, N$$

where $x = x(t) = \{x_1(t), x_2(t), \ldots, x_N(t)\}$. In general, the right-hand side of the ordinary differential equation of the Mth order can depend on derivatives of the different orders of all objects' states

$$F\left(t, x, \frac{dx}{dt}, \frac{d^2x}{dt^2}, \ldots, \frac{d^Mx}{dt^M}\right) = 0 \quad i = 1, \ldots, N$$

To calculate system dynamics, one should provide initial condition of the system (values for all object states at starting time point $t = t_0$)

$$x(t_0) = X^0$$

The above-mentioned general differential equation is said to be linear if F is a linear function of the variables $x, \frac{dx}{dt}, \frac{d^2x}{dt^2}, \ldots, \frac{d^Mx}{dt^M}$ and can be read as

$$a_M(t)x + a_{M-1}(t)\frac{dx}{dt} + \cdots + a_0(t)\frac{d^Mx}{dt^M} = f(t)$$

where: $a_i(t)$, $i = 1, \ldots, M$ and $f(t)$ some functions.

When the right-hand function is complex and does not allow an analytical solution, the system dynamics can be calculated numerically. There are many tools for solving this problem.

The equilibrium state of the system corresponds to the system state when there are no changes in all states with time. Therefore, the condition of equilibrium state can be found by solving the system of equations

$$F_i(t, x) = 0 \quad i = 1, \ldots, N$$

For instance, the simple model of exponential loss of the gene A expression with time can be expressed as

$$\frac{dA}{dt} = -kA \quad A(0) = A^0$$

where k is the rate of loss of gene expression with time and A^0 is the initial expression of gene A at time $t = 0$. For this simple equation, a solution can be found easily

$$A(t) = A_0 \exp(-kt)$$

and the equilibrium state for this system is trivial: $\overline{A} = 0$.

5.6.4. Partial Differential Equation Models

Partial differential equations can be used for the description of the system states if they change not only with time but with respect to other parameters (e.g., the object's size and location). As an example, let us describe the diffusion of expression of a particular gene in a cell. The equation can be read as follows

$$\frac{dA}{dt} = \alpha \frac{d^2A}{ds^2}$$

Here $A = A(t, s)$ is gene expression distributed in cells with respect to time t ($t \geq 0$) and coordinate s, $0 \leq s \leq S$. Having initial distribution of the expression of this gene in the cell

$$A(0, s) = g(s)$$

and boundary conditions

$$A(t, 0) = p_0(t) \quad A(t, S) = p_S(t)$$

where $g(s)$, $p_0(t)$, and $p_S(t)$ are given functions, one can calculate $A(t, s)$ at any time and coordinate point.

5.6.5. Stochastic Models

Stochastic models deal with the dynamic history of each object of the model. In other words, for each object, the next state must be calculated using the set of probabilistic rules. Each rule shows the probability of the object being changed in a particular interval of time, and the probability of it coming to each state. Therefore, the change of state in this type of model is probabilistic, not deterministic.

Let us assume that object x in the system has a finite state space with L states (as in the kinetic logic model): $\{X_1, X_2, \ldots, X_L\}$. For each time step t_{k+1} there is a transition probability $P(x_{k+1}|x_0, \ldots, x_k)$; and chain $x_0, \ldots, x_k$ represents the history of the system. Variables x_k form a Markov chain if and only if for any k

$$P(x_{k+1}|x_0, \ldots, x_k) = P(x_{k+1}|x_k)$$

In other words, the future state depends on the only present state. All probability values $[P(X_i|X_j)$, the probability of the system jumping from the ith to the jth state] form a transition matrix.

Suppose that the system can jump into state X_i at time t_k with transition rate λ_k^i. After calculating the probability of coming into the ith state at time t_k

$$p_k^i = \frac{\lambda_k^i}{\lambda}, \quad \lambda = \sum_{i=1}^{L} \lambda_k^i$$

one can easily calculate the next state of the system. The formula for calculating the next time point depends on the distribution of the jumps $t_{k+1} - t_k$; for instance, for the case of the exponential process, the next time point is $t_{k+1} = t_k - \ln(r)/\lambda$, where r is a random value uniformly distributed in (0, 1).

Figure 18 shows the examples of dynamics calculated for the model of exponential decay for differential equation and stochastic models.

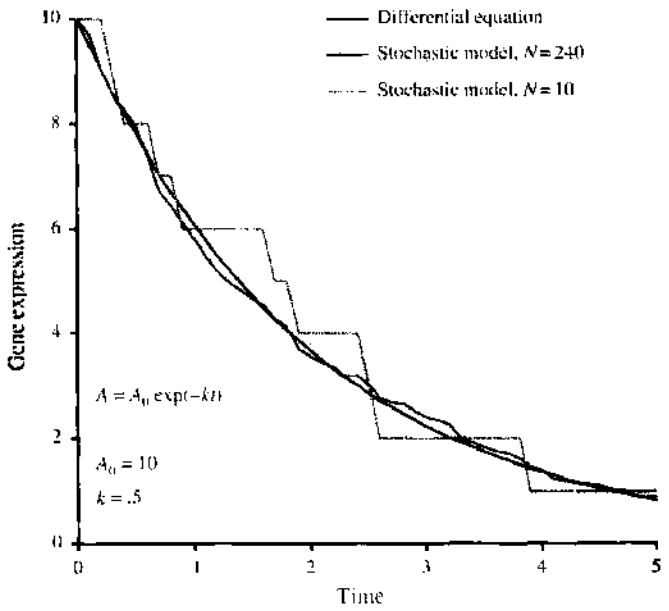


Figure 18. Examples of exponential decay calculated with the use of a differential equation and stochastic model. Two numbers of objects (N) were used for the stochastic model. One can see that with the increase of N, the solution of the stochastic model is the approximate better solution of the differential equation model.

5.6.6. Neural Network Models

Neural networks provide a model of computation that is different from traditional algorithms. Typically, they are not explicitly programmed to perform a given task; rather, they learn to do the task from examples of desired input/output behavior. The networks automatically generalize their processing knowledge into previously unseen situations, and they perform well for the noisy, incomplete, or inaccurate input data.

In general, the artificial neural network is a model consisting of interconnected units evolving in time. Connection between units i and j is usually characterized by the weight, denoted by w_{ij}. There are three important architectures of the neural network, based on the connectivity: recurrent (contains direct loops), feed-forward (contains no direct loops); and layered (units are organized into layers, and connections are between layers).

The behavior of each unit in time can be described by the time-dependent functions, stochastical process, Bayesian network, and so forth. Therefore, the ith unit receives total input x_i from the units connected to it and generates the response

$$f_i(x_i) = \sum_{j \in C} w_{ij} y_i$$

where C is the set of units having connection to the ith unit. When the response is represented as a threshold function

$$f(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases}$$

the unit is called a threshold gate and can generate only binary decisions.

The most usual application of the neural network is classification (to arrange input into number of classes). The most important feature of the neural network is learning from examples. It consists of the model fitting and parameter estimation, with the use of the training subset of input data, and validation of the model output, with the use of the validation subset.

5.6.7. Hybrid Models

Some models can combine features of different types of models. Among these models are

- Boolean/differential-equation models: Parameters of the differential equation can depend on discrete system states.

- Differential equations/stochastic models: Introduction in the right-hand side of the differential equations' statistical terms describing, for example, the noise in the system values. It should be noted that some differential models can show stochastic behavior for some values of the parameters. For instance, the Mackey–Glass differential equation

$$\frac{dx}{dt} = \frac{ax(t-\tau)}{1+x^{10}(t-\tau)} - bx$$

behaves as a chaotic time series when the parameters have the following values: $x(0) = 1.2$, $\tau = 17$, $a = 0.2$, $b = 0.1$, and $x(t) = 0$, for $t < 0$.

The next step in dynamic cell modeling would be to try and model the regulation of more genes, and hopefully a large set of genes (see Ref. [202]). Patterns of collective regulation of genes, such as chaotic attractors, are observed in the above reference. Mutual information/entropy of clusters of genes can be evaluated.

5.7. Gene Network Modeling

In living systems, many dynamic, adaptive, evolving processes are observed at different levels, and at different stages of the development, that are involved in a complex interaction. At a molecular level and a cell level, the DNA, the RNA, and the protein molecules evolve and interact in a continuous way. The genes form dynamic gene networks (GNs) that define the complexity of the living organism [204]. It is not just the number of the genes in a genome, however, but the interaction between the genes that makes one organism more complex than another.

Many functions are associated with a neuronal cell and with neural networks in the brain [205]. An ensemble of cells (neurons) operates in concert, defining the function of the neural network (e.g., perception of a sound, or a brain disease such as epilepsy [206]). At the level of the whole brain, a complex dynamic interaction is observed, and certain cognitive functions are performed (e.g., speech and language learning, visual pattern recognition).

The genes, encoded in the DNA, which are transcribed into RNA and then translated into proteins in each cell, contain important information related to the brain activities. A specific gene from the genome relates to the activity of a neuronal cell in terms of a specific function, but the functioning of the brain is much more complex than that. The interaction between the genes is what defines the functioning of a neuron. Even in the presence of a mutated gene in the genome that is known to cause a brain disease, the neurons can still function normally provided a certain pattern of interaction between the genes is maintained—a certain state of the GRN [207]. However, if there is no mutated gene in the genome, certain abnormalities in brain functioning can be observed, as defined by a certain state of the interaction between the genes [206]. The above-cited and many other observations point to the significance of modeling a neuron and a neuronal ensemble at the gene level to predict the state of the ensemble. The process of modeling the gene interaction for the purpose of brain understanding is a significant challenge to biologists, mathematicians, information and computer scientists, brain scientists, and researchers from many other areas.

Models of GRN, derived from gene-expression RNA data, have been developed using different mathematical and computational methods, such as statistical correlation techniques [208, 209], evolutionary computation [210, 211], neural networks [212, 213], differential equations [214], and others [133, 215]. In Ref. [216], a simple GN model of five genes and gene clusters is derived from the time-course gene-expression data of a leukemia cell line U937 treated with retinoic acid with two phenotype states—cancer and normal. The model uses adaptive artificial neural networks—evolving connectionist systems, trained on data in an adaptive mode [15].

A simple GRN of four genes is given in Fig. 19.

6. IMPLICATIONS FOR MEDICINE

Profiling gene and protein expression using DNA and protein arrays has a tremendous effect in molecular-based classifications of diseases. There are two important tasks, among others, in this area: finding the correlation between subsets of genes/proteins and disease features

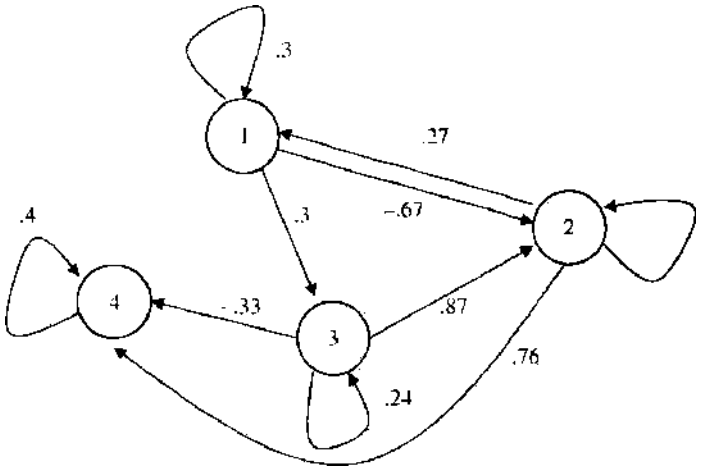


Figure 19. A simple gene regulatory network representing only four genes (the nodes) and their relative interaction strength (the arcs). Four functions are used to calculate the activity of each gene, depending on the activity of other genes in the network; these functions are not shown.

(progression, localization, etc.), and identifying the smallest informative set of genes/proteins associated with specific disease features.

Microarray technology offers an opportunity to screen thousands of genes simultaneously, to be monitored in parallel. New disease subtypes or molecularly distinct forms of the disease can be identified with the use of this technology: B-cell lymphoma [217], two molecularly distinct forms of diffuse large B-cell lymphoma with gene-expression patterns indicative of different stages of B-cell differentiation were identified; breast tumors [218, 219], gene-expression patterns provided a distinctive molecular portrait of each tumor in a set of 65 surgical specimens of human breast tumors from 42 different individuals; human acute leukemia [128], automatic discovery of the distinction between AML and ALL. In Ref. [220], cDNA microarray technology was used to explore variation in gene expression in 60 cell lines of human cancer, and a consistent relationship between the gene expression patterns and the tissue of origin was found. Specific features of these gene expression patterns appeared to be related to physiological properties of the cell lines (doubling time in culture, drug metabolism, and interferon response).

For some cases, DNA microarray technology is an inadequate method, as was noted in Ref. [221] for autoimmune diseases: the disease may manifest not in itself at the RNA level but, rather, at the protein one; protein function can be regulated by posttranslational modifications—phosphorylation, glycosylation, and sulfation, as well as many other modifications, are extremely important for protein function, as they can determine activity, stability, localization, and turnover; and there may be nonpredictive correlations between RNA expression and protein expression and function.

The correlation between levels of mRNA measured in oligonucleotide microarrays and protein is an important issue in DNA microarray technology. The lack of this correlation means that the predictive property of the gene expression is independent of gene function. In Ref. [222], two-dimensional PAGE, mass spectrometry, and Affymetrix oligonucleotide microarrays were used to identify proteins showing increased expression in lung adenocarcinoma and to examine whether the changes in protein expression may be attributable to transcriptional or other mechanisms of regulation. A comparison of the mRNA expression values and the protein expression values within the same tumor samples shows that expression of only two of 14 genes correlated significantly with levels of proteins. The lack of correlation between mRNA and protein level was also noted in Refs. [223, 224]. Also in some cases, using chips from different companies may give results that are not significantly correlated [225].

There is no strict linear relationship between genes and the "proteome" of a cell. Proteomics is complementary to genomics because it focuses on the gene products, and for

this reason proteomics directly contributes to drug development, as almost all drugs are directed against proteins.

Proteomics is a promising tool for the identification of new diagnostic tools (identification of disease markers or proteins that appear or disappear during the course of a disease), development of drugs, improvement of the efficiency of clinical trials (availability of biologically relevant markers for drug efficacy and safety), and clinical diagnostic testing. These approaches include [226] the analysis of protein expression in normal and disease tissue, analysis of secreted proteins in cell lines and primary cultures, and direct serum protein profiling. Aberrantly expressed proteins might represent new markers. MS allows yielding comprehensive profiles of peptides and proteins without the need of first separating them, and it is highly suited for marker identification.

The changes in protein expression that enable tumors to initiate and progress in the local tissue microenvironment were analyzed in Ref. [227] with the use of an antibody microarray. It was demonstrated that quantitative, and potentially qualitative, differences in expression patterns of multiple proteins within epithelial cells reproducibly correlate with tumor progression.

A reverse-phase protein array approach with immobilization of the tissue's proteins was reported in Ref. [228]. These arrays were used for the screening of molecular markers and pathway targets in patient-matched human tissue during disease progression. In contrast to previous protein arrays that immobilize the probe, reverse-phase protein arrays immobilize the whole repertoire of patient proteins that represent the state of individual tissue cell populations undergoing disease transitions. A high degree of sensitivity, precision, and linearity was achieved, making it possible to quantify the phosphorylated status of signal proteins in human tissue cell subpopulations.

APPENDIX: GLOSSARY

Artificial neural networks are biologically inspired computational models that consist of processing elements (called neurons) and the connections between them, with coefficients (weights) bound to the connections, which constitute the neuronal structure. To the structure are also attached training and recall algorithms. One of the most popular training algorithms is the backpropagation algorithm for adjusting the connection weights in a neural network, where the gradient descent rule is used for finding the optimal connection weights w_{ij} that minimize a global error E. A change of weight Δw_{ij} at a cycle $(t + 1)$ is in the direction of the negative gradient of the error E.

Bayesian probability The following formula, which represents the conditional probability between two events C and A, is known as the Bayes Formula (Thomas Bayes, eighteenth century mathematician)

$$p(A|C) = \frac{p(C|A)p(A)}{p(C)}$$

Using the Bayes formula involves difficulties, mainly concerning the evaluation of the prior probabilities $p(A)$, $p(C)$, $p(C|A)$. In practice (e.g., in statistical pattern recognition), the latter is assumed to be of a Gaussian type. The Bayes theorem assumes that if the condition C consists of condition elements $C_1, C_2, \ldots, C_k$ they are independent (which may not be the case in some applications).

Clustering Based on a measured distance between instances (objects, points, vectors) from the problem space, subareas in the problem space of closely grouped instances can be defined. These areas are called clusters. They are defined by their cluster centers and by the membership of the data points to them. A center c_i of a cluster C_i is defined as a point having minimum mean distance for all instances in the cluster. Let us have a set X of p data items represented in an n-dimensional space. A clustering procedure results in defining k disjoint subsets (clusters), such that every data item (n-dimensional vector) belongs to one only cluster. A cluster membership function M_i is defined for each of the clusters

$C_1, C_2, \ldots, C_k$:

$$M_i : X \rightarrow \{0, 1\}$$

$$M_i(x) = \begin{cases} 1 & x \in C_i \\ 0 & x \notin C_i \end{cases}$$

where x is a data instance (vector) from X. In fuzzy clustering, one data vector may belong to several clusters to certain degree of membership, with all of the degrees summing up to 1.

Data normalization is a transformation of data from its original scale into another, predefined scale (e.g., [0, 1]). Normalization is linear when uses the following formula (for the case of a targeted scale of [0, 1])

$$v_{\text{norm}} = \frac{v - x_{\min}}{x_{\max} - x_{\min}}$$

where v is a current value of the variable x; $x_{\min}$ is the minimum value for this variable, and $x_{\max}$ is the maximum value for that variable x in the data set.

Distance between data vectors A way of measuring difference between data vectors. The distance between two data points in an n-dimensional geometrical space can be measured in several ways, for example

Hamming:

$$D_{ab} = \sum |a_i - b_i|$$

Euclidean distance:

$$E_{ab} = \sqrt{\frac{1}{n} \sum (a_i - b_i)^2}$$

Fuzzy clustering is a procedure of clustering data into possibly overlapping clusters, such that each of the data examples may belong to each of the clusters to a certain degree. The procedure aims at finding the cluster centers V_i ($i = 1, 2, \ldots, c$) and the cluster membership functions μi, which define to what degree each of the n examples belong to the ith cluster. The number of clusters c is either defined *a priori* (supervised type of clustering) or chosen by the clustering procedure (unsupervised type of clustering). The result of a clustering procedure can be represented as a fuzzy relation $\mu_{i,k}$ such that

(i) $\sum \mu_{i,k} = 1$, for each $k = 1, 2, \ldots, n$; (the total membership of an instance to all the clusters equals 1)

(ii) $\sum_{\mu_{i,k}}$ for each $i = 1, 2, \ldots, c$ (there are no empty clusters)

Information is the ordered, structured, interpreted data—the news.

Knowledge is the theoretical or practical understanding of a subject: gained experience, true and justified belief, the way we do things.

Knowledge-based neural networks (KBNNs) These are prestructured neural networks allow for data and knowledge manipulation, including learning from data, rule insertion, rule extraction, adaptation, and reasoning. KBNNs have been developed either as a combination of symbolic AI systems and NNs, as a combination of fuzzy logic systems and NNs, or as other hybrid systems. Rule insertion and rule extraction operations are typical operations for a KBNN to accommodate existing knowledge along with data and to produce an explanation of what the system has learned.

Kohonen Self-Organizing Map (SOM) A self-organized map neural network for unsupervised learning invented by Professor Teuvo Kohonen and developed by him and other researchers [229, 230].

Multilayer perceptron network (MLP) is a neural network that consists of an input layer, at least one intermediate or "hidden" layer, and one output layer, with the neurons from each layer being fully connected (or, in some particular applications, partially connected) to the neurons from the next layer.

Multiple sequence alignment is the procedure of comparing sequences by searching for the similarity in the subsets that are in the same order in the sequences. Each subset can consist of one or more characters of the sequence and gaps between them.

Principle component analysis (PCA) Finding a smaller number of m components $Y = (y_1, y_2, \ldots, y_m)$ (aggregated variables) that can represent the goal function $F(x_1, x_2, \ldots, x_n)$ of n variables, $n > m$ to a desired degree of accuracy Θ (i.e., $F = MY + \Theta$, where M is a matrix that has to be found through the PCA).

Probability theory is based on the following three axioms:

Axiom 1. $0 \leq p(E) \leq 1$ The axiom defines the probability $p(E)$ of an event E as a real number in the closed interval [0, 1]. A probability $p(E) = 1$ indicates a certain event, and $p(E) = 0$ indicates an impossible event.

Axiom 2. $\sum p(E_i)\ E_1 \cup E_2 \cup \cdots \cup E_k = U$, where U is a problem space (universum);

Axiom 3. $p(E_1 \vee E_2) = p(E_1) + p(E_2)$, where E_1 and E_2 are mutually exclusive events. This axiom indicates that if the events E_1 and E_2 cannot occur simultaneously, the probability of one or the other happening is the sum of their probabilities.

REFERENCES

1. J. C. Wooley and M. N. Varma, *Basic Life Sci.* 63, 1 (1994).
2. S. Brenner, *Novartis Found. Symp.* 213, 106 (1998).
3. D. S. Roos, *Science* 291, 1260 (2001).
4. J. D. Watson and F. H. C. Crick, *Nature* 171, 737 (1953).
5. B. A. Shapiro and W. Kasprzak, *J. Mol. Graphics* 14, 194 (1996).
6. B. A. Shapiro, W. Kasprzak, J. C. Wu, and K. Currey, in "Pattern Discovery in Biomolecular Data" (J. T. L. Wang, B. A. Shapiro, and D. Shasha, Eds.), p. 183. Oxford University Press, New York, 1999.
7. J. S. Richardson, *Biophysics* 63, 1186 (1992).
8. J. L. McClelland and D. E. Rummelhart, in "Explorations in Parallel Distributed Processing," Vol. 3, p. 318. MIT Press, Cambridge, MA, 1988.
9. D. G. Kneller, F. E. Cohen, and R. Langridge, *J. Mol. Biol.* 214, 171 (1990).
10. B. Rost and C. Sander, *Proc. Nat. Acad. Sci.* 90, 7558 (1993).
11. B. Rost and C. Sander, *J. Mol. Biol.* 232, 584 (1993).
12. B. Rost and C. Sander, *Protein* 19, 55 (1994).
13. J. A. Cuff and G. J. Barton, *Proteins* 34, 508 (1999).
14. N. Qian and T. Sejnowski, *J. Theoret. Biol.* 202, 865 (1988).
15. N. Kasabov, "Evolving Connectionist Systems—Methods and Applications in Bioinformatics, Brain Study and Intelligent Machines." Springer, New York, 2002.
16. N. Guex and M. C. Peitsch, *Electrophoresis* 18, 2714 (1997).
17. M. C. Peitsch, *Biochem. Soc. Trans.* 24, 274 (1996).
18. O. Lund, K. Frimand, J. Gorodkin, H. Bohr, J. Bohr, J. Hansen, and S. Brunak, *Protein Eng.* 10, 1241 (1997).
19. P. A. Bates, L. A. Kelley, R. M. MacCallum, and M. J. E. Sternberg, *Prot. Struct. Funct. Genet.* 5, 39 (2001).
20. P. A. Bates and M. J. E. Sternberg, *Prot. Struct. Funct. Genet.* 3, 47 (1999).
21. B. Contreras-Moreira and P. A. Bates, *Bioinformatics* 18, 1141 (2002).
22. I. N. Shindyalov and P. E. Bourne, Improving alignments in HM protocol with intermediate sequences, in "Fourth meeting on the critical assessment of techniques for protein structure prediction," p. A-92, 2000.
23. I. N. Shindyalov and P. E. Bourne, *Nucl. Acids Res.* 29, 228 (2001).
24. A. J. Gibbs and G. A. McIntyre, *Eur J Biochem.* 16, 1 (1970).
25. J. V. J. Maizel and R. P. Lenk, *Proc. Natl. Acad. Sci.* 78, 7665 (1981).
26. D. J. States and M. S. Boguski, in "Sequence Analysis Primer" (M. Gribskov and J. Devereux, Eds.), p. 92. Stockton Press, NY, 1991.
27. G. Grillo, F. Licciulli, S. Liuni, E. Sbisa, and G. Pesole, *Nucl. Acids Res.* 31, 3608 (2003).
28. V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D. U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender, *Nucl. Acids Res.* 31, 374 (2003).
29. K. Quandt, K. Frech, H. Karas, E. Wingender, and T. Werner, *Nucl. Acids Res.* 23, 4878 (1995).
30. S. F. Altshul, *J. Mol. Biol.* 219, 555 (1991).
31. D. J. States, W. Gish, and S. F. Altshul, *Methods* 3, 66 (1991).
32. M. O. Dayhoff, in "Atlas of Protein Sequence and Structure," Vol. 5, suppl. 3. National Biomedical Research Foundation, Georgetown University, Washigton, DC, 1978.
33. G. H. Gonnet, M. A. Cohen, and S. A. Benner, *Science* 256, 1443 (1992).
34. D. T. Jones, W. R. Taylor, and J. M. Thornton, *Comput. Appl. Biosci.* 8, 275 (1992).
35. S. A. Benner, M. A. Cohen, and G. H. Gonnet, *Protein Eng.* 7, 1323 (1994).
36. S. Henikoff and J. G. Henikoff, *Proc. Natl. Acad. Sci.* 89, 10915 (1991).
37. G. H. Gonnet, M. A. Cohen, and S. A. Benner, *Biochem. Biophys. Res. Commun.* 199, 489 (1994).

38. D. T. Jones, W. R. Taylor, and J. M. Thornton, *FEBS Lett.* 339, 269 (1994).
39. D. W. Mount, "Bioinformatics," Cold Spring Harbor Laboratory Press, Cold Sprin Harbor, NY, 2001.
40. D. J. Lipman and W. R. Pearson, *Science* 2277, 1435 (1985).
41. W. R. Pearson and D. J. Lipman, *Proc. Natl. Acad. Sci.* 85, 2444 (1988).
42. A. Klingenhoff, K. Frech, K. Quandt, and T. Werner, *Bioinformatics* 15, 180 (1999
43. K. Frech, J. Danescu-Mayer, and T. Werner, *J. Mol. Biol.* 270, 674 (1997).
44. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, *J. Mol. Biol.* 21, 403 (1990).
45. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, andD. J. Lipman, *Nucl. Acids Res.* 25, 3389 (1997).
46. V. C. Wasinger, S. J. Cordwell, A. Cerpa-Poljak, J. X. Yan, A. A. Gooley, M. F. Wilkins, M. W. Duncan, R. Harris, K. L. Williams, and I. Humphery-Smith, *Electrophoresis* 16, 1090 (1995)
47. S. Fields and O. Song, *Nature* 340, 245 (1989).
48. V. Schachter, *Comp. Proteomics Suppl.* 32, S16 (2002).
49. J. S. Albala, *Expert Rev. Mol. Diagn.* 1, 145 (2001).
50. H. Zhu, J. F. Klemic, S. Chang, P. Bertone, A. Casamayor, K. G. Klemic, D. Smith, M. Gerstein, M. A. Reed, and M. Snyder, *Nat. Genet.* 26, 283 (2000).
51. G. MacBeath and S. L. Schreiber, *Science* 289, 1760 (2000).
52. P. Arenkov, A. Kukhtin, A. Gemmell, S. Voloshchuk, V. Chupeeva, and A. Mirzabekov, *Anal. Biochem.* 278, 123 (2000).
53. Y. Lee, E. K. Lee, Y. W. Cho, T. Matsui, I.-C. Kang, T.-S. Kim, and M. H. Han, *Proteomics* 3, 2289 (2003).
54. M. F. Templin, D. Stoll, J. M. Schwenk, O. Potz, S. Kramer, and T. O. Joos, *Proteomics* 3, 2155 (2003).
55. T. S. Lewis, J. B. Hunt, L. D. Aveline, K. R. Jonscher, D. F. Louie, J. M. Yeh, J. S. Nahreini, K. A. Resing, and N. G. Ahn, *Mol. Cell* 6, 1343 (2000).
56. J. H. McKerrow, V. Bhargava, E. Hansell, S. Huling, T. Kuwahara, M. Matley, I. Coussens, and R. Warren, *Mol. Med* 6, 460 (2000).
57. M. J. Han and S. Y Lee, *Proteomics* 3, 2317 (2003).
58. D. N. Chakravarti, B. Chakravarti, and I. Moutsatsos, *Comp. Proteomics Suppl.* 3, S4 (2002).
59. R. C. Beavis and D. Fenyo, in "Proteomics: A Trends Guide" (W. Blackstock and M. Mann, Eds.), p. 22. Elsevier, Amsterdam, 2000.
60. D. Fenyo, *Curr. Opin. Biotechnol.* 11, 391 (2000).
61. N. L. Anderson, J. Taylor, A. E. Scandora, B. P. Coulter, and N. G. Anderson, *Clin. Chem.* 27, 1807 (1981).
62. J. I. Garrels, *J. Biol. Chem.* 254, 7961 (1979).
63. P. F. Lemkin and L. E. Lipkin, *Comput. Biomed. Res.* 14, 272 (1981).
64. R. Appel, D. Hochstrasser, C. Roch, M. Funk, A. F. Muller, and C. Pellegrini, *Electrophoresis* 9, 136 (1988).
65. R. D. Appel, D. F. Hochstrasser, M. Funk, J. R. Vargas, C. Pelegrini, A. F. Muller, and J. R. Scherrer, *Electrophoresis* 12, 722 (1991).
66. T. Pun, D. F. Hochstrasser, R. D. Appel, M. Funk, V. Villars-Augsburger, and C. Pelegrini, *Appl. Theor. Electrophor.* 1, 3 (1988).
67. D. G. Rowlands, A. Flook, P. I. Payne, A. van Hoff, T. Niblett, and S. McKee, *Electrophoresis* 9, 820 (1988).
68. J. I. Garrels, *J. Biol. Chem.* 264, 5269 (1989).
69. W. J. Henzel, T. M. Bulleci, J. T. Stults, S. C. Wong, C. Grimley, and C. Watanabe, *Proc. Natl. Acad. Sci.* 90, 5011 (1993).
70. P. James, M. Quadroni, E. Carafoli, and G. Gonnet, *Biochem. Biophys. Res. Commun.* 195, 58 (1993).
71. M. Mann, P. Hojrup, and P. Roepstorff, *Biol. Mass Spectrom.* 22, 338 (1993).
72. D. J. C. Pappin, P. Hojrup, and A. J. Bleasby, *Curr. Biol.* 3, 327 (1993).
73. D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell, *Electrophoresis* 20, 3551 (1999).
74. J. K. Eng, A. L. McCormack, and J. R. I. Yates, *J. Am. Soc. Mass Spetrom.* 5, 976 (1994).
75. M. Mann and M. Wilm, *Anal. Chem.* 66, 4390 (1994).
76. P. D. Von Haller, E. Yi, S. Donohoe, K. Vaughn, A. Keller, A. I. Nesvizhskii, J. Eng, X. J. Li, D. R. Goodlett, R. Aebersold, and J. D. Watts, *Mol. Cell Proteomics* 2, 428 (2003).
77. P. D. Von Haller, E. Yi, S. Donohoe, K. Vaughn, A. Keller, A. I. Nesvizhskii, J. Eng, X. J. Li, D. R. Goodlett, R. Aebersold, and J. D. Watts, *Mol. Cell Proteomics* 2, 426 (2003).
78. A. Keller, A. I. Nesvizhskii, E. Kolker, and R. Aebersold, *Anal. Chem.* 74, 5383 2002).
79. D. K. Han, J. Eng, H. Zhou, and R. Aebersold, *Nat. Biotechnol.* 19, 946 (2001).
80. A. I. Nesvizhskii, A. Keller, E. Kolker, and R. Aebersold, *Anal. Chem.* 75, 4646 2003).
81. J. A. Taylor and R. S. Johnson, *Anal. Chem.* 73, 2594 (2001).
82. M. R. Wilkins, E. Gasteiger, A. A. Gooley, B. R. Herbert, M. P. Molloy, P. A. Binz, K. Ou, J. C. Sanchez, A. Bairoch, K. L. Williams, and D. F. Hochstrasser, *J. Mol. Biol.* 289, 645 (1999)
83. A. Sali, R. Glaeser, T. Earnest, and W. Baumeister, *Nature* 422, 216 (2003).
84. E. Krieger, S. B. Nabuurs, and G. Vriend, *Methods Biochem. Anal.* 44, 509 (2003).
85. A. Godzik, *Methods Biochem. Anal.* 44, 525 (2003).
86. W. J. Browne, A. C. North, D. C. Phillips, K. Brew, T. C. Vanaman, and R. L. Hill, *J. Mol. Biol.* 42, 65 (1969).
87. J. Greer, *Proc. Natl. Acad. Sci.* 77, 3393 (1980).
88. J. Greer, *Proteins* 7, 317 (1990).
89. R. Sanchez, U. Pieper, F. Melo, N. Eswar, M. A. Marti-Renom, M. S. Madhusudhan, N. Mirkovic, and A. Sali, *Nat. Struct. Biol.* 7, 986 (2000).
90. D. S. Dimitrow, *Cell* 101, 697 (2000).

91. P. Prabakaran, X. Xiao, and D. S. Dimitrov, *Biochem. Biophys. Res. Commun.* 314, 235 (2004).
92. W. Li, M. J. Moore, N. Vasilieva, J. Sui, S. K. Wong, M. A. Berne, M. Somasundaran, J. L. Sullivan, K. Luzuriaga, T. C. Greenough, H. Choe, and M. Farzan, *Nature* 426, 450 (2003).
93. X. Xiao, S. Chakraborti, A. S. Dimitrov, K. Gramatikoff, and D. S. Dimitrov, *Biochem. Biophys. Res. Commun.* 312, 1159 (2003).
94. M. Donoghue, F. Hsieh, E. Baronas, K. Godbout, M. Gosselin, N. Stagliano, M. Donovan, B. Woolf, K. Robison, R. Jeyaseelan, R. E. Breitbart, and S. Acton, *Circ. Res.* 87, E1 (2000).
95. S. R. Tipnis, N. M. Hooper, R. Hyde, E. Karran, G. Christie, and A. J. Turner, *J. Biol. Chem.* 275, 33238 (2000).
96. M. A. Crackower, R. Sarao, G. Y. Oudit, C. Yagil, I. Kozieradzki, S. E. Scanga, A. J. Oliveira-dos-Santos, J. da Costa, L. Zhang, Y. Pei, J. Scholey, C. M. Ferrario, A. S. Manoukian, M. C. Chappell, P. H. Backx, Y. Yagil, and J. M. Penninger, *Nature* 417, 822 (2002).
97. R. Natesh, S. L. Schwager, E. D. Sturrock, and K. R. Acharya, *Nature* 421, 551 (2003).
98. H. M. Kim, D. R. Shin, O. J. Yoo, H. Lee, and J. O. Lee, *FEBS Lett.* 538, 65 (2003).
99. D. S. Dimitrov, *Nat. Rev. Microbiol.* 2, 109 (2004).
100. J. D. Thompson, D. G. Higgins, and T. J. Gibson, *Nucl. Acids Res.* 22, 4673 (1994).
101. M. J. Suteliffe, F. R. Hayes, and T. L. Blundell, *Protein Eng.* 1, 385 (1987).
102. M. J. Suteliffe, I. Haneef, D. Carney, and T. L. Blundell, *Protein Eng.* 1, 377 (1987).
103. R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton, *J. Appl. Cryst.* 26, 283 (1993).
104. C. J. Sigrist, L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch, and P. Bucher, *Brief. Bioinform.* 3, 265 (2002).
105. B. Lee and F. M. Richards, *J. Mol. Biol.* 55, 379 (1971).
106. A. Nicholls, K. A. Sharp, and B. Honig, *Proteins* 11, 281 (1991).
107. J. Kyte and R. F. Doolittle, *J. Mol. Biol.* 157, 105 (1982).
108. J. L. Guy, R. M. Jackson, K. R. Acharya, E. D. Sturrock, N. M. Hooper, and A. J. Turner, *Biochemistry* 42, 13185 (2003).
109. O. Spiga, A. Bernini, A. Ciutti, S. Chiellini, N. Menciassi, F. Finetti, V. Causarono, F. Anselmi, F. Prischi, and N. Niccolai, *Biochem. Biophys. Res. Commun.* 310, 78 (2003).
110. A. Levchenko, *Mol. Biol. Rep.* 28, 83 (2001).
111. A. J. Lotka, *J. Amer. Chem. Soc.* 42, 1595 (1920).
112. V. Volterra, *Mem. Acad. Lincei* 2, 31 (1926).
113. A. L. Hodgkin and A. F. Huxley, *J. Physiol. (Lond.)* 117, 500 (1952).
114. M. Eigen, *Naturwissenschaften* 58, 465 (1971).
115. A. Gierer and H. Meinhardt, *Kybernetik* 12, 30 (1972).
116. S. R. Neves and R. Iyengar, *BioEssays* 24, 1110 (2002).
117. H. Kitano, *Nature* 420, 206 (2002).
118. U. Bhalla and R. Iyengar, *Science* 283, 381 (1999).
119. B. N. Kholodenko, O. V. Demin, G. Moehren, and J. B. Hoek, *J. Biol. Chem.* 274, 30169 (1999).
120. I. A. Sidorov and A. A. Romanyukha, *Math. Biosci.* 115, 187 (1993).
121. G. I. Marchuk, R. V. Petrov, A. A. Romahyukha, and G. A. Bocharov, *J. Theoret. Biol.* 151, 1 (1991).
122. L. W. Loew, A. Cowan, and I. Moraru, "Computational Cell Biology—challenges and oportunities for an emerging field. A report based on Round table discussion at the First International Symposium on Computational Cell Biology," Cranwell Resort, Lenox, MA, 2001.
123. I. Cloete and J. Zurada, "Knowledge-Based Neurocomputing," MIT Press, Cambridge, MA, 2000.
124. "Microarray Biochip Technology," Eaton Publishing, Natick, MA, 2001.
125. M. Futschik and N. Kasabov, Evolving fuzzy neural networks for knowledge discovery from gene expression data—a case study. In "RECOMB'2001 Proceedings—Currents in Computational Molecular Biology 2001" (T. Lengauer and D. Sankoff, Eds.), Montreal, 2001, p. 175.
126. U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, *Proc. Nat. Acad. Sci.* 96, 6745 (1999).
127. N. Kasabov, Adaptive Learning Method and System, patent 503882, New Zealand (2000).
128. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, *Science* 286, 531 (1999).
129. J. Hasty, D. McMillen, F. Farren, and J. J. Collins, *Nat. Rev. Genet.* 2, 268 (2001).
130. G. Weng, U. S. Bhalla, and R. Iyengar, *Science* 284, 92 (1999).
131. P. Smolen, D. A. Baxter, and J. H. Byrne, *Neuron* 26, 567 (2000).
132. H. Jeff, F. Farren, D. Milos, M. David, and J. J. Collins, *Chaos* 11, 207 (2001).
133. H. de Jong, *J. Comput. Biol.* 9, 67 (2002).
134. S. Huang, *Pharmacogenomics* 2, 203 (2001).
135. H. Bolouri and E. H. Davidson, *BioEssays* 24, 1118 (2002).
136. J. Khan, J. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westerman, F. Berthold, M. Schwab, C. R. A. Nioneseu, C. Peterson, and P. S. M. Meltzer, *Nat. Med.* 7, 673 (2001).
137. A. Metcalfe, "Statistics in Engineering—A Practical Approach," Chapman & Hall, London, 1994.
138. R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis, *Mol. Cell* 2, 73 (1998).
139. P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, *Mol. Biol. Cell* 9, 3273 (1998).

140. J. L. deRisi, V. R. Iyer, and P. O. Brown, *Science* 275, 680 (1997).
141. S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. O. Brown, and I. Herskowitz, *Science* 282, 699 (1998).
142. N. Pal and J. C. Bezdek, *IEEE Trans. Fuzzy Syst.* 3, 370 (1995).
143. L. von Bertalanffy, "General System Theory, Foundations, Development, Applications." George Braziller, New York, 1969.
144. J. Gibbs, *Science* 287, 1969 (2000).
145. C. Sander, *Science* 287, 1977 (2000).
146. D. Noble, *Science* 295, 1678 (2002).
147. H. Kitano, in "Foundations of Systems Biology," MIT Press, Cambridge, MA, 2001.
148. N. Weiner, "Cybernetics or Control and Communication in the Animal and the Machine." MIT Press, Cambridge, MA, 1948.
149. L. von Bertalnffy, "Modern Theories of Development: An Introduction to Theoretical Biology." Oxford University Press, New York, 1933.
150. M. A. Savageau, "Biochemical Systems Theory." Addison-Wesley, Reading, MA, 1976.
151. K. C. Chen, A. Csikasz-Nagy, G. B. ., J. Val, B. Novak, and J. J. Tyson, *Mol. Bol. Cell* 11, 369 (2000).
152. M. T. Borisuk and J. J. Tyson, *J. Theoret. Biol.* 195, 69 (1998).
153. K. J. Kauffman, P. Prakash, and J. S. Edwards, *Curr. Opin. Biotechnol.* 14, 491 (2003).
154. J. S. Edwards, R. U. Ibarra, and B. O. Palsson, *Nat. Biotechnol.* 19, 125 (2001).
155. M. A. Savageau, *Curr. Topics Cell. Regulation* 6, 63 (1972).
156. H. Kacser and J. A. Burns, *Symp. Soc. Exp. Biol.* 27, 65 (1973).
157. U. Alon, M. G. Surette, N. Barkai, and S. Leibler, *Nature* 397, 168 (1999).
158. G. von Dassow, E. Mier, M. Munro, and M. Odell, *Nature* 406, 188 (2000).
159. T-M. Yi, Y. Huang, M. I. Simon, and J. Doyle, *Proc. Natl. Acad. Sci.* 97, 4649 2000).
160. H. Kurata and K. Taira, in "Systematic Analysis of the Robustness in Complex Reaction Networks of Bacteria," Proceedings of the Fourth Annual International Conference on Computational Molecular Biology, Tokyo, Japan, 2000, Vol. 36.
161. M. E. Csete and J. C. Doyle, *Science* 295, 1664 (2002).
162. J. H. Brown and G. B. West, Eds., "Scaling in Biology." Oxford University Press, New York, 2000.
163. J. Whitfield, *Nature* 413, 342 (2001).
164. P. Smolen, D. A. Baxter, and J. H. Byrne, *AJP--Cell Physiol.* 274, C531 (1998).
165. D. E. Koshland, Jr., A. Goldbeter, and J. B. Stock, *Science* 217, 220 (1982).
166. A. Goldbeter and D. E. Koshland, Jr., *Proc. Natl. Acad. Sci.* 78, 6840 (1981).
167. J. E. Ferrell, Jr. and E. M. Machleder, *Science* 280, 895 (1998).
168. J. E. Ferrell and X. Wen, *Chaos* 11, 227 (2001).
169. U. S. Bhalla and I. Ravi, *Chaos* 11, 221 (2001).
170. M. B. Elowitz and S. Leibler, *Nature* 403, 335 (2000).
171. J. Koshland, *Science* 295, 2215 (2002).
172. M. Hucka, A. Finney, H. Sauro, H. Bolouri, J. Doyle, and H. Kitano, in "Foundations of Systems Biology." MIT Press, Cambridge, MA, 2001.
173. A. P. Arkin, http://gobi.lbl.gov/~aparkin/Stuff/Software.html (2001).
174. I. Goryanin, T. C. Hodgman, and E. Selkov, *Bioinformatics* 15, 749 (1999).
175. I. Goryanin, http://biosim.genebee.msu.su/models_3_en.php.
176. M. Tomita, Y. Nakayama, Y. Naito, T. Shimizu, K. Hashimoto, K. Takahashi, Y. Matsuzaki, K. Yugi, F. Miyoshi, Y. Saito, A. Kuroki, T. Ishida, T. Iwata, M. Yoneda, M. Kita, Y. Yamada, E. Wang, S. Seno, M. Okayama, A. Kinoshita, Y. Fujita, R. Matsuo, T. Yanagihara, D. Watari, S. Ishinabe, and S. Miyamoto, http://www.e-cell.org (2001).
177. M. Tomita, K. Hashimoto, K. Takahashi, T. S. Shimizu, Y. Matsuzaki, F. Miyoshi, T. Saito, S. Tanida, K. Yugi, J. C. Venter, and C. Hutchison, *Bioinformatics* 15, 72 (1999).
178. P. Mendes, Trends *Biochem. Sci.* 22, 361 (1997).
179. P. Mendes, http://www.gepasi.org (2001).
180. H. M. Sauro, *Math. Comput. Modelling* 15, 15 (1991).
181. H. Sauro and D. A. Fell, in "Animating the Cellular Map 9th International BioThermoKinetics Meeting" (J.-H. S. Hofmeyr, J. M. Rohwer, and J. L. Snoep, Eds.), Stellenbosch University Press, Stellenbosch, South Africa, 2000.
182. D. Bray, C. Firth, N. Le Novère, and T. Shimizu, http://www.anat.cam.ac.uk/comp-cell (2001).
183. C. J. Morton-Firth and D. Bray, *J. Theoret. Biol.* 192, 117 (1998).
184. J. Schaff, B. Slepchenko, and L. M. Loew, in "Methods in Enzymology," Vol. 321, p. 1. Academic Press, San Diego, CA, 2000.
185. J. C. Schaff, B. Slepchenko, F. Morgan, J. Wagner, D. Resasco, D. Shin, Y. S. Choi, L. Loew, J. Carson, A. Cowan, I. Moraru, J. Watras, M. Teraski, and C. Fink, http://www.nrcam.uchc.edu/ (2001).
186. T. Bray, J. Paoli, and C. M. Sperberg-McQueen, www.w3.org/TR/1998/REC-xml-19980210 (1998).
187. M. Hucka, A. Finney, H. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J.-H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer,

N. Le Novère, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang, *Bioinformatics* 19, 524 (2003).
188. W. J. Hedley, M. R. Nelson, D. P. Bullivant, and P. F. Nielsen, *Philos. Trans. R. Soc. Lond. A* 359, 1073 (2001).
189. K. W. Kohn, *Mol. Biol. Cell.* 10, 2703 (1999).
190. K. Kohn, *Chaos* 11, 84 (2001).
191. I. Pirson, N. Fortemaison, C. Jacobs, S. Dremier, J. E. Dumont, and C. Maenhaut, *Trends Cell Biol.* 10, 404 (2000).
192. D. L. Cook, J. F. Farley, and S. J. Tapscott, *Genome Biol.* 2, researeh0012.1 (2001).
193. H. Kitano, *BIOSILICO* 1, 169 (2003).
194. A. Funahashi, M. Morohashi, and H. Kitano, *BIOSILICO* 1, 159 (2003).
195. B. Mishra, R. S. Daruwala, Y. Zhou, N. Ugel, A. Policriti, M. Antoniotti, S. Paxia, M. Rejali, A. Rudra, V. Cherepinsky, N. Silver, W. Casey, C. Piazza, M. Simeoni, P. Barbano, M. Spivak, J. Feng, O. Gill, M. Venkatesh, F. Cheng, B. Sun, I. Ioniata, T. Anantharaman, E. J. Hubbard, A. Pnueli, D. Harel, V. Chandru, R. Hariharan, M. Wigler, F. Park, S. C. Lin, Y. Lazebnik, F. Winkler, C. R. Cantor, A. Carbone, and M. Gromov, *OMICS* 7, 253 (2003).
196. B. M. Slepchenko, J. C. Schaff, I. Macara, and L. M. Loew, *Trends Cell Biol.* 13, 570 (2003).
197. P. Mendes, *Comput. Applic. Biosci.* 9, 563 (1993).
198. P. Mendes, *Trends Biochem. Sci.* 22, 361 (1997).
199. P. Mendes and D. B. Kell, *Bioinformatics* 14, 869 (1998).
200. M. Tomita, *Trends Biotechnol.* 19, 205 (2001).
201. R. Zacks, *MIT Technol. Rev.* 37 (2001).
202. M. A. Gibson and E. Mjolsness, in "Computational Modeling of Genetic and Biochemical Networks" (J. M. Bower and H. Bolouri, Eds.), p. 1. MIT Press, Cambridge, MA, 2001.
203. R. Somogyi, S. Fuhrman, and X. Wen, in "Computational Modelling of Genetic and Biochemical Network" (J. M. Bower and H. Bolouri, Eds.), p. 120. MIT Press, Cambridge, MA, 2001.
204. P. Baldi and S. Brunak, "Bioinformatics—A Machine Learning Approach." MIT Press, Cambridge, MA, 2001.
205. M. Arbib, "The Handbook of Brain Theory and Neural Networks." MIT Press, Cambridge, MA, 2003.
206. V. Crunelli and N. Lereshe, *Nat. Rev. Neurosci.* 3, 371 (2002).
207. R. Morita, E. Miyazaki, C. G. Fong, X. N. Chen, J. R. Korenberg, A. V. Delgado-Escueta, and K. Yamakawa, *Biochem. Biophys. Res. Commun.* 248, 307 (1998).
208. P. D'Haeseleer, S. Liang, and R. Somogyi, *Bioinformatics* 16, 707 (2000).
209. S. Liang, S. Fuhrman, and R. Somogyi, Pacific Symposium on Biocomputing, Proceedings, Hawaii, 1998, p. 3.
210. S. Ando, E. Sakamoto, and H. Iba, "Proceedings of 6th Joint Conference on Information Sciences." Research Triangle Park, NC, 2002, p. 1249.
211. G. Fogel and D. Corne, "Evolutionary Computation for Bioinformatics." Morgan Kaufmann, University of Reading, UK, 2003.
212. J. Vohradsky, *J. Biol. Chem.* 276, 36168 (2001).
213. J. Vohradsky, *FASEB J.* 15, 846 (2001).
214. K. W. Kohn and D. S. Dimitrov, in "Computer Modeling and Simulation of Complex Biological Systems," (S. S. Iyengar, Ed.). CRC Press, Boca Raton, FL, 1998, pp. 101–123.
215. J. Bower and H. E. Bolouri, "Computational Modelling of Genetic and Biochemical Networks." MIT Press, Cambridge, MA, 2001.
216. N. Kasabov and D. Dimitrov, A method for gene regulatory network modeling with the use of evolving connectionist systems, "ICONIP'2002—International Conference on Neuro-Information Processing." IEEE Press, Singapore, 2002.
217. A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, and L. M. Staudt, *Nature* 403, 503 (2000).
218. C. M. Perou, T. Sorlie, M. B. Eisen, M. van de Rijn, S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Jonsen, L. A. Akslen, O. Fluge, A. Pergamenshikov, C. Williams, S. X. Zhu, P. E. Lonning, A.-L. Borresen-Dale, P. O. Brown, and D. Botstein, *Nature* 406, 747 (2000).
219. T. Sorlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lonning, and A. L. Borresen-Dale, *Proc. Natl. Acad. Sci.* 98, 10869 (2001).
220. D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, R. M. Van de, M. Waltham, A. Pergamenschikov, J. C. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein, and P. O. Brown, *Nat. Genet.* 24, 227 (2000).
221. W. H. Robinson, L. Steinman, and P. J. Utz, *Arthritis Rheumatism* 46, 885 (2002).
222. G. Chen, T. G. Gharib, C. C. Huang, D. G. Thomas, K. A. Shedden, J. M. G. Taylor, S. L. R. Kardia, D. E. Misek, T. J. Giordano, M. D. Iannettoni, M. B. Orringer, S. M. Hanash, and D. G. Beer, *Clin. Cancer Res.* 8, 2298 (2002).
223. S. Gygi, Y. Rochon, B. R. Franza, and R. Aebersold, *Mol. Cell. Biol.* 19, 1720 (1999).
224. B. Futcher, G. I. Latter, P. Monardo, C. S. McLaughlin, and J. I. Garrels, *Mol. Cell. Biol.* 19, 7357 (1999).

225. P. K. Tan, T. J. Downey, E. L. Spitznagel, Jr., P. Xu, D. Fu, D. S. Dimitrov, R. A. Lempicki, B. H. Raaka, and M. C. Cam, *Nucl. Acids Res.* 31, 5676 (2003).
226. S. Hanash, *Nature* 422, 226 (2003).
227. V. Knezevic, C. Leethanakul, V. E. Bichsel, J. M. Worth, V. V. Prabhu, J. S. Gutkind, L. A. Liotta P. J. Munson, E. F. Petricoin, and D. B. Krizman, *Proteomics* 1, 1271 (2001).
228. C. P. Paweletz, L. Charboneau, V. E. Bichsel, N. L. Simone, T. Chen, J. W. Gillespie, M. R. Emmert-Buck, M. J. Roth, E. F. Petricoin, and L. A. Liotta, *Oncogene* 20, 1981 (2001).
229. T. Kohonen, *IEEE* 78, 1464 (1990).
230. T. Kohonen, "Self-Organizing Maps." Springer, 1997.

CHAPTER 2

Computational Studies of Protein Folding

Michiel van Lun, David van der Spoel

Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden

CONTENTS

1. INTRODUCTION

The protein folding problem is one of the grand challenges in science. The problem appeals to biologists, chemists, physicists, and mathematicians alike and has been studied extensively by specialists from all of these disciplines. The problem is particularly appealing because it

ISBN: 1-58883-048-9

Handbook of Theoretical and Computational Nanotechnology
Edited by Michael Rieth and Wolfram Schommers
Volume 6: Pages (47–72)

can be formulated in simple terms, although a generic solution still seems elusive. A solution to the problem would not only further the fundamental understanding of "how life works" but also have a tremendous impact on all applied life sciences. This includes medicine, for example, understanding the molecular basis of diseases like Alzheimer disease (which is due to protein misfolding in the cell); pharmaceutical applications, such as structure-based drug design; technologically oriented applications, like the use of proteins as molecular switches; and many more. It is the combination of a problem that is simple to formulate, difficult to solve, but has huge potential for applications that makes the protein folding problem attract researchers from every scientific discipline. In it's simplest form the problem can be stated as: "Given the amino acid sequence of a polypeptide chain, how does it fold to it's native conformation?"

Some historical background is necessary to appreciate the problem. In the 1960s, Levinthal stated what is now know as the Levinthal paradox: "How can a protein fold to its native state in a finite time, given the enormous number of possibilities?". He reasoned [1] that every amino acid in a protein chain can have roughly three different conformations (as deduced from the peptide backbone models of Ramachandran et al. [2–4]. Fig. 1). Therefore, if a protein has 100 amino acids, there are 3^{100} (5×10^{47}) discrete possibilities. Obviously, because a protein can fold in finite time, it cannot work by enumeration, that is, random diffusion. Protein folding research made a big leap forward since Anfinsen [6] first demonstrated that a protein can refold *in vitro* to its native structure after being denatured, a finding that earned him the Nobel prize in chemistry 1972 (and Anfinsen's Nobel lecture [7] still makes for a good introduction into the topic). Subsequently, this has been shown for many different proteins by other researchers. The importance of this finding lies in the fact that it proves that the information needed to determine the three-dimensional structure of a protein is contained in the amino-acid sequence, at least for the class of water-soluble proteins. It does not seem far-fetched to speculate that the same holds for membrane proteins in their native environment, although experimental evidence is scarce in this field [8–11]. Anfinsen supports the "thermodynamic hypothesis" [7], that is, that the native conformation is determined by the minimum of the Gibbs free energy of the protein *in its environment*. In this light, the protein folding problem should therefore be rephrased as: "Given the amino acid

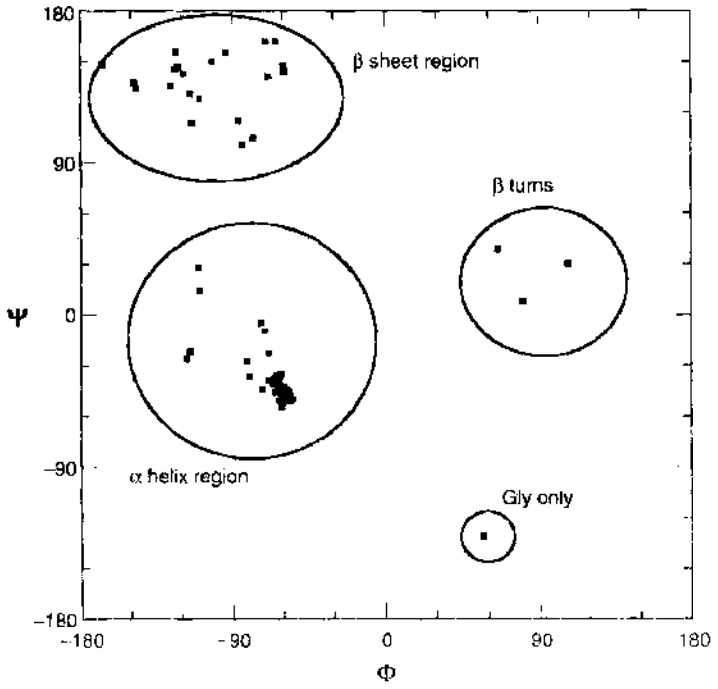


Figure 1. Ramachandran plot of the distribution of Φ/Ψ torsion angles in the C-terminal fragment of L7/L12 ribosomal protein (protein data bank entry 1CTF [5]). There are three regions that are energetically accessible to all amino acids and one further region that is only possible for Gly, as it does not have a side chain.

sequence of a polypeptide chain, and it's native environment, how does it fold to it's native conformation?"

For many practical purposes it would suffice to know just the final result of the folding (i.e., the native structure). This, limited, version of the problem is the realm of protein structure prediction, for which many methods have been developed, mainly based on knowledge of other protein structures, and there even is a (bi)annual contest [12]. In this review, we concentrate on roads to solve the entire problem, as solving will yield the native conformations of target sequences, while simultaneously enhancing our understanding of one of the most fundamental processes in the cell. A number of partial problems have been tackled over the years, stimulated by a number of fascinating findings. The title of Levinthal's famous paper [1] is: "Are there pathways for protein folding?". This seemed, at the time, the only logical way around the Levinthal paradox. If a protein can not sample all possible conformations, it must, somehow, follow a programmed "pathway" from the denatured (unfolded) to the native state, which can be written as a simple chemical reaction:

$$U \leftrightarrow N \tag{1}$$

As any chemical reaction, this one has an equilibrium constant k_{eq} related to the fraction native *versus* unfolded protein *under specified environmental conditions*:

$$k_{eq} = \frac{[N]}{[U]} = \exp -\frac{\Delta G_{eq}}{(k_B T)} \tag{2}$$

where $[U]$ and $[N]$ are the fractions of unfolded and folded protein, respectively, ΔG_{eq} is the Gibbs free energy of folding, k_B is the Boltzmann constant and T is the temperature. Furthermore, the reaction has a rate constant k_{act} related to the activation energy ΔG_{act} (Fig. 2).

This model was backed by the finding that folding, at least for small single domain proteins, is cooperative [13, 14], that is, it seems to fold at once. The evidence for this is mainly that the kinetics of folding, as monitored by different probes, could be described by a single exponential [15]. Subsequently, sensitive, time-resolved experimental techniques, such as nuclear magnetic resonance (NMR) H/D exchange were introduced [16] to study proteins *during* folding. One of the main incentives for this work was the discovery of metastable intermediate (I) "conformations." Because these intermediates were neither correctly folded (N) nor completely denatured (U), they were thought to be compact structures, with incorrect internal packing, and hence names "molten globules." At one stage it was postulated that such intermediates might be obligatory structures [17, 18], that is, each folding reaction would have to pass through the intermediate in order to reach the native state

$$U \leftrightarrow I \leftrightarrow N \tag{3}$$

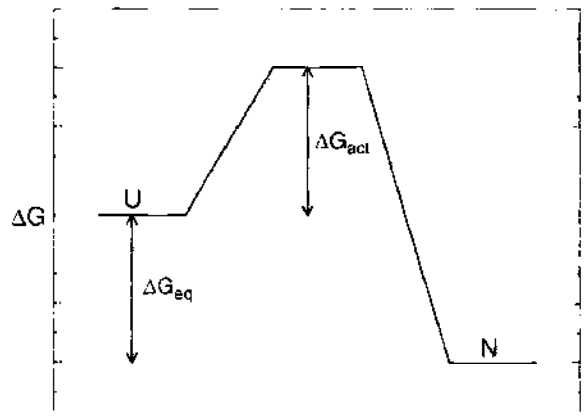


Figure 2. Schematic diagram of energetics of protein folding, according to Eq. (1). The activation energy ΔG_{act} determines the rate of folding, while the equilibrium energy ΔG_{eq} determines the ratio of unfolded and folded proteins.

Such a reaction can obviously described by multiple energies and rate constants (Fig. 3), and, moreover, be generalized to

$$U \leftrightarrow I_1 \leftrightarrow I_2 \ldots I_n \leftrightarrow N \tag{4}$$

An excellent review of protein folding kinetics is given by Schmid [15], while a review of the importance of the molten globule was published in the same book by Ptitsyn [19].

During the 1990s, a so-called new view of protein folding became popular (see [20] for a comparison of the classical and the new view), in part due to theoretical insights [21–23]. More and more researchers are now accepting the position that there may actually be multiple ways for a protein to fold [24]. The theoretical contribution referred to above, is the so-called folding funnel (described in somewhat more detail in Section 2.2). This comprises nothing more than a visualization of a simplified (two-dimensional) energy landscape, with a central well representing the energy minimum, corresponding to the native state of the protein. Hence, the "new view" consists of nothing more than a restatement of the "thermodynamic hypothesis" supported by Anfinsen [7], using the obvious fact that the energy is a continuous function of the coordinates. Finally it should be noted that protein folding *in vitro* is not the same as protein folding *in vivo* [25]. In the cell there are many different factors involved in protein folding, such as chaperon proteins. Another difference lies in the chemical environment, that is, the cell interior is a big "soup" with many different proteins, and many small organic molecules diffusing around. In this context, it is interesting that new studies are appearing that investigate the so-called crowding effects on folding and stability of proteins [26, 27].

The protein folding problem has attracted attention of theoreticians from early on. One of the first published simulations was that by Levitt and Warshel [28] on the folding of BPTI, which was modeled using a chain of beads, with each bead representing an amino acid. After that, many different methods have been used, some of which will be described in more detail below. However, the main focus of this review is on atomistic scale modeling of protein folding. The protein folding problem has been called "the holy grail of biocomputing" [29], as its solution seems to come closer all the time yet remains far away. Nevertheless, classical molecular dynamics simulations on the atomic scale currently seem to be the only possible compromise between accuracy and computer time requirements, that will enable complete simulations of protein folding. More accurate methods, including any quantum chemical method, are too expensive computationally, while more crude methods (e.g., using a grid representation of a protein) most likely are not accurate enough to be generally applicable. Before we set out to review the current literature in the field, it may be beneficial to write down what we see as requirements for a "real" solution of the protein folding problem:

1. The end point of the folding process should be the experimentally determined native structure, under the specified environmental conditions.

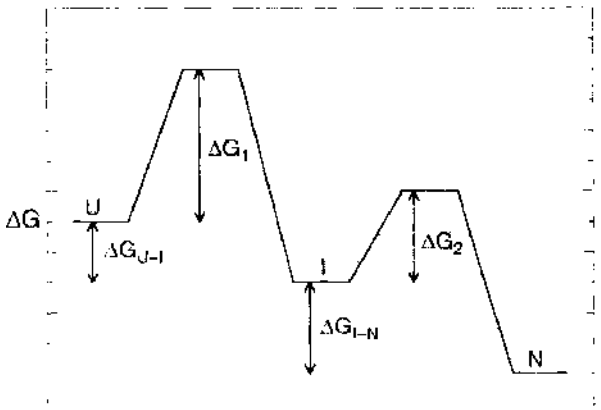


Figure 3. Schematic diagram of energetics of protein folding through an intermediate, according to Eq. (3). The activation energies ΔG_1 and ΔG_2 together determine the rate of folding, while the equilibrium energy ΔG_{eq} determines the ratio of unfolded and folded proteins.

2. The rate of folding should correspond to the experimental kinetic data, if available.
3. Folding *in silico* should be reproducible.
4. The correct structure should be predicted without prior knowledge, (i.e., based on an energy criterion) [7].

2. METHODS AND MODELS

Many different methods and models have been applied to obtain insight in aspects of the protein folding problem. Even some very crude models, based on discrete positions, have been shown to yield new insight. We review here, very briefly, two of these methods (Sections. 2.1 and 2.2) before we proceed with what is the main topic of this review, atomistic modeling of protein folding (Section 2.3). Examples of the latter method are given in Section 3, followed by conclusions and outlook (Section 4).

2.1. Grid-Based Simulations and Complete Enumeration Methods

In order to simplify the search problem, proteins have been described using amino acids as the fundamental unit, on a discrete lattice. To also reduce the combinatorial problem, Dill et al. have worked with proteins containing just two residue types, H (hydrophobic) and P (polar). Obviously, H and P residues attract their own kind, while H and P do not like each other. Furthermore, P residues are happy to be located at the outside of the molecule, whereas H residues are not. These models have been implemented initially on two-dimensional grids, and later on three-dimensional grids as well. One of the most interesting findings coming from this work is the hydrophobic zipper [30], a proposal in which hydrophobic residues in close contact sequence-wise, aggregate, and hence can form a folding nucleus, or scaffold upon which other residues can fold. This mechanism provides for *a* way around the Levinthal paradox [1]. Further work, like the prediction of a four-helix bundle (also known as a Leucine zipper) [31], and the study of side chain packing [32, 33] are at least in part based on the insights from the simple grid models. Some of the grid work, based on simulations of 16-residue HP polypeptide on a $3\times3\times3$ cubic lattice, has been used to implicate (among other) that the native state of a protein is characterized by a well-defined energy minimum [34–36]. This has however been proven wrong by other complete enumeration methods [37], and it is also in contrast with molecular dynamics simulations which have suggested that the global energy minimum consists of a "wide shallow" basin (i.e., not at all well defined minimum) [38]. Much of the grid-model related work has been reviewed some time ago [39–41], and this need not be repeated here.

2.2. The Funnel Approach

In the early 1990s, Wolynes et al. introduced the concept of the *folding funnel* [21–23]. This comprises a simplified description of the energy landscape of a protein in solution, referring back to the "thermodynamic hypothesis," strongly supported by Anfinsen, that the native state of a protein is identical to the Gibbs energy minimum, given the environmental conditions (pH, salt, membrane/solution etc.) [7]. The main conceptual breakthrough of the funnel approach lies therefore in the realization that there must be multiple pathways for a protein to fold to the native state [24]. Although it is quite obvious that there must be many unfolded protein conformations with identical energies, the native state of HPr (the histidine-containing phosphocarrier protein [42]) was described to be a shallow basin, rather than a deep funnel [38]. This points to a major weakness in the funnel model: that of oversimplification of a complicated problem. The appealing, but schematic, figures of energy landscapes can be held for more revealing than they in fact are. On the other hand, a correct description of an energy landscape (i.e., in 3N dimensions for a protein consisting of N atoms), immediately yields the equilibrium populations of the respective energy wells, but is impossible to generate due to the Levinthal paradox. Nevertheless, it is good to see that a number of researchers have now started to generate more quantitative energy landscapes, although approximations are still necessary. Some of these attempts have been covered below in Sections 3.2, 3.4, and 3.7.

2.3. All Atom Molecular Dynamics Simulations of Protein Folding

Early computer simulations used very crude models of proteins, using amino acids represented by just one or two beads [30–36, 39–41, 43]. Computer simulations of protein folding were focused on the protein itself, and limitations in computational power urged the need for simple models. Therefore, water was not taken into account in these simulations. With increasing computational capacity, the use of atomic scale models including water is the obvious route to higher accuracy.

Water is the natural environment for all proteins, also partly for membrane proteins, and its role in protein folding is not trivial. However, an all atom molecular dynamics simulation of the folding of a complete protein is still an expensive operation and the use of simpler, but sufficient models is desirable. For each simulation, an appropriate solvent representation can be taken [44]. Solvent can be treated as a bulk medium, affecting the behavior of the protein through its dielectric properties. Or, it might have an influence on, for example, a protein embedded in a lipid membrane. In the first case, water can be treated implicitly if one wants to save computer time in order to concentrate on the solute. The protein is treated as if being in the gas phase; water is present, but not participating in, for example, hydrogen bonding. This implicit solvent treatment has been developed in many varieties, but the most commonly used model is the generalized born/surface area (GB/SA) model [44–48].

A number of studies have adapted this all atom protein with continuum solvent model. The object of study is usually a small (part of a) protein that represents not more than one domain, or a designed mini protein that contains a typical fold for proteins like an α-helix or β-sheet (or not even that). Although the proteins under study are single domain structures, unlike most proteins, studying protein folding by looking only at a small part of it is justified if one realizes that the basic structural unit of a protein is formed not by the polypeptide chain, but by its sub domains. These geometrically separate entities are in most cases less than 20 kD in weight [49]. Often, factors like a cysteine bridge or a metal ion fixate the structure, in other cases binding of a ligand is required. For molecular dynamics simulations of protein folding, structures without these cofactors are preferred. The smallest natural domains that fold autonomously are 32 to 40 residues, but even smaller proteins have been engineered [50, 51].

It has been deduced from experiments that the upper limit for secondary structure formation of an alpha helix is 80–200 ns [52]. It must be noted, however, that it can be longer or shorter before a final structure appears. Helix formation is substantially faster than β-hairpin formation, which has been measured in the microsecond range: a three stranded β-sheet peptide was folded in 10 μs [53]. It was demonstrated as well, that folding kinetics of an α-helix rich peptide is less dependent on the viscosity of the solvent than β-hairpin formation [54, 55]. A general remark on experiments needs to be made here: protein folding kinetics seems to become faster with the improvement in accuracy of measurement, thus proteins are now believed to fold faster than what was thought 10 years ago.

Computational studies of protein folding have used different approaches. In what follows, the focus will be on those simulations that deal with protein folding and the emphasis will be on folding events. Terms like radius of gyration (molecule size), C_α root mean square deviation (RMSD; equality to native structure), number of native contacts, and secondary structure formation are best used for such a structural description (for an example of all these measures, comparing the stability of potential folding nuclei, see [56]).

Molecular dynamics simulations are done in different manners. Starting from extended structure, one can simply follow what happens below the melting temperature using a modern parallel computer or cluster and appropriate software. The timescale of folding, however, is still too long to witness complete protein structure formation. Moreover, molecular dynamics simulations rely on many parameters, the suitability of which for folding simulations was not tested until very recently [57]. One can also work the other way around. Proteins un- and refold faster at higher temperatures. An approach to study protein folding is to probe unfolding of proteins and (limited) refolding events. The assumption is that by clustering conformations you can characterize transition and intermediate states. Sampling proteins that are folded to a different degree may give an idea of possible folding pathways at a low CPU-cost. This way, the folding behavior of several protein has been studied.

In order to get an idea of the complete folding pathway, energy landscapes can be reconstructed by sampling lots of differently folded states. The assumption is that folding passes the easiest energy barriers from extended (high energy) structure to the final native (lowest energy) structure. With the map of the energy landscape in hand one can deduce a folding pathway. Running a single structure simulation may end up being trapped in a low energy—but not the native—state. The replica exchange molecular dynamics (REMD) technique tries to circumvent this by running a set of simulations at varying temperature and at certain times exchanging structures from one run to the other. Any system that might be trapped in a low energy conformation can get enough energy to escape from it and go on folding toward the native structure. A third variation involves the use of an immense amount of processors. Proteins as an ensemble fold too slow to simulate right now, but a very tiny percentage of proteins may fold very fast. This is an unlikely event, but given enough processors, it is possible to capture such an event. Selecting the fastest folding structures for further simulation, successively searching again for the fastest folding strands and repeating this simulation/selection-cycle several times, can generate a complete folding trajectory in the equivalent order of microseconds [58]. Because the folding is not witnessed in one run, it is better to speak of generations than folding time, but one can estimate how much time one generation is and convert that to a better interpretable microseconds scale. Whichever technique is used, basic guidelines for proper simulations have been published [59].

All approaches have their advantages and their drawbacks, but each of them can clarify something about the nature of protein folding. First some basic peptide systems will be discussed, like the α-helix and β-hairpin. Folding of bigger structures is being discussed afterward. The various approaches in simulation and the use of implicit versus explicit solvent will be evaluated.

3. EXAMPLE PROTEINS

3.1. α-Helices

The α-helix is the fastest folding secondary structure element of a protein. The relatively small size and clear definition of native structure made this an early object of study for protein folding [60–64]. There are several α-helix peptides that are particularly suitable for molecular dynamics studies. Alanine-rich peptides are an excellent object of study, because of their high intrinsic helical propensity in water [65–67]. The folding behavior of the peptides A_{21} and F_s was studied using molecular dynamics [48, 58, 68]. F_s is a peptide with the sequence Ac-A_5-$(A_3RA)_3$A-NH, which is known to form a highly stable α-helix structure in water [69, 70]. The folding of F_s peptide in implicit water has been compared to that of a pure poly-A chain of 21 residues [58]. The reason for this is the assumption that Arg is a nucleation blocker when compared to alanine. In implicit solvent, poly-Ala folds—as to be expected—very fast into an α-helix: 18 ($\pm$8) ns and has a large helical content of 17.3 residues. The F_s peptide on the other hand has a slightly slower folding time (82 ± 60 ns) and has a lower helical content (15.1 residues). The alanine-rich N-terminus of F_s folds much quicker: in 15 ± 10 ns. It seems that Arg residues are responsible for differences in folding rates.

The folding equilibrium of these α-helical peptides has been studied in explicit solvent as well [68]. Applying replica exchange molecular dynamics, extensive simulations (1.5 μs) were performed over a wide range of temperatures (270 to 550 K). The explicit treatment of water shows that peptides containing Arg are considerably more stable than peptides containing only alanine. The explanation for this is in the shielding of backbone hydrogen bonds by the side-chains of Arg. They make it a lot more difficult for the solvent molecules to attack and break the main chain hydrogen bonds [71, 72]. Local opening and closing of the backbone hydrogen bonds can be caused by thermal fluctuation. The Arg side chain partially shields the carbonyl oxygen of the fourth amino acid upstream from the water. Furthermore, the guanidinium side chain interacts with the carbonyl oxygen atom and stabilizes the folded conformation [68].

Stabilization of α-helix by side chains is not limited to Arg residues. In a study of helices from the central domain of smooth muscle caldesmon, side-chain interaction also plays an

important role [73]. The protein contains 10 repeats of a 13-residue motif, with many charged side-chains. From circular dichroism spectroscopy studies, it had been suggested that these helix structures are stabilized by salt bridges between positively and negatively charged side chains of residues four residues down in the sequence [74]. In up to 100 ns of simulation in explicit solvent of a single, double, and triple α-helix repeat of this protein, it was indeed found that this type of salt bridge is formed [73]. Out of the possible four salt bridges, two are preferred. The N-terminal salt bridges are more stable, whereas the C-terminus has less helical content throughout the simulation. Notably, the strongest salt bridge is formed by Glu3-Lys7, which is the only side chain pair with an N-C acid/base orientation, allowing for stabilization of the α-helix dipole [75]. To a lesser extent, α-helix stabilizing salt bridges are formed with the third residue. Side-chain interactions with the fifth consecutive residue are found as well, but this plays a role in π-helix stabilization. α to π-helix transition in a single α-helix structure can occur upon the insertion of water into α-helical hydrogen bonds. This has been identified before in molecular dynamics simulations of synthetic peptides [76], but could be the result of artifacts due to force fields [77].

The detailed simulations in implicit and explicit solvent incited a comparison of folding behavior in GB/SA and explicit solvent [48]. Replica exchange molecular dynamics for the formerly mentioned peptides F_s and A_{21} was done with two different AMBER force fields (pram94 [78] and pram-mod [68]). The latter reproduced the thermal denaturation curves in better agreement with experiment. Simulation shows that F_s peptide is more helical than A_{21} in explicit water, whereas in GB/SA simulations the F_s peptide is less helical than A_{21}. Changing force field from parm94 to parm-mod decreases the helix content in explicit solvent, whereas it increases helix content in continuum solvent.

Though the aim of most simulations of α-helices is to understand protein folding better, it also highlights the importance of using a correct native-like environment. The final structure may be reproducible in implicit solvent, but there is a dependency on the amino acid sequence. For *ab initio* prediction of protein folding, this is not enough. The force field and solvent model together should be able to yield a correct protein folding pathway in order to obtain dependable results for the final structure.

3.2. The β-Hairpin

Like the α-helix, the β-hairpin is a convenient system to study, as it is fairly small and the correct structure is easy to identify. Hence, this is a good opportunity to study another part of the protein folding problem. The structure of the C-terminal β-hairpin of protein G has been investigated experimentally to a great extent, allowing detailed comparison with molecular dynamics studies [54, 79–85].

This β-hairpin consists of 16 residues and one turn. The plane of parallel strands shows a slight twist, and hydrophobic residues tend to be located at one side, which can be seen as a kind of packing (Fig. 4). Thermal unfolding of this peptide showed that the kinetics behave as in a two-state system. It was also shown that the complete folding happens on a scale of several microseconds [54]. This timescale is still beyond reach of normal molecular dynamics simulations. A full atomic-scale implicit solvent simulation was first done in 2001 [87]. In this study, several important questions were addressed, like the nature of the on-pathway intermediate as observed in some of the studies of the peptide; the existence of off-pathway intermediates that may hinder hairpin formation; the relative importance of interstrand hydrogen bonds in comparison to hydrophobic core formation; the order of structure formation; and the heterogeneity of the folded state [87, 88]. The method of sampling—ensemble dynamics—means that two kinds of structures are generated: a good number (40) of structures in folded conformation and a large group of incompletely folded structures. This large group contains almost any kind of imaginable structures, including some short-lived α-helices and mixed α-β-structures. It has to be mentioned that these last ones were formed in really small numbers and are most likely a random event. They do not seem to be relevant on the folding pathway. The semihelical structures could, on the other hand, lead to a final folded conformation. The helical turn formation appeared to be independent of residue identity in this fragment. All structures with more than five residues in helical conformation did not

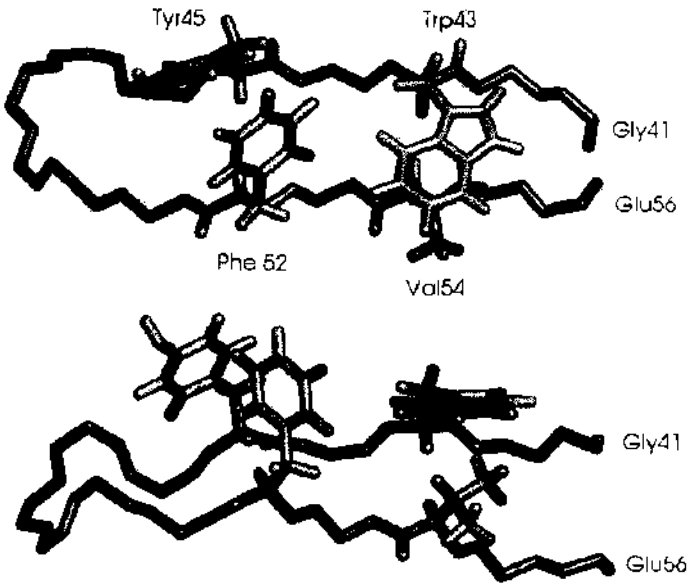


Figure 4. The second β-hairpin of the B1 fragment from Protein G from *Staphylococcus aureus* is a well studied peptide structure. The backbone of the molecule is shown from two sides. It has four residues that form a hydrophobic core (shown as sticks): Trp43, Tyr45, Phe52, and Val54. Images created with Pymol [86] from Protein Data Bank entry 1GB1.

have any β-sheet character. In total, the simulations gave eight different, completely independent folding trajectories [87]. Focusing on one of them, the first thing that happens is a collapse from extended conformation to a more globular form. At this moment, three temporary hydrogen bonds are formed, and a first small hydrophobic interaction occurs. This hydrophobic interaction is a key feature for further hairpin formation. The collapse changes the radius of gyration of the peptide (close) to its final value and the RMSD reaches a plateau that exists for about 300 ns (30 generations, to be precise, where 1 generation in simulation corresponds to approximately 10 ns). This gives the opportunity to create and break hydrogen bonds in various combinations, mainly close to the hydrophobic interaction. After more than 41 generations (corresponding to approximately 400 ns), the full hydrophobic core is formed and the key hydrogen bonds are present. This marks the completion of the hairpin formation. From now on the number of hydrogen bonds may fluctuate, but the two key bonds rarely break and the structure has all the features of a β-hairpin. The other complete folding pathways show mostly the same characteristics: early hydrophobic contact with the formation of two central hydrogen bonds, followed by strand parallelization and a fluctuating number of additional hydrogen bonds.

Other ways of β-hairpin folding have been reported. Folding, starting with the formation of the β-hairpin, followed by zipping of the strands, has been observed in Monte Carlo and molecular dynamics simulations [54, 89, 90]. A second mechanism involves the previously described hydrophobic collapse, followed by hydrophobic core optimization and forming of the hydrogen bond network [91]. A third mechanism can be described as the sliding of one strand over another [92]. This too starts with a rapid collapse, followed by appearance of temporary hydrogen bonds that are not present in the final structure. The number of non-final hydrogen bonds then increases and several β-loop structures are tried. The hydrophobic interactions disrupt the non-native hydrogen bonds and force two strands toward the native form by sliding along each other and then almost instantly forming the native hydrogen bonds.

An important point is whether the GB/SA model can be applied to study these kind of systems. The answer to that question is not straightforward. An interesting comparison study has been done addressing this issue, in which the free energy landscape of the same β-hairpin was explored and compared in implicit and explicit water [93, 94]. Implicit and explicit solvent simulations were carried out with replica exchange molecular dynamics under the same conditions. At a range of temperatures, replicas of the system run parallel and on a regular basis the conformations of neighboring systems are exchanged. In the case of

implicit solvent, a GB/SA model was used. The implicit solvent simulations had 18 replicas simulated from 270 K to 690 K. In the energy landscape of these simulations the native state does not have the lowest energy: the amount of β-hairpin conformations is nearly half the amount found in the explicit solvent simulations or experiments [93]. The overall energetic map shape has a similar contour, but in implicit solvent fewer hydrogen bonds are preferred. This results, for example, in a hydrophobic residue lying on the outside of the chain, whereas in explicit solvent the hydrophobic residues form a well packed core (Fig. 5).

Hydrophobic packing turns out to be the key difference between implicit and explicit solvent, in the β-hairpin: in continuum solvent, one of the four hydrophobic residues—Phe52—is expelled from the hydrophobic core. At the same time, a salt bridge is formed between charged residues in the continuum solvent model, whereas in explicit solvent these residues are extended into the solvent. Therefore, it seems that the balance between electrostatic forces and hydrophobic interaction is not correct in the continuum solvent. A correct description of the interaction with the environment seems to be a prerequisite for correct folding of this system. In addition, the existence of salt bridges in the implicit solvent stabilizes the hairpin structure such that unfolded states are prohibited [88].

3.3. Other Short Peptides

For a long time, short peptides in solution were thought not to have any measurable structure [95, 96], even so they were studied quite extensively, because they were suspected to be important as folding nucleation sites [97, 98]. However, not many were found with detectable structure. The possibly shortest peptide that does have a distinct conformation, as determined by NMR, is only four residues long, with sequence Tyr-Thr-Gly-Pro. The main interaction is between the side chain of the Tyrosine and the amide proton of the Gly [99]. Interestingly, the results, including the folding kinetics (which consists of a mere opening and closing motion) could be reproduced in a 1-ns molecular dynamics simulaton [100]. Another example of a short peptide with a long history of structural studies is enkephalin. Actually, there are two naturally occurring sequences consisting of five amino acids: Met- and Leu-enkephalin (sequence Tyr-Gly-Gly-Phe-Met/Leu, respectively). Computer simulations of Leu-enkephalin in different solvents (water and dimethyl sulfoxide) show that there is no detectable structure in water, while there is a clear hydrogen bond or salt bridge (depending on the protonation state) between the N- and C- terminus of the pepide [101]. The difference in structure of the peptide in different solvents demonstrates that it is crucial to model the solvent correctly. Similar results had been found in earlier simulation of α-helices in trifluoroethanol, a helix-stabilizing agent [102, 103]. The topic of structure in very

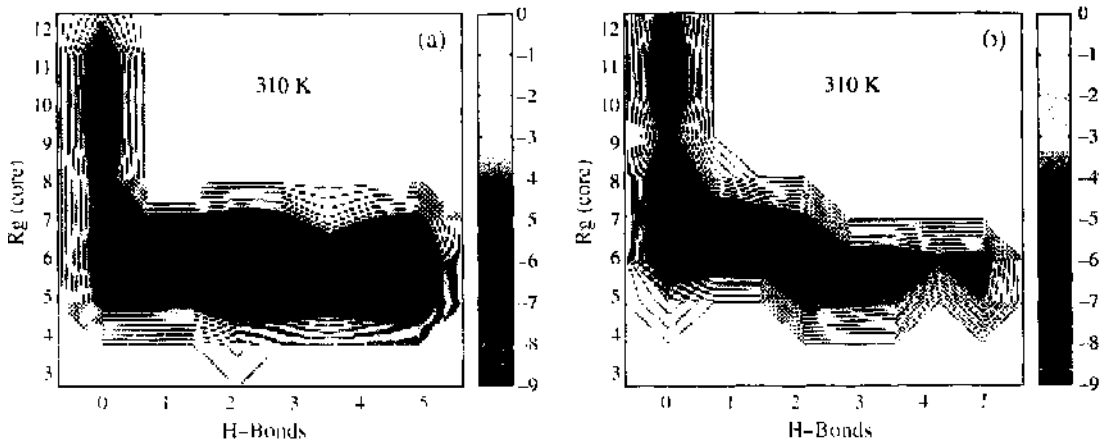


Figure 5. Comparison for the β-hairpin of the free energy contour maps *versus* the number of β-sheet hydrogen bonds and the hydrophobic core radius of gyration R_g for explicit (a) and implicit (b) solvent simulations at 310 K. A hydrogen bond is counted if the distance between two heavy atoms (N and O in this case) is less than 3.5 Å and the angle N-H...O is larger than 150°. The overall energy contour is similar in both models, but in implicit solvent fewer hydrogen bonds are preferred [93]. Reprinted with permission from [93]. R. Zhou et al., *Proc. Natl. Acad. Sci. U.S.A.* 98, 14931 (2001), © 2001, National Academy of Sciences, U.S.A.

short peptides is naturally related to that of the structure of unfolded proteins. Although it has been problematic to find residual structure in unfolded/denatured proteins [104, 105] it is nevertheless crucial to understand this "other half of the folding equation" [106]. This renewed interest in unfolded proteins has also sparked new measurements of basic properties like random-coil chemical shifts [107, 108], which can be reproduced reasonably well in simulations [109].

3.4. β-Sheet Peptides

β-Sheets are an important structural feature of proteins. While complete folding of large β-sheets is still not feasible *in silico*, work has been done on smaller systems, like the β-hairpin and mini-peptides containing a small β-sheet (e.g., [110]). Often, it involves a designed structure, or at least a part of a bigger structure. One such system is a small three-stranded β-sheet peptide consisting of 20 residues [53]. This protein is monomeric in aqueous solution (as proven by NMR experiments) and it contains two Gly residues that are part of the turns in its β-sheet conformation. This GS-peptide exists in a single stable form at 283 K. Molecular dynamics studies of this protein focus on sampling unfolding and refolding events using implicit solvent [111, 112]. Simulations were carried out at both 360 K and 330 K. Starting from the folded, three-stranded β-sheet structure, several unfolding and refolding events have been sampled. There appear to be two folding pathways, similar at both temperatures. Both went through a crucial β-hairpin formation stage, followed by zipper-like β-strand formation. Because the NMR structures were not well defined, starting structures for simulation were obtained by first doing a 100-ns simulation at 300 K (at which temperature a three stranded β-sheet was formed within 75 ns). A second 200-ns simulation at the same temperature was used for defining native contacts. Thus, this is not a true *ab initio* folding simulation, but it gives an impression of possible folding mechanisms.

Water was modeled, based on solvent-accessible surface where the surface is calculated by assuming a solvent radius of 1.4 Å. Further adaptations included a linear distance-dependent dielectric function for electrostatic interactions, and charged side chains were neutralized. The speed of (un)folding would be significantly slower (and possibly closer to the in vitro observed folding time) if the viscosity of water had been considered [112] and if the simulations had been done at lower temperature. Still, this protocol correctly reproduced the native structure.

In a distributed computing approach in implicit solvent, many folding events of the GS peptide have been sampled [113]. At its equilibrium temperature of 330 K, 72 folding and 73 unfolding events were observed in a total of 12.6 μs simulation time. Average folding happened in 83 ns. Of the 14,300 simulations of 1 ns at 330 K starting from extended structure, 1,300 were continued to 6 ns. Native contacts were defined as present at least 50% of the time in a 200 ns/300 K simulation. The fastest folding trajectories were characterized by almost instant collapse to correct native contacts, unlike the equilibrium simulations, where metastable states occur and unfolding happens before final folding. Interestingly, the very fast folding pathways traverse the highest free energy region on the free energy surface. There seems to be no preference for a specific folding pathway in the 1-ns simulations.

A similar small peptide, but with a different amino acid sequence is the designed three-stranded β-sheet structure betanova [114]. The structure has been characterized by NMR and its folding kinetics are known through circular dichroism and fluorescence spectroscopy. Thermodynamic stability is optimal at 278 K. Molecular dynamics simulations have been done with explicit and implicit water usage in order to compare both solvent models [115, 116]. The native state was defined by two molecular dynamics simulations of 2 ns at 275 K of the native structure. There is a large overlap in native contacts and the backbone root mean square deviation is fairly stable in both solvents at 2–2.5 Å (1.9 and 2.2 Å compared to NMR structure respectively), which means that the GB/SA model is maintaining the native structure. The free energy surface was reconstructed for both solvent models (Fig. 6).

Though similar at first sight, there are some differences. In explicit solvent, there are three basins (unfolded, collapsed and native state) of which the lowest energy one corresponds to the native state, having a high share of native contacts (60% to 85%) and low radius of

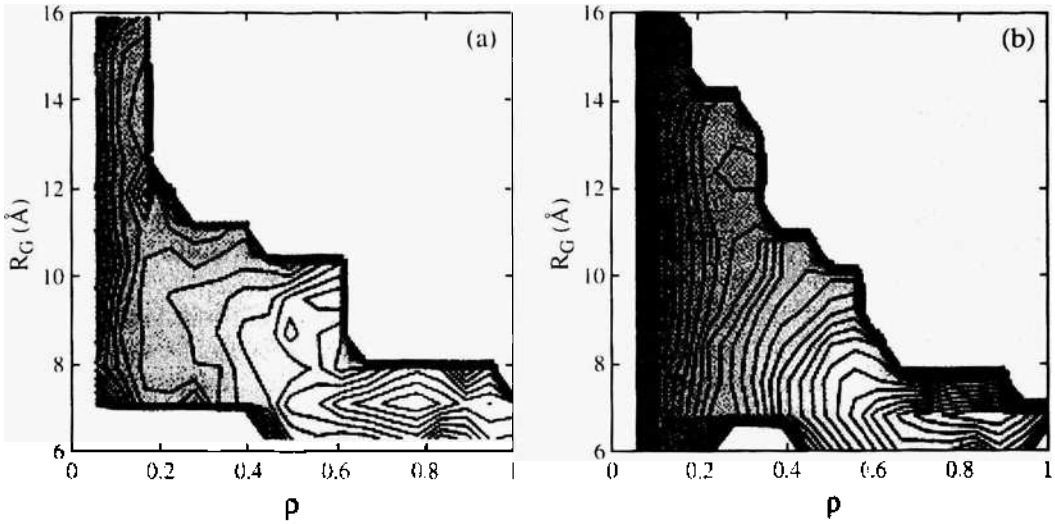


Figure 6. Potential of mean force as a function of radius of gyration, R_g, and fraction of native contacts, ρ in β-sheet peptide betanova. The contours are drawn every 0.5 kcal/mol: (a) with explicit solvent at 275 K; (b) with GB solvent at 275 K. Reprinted with permission from [116], B. D. Bursulaya and C. L. Brooks, *J. Phys. Chem.* 104, 12378 (2000). © 2000, American Chemical Society.

gyration. The folding of betanova in explicit solvent happens in two stages: first there is a collapse to a structure with a medium radius of gyration and some native like contacts, then it enters the native basin. In implicit solvent at the same temperature, the free energy landscape of betanova has a different shape. It is much more diagonal, suggesting a smooth transition from unfolded to native state. There seems to be no intermediate state and partially folded states are easily disrupted. The stability of the native state is overemphasized compared to the explicit solvent simulations.

3.5. The Villin Headpiece

α-Helices and β-sheets form the majority of structure elements present in proteins. Though simulations of normal-sized proteins are not yet feasible at the time of writing, there are some proteins that are on a mini-scale. The Villin headpiece subdomain is a native sequence of only 36 amino acids. It was (and is) the first protein for which 1 μs of true folding was simulated in an explicit water environment [117]. It contains three helices at residues 4–8, 15–18, and 23–30. Between these helices, there is a loop from residues 9 to 14 and a turn from residues 19 to 22. Its main feature is a closely packed hydrophobic core (Fig. 7). In Duan and Kollman's simulation [117], a two-phase folding process was observed. Starting from an extended structure the folding involves a "burst" phase in which there is a steady rise in native contacts and helical content, up to 45% and 60% respectively. Also, the solvation component of the free energy is decreasing to the level of the native structure. This is happening on a time scale of 60 ns. The radius of gyration varied from 16 Å, representing the extended structure, to 8.7 Å, for the most compact structures. The native structure has a radius of gyration of 9.4 Å. The radius of gyration reaches a stable level after 150 ns, while the main-chain RMSD varied between 12.4 and 4.5 Å. On the whole, the first phase of folding can be seen as a hydrophobic collapse. The second phase is a slower process, with helical content dropping at first and the radius of gyration decreasing a little. The number of native contacts stays more or less on the same level for some time. Then they gradually rise after 200 ns of simulation. The core of the protein also becomes less accessible during the second phase. In this simulation, the side chains of the amino acids do not reach complete native packing, but on the whole, the Villin headpiece is in a native-like state halfway to 1 μs. Duan and Kollman speculated that their simulation might well result in one of many intermediates of which one could give the final optimal configuration for side-chain packing. An interesting observation is that the formation of native contacts, more than five residues apart in sequence, is very infrequent in 90% of the simulation time, whereas native

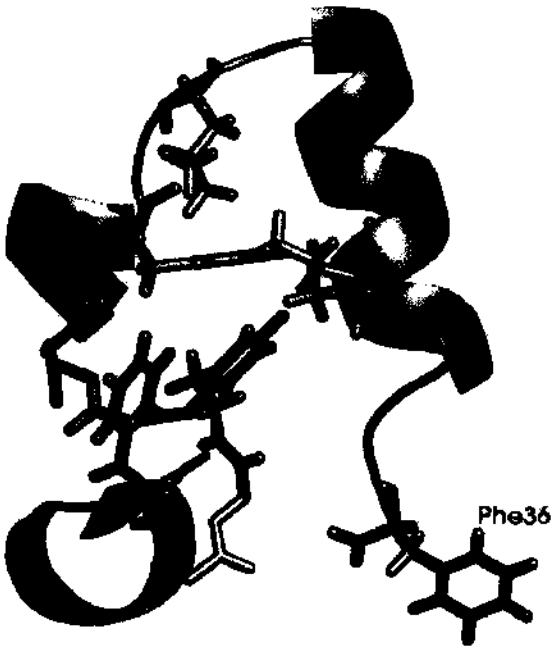


Figure 7. The Villin headpiece subdomain contains 36 amino acids. There are three helices and its main feature is a densely packed hydrophobic core—some of the hydrophobic core residues are shown in medium gray sticks. Phenylalanine 36 is depicted in light gray sticks (bottom right) Source for the figure: entry 1VII [118] from the Protein Data Bank; image created with Pymol [86].

contacts less than four residues apart were formed during more than half the simulation time. The formation of tertiary contacts is therefore likely to be the bottleneck of the folding process [117]. However, this finding could also be explained by the fact that Villin is mainly α-helical, and hence i, i + 4 contacts are the most common ones.

The simulation time of 1 μs is unprecedented, using an explicit solvent representation, but there are some drawbacks about this study. The first is that there is only one simulation and therefore statistical analysis is somewhat difficult. Another is that the cutoff of 8 Å for electrostatic interaction is somewhat below standard [29]. This is of some concern, since we do not know how a cutoff of this magnitude influences this type of folding process. There are, however, many papers in the literature describing the artifacts in simulations due to cutoffs (for example [119–124]).

The folding time of the Villin headpiece subdomain was determined experimentally at 4.3($\pm$0.6) μs [125]. In order to analyze the folding, a true atomic-scale simulation showing the complete folding trajectory would be ideal but requires too much computational time at the moment. In another approach, folding trajectories of Villin were calculated in a world-wide distributed computing project, generating much more folding time [126]. Thousands of relatively short folding trajectories were computed in this manner, using implicit water. This allows for faster calculation of folding events while still simulating some aspects of water behavior, such as frictional effects. Starting from an extended state, the radius of gyration and solvent accessible surface decrease to native state levels within 20 ns. The conformations in the ensemble of collapsed structures are partially with native secondary structure, but with incorrect backbone topology and side-chain packing. A feature that distinguishes these simulation structures from experimental native structure is the Phe36 residue. In simulation it packs with the other phenylalanine residues in the hydrophobic core, whereas in the experimental structure the hydrophobic side-chain is exposed to the solvent. It was predicted on the basis of this observation, assuming this packing is some sort of folding trap from which the Phe36 needs to break loose, that removing the Phe36 side-chain would increase folding speed [126]. In the Duan and Kollman 1-μs simulation, the initial wrong packing of Phe36 is seen in the early stage of the Villin headpiece folding as well [117].

It may well be, however, that this is a necessary stage in correct folding. It was suggested from experiments that non-native hydrophobic interactions are crucial for stabilizing

intermediates on the folding pathway [127]. This can avoid premature aggregation and proteolytic degradation, thus giving the protein time to fold correctly. Following the reasoning of [126], replacing the Phe36 side-chain in the Villin headpiece should, under cellular conditions, lead to poorly formed protein and/or lower active protein evels. Phe36 has been replaced in experiments by Alanine [125]. A decrease in folding ratewas not observed however. Several problems arise when comparing folding rates in vitro and *in silico*. One of them is the arbitrary definition of a folded state in simulations. A more basc difficulty is the usage of the GB/SA model [128].

Åqvist has shown that an electrostatic cutoff of 8 Å fundamentall influences the folding process [129]. Here, the difference between an 8 Å and a 14 Å cutoff resulted in the peptide being much more unfolded. The importance of the relatively short cutoff used by Duan and Kollman may be illustrated by a 200-ns all atom simulation of the Villin headpiece using implicit solvent [130]. The simulation produced ≈4 ns of folding trajectory per day, about the same as the Duan and Kollman supercomputer simulation 4 years before (5 ns), but now on a single PC. Like the former simulation, this one started with a 50-ns equilibrium step at 300 K starting from NMR structure, followed by the same denaturation scheme. Within 20 ns of 300 K simulation, two hydrophobic clusters are formed, creating a pocket where several hydrophilic and polar residues are unfavorably trapped. This results in a rearrangement at 29 ns, which brings the hydrophobic residues to form a loose, native-like, hydrophobic core. Over the course of the next 30 ns, the hydrophobic core tightens and "settles". This "burst phase" time of ≈60 ns agrees remarkably well with Duan and Kollman's simulation. Between 60 and 200 ns, the protein reaches a quasi steady-state. In terms of RMSD, the implicit solvent simulation reaches a lower value than found by Duan and Kollman (3.36 Å *vs* 4.8 Å). Between 30 and 90 ns, the RMSD has a larger fluctuation. This marks the commencement of structural rearrangements. After 170 ns, a third rearrangement takes place, but this is not shown by shifts in RMSD. Two hundred nanoseconds is still far from the calculated folding time of 1 to 10 μs [131]. Nevertheless, the structure shows a native-like folding, and the parallel with the 1-μs simulation is evident. There is a difference in the time needed for folding events. In implicit water simulations, no structural rearrangement are required to expel water. As a consequence, these simulations exhibit smaller short time scale structural fluctuations. Another reason for this is the stabilizing effect of implicit solvent on internal hydrogen bonding. In explicit water on the other hand, the water-exposed hydrogen bonds are frequently attacked and broken, which leads to more significant structural fluctuations. Ideally, the implicit solvent should correctly describe the friction sources and random forces, free energy of solvation and act like a dielectric medium, modeling the electrostatic screening due to solvent particles. This ideal combination rarely happens [130]. Finally it has been shown in native state simulations of Villin, in explicit water and with the particle mesh Ewald algorithm [132, 133] for electrostatic interactions, that most NMR parameters can be reproduced and, moreover, the agreement of simulation trajectory with NMR data was better than that of the NMR structure ensemble itself [57]. This indicates that the force fields are most likely good enough for folding as well, at least when using explicit water and when electrostatic interactions are taken into account properly.

3.6. Trp-Cage

The 39 residue peptide exendin-4 was artificially truncated to the smallest size (20 residues) that still contained a typical protein structure as shown by NMR structure determination [50]. The designed mini-protein is accurately described by its name Trp-cage. Trp-cage consists of 20 residues, and in its center is a hydrophobic core comprising mainly of a tryptophan side chain (Fig. 8).

A folding time of 4 μs has been reported for this peptide, based on laser temperature-jump spectroscopy [134]. There are several computational studies done on this small system [135–138]. In a worldwide distributed folding simulation of the Trp-cage, it was found that a compact unfolded state could be reached in 10 ns [138]. Based on thousands of simulations, complete folding was estimated to take 1.5 to 8.7 μs. A 100-ns simulation in implicit solvent at 300 K, using a modified version of the AMBER force field, resulted only in kinetically

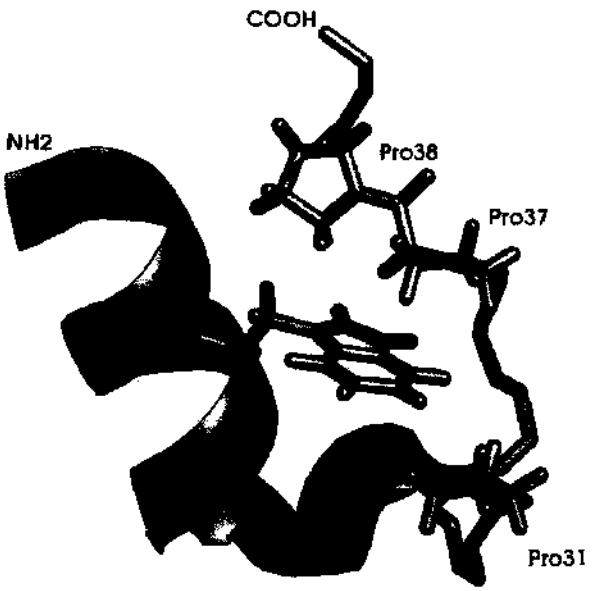


Figure 8. The Trp-cage is a very small and fast-folding protein. The residues involved in Trp-cage core formation are depicted as sticks, with the Trp25 residue in the center of the molecule. Image created with Pymol [86], from NMR structure 1L2Y [50] obtained from the Protein Data Bank.

trapped conformations [135]. These simulation results were very dependent on the starting structure. In order to resolve this issue, the temperature was raised to 325 K and the two subsequent simulations converged to comparable structures when checked with RMSD from the folded structure. The "folding," however, is not easily deducible by monitoring the potential energy change over the 20-ns range. In another 100-ns simulation in implicit solvent, the folding of the Trp-cage could be described at 300 K [136]. The biggest difference with the previous simulation was the use of a different force field. Simulation in [136] started from an extended conformation. There, the RMSD decreased in the first 10 ps from 10 to 4 Å and stayed like this for the next 23 ns. Between 23 and 29 ns, a series of changes in this metastable structure took place, and an almost NMR-like structure appears (RMSD of 1 Å). The folding was also tracked using Nuclear Overhauser effect (NOE) violations. The fraction of NOE violations is 60% initially, when the protein is in its fully extended state. After a steady decrease to 48%, the fraction reduces sharply between 2 ns and 4 ns. To 23 ns it remains fairly the same, and then there is another sharp decrease to below 20%, which at the end of the simulation is about 14%. When increasing the NOE distance restraints by 0.4 Å, the average NOE violation drops to 3–7%. If this is compared to the NMR ensemble, which has an average of 3.6% of NOE distance violations, it can be concluded that the simulated structure is very much like the NMR structure.

Using replica exchange molecular dynamics and an implicit solvent model, an aggregate of 92 ns of Trp-cage folding time was simulated [137]. Starting from an extended structure, the overall topology of the Trp-cage is reproduced. There are, however, severe NOE violations. Five out of eight violations are related to Leu7, which is supposed to be tightly packed against the hydrophobic core, but is more solvent orientated toward the end of the simulation. The N-terminal Asn1 often forms a hydrogen bond with Ile4 in the simulation, whereas in the NMR structure the Asn1 side chain sticks into solution. Further, Ser14 forms an internal hydrogen bond to Gly10 and Gly11 instead of to Asp9. Also, in simulated structures the Arg16 side chain does not pack as close to the surface as in NMR. Apart from the violations, the folded structure is also overly stable. Its melting point is too high (400 K) compared to experiment (315 K). This, however, is to be expected as the implicit solvent model has no temperature dependence. The force field in use [78] is likely to play a part in anomalous stability—especially in α-helices [116]. The GB/SA solvent model is known to overstabilize solvent-exposed salt bridges [94]. How this relates to the mostly hydrogen bond related violations needs to be sorted out, but interaction with the solvent is clearly lacking.

3.7. Protein A

The B fragment of protein A from *Staphylococcus aureu* is a three-helix bundle of 46 residues [139] (Fig. 9). Because a full-scale *ab initio* folding simulation is still too computationally expensive, investigations into the folding mechanism have been done by artificially constructing the free energy landscape [140]. The system was placed in a periodic boundary box with explicit solvent (more than 16,000 atoms), using a modified AMBER force field. A total of 82 simulations applying replica exchange molecular dynamics at a range of temperatures (from 277 to 548 K) were performed, starting from different conformations. The temperature variation in the replica exchange method implies that the time history of the replicas is not directly related to the folding pathway(s) at a single temperature. However, it can give a description of the order of events during folding. At a low temperature, the energy surface shows a large folding basin where two minima are close to each other. One of these corresponds to the native state, the other one is nearly folded and has a hydrated core. These areas are equally populated at 387 K. At very high temperature, there is only one energy minimum, corresponding to the unfolded state. At an intermediate temperature, there are two low energy basins separated by a free energy barrier. The melting temperature (defined as the temperature where the folded and unfolded state are equally populated) lies at the relatively high temperature of 421 K.

It appears that unfolded states are not completely devoid of native structure. Notably, some of the early formed native contacts in the loops have to be rearranged when the system is near the transition (melting) region. Secondary structure formation is coupled with solvent exclusion and core formation.. The third helix forms early. Unfolding happens first for helix I, then helix II, and finally helix III. The unfolded ensemble has a stable β-hairpin at helix I. The stability of helix III seems to come from strongly interacting side chains. The formation of α-helices is not gradually building up: turn formation occurs quite suddenly and the turns partially reverse their formation near the transition state to allow tertiary structure to form through packing.

3.8. Chymotrypsin Inhibitor 2

Chymotrypsin inhibitor 2 has one α-helix and three β-strands in one sheet, where the helix is packed against the sheet (Fig. 10). It has been determined experimentally that it folds according to a two-state model [141, 142]. Molecular dynamics simulations of unfolding of chymotrypsin inhibitor 2 at 348 to 498 K yields similar unfolding events [143]. If taken in reverse this shows that—independent of temperature—first there is a nucleus forming, with parts of the helix and β-sheet in it. The structure then collapses to it final form. Its "folding" was deemed similar to nucleation-condensation [144].

Of course one would like to see the folding events happening in chronological order. The size of chymotrypsin inhibitor 2 (64 residues) has been a severe barrier in folding simulation

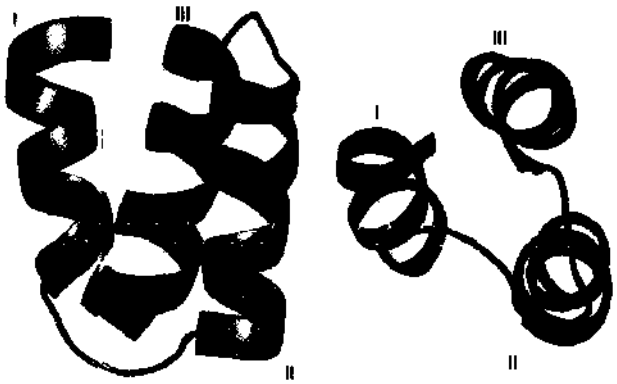


Figure 9. NMR structure of the three helix bundle in fragment B from protein A from *Staphylococcus aureus* in a cartoon representation. Helices are numbered I, II, and III in side- (left) and top-view (right). Almost all hydrophobic side chains are packed together in the space between the helices. Image created with Pymol [86] from Protein Data Bank entry 1BDD [139].

Figure 10. Chymotrypsin inhibitor 2 has four strands in a β-sheet and one α-helix. View from two different angles. Image created with Pymol [86] from Protein Data Bank entry 3CI2 [145].

experiments. In an implicit solvent simulation [146] a special approach has been taken by applying targeted molecular dynamics. With the help of a constraint, reaction paths are calculated between two different conformations, thereby continuously decreasing the distance to the target distance. The native structure can successfully be generated this way (Fig. 11). Typically, the α-helix is formed first, followed by the formation of $\beta 3$-$\beta 4$ contacts. In this transition state, the secondary structure is still mostly absent. The α-helix is not in a stable conformation yet and is disrupted before the final helix shape appears; the β-sheet is completed afterward. The main driving force in the beginning of the folding process is the burial of hydrophobic residues, as illustrated by a rapid decrease in hydrophobic solvent

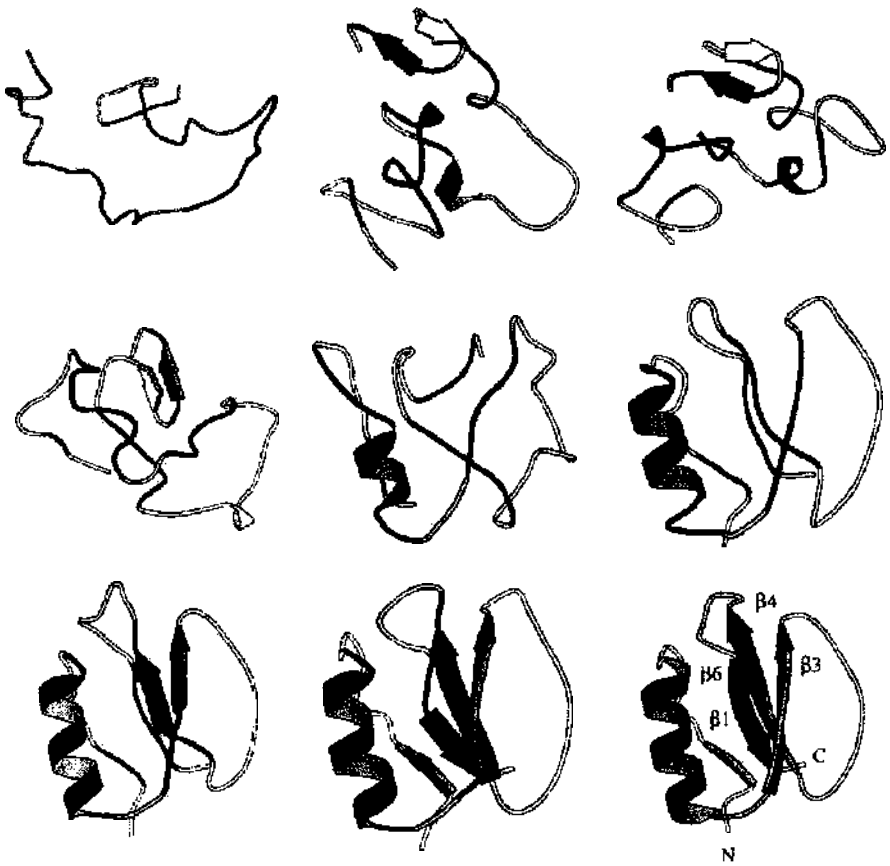


Figure 11. Snapshots from one of the folding trajectories of chymotrypsin inhibitor 2 at 0, 200, 400, 550, 700, 1000, 1050, 1100 and 1200 ps. The rise of the α-helix is followed by the formation of $\beta 3$-$\beta 4$ contacts. Sheet formation happens gradually, but only after disruption and subsequent reformation of the α-helix. The structure formation was accelerated by the use of an artificial constraint. Reprinted with permission from [146], P. Ferrara et al., *Proteins: Struct. Funct. Gen.* 39, 252 (2000). © 2000, Wiley-Liss.

accessible surface area. This simulation agrees well with the nucleation-condensation theory, as obtained from thermal denaturation studies.

3.9. Barnase

Barnase is a 110-residue ribonuclease from *Bacillus amyldiquefaciens*. It is a multidomain protein, with three α-helices, followed by five antiparallel β-strands (Fig. 12). Folding and unfolding properties have been studied experimentally to a great extend [147]. This protein was chosen for molecular dynamics simulations of thermal unfolding by Dagget and coworkers [148]. In an all atom representation with implicit water, the unfolding of barnase was simulated during 4 ns at 498 K, starting from the NMR structure. Denaturation does not lead to a completely extended structure; the α helix content stays relatively high, although the β-sheet is fully disrupted. The residual structure is mainly accounted for by helix α1 and α2. The average length for these first two residual helices is 5 to 7 residues (native structure 7 to 11). Helix α1 disappears in 2 ns but reestablishes itself by forming most of its native hydrogen bonds again. Also, two hydrophobic clusters appear: between helix α1 and residues 88-89 and around Trp94, which is in β-strand 4. This second hydrophobic cluster stabilizes the turn between Tyr90 and Trp94. The remainder of the early structure corresponds to the experimental intermediate and transition sate [149], supporting the idea of a folding pathway where early events are initiated in regions of local structure. If taken in reverse, secondary and tertiary structure are formed succesively following initiation. During this stage, α1 and β(3-4) undergo conformational changes. These secondary structure elements form an intermediate scaffold on which the remainng β-strands pack [150]. Further collapse and structure consolidation takes place as the protein progresses to the transition state, and residues in the folding nucleus are brought in cose proximity to each other. The existence of the intermediate and transition state is experimentally confirmed by Φ-value analysis [151, 152]. The Φ-value analysis involves making mutations throughout the protein and measuring the effect on native (N), transition (T), and denatured (D) state. The ratio of the resulting free energy changes are referred to as Φvalues: $\Phi = \Delta\Delta G_{TS-D}/\Delta\Delta G_{N-D}$. Ratios (between 0 and 1) reflect amount of structure at the mutation site.

3.10. SH3

The amino acid sequence of a protein determines the three-dimensional structure. The Src homology 3 (SH3) protein family is big enough to allow aninvestigation of *how* the sequence plays a part in protein folding. There are more than 400 family members, most of them β-sheet proteins. Src-SH3 has 5 β-strands and a 3_{10}-helix in just 57 residues (Fig. 13).

It is known that this protein folds rapidly, without detecable intermediates [155]. In order to study the folding mechanism of src-SH3, several approaches have been used. Among them are importance-sampling molecular dynamics, for reconstructing the free energy landscape [156], and thermal unfolding studies [157]. Thermal unfolding results in a large number

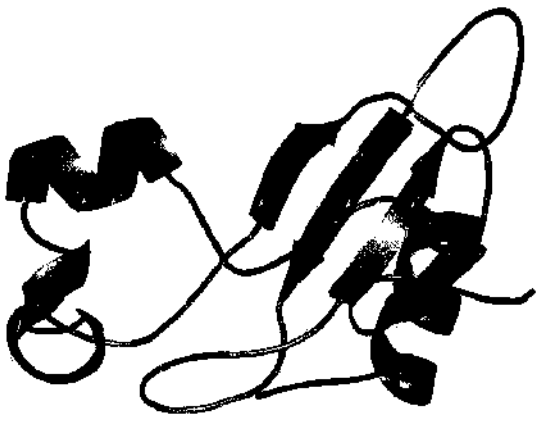


Figure 12. Barnase has three α-helices, followed by an antiparallel, our-stranded β-sheet. Image created with Pymol [86] from Protein Data Bank entry 1A2P [153].

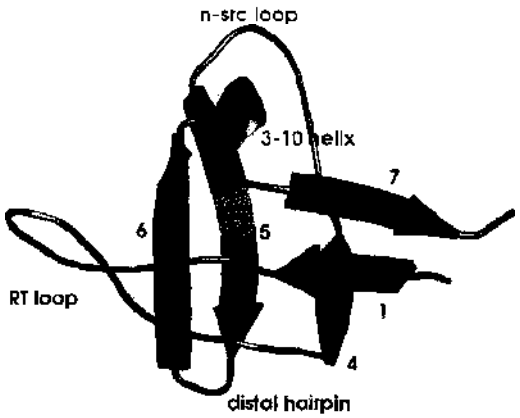


Figure 13. The src-SH3 domain in cartoon representation. Several important regions, which are mentioned in the text, are labeled. For a clearer view on the structure, the small β-strands 2 and 3, that are close to the RT loop, are not shown as cartoons. Image created with Pymol [86] from Protein Data Bank entry 1SHG [154].

of structures that are quite dissimilar. The trajectories were first grouped according to C_α RMSD, but only clustering with respect to the percentage of native hydrophobic contacts and hydrogen bonds made it possible to draw conclusions about the unfolding process, probably because these measures are time independent and directly correlated to the loss of native structure. It was found that there is a hierarchy in loss of native structure. The part of the structure that is maintained for the longest time is the three-stranded β-sheet, composed of the n-src loop and the distal β-hairpin. The hydrophobic contacts and the inner sheet hydrogen bonding stabilize this conformation. Paring between strand 1 and 7, the 3_{10}-helix, and interactions in the proteins core as well as the RT loop are all lost earlier in thermal unfolding [157]. From mutation experiments, it is known that the preserved structure that appears in molecular dynamics unfolding studies is largely formed in the transition state ensemble [158]. Deducing the folding mechanism from a map of the energy landscape leads to the notion that the sheet between β-strands 1 and 7 is unlikely to form early in the folding process [156]. There, early secondary structure formation is preferred in β-strands 2 and 3. As folding continues, the radius of gyration decreases and β-strands 2, 3, and 4 are formed mostly, as are the contacts between β-strands 2 and 3 and in the distal β-hairpin. The lowest contact probability lies between the RT loop and and β-strands 2, 3, and 4. The last folding event happens as the final contacts are made between the hydrophobic part of the sheets and the terminal strands, thereby expelling the water molecules and completing the hydrophobic core. The fact that that there is a distinct sequence of events in structure loss upon heating and that there is a preferred order in native contact formation, as deduced from the energy landscape, supports the idea that folding of src-SH3 is mostly governed by its topology.

3.11. β-β-α Structures

BBA5 is a 23-residue artificial peptide, based on the structure of the zinc finger domain, without an ion [159, 160]. It contains three essential elements of secondary structure. A D-proline favors the formation of a type II' β-turn, stabilizing a hairpin. The carboxy-terminal end holds an α-helix that interacts with the amino-terminal β-hairpin and forms a small hydrophobic core. Based on this molecule, Pande and coworkers created two mutants, both of which were created also by homology modeling [161]. In the first, Phe8 is replaced by a tryptophan residue. The second, double mutant, replaces Val3 with a Tyrosine. The experimentally derived folding time of this mini-protein is 7.5 μs ($\pm$3.5 μs). Using worldwide distributed computing and an implicit solvent model, a total of $\approx$700 μs of simulated folding time was gathered. Molecular dynamics simulations were done at 278 and 298 K for the single mutant and 278, 378, and 478 K for the double mutant. Out of the 32,500 trajectories, the β-hairpin was detected in 1100, the α-helix in more than 21,000 trajectories. Correctly folded structures were discriminated on the fact of having native secondary structure and a low RMSD of the backbone (RMSD-Cα). The 298 K single mutant simulation has a folding

time of 16 μs (7–43 μs). At 298 K the interpolated folding time for the double mutant is 6 μs (3–13 μs). These times are in agreement with experimental observations [161]. Although the BBA5 native structure seems to be reproduced correctly, one can comment on the definition of correct final structure. Pande et al. admit that the criteria for definition of folded structure are somewhat arbitrary. There is a more fundamental issue concerning selection of correct folds: it heavily depends on knowing the outcome. Obviously, a measure that is not directly structure dependent, like an energy evaluation, is preferred.

3.12. Protein G Segment B1

The β-hairpin mentioned earlier [88, 162] is part of protein G. Some studies have been done on the folding of the entire protein domain [163–166]. Segment B1 of streptococcal protein G is an IgG binding domain of 56 residues. It has an antiparallel β-sheet of four strands, and halfway in the sequence is an α-helix that lies on top of the sheet, covering the hydrophobic core. Ninety percent of the residues participate in secondary structure, and its melting temperature is 360 K (Fig. 14).

Explicit solvent simulations up to 2 ns, starting from the average NMR and x-ray structure, have been performed in order to shed a light on the protein's dynamics structure and especially the behavior of water molecules [163]. The reason for this is that there are two water molecules bound to the NMR structure [167]. They are not buried, nor are there any extra groups on the surface that have a place for hydrogen bonding. The α-helix in the NMR structure has a slight distortion, suggested to be stabilized by a water molecule and therefore packing even tighter to the β-sheet. Molecular dynamics simulation did, however, not confirm this. A slightly rotated but straight helix is preferred. The other water molecule is found to be present close to an existing hydrogen bond, between residues Ala29 and Tyr33 [163]. The difference between NMR and molecular dynamics was explained by the force field used.

The free energy landscape for the B1 segment of protein G has been studied extensively [164, 168]. From this work, it was deduced that structure originates in two regions. Initial general collapse yields as much as 35% of the native contacts and there seem to be no high energy barriers. The α-helix forms first—as expected—while the β-sheet forms only gradually. Water is present in the core of the molecule until late in the folding process and only expelled when 80% of the native contacts is formed. In an experimental mutational study of the B1 segment, the sequence of the helix was replaced with the sequence of the second β-hairpin [166]. Up to 10 mutations in the helix were allowed before the helix structure became unmaintainable. Additional mutations resulted in loss of fold, as confirmed by circular dichroism and NMR experiments. Though the simulations [164, 168] were too short to observe β-hairpin formation (2 ns), it is sufficient to see the preference for the α-helix structure on top of the β-sheet.

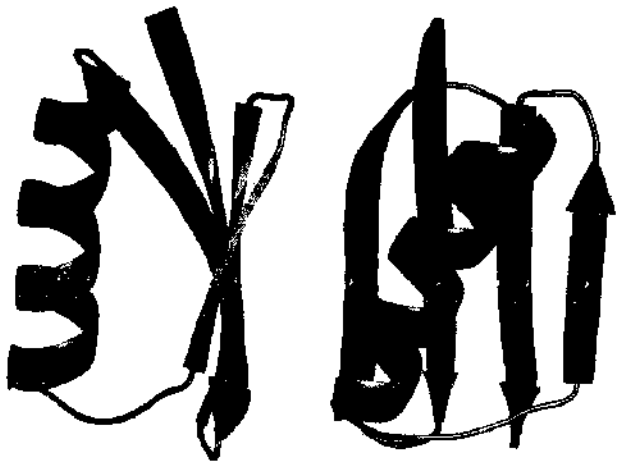


Figure 14. Almost all residues of the B1 segment of Proteins G from *Staphylococcus aureus* participate in secondary structure as shown in this cartoon representation. Two sided view, created from the Protein Data Bank entry 1PGA [79] with Pymol [86].

4. CONCLUSIONS

To learn about protein folding in atomistic detail almost inevitably means doing molecular dynamics simulations. It is crucial, however, to compare the results to as much of the available experimental data as possible [57]. The studies reported in the literature hitherto have demonstrated several things. First about protein folding itself: We are now able to reproduce correctly various kinds of structures, at least the basic secondary structure elements. Successful folding simulations have been performed with implicit and explicit solvent, although the definition of successful is sometimes stretched. The timescale of folding of α-helices and β-hairpins is simulated with rates comparable to current experimental rates. But in implicit solvent, at least those without stochastic friction terms to compensate for the viscosity of the solvent, folding goes one to two orders of magnitude faster: an α-helix in 1 ns [169]; a β-hairpin in 10 ns [169]; a three-stranded β-sheet in 100 ns [112] (compared to 100 ns [170], 1 μs [170], 1 μs [53], respectively, in experiment). The folding of mini-proteins is a logical step with the increasing computational power that is becoming available. Formation of even bigger domains and quarternary structure is something for the future. And when this will be possible, demand for even bigger simulations will arise, containing several proteins to mimic, for example, chaperons and crowding effects. There will never be enough time [171].

The second subject that recent molecular dynamics simulations have taught us about is solvent treatment. Water is the main substance of all atom protein folding simulations. The choice for using a GB/SA model for the solvent or an explicit treatment of water like TIP3 is mainly dependent on the available computational power and efficiency of software [172], because the bulk of atoms in simulations is water. There have been numerous studies evaluating both models [93, 128, 173]. For the betanova peptide, it was found that the presence of implicit water did not alter the folded state significantly, but there are some discrepancies, especially in the process of folding. In the folding of the complete protein G, water does play a role. Even the C-terminal β-hairpin of protein G clearly shows a different folding preference in implicit solvent [94]. Explicit water allows for easier disruption of hydrogen bonds [174, 175] and solvation of hydrophilic side chains. The omission of explicit solvent in the worldwide distributed simulation of the β-hairpin [87] could henceforth be regarded as a missed opportunity. Otherwise, the worldwide distributed computing simulations by the Stanford group of V. J. Pande represent a major leap in breaking the folding time barrier [58, 87, 91, 126, 138, 176, 177]. Increasing the probability of catching rare but rapid folding events by simply using a huge amount of processors (running a screensaver program [178, 179]) is a clever and exciting approach. In order to register an energy barrier crossing, several tens of nanoseconds of simulation is done per processor. Because of the limited time per processor, the use of an implicit solvent model makes sense, but this also represents the main drawback to these studies. The use of faster simulation software [172, 180] and explicit solvent will be the next step in worldwide distributed computing in studies of protein folding. As pointed out recently, it is not so sure that selecting the fastest folding trajectories gives a representative of the folding pathway(s) [113]. The simulated time per processor must be enough to capture relevant folding events. A study on the minimal time needed to get accurate folding events still has to be performed. Another issue to resolve is the determination of folding correctness. The valid method at the moment is to compare to the existing structure in the Protein Data Bank. This tells us something about folding of known proteins, but we would like to be able to predict unknown proteins as well. In addition, until now, mainly water soluble proteins have been tackled. The vast majority of proteins is, however, partly buried in a lipid membrane. *Ab initio* determination of the native structure and function of unknown proteins will turn out to be *the* challenge of the twenty-first century. We therefore have to be able to discern native states by an objective manner, like a free energy evaluation.

For a number of proteins, energy contour maps have been generated and used to show that several conformational states are accessible. In the energy contour maps (landscapes), the existence of basins indicates that some states are more likely to be occupied. The inherent assumption then is that the structure hops from basin to basin, but demonstrating that it is possible, is not the same as proving it happens in chronological order. Given enough time, any state may be occupied at some point. In this respect it is interesting to know that

a low-energy structure does not have to be the expected native (functional) strucure *per se*. It was found for the extracellular bacterial α-lytic protease that its unfolded state was actually more stable than its folded state [181]. When denatured, it requires millena to refold ($t_{1/2}$ = 1800 years) [182]. It is even not uncommon for naturally occurring proteins to be misfolded from the start [183, 184]. In nature, it does not matter whether a proten is folded correctly, only whether it can perform its function. It is worth to keep in mind that overly stable proteins can severely disrupt the delicate metabolic balance in a cell and are indeed related to, for example, cancerous growth. Every protein has a negative feedbick mechanism for genetic expression, which guarantees a limited life-time. Nature delibeiately limits the thermodynamics stability of proteins (i.e., the funnel [21] should *not* be very deep). In biological systems, the term *folded* refers more to the functional state of a protein, whereas in molecular dynamics simulations it always means NMR-like or lowest energy structure, because these are the things that can be measured in the respective fields. In the distant future the gap between functionally folded and energetically stable might he closed by conducting long-time folding simulations of effector and acceptor proteins as one system and sampling all conformations that result in interaction. Selection of structures based on this "functional folding sampling" is not yet feasible, although some effort has been made in this direction [185].

The microsecond-simulation of the Villin headpiece is still the only real-time simulation with explicit solvent in atomistic detail [117]. Consequently, it revealed an unprecedented insight in protein folding. The short cutoff for electrostatic interaction and the fact that it was one run are the important drawbacks, however. Apart from this, the requirement of an extensive computational facility is a barrier for most scientific groups.

Though not so much discussed in this review, the issue of choosing the right force field is not completely solved yet. As demonstrated recently, it is possible now to design structures using molecular dynamics techniques and correctly predict a buildable protein [186]. Although only the most stable structure was selected for further development, it shows the start of exploration of protein structures not yet observed in nature. The important thing about the method used is that it is a real test of force field precision on protein behavior. The great goal of protein folding studies, to predict correctly the native protein structure of a given sequence, is, among other things, very dependent on the force field parameters. As demonstrated for the α-helical peptides and the β-sheet betanova, the force fields can have an influence on the outcome of the simulation [48, 116]. This is, however, not always the case. A comparison of the effect on protein folding of various force fields in combination with explicit and implicit solvent is given in [187]. A lot of adjustments and improvements are made in this field, and, as computational power increases, long-term behavior of solvated amino acid sequences can be observed, while simultaneously the force field is being put to test on its long-term accuracy [57]. There is definitely reason for optimism, and it is possible that the current force fields are good enough for protein folding simulations. However, subtle effects like amino-aromatic interactions [100] may require polarizability to be included in force fields [188]. Moreover, even seemingly simple effects, like the excess energy of mixing of alcohol and water, are seriously underestimated by state of the art (OPLS [189]) force fields [190]. Further development is needed, but as Wilfred van Gunsteren put it recently, "We are getting there."[1]

The protein folding problem is not solved yet. The first requirement in the solution of the problem is that the end point of the folding process should be the experimentally determined native structure, under the specified environmental conditions. Some simulations have reached that point or come very close to it, but results are certainly not consistent at the moment. This may have various causes, including simulation length, water treatment, and force field used, in combination with a specific amino acid sequence. The second part of the solution—folding rate should correspond to experimental kinetic data—is met only in explicit solvent, demonstrating that the usefulness of implicit solvent models is limited.

Reproducibility is the scientific basis of molecular dynamics simulations. Having multiple runs is computationally expensive, but necessary. The 1-μs run of the Villin headpiece was

[1] Structural Biology Network 7th annual conference, Tällberg, Sweden, June 2003.

only repeated for a much shorter time, and in that sense it is statistically not relevant [117]. On the other hand, the generation of thousands of folding trajectories, as used in the world-wide distributed computing approach, seems better, but one has to keep in mind that each run is very short, and the observed folding pathway may therefore be biased compared to real-time folding.

Structure prediction without prior knowledge is something that is beyond the aim of all atom folding studies at the moment. Most simulations are reproducing existing structure in order to verify the correct outcome of the simulations. Real prediction based on energy criteria is possible [186] but is more difficult for biological proteins. Nevertheless, we are confident that, with the continuing increase in computational power, the protein folding problem will be solved eventually.

REFERENCES

1. C. Levinthal, *J. Chim. Phys.* 65, 44 (1968).
2. G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, *J. Mol. Biol.* 7, 95 (1963).
3. G. N. Ramachandran, C. M. Venkatachalam, and S. Krimm, *Biophys. J.* 6, 849 (1966).
4. G. N. Ramachandran and V. Sasisekharan, *Adv. Protein Chem.* 23, 283 (1968).
5. M. Leijonmarck and A. Liljas, *J. Mol. Biol.* 195, 555 (1987).
6. C. B. Anfinsen, *Science* 181, 223 (1973).
7. C. B. Anfinsen, *Nobel lecture* (1972). Available at http://www.nobel.se.
8. S. H. White and W. C. Wimley, *Ann. Rev. Biophys. Biomol. Struct.* 28, 319 (1999).
9. J. L. Popot and D. M. Engelman, *Ann. Rev. Biochem.* 69, 881 (2000).
10. J. H. Kleinschmidt, *Cellul. Mol. Life Sci.* 60, 1547 (2003).
11. P. J. Booth, *Biochim. Biophys. Acta-Biomembranes* 1610, 51 (2003).
12. E. E. Lattman, *Proteins: Struct. Funct. Gen.* 45, 1 (2002).
13. T. C, K. C. Aune, and A. Ikai, *J. Mol. Biol.* 73, 185 (1973).
14. P. Privalov and N. N. Khechinasvili, *J. Mol. Biol.* 86, 665 (1974).
15. F. X. Schmid, in "Protein Folding" (T. E. Creighton, Ed.), p. 197. Freeman, New York, 1992.
16. R. L. Baldwin, *Curr. Opin. Struct. Biol.* 3, 84 (1993).
17. K. Kuwajima, *Proteins: Struct. Funct. Gen.* 6, 87 (1989).
18. O. B. Ptitsyn, R. H. Pain, G. V. Semisotnov, E. Zerovnik, and O. I. Razgulyaev, *FEBS Lett.* 262, 20 (1990).
19. O. B. Ptitsyn, in "Protein Folding," (T. E. Creighton, Ed.), p. 243. Freeman, 1992.
20. R. L. Baldwin, *J. Biomol. NMR* 5, 103 (1995).
21. H. Frauenfelder, S. G. Sligar, and P. G. Wolynes, *Science* 254, 1598 (1991).
22. H. Frauenfelder and P. G. Wolynes, *Physics Today* 47, 58 (1994).
23. J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, *Proteins: Struct. Funct. Gen.* 21, 167 (1995).
24. J. S. Weissman, *Curr. Biol.* 2, 255 (1995).
25. R. B. Freedman, in "Protein Folding" (T. E. Creighton, Ed.), p. 455. Freeman, 1992.
26. O. Medalia, I. Weber, A. S. Frangakis, D. Nicastro, G. Gerisch, and W. Baumeister, *Science* 298, 1209 (2002).
27. R. J. Ellis and A. P. Minton, *Nature* 425, 27 (2003).
28. M. Levitt and A. Warshel, *Nature* 253, 694 (1975).
29. H. J. C. Berendsen, *Science* 282, 642 (1998).
30. K. A. Dill, K. M. Fiebig, and H. S. Chan, *Proc. Natl. Acad. Sci. U.S.A.* 90, 1942 (1993).
31. M. Vieth, A. Kolinski, C. L. Brooks III, and J. Skolnick, *J. Mol. Biol.* 237, 361 (1994).
32. S. Bromberg and K. A. Dill, *Prot. Sci.* 3, 997 (1994).
33. J. Liang and K. A. Dill, *Biophys. J.* 81, 751 (2001).
34. E. Shakhnovich, G. M. Farztdinov, and A. M. G. Karplus, *Phys. Rev. Lett.* 67, 1665 (1991).
35. A. Sali, E. Shakhnovich, and M. Karplus, *Nature* 369, 248 (1994).
36. A. Dinner, A. Sali, M. Karplus, and E. Shakhnovich, *J. Chem. Phys.* 101, 1444 (1994).
37. G. M. Crippen and Y. Z. Ohkubo, *Proteins: Struct. Funct. Gen.* 32, 425 (1998).
38. B. L. de Groot, A. Amadei, R. M. Scheek, N. A. J. van Nuland, and H. J. C. Berendsen, *Proteins: Struct. Funct. Gen.* 26, 314 (1996).
39. K. A. Dill, *Curr. Opin. Struct. Biol.* 3, 99 (1993).
40. K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomes, and H. S. Chan, *Protein Sci.* 4, 561 (1995).
41. M. Karplus and A. Sali, *Curr. Opin. Struct. Biol.* 5, 58 (1995).
42. N. van Nuland, F. W. Hangyi, R. C. van Schaik, H. J. C. Berendsen, W. F. van Gunsteren, R. M. Scheek, and G. T. Robillard, *J. Mol. Biol.* 237, 544 (1994).
43. J. Skolnick and A. Kolinski, *J. Mol. Biol.* 212, 787 (1990).
44. A. R. Leach, "Molecular Modelling: Principles and Applications." Addison Wesley Longman, Harlow, UK, 1996.
45. W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, *J. Am. Chem. Soc.* 112, 6127 (1990).
46. B. Lee and F. M. Richards, *JMB* 55, 379 (1971).

47. R. Constanciel and R. Contreras, *Theor. Chim. Acta* 65, 1 (1984).
48. H. Nymeyer and A. E. Garcia, *Proc. Natl. Acad. Sci. U.S.A.* 100, 13934 (2003).
49. G. E. Schulz and R. H. Schirmer, "Principles of Protein Structure." Springer-Verlag, Berlin, Heidelberg, New York, 1998.
50. J. W. Neidigh, R. M. Fesinmeyer, and N. H. Andersen, *Nat. Struct. Biol.* 9, 425 (2002).
51. C. J. McKnight, D. S. Doering, P. T. Matsudaira, and P. S. Kim, *J. Mol. Biol.* 260, 126 (1996).
52. N. Ferguson and A. R. Fersht, *Curr. Opin. Struct. Biol.* 13, 75 (2003).
53. E. de Alba, J. Santoro, M. Rico, and M. A. Jimenez, *Protein Sci.* 8, 854 (1999).
54. V. Munoz, P. A. Thompson, J. Hofrichter, and W. A. Eaton, *Nature* 390, 196 (1997).
55. G. S. Jas, W. A. Eaton, and J. Hofrichter, *J. Phys. Chem. B.* 105, 261 (2001).
56. D. van der Spoel, H. J. Vogel, and H. J. C. Berendsen, *Proteins: Struct. Funct. Gen.* 24, 450 (1996).
57. D. van der Spoel and E. Lindahl, *J. Phys. Chem. B.* 107, 11178 (2003).
58. V. S. Pande, I. Baker, J. Chapman, S. P. Elmer, S. Khaliq, S. M. Larson, Y. M. Rhee, M. R. Shirts, C. D. Snow, E. J. Sorin, and B. Zagrovic, *Biopolymers* 68, 91 (2000).
59. W. F. van Gunsteren and A. E. Mark, *J. Chem. Phys.* 108, 6109 (1998).
60. J. Tirado-Rives and W. L. Jorgensen, *Biochemistry* 30, 3864 (1991).
61. V. Daggett and M. Levitt, *J. Mol. Biol.* 223, 1121 (1992).
62. J. Hermans, *Curr. Opin. Struct. Biol.* 3, 270 (1993).
63. C. L. Brooks III and L. Nilsson, *J. Am. Chem. Soc.* 115, 11034 (1993).
64. K. V. Soman, A. Karimi, and D. A. Case, *Biopolymers* 33, 1567 (1993).
65. P. C. Lyu, M. I. Liff, L. A. Marky, and N. R. Kallenbach, *Science* 250, 669 (1990).
66. M. Blaber, X. J. Zhang, and B. W. Mathews, *Science* 262, 917 (1993).
67. E. J. Spek, C. A. Olson, Z. Shi, and N. R. Kallenbach, *J. Am. Chem. Soc.* 121, 5571 (1999).
68. A. E. Garcia and K. Sanbonmatsu, *Proc. Natl. Acad. Sci. U.S.A.* 99, 2782 (2002).
69. D. J. Lockhart and P. S. Kim, *Science* 260, 198 (1993).
70. S. Williams, T. P. Casgrove, R. Gilmanshin, K. S. Fang, R. H. Callender, W. H. Woodruff, and R. B. Dyer, *Biochemistry* 35, 691 (1996).
71. F. M. DiCapua, S. Swaminathan, and D. L. Beveridge, *J. Am. Chem. Soc.* 112, 6768 (1990).
72. P. Doruker and I. Bahar, *Biophys. J.* 72, 2445 (1997).
73. C. M. Shepherd, D. van der Spoel, and H. J. Vogel, *J. Biomol. Struct. Dyn.* 21, 80 (2004).
74. E. Wang and C. L. Wang, *Arch. Biochem. Biophys.* 329, 156 (1996).
75. P. V. D. W. G. J. Hol and H. Berendsen, *Nature* 273, 443 (1978).
76. K. H. Lee, D. R. Benson, and K. Kuczera, *Biochemistry* 39, 13737 (2000).
77. M. Feig, A. D. MacKerell, and C. L. Brooks, *J. Phys. Chem. Mar.* 107, 2831 (2003).
78. W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz-Jr, D. M. Ferguson, D. C. Spellmeyer, T. Fox., J. W. Caldwell, and P. A. Kollman, *J. Am. Chem. Soc.* 117, 5179 (1995).
79. A. M. Gronenborn, D. L. F. N. Z. Essig, A. Achari, M. Whitlow, and P. T. Wingfield, *Science* 253, 657 (1991).
80. F. J. Blanco, M. A. Jimenez, A. Pineda, M. Rico, J. Santoro, and J. L. Nieto, *Biochemistry* 33, 6004 (1994).
81. F. J. Blanco and L. Serrano, *Eur. J. Biochem.* 230, 634 (1995).
82. S. Honda, N. Kobayashi, and E. Munekata, *J. Mol. Biol.* 295, 269 (2000).
83. N. Kobayashi, S. Honda, H. Yoshii, and E. Munekata, *Biochemistry* 39, 6564 (2000).
84. B. Ma and R. Nussinov, *Protein Sci.* 12, 1882 (2003).
85. P. G. Bolhuis, *Proc. Natl. Acad. Sci. U.S.A.* 100, 12129 (2003).
86. W. L. DeLano, "The Pymol Molecular Graphics System." Available at http://www.pymol.org, 2002.
87. B. Zagrovic, E. J. Sorin, and V. Pande, *J. Mol. Biol.* 313, 151 (2001).
88. J. Tsai and M. Levitt, *Bioph. Chem.* 101, 187 (2002).
89. A. Kolinski, B. Ilkowski, and J. Skolnick, *Biophys. J.* 77, 2942 (1999).
90. H. Wang, J. Varaday, L. Ng, and S. S. Sung, *Proteins: Struct. Funct. Gen.* 37, 325 (1999).
91. V. J. Pande and D. S. Rokshar, *Proc. Natl. Acad. Sci. U.S.A.* 96, 9062 (1999).
92. G. Wei, P. Derreumaux, and N. Mousseau, *J. Chem. Phys.* 119, 6403 (2003).
93. R. Zhou, B. J. Berne, and R. Germain, *Proc. Natl. Acad. Sci. U.S.A.* 98, 14931 (2001).
94. R. Zhou and B. J. Berne, *Proc. Natl. Acad. Sci. U.S.A.* 99, 12777 (2002).
95. H. J. Dyson and P. E. Wright, *Annu. Rev. Biophys. Biophys. Chem.* 20, 519 (1991).
96. H. J. Dyson and P. E. Wright, *Curr. Opin. Struct. Biol.* 3, 60 (1993).
97. P. E. Wright, H. J. Dyson, and R. A. Lerner, *Biochemistry* 27, 7167 (1988).
98. J. Kemmink and T. E. Creighton, *J. Mol. Biol.* 234, 861 (1993).
99. J. Kemmink, C. P. M. van Mierlo, R. M. Scheek, and T. E. Creighton, *J. Mol. Biol.* 230, 312 (1993).
100. D. van der Spoel, A. R. van Buuren, D. P. Tieleman, H. J. C. Berendsen, *J. Biomol. NMR* 8, 229 (1996).
101. D. van der Spoel and H. J. C. Berendsen, *Biophys. J.* 72, 2032 (1997).
102. H. de Loof, L. Nilsson, and R. Rigler, *J. Am. Chem. Soc.* 114, 4028 (1992).
103. A. R. van Buuren and H. J. C. Berendsen, *Biopolymers* 33, 1159 (1993).
104. K. A. Dill and D. Shortle, *Annu. Rev. Biochem.* 60, 795 (1991).
105. D. R. Shortle, *Curr. Opin. Struct. Biol.* 6, 24 (1996).
106. D. Shortle, *FASEB J.* 10, 27 (1996).
107. D. Wishart, C. G. Bigam, A. Holm, R. S. Hodges, and D. Sykes, *J. Biomol. NMR* 5, 67 (1995).
108. K. W. Plaxco, C. J. Morton, S. B. Grimshaw, J. A. Jones, M. Pitkeathly, I. D. Campbell, and C. M. Dobson, *J. Biomol. NMR* 10, 221 (1997).

109. D. van der Spoel, *Biochem. Cell Biol.* 76, 164 (1998).
110. D. J. Tobias, S. F. Sneddon, and C. L. Brooks III, *J. Mol. Biol.* 227, 1244 (1992).
111. P. Ferrara, J. Apostakolis, and A. Caflisch, *Proc. Natl. Acad. Sci. U.S.A.* 97, 10780 (2000).
112. A. Cavalli, P. Ferrara, and A. Caflisch, *Proteins: Struct. Funct. Gen.* 47, 305 (2002).
113. E. Paci, A. Cavalli, M. Vendruscolo, and A. Caflisch, *Proc. Natl. Acad. Sci. U.S.A.* 100, 8217 (2003).
114. T. Kortemme, M. Ramirez-Alvardo, and L. Serrano, *Science* 281, 253 (1998).
115. B. D. Bursulaya and C. L. Brooks III, *J. Am. Chem. Soc.* 121, 9947 (1999).
116. B. D. Bursulaya and C. L. Brooks III, *J. Phys. Chem.* 104, 12378 (2000).
117. Y. Duan and P. A. Kollman, *Science* 282, 740 (1998).
118. C. J. McKnight, P. T. Matsudaira, and P. S. Kim, *Nat. Struct. Biol.* 4, 180 (1997).
119. R. Walser, A. E. Mark, and W. F. van Gunsteren, *Biophys. J.* 78, 2752 (2000).
120. M. Patra, M. Karttunen, M. T. Hyvonen, E. Falck, P. Lindqvist, I. Vattulainen, *Biophys. J.* 84, 3636 (2003).
121. R. J. Loncharich and B. E. Brooks, *Proteins* 6, 32 (1989).
122. J. G. Guenot and P. A. Kollman, *J. Comp. Chem.* 14, 295 (1993).
123. O. Steinhauser and H. Schreiber, *J. Mol. Biol.* 228, 909 (1992).
124. M. Saito, *J. Chem. Phys* 101, 4055 (1994).
125. J. Kubelka, W. E. Eaton, and J. Hofrichter, *J. Mol. Biol.* 329, 625 (2003).
126. B. Zagrovic, C. D. Snow, M. R. Shirts, and V. S. Pande, *J. Mol. Biol.* 323, 927 (2002).
127. J. Klein-Seetharaman, M. Oikawa, S. B. Grimshaw, J. Wirmer, E. Duchardt, T. Ueda, T. Imoto, L. J. Smith, C. M. Dobson, and H. Schwalbe, *Science* 295, 1719 (2002).
128. M. Y. Shen and K. F. Freed, *Biophys. J.* 82, 1791 (2002).
129. J. Åqvist, *FEBS Lett.* 457, 414 (1999).
130. M.-Y. Shen and K. F. Freed, *Proteins: Struct. Funct. Gen.* 49, 439 (2002).
131. M. Lee, Y. Duan, and P. A. Kollman, *Proteins: Struct. Funct. Gen.* 39, 309 (2000).
132. T. Darden, D. York, and L. Pedersen, *J. Chem. Phys.* 98, 10089 (1993).
133. U. Essman, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, *J. Chem. Phys.* 103, 8577 (1995).
134. L. Qiu, S. A. Pabit, A. E. Roitberg, and S. J. Hagen, *J. Am. Chem. Soc.* 124, 12952 (2002).
135. C. Simmerling, B. Strockbine, and A. E. Roitberg, *J. Am. Chem. Soc.* 124, 11258 (2002).
136. S. Chowdhury, M. C. Lee, G. Xiong, and Y. Duan, *J. Mol. Biol.* 327, 711 (2003).
137. J. W. Pitera and W. Swope, *Proc. Natl. Acad. Sci. U.S.A.* 100, 7587 (2003).
138. C. D. Snow, B. Zagrovic, and V. S. Pande, *J. Am. Chem. Soc.* 124, 14548 (2002).
139. H. Gouda, H. Torigoe, A. Saito, M. Sato, Y. Arata, and I. Shimida, *Biochemistry* 40, 9665 (1992).
140. A. E. Garcia and J. N. Onuchic, *Proc. Natl. Acad. Sci. U.S.A.* 100, 13898 (2003).
141. S. E. Jackson and A. R. Fersht, *Biochemistry* 30, 10436 (1991).
142. L. S. Itzhaki, D. E. Otzen, and A. R. Fersht, *J. Mol. Biol.* 254, 260 (1995).
143. R. Day, B. J. Bennion, and V. Daggett, *J. Mol. Biol.* 322, 189 (2002).
144. S. L. Kazmirski, K. B. Wong, S. M. Freund, Y. J. V. Tan, and A. R. F. Daggett, *Proc. Natl. Acad. Sci. U.S.A.* 98, 4349 (2001).
145. S. Ludvigsen, H. Y. Shen, M. Kjaer, J. C. Madsen, and F. M. Poulsen, *J. Mol. Biol.* 222, 621 (1991).
146. P. Ferrara, J. Apostolakis, and A. Caflisch, *Proteins: Struct. Funct. Gen.* 39, 252 (2000).
147. A. R. Fersht, *FEBS Lett.* 325, 5 (1993).
148. C. J. Bond, K.-B. Wong, J. Clarke, A. R. Fersht, and V. Daggett, *Proc. Natl. Acad. Sci. U.S.A.* 94, 13409 (1997).
149. J. M. Matthews and A. R. Fersht, *Biochemistry* 34, 6805 (1995).
150. V. Daggett and A. R. Fersht, *Nature Rev.* 4, 497 (2003).
151. A. Li and V. Daggett, *J. Mol. Biol.* 275, 677 (1998).
152. V. Daggett, A. Li, and A. R. Fersht, *J. Am. Chem. Soc.* 120, 12740 (1998).
153. Y. Mauguen, R. W. Hartley, E. J. Dodson, G. G. D. G. Bricogne, C. Chothia, and A. Jack, *Nature* 297, 162 (1982).
154. A. Musacchio, M. Noble, R. Paupit, R. Wieringa, and M. Saraste, *Nature* 359, 851 (1992).
155. K. W. Plaxco, K. T. Simons, and D. Baker, *J. Mol. Biol.* 277, 985 (1998).
156. J.-E. Shea, J. N. Onuchic, and C. L. Brooks III, *Proc. Natl. Acad. Sci. U.S.A.* 99, 16064 (2002).
157. J. Tsai, M. Levitt, and D. Baker, *J. Mol. Biol.* 291, 215 (1999).
158. V. P. Grantcharova, D. S. Riddle, and D. Baker, *Proc. Natl. Acad. Sci. U.S.A.* 97, 7084 (2000).
159. M. D. Struthers, R. P. Cheng, and B. Imperiali, *Science* 271, 342 (1996).
160. M. Struthers, J. J. Ottesen, and B. Imperiali, *Fold. Des.* 3, 95 (1998).
161. C. D. Snow, H. Nguyen, and V. S. Pande, *Nature* 420, 102 (2002).
162. R. H. Zhou, B. J. Berne, and R. Germain, *Proc. Natl. Acad. Sci. U.S.A.* 98, 14931 (2001).
163. F. B. Sheinerman and C. L. Brooks III, *Proteins: Struct. Funct. Gen.* 29, 193 (1997).
164. F. B. Sheinerman and C. L. Brooks III, *J. Mol. Biol.* 278, 439 (1998).
165. N. Calimet, M. Schaefer, and T. Simonson, *Proteins: Struct. Funct. Gene.* 45, 144 (2001).
166. D. Cregut and L. Serrano, *Protein Sci.* 8, 271 (1999).
167. G. M. Clore and A. M. Gronenborn, *J. Mol. Biol.* 223, 853 (1992).
168. F. B. Sheinerman and C. L. Brooks III, *Proc. Natl. Acad. Sci. U.S.A.* 95, 1562 (1998).
169. P. Ferrara, J. Apostolakis, and A. Caflisch, *J. Phys. Chem.* 104, 5000 (2000).
170. W. A. Eaton, V. Munoz, S. J. Hagen, G. S. Jas, L. J. Lapidus, E. R. Henry, and J. Hofrichter, *Annu. Rev. Biophys. Biomol. Struct.* 29, 327 (2000).

171. T. Hansson, C. Oostenbrink, and W. F. van Gunsteren, *Curr. Opin. Struct. Biol.* 12, 190 (2002).
172. E. Lindahl, B. A. Hess, and D. van der Spoel, *J. Mol. Mod.* 7, 306 (2001).
173. B. D. Bursulaya, J. Jeon, C. N. Yang, and H. J. Kim, *J. Phys. Chem. A.* 104, 45 (2000).
174. L. D. Barron, L. Hecht, and G. Wilson, *Biochemistry* 36, 13143 (1997).
175. F. Xu and T. A. Cross, *Proc. Natl. Acad. Sci. U.S.A.* 96, 9057 (1999).
176. B. Zagrovic, C. D. Snow, S. Khaliq, M. R. Shirts, and V. S. Pande, *J. Mol. Biol.* 323, 153 (2002).
177. B. Zagrovic and V. J. Pande, *J. Comput. Chem.* 24, 1432 (2003).
178. M. R. Shirts and V. S. Pande, *Science* 86, 4983 (2000).
179. M. R. Shirts and V. S. Pande, *Phys. Rev. Lett.* 86, 4983 (2001).
180. H. J. C. Berendsen, D. van der Spoel, and R. van Drunen, *Comp. Phys. Comm.* 91, 43 (1995).
181. S. S. Jaswal, J. L. Sohl, J. H. Davis, and D. A. Agard, *Nature* 415, 343 (2002).
182. J. L. Sohl, S. S. Jaswal, and D. A. Agard, *Nature* 395, 817 (1998).
183. U. Schubert, L. C. Anton, J. Gibbs, C. G. Norbury, J. W. Yewdell, and J. R. Bennink, *Nature* 404, 709 (2000).
184. G. C. Turner and A. Varshavsky, *Science* 289, 21117 (2000).
185. D. A. Debe, M. J. Carlson, and W. A. Goddard III, *Proc. Natl. Acad. Sci. U.S.A.* 96, 2596 (1999).
186. B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, D. Baker, *Science* 302, 1364 (2003).
187. R. Zhou, *Proteins: Struct. Funct. Gen.* 53, 148 (2003).
188. P. J. van Maaren and D. van der Spoel, *J. Phys. Chem. B.* 105, 2618 (2001).
189. W. L. Jorgensen, "OPLS Force Fields," Vol. 3, p. 1986, Wiley: New York, 1998.
190. E. J. Wensink, A. C. Hoffmann, P. J. van Maaren, and D. van der Spoel, *J. Chem. Phys.* 119, 7308 (2003).

CHAPTER 3

Biomolecular Machines

Artur Baumgaertner

Forschungszentrum, Institut für Festkorperforschung, Jülich, Germany

CONTENTS

1. INTRODUCTION

Nanotechnology is perfectly realized in biological systems. Cells are essentially biological assemblers that build thousands of custom-designed molecules and construct new assemblers. This view [1] was pioneered by Richard Feynman [2] and popularized by Eric Drexler's [3, 4] evocative idea of a self-replicating assembler building nanoscale devices atom by atom. Bionanotechnology, being a synonym for nanobiotechnology, is a rapidly growing field that encompasses contributions from various disciplines, ranging from engineering and computational sciences to physics, chemistry, and biology. Hence, there is an increasing need

ISBN: 1-58883-048-9

Handbook of Theoretical and Computational Nanotechnology
Edited by Michael Rieth and Wolfram Schommers
Volume 6: Pages (73–161)

for reviews and textbooks that provide an introduction to biomolecular sciences and their impact on nanotechnology.

Most functions in the cell are not carried out by single-protein enzymes, colliding randomly within the cellular jungle, but by macromolecular complexes containing multiple subunits with specific functions [5]. Living cells are made up of these complexes, which carry out many of the functions essential for their existence, differentiation, and reproduction. In many cases the malfunction of these proteins can be a source of disease; for example, myosin mutations, particularly in the head and neck region of the molecule, can result in inherited diseases such as familial hypertrophic cardiomyopathy. An understanding of the mechanisms of these proteins may provide a guide for therapy.

Many of these complexes can be described as "molecular machines" or "molecular motors" or "molecular devices," depending on their sizes, complexity, and tasks. Indeed, this designation captures many of the aspects characterizing these biological complexes: modularity, complexity, cyclic function, and, in most cases, the consumption of energy. In molecular machines or motors, a rotary or linear movement is used for motility, nucleic acid processing, folding or unfolding, or as transducers of light or chemical energy. Examples of such molecular machines are the replisome, the transcriptional machinery, the spliceosome, the ribosome, but also smaller machines, residing in the plasma and organellar membranes, are known as ion and protein transporters or as the bacterial flagellar motor.

The size of these macromolecular complexes, alone, often makes them inaccessible to X-ray crystallography. The structure determination of large, complex protein ensembles will pose a particularly difficult problem for structural biologists. As a result, many efforts have concentrated on determining the structures of individual subunits and domains within the machines. Consequently, information on the organization of the assembly subunits, their interactions, and sometimes their precise function within the context of the fully functional complex is often lost.

Computational and Theoretical Nanobiology. In the postgenomic era, the recognition that computational methodologies will play a critical role in biology is widespread. Cell and molecular biologists have realized the impracticality of trying to successfully predict complex molecular mechanisms using intuition. Accordingly, the molecular and cell biology communities are now seeking suitable avenues for enabling them to add computational tools to their research kits.

The essential question in understanding biomolecular machines is concerned with the explanation of the macroscopic phenomenology in terms of the atomic structures and forces involved. Although a complete description is not yet available even for the best-characterized system, F_1-ATPase, which is discussed in Section 4.1, considerable progress has been made recently, not only from an experimental point of view, but also with respect to computational and theoretical achievements. Some of these machines, which have been studied by simulations and mathematical methods are listed in Table 1. An essential element of the progress has been improvements in the experimental methodology, of which single-molecule experiments are arguably the most important. However, a full understanding, in particular, of the kinetics of the dynamical mechanism requires theoretical and computational techniques that make use of atomic-resolution structures and detailed calculations of the molecular energetics and dynamics. For example, how are the conformational changes that engender the motion of rotary motors such as ATPases produced by ATP as the energy source?

Considerable progress has been made in the past two years by a combination of biophysical techniques and theoretical analysis. Single-molecule studies have played an important role for a variety of motors including kinesin, myosin, and polymerases. The understanding of various ATPase complexes, among the smallest biomolecular rotary motors, has made particular progress by the interplay of experimental and theoretical studies.

Energy Source. The term *motor* or *machine* is used to describe some biomolecular complexes because they transduce one form of energy to another (e.g., chemical binding to mechanical work). These machines make use of chemical energy from a variety of sources, of which the most common is the binding energy of ATP, H_2O, and its hydrolysis products ADP, H_2PO_4:

$$ATP^{4-} + H_2O \rightarrow ADP^{3-} + H_2PO_4^-$$

Table 1. List of some molecular machines which have been studied theoretically and by simulations. Abbreviations: ATP = hydrolysis of adenosine triphosphate, n-phos = nucleotide phosphorylation, $\Delta\Psi$ = membrane potential and ion gradient.

Biomolecular Devices	Function	Energy source
Membrane transducers		
GPCR	Signal relay	Ligands
Rhodopsin		
CCR5, CXCR4		
Catalytic receptors	Signal relay	Photons, ligands
Insulin, EGF		
Chemotactic sensors	Signal relay	Photons, ligands
sR		
Receptor-ligand	Signal relay	Photons, ligands
Tition, GroEL, etc.	Unfolding, refolding	ATP
Membrane channels		
Porins	Valve for ion, water	Gradients
OmpF, OmpT, OmpA, FhuA		
AQP		
Maltoporin		
Ion channels	Selective ion valve	Gating by
AChR		Ligand
ClC		
gA		$?\Delta\Psi$
GluR		Ligand
KcsA		pH
KvAP		Voltage
MscL, MscS		Strain
Membrane transporters		
Ion Translocons	Ion pumps	Photons
bR	H^+	
hR	Cl^-, anion	
Ion translocases	Ion pumps	
V-ATPase	H^+	ATP
Ca^{2+}-ATPase	Ca^{2+}	ATP
Na^+-K^+-ATPase	Na^+, K^+	ATP
Solute Translocases	Solute transport	ATP
HisP, MsbA, GLUT3		ATP
Protein translocases	Protein transport	ATP
Mitochondrial pore		ATP, $\Delta\Psi$
ER pore		ATP, GTP
Nuclear pore		
Membrane Motors		
ATP synthase	ATP synthesis or H^+ pump	$\Delta\Psi$ or ATP
Flagellar motor	Bacterial motility	$\Delta\Psi$
Motor Proteins		
Track Motors		ATP
Myosin II	Muscle action	ATP
Myosin V, VI	Cargo transport	ATP
Kinesin	Cargo transport	ATP
Dynein	Flagellar motion	ATP
Nucleotidases		ATP
Polymerase	Catalyze synthesis	ATP, n-phos
Helicase	Unwind dDNA	ATP

Proton and ion gradients, as well as redox potential differences, also serve as the energy source in some cases.

Structure. Biomolecular machines range from single subunits (e.g., DNA polymerases) through the smallest rotary motor, F_1-ATPase, composed of nine subunits in mitochondria, to the flagella motor of bacteria, which can be composed of several hundred subunits of a number of different proteins.

Cellular Function. Each protein machine possesses its specific function and often it forms an element of the chemical network of which the cell is composed. The motors

have a wide range of functions, including chemical (e.g., ATP) synthesis, organelle transport, muscle contraction, protein folding, and translocation along DNA/RNA and protein filaments. They play an important role in cellular signaling, cell division, and cellular motion.

Classification. Table 1 lists some of the known biomolecular machines with their functions and energy sources. There exists others and that all the functions of the known motors are not yet recognized. Table 1 includes the class of membrane channels. In a strict sense, these proteins are not motors, but, rather, "valves" whose gating are controlled by some form of external free energy. However, because they are one of the most intensively studied biomolecular devices, they are discussed in a special chapter.

2. MEMBRANE TRANSDUCERS

In the following a group of membrane proteins is considered that transduct signals across membranes where transport of matter (ions, proteins, etc.) is not involved. Instead, these "transducer" relay a signal across the plasma membrane, and the influence they exert on events in the cytosol or nucleus, generally depends on the production of new intracellular signals. The major goal of theoretical and computational studies is the elucidation how a conformational change induced in a cell surface receptor protein by binding of an extra-cellular ligand enables the receptor to act directly or indirectly as a transducer, converting the extracellular signal into a signal inside the cell.

2.1. G-Protein-Coupled Receptors

G proteins are GTP-binding regulatory proteins. They are intracellulary coupled to membrane-spanning protein receptors. These G-protein-coupled receptors (GPCRs) respond to a variety of different external stimuli and activate G proteins. GPCRs activate signaling paths in response to stimuli such as Ca^{2+}, amines, hormones, peptides, and even large proteins. GPCRs share a conserved transmembrane structure comprising seven helices connected by alternating extracellular and intracellular loops. Binding of specific ligands to the extracellular or transmembrane domains causes conformational changes that act as a switch to relay the signal to G proteins that in turn evoke further intracellular responses. A GPCR-mediated biological response involves a series of events (i.e., receptor activation, receptor-G protein interaction, and receptor-induced G protein activation) for which a detailed mechanism still remains elusive at the molecular level. Whereas residues located in the helical bundle and at the boundary between the membrane and the cytosol may play a role in the 'conformational switch' underlying receptor activation (i.e., the transition from the inactive [R] to active [R*] state), amino acids in the intracellular loops are believed to be more directly involved in receptor-G protein interaction and/or receptor induced G protein activation. The combination of these two latter events, which cannot be unequivocally separated experimentally, is generally indicated as receptor-G protein coupling.

2.1.1. Rhodopsin

Rhodopsins are members of the largest subfamily, constituting 90% of all GPCRs. These are activated by light and turn on the signaling pathway that leads to vision. Mutations in the rhodopsin gene lead to human retinal pathologies. Recently the structure of bovine rhodopsin from diffraction data have been reported [6]. The structure of bovine rhodopsin at 2.8 Å (Fig. 1) shows that the highly organised structure of the extracellular region, including a conserved disulfide bridge, forms a basis for the arrangement of the seven-helix transmembrane motif. The identification of a set of residues that mediate interactions between the transmembrane helices and the cytoplasmic surface, where G protein activation occurs, suggests a possible structural change upon photoactivation. It was a surprise that this hydrophobic membrane protein was crystallised directly from a detergent-solubilised sample, without the need for any alternative techniques such as lipidic cubic phase crystallisation. This structure will certainly be used for the modeling of various GPCRs, which are the targets of many drugs.

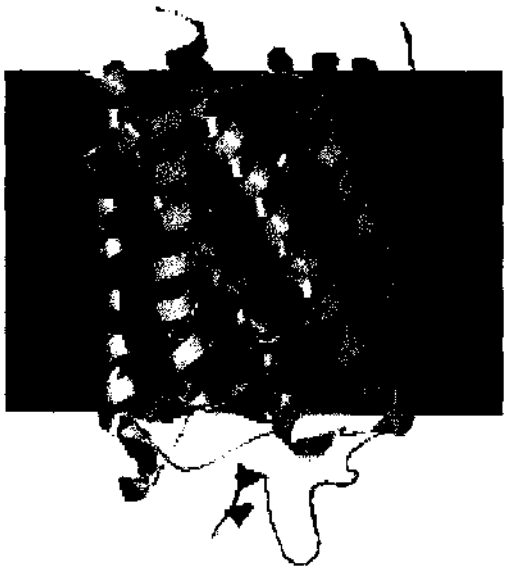


Figure 1. Ribbon drawing of bovine rhodopsin (PDB code 1JSQ).

The very first steps of the dynamics of the related conformational transformations have been studied by molecular dynamics simulations [7]. The retinal cis-trans isomerization and early relaxation steps had been studied in a 10-ns molecular dynamics simulation of a fully hydrated model of membrane-embedded rhodopsin. The isomerization, induced by transiently switching the potential energy function governing the C11=C12 dihedral angle of retinal, completes within 150 fs and yields a strongly distorted retinal. The most significant conformational changes in the binding pocket are straightening of retinal's polyene chain and separation of its β-ionone ring from Trp-265. In the following 500 ps, transition of 6s-cis to 6s-trans retinal and dramatic changes in the hydrogen bonding network of the binding pocket involving the counterion for the protonated Schiff base, Glu-113, occur. The energy initially stored internally in the distorted retinal is transformed into nonbonding interactions of retinal with its environment. During the following 10-ns, increased mobilities of some parts of the protein, such as the kinked regions of the helices, mainly helix VI, and the intracellular loop 12, were observed, as well as transient structural changes involving the conserved salt bridge between Glu-134 and Arg-135. These features prepare the protein for major structural transformations achieved later in the photocycle. Retinal's motion, in particular, can be compared to an opening turnstile freeing the way for the proposed rotation of helix VI. This was demonstrated by a steered molecular dynamics simulation [7] in which an applied torque enforced the rotation of helix VI. The molecular model was constructed using the refined crystal structure of Rh from PDB entry 1 HZX. The disk membranes of the rod cells in which Rh resides in the eye were composed of unsaturated lipids. The system was generated from a hexagonal lattice of POPC molecules with a lateral distance of 9 Å and a thickness of 38 Å (defined as the distance between the planes formed by the phosphorus atoms of the two leaflets). The position of the protein in the membrane was chosen according to an optimal matching of the protein's hydrophobic surface, including the two palmitoyl moieties connected to Cys-322 and Cys-323, with the hydrophobic part of the membrane. After superposition of membrane patch and protein, the lipids clashing with the protein were removed, except for a few lipids that overlapped only with a few protruding atoms without intersecting chains and that had otherwise enough space to avoid protein atoms. The model then contained 108 POPC molecules.

2.1.2. Adrenergic Receptor

The adrenergic receptor (AR) belongs to the rhodopsin family of GPCRs. Computer simulations on a new model of the 1b-adrenergic receptor [8] based on the crystal structure of rhodopsin have been combined with experimental mutagenesis to investigate the role of residues in the cytosolic half of helix 6 in receptor activation. The results support the

hypothesis that a salt bridge between the highly conserved arginine (R1433.50) of the E/DRY motif of helix 3 and a conserved glutamate (E2896.30) on helix 6 constrains the 1b-AR in the inactive state. In fact, mutations of E2896.30 that weakened the R1433.50-E2896.30 interaction constitutively activated the receptor. The functional effect of mutating other amino acids on helix 6 (F2866.27, A2926.33, L2966.37, V2996.40, V3006.41, and F3036.44) correlates with the extent of their interaction with helix 3 and in particular with R1433.50 of the E/DRY sequence.

2.1.3. CCR5

The G-protein coupled receptor CCR5 is the main co-receptor for macrophage-tropic HIV-1 strains. Peterlini [9] has built a structural model of the chemokine receptor CCR5 and used it to explain the binding and selectivity of the antagonist TAK779. Models of the extracellular (EC) domains of CCR5 have been constructed and used to rationalize current biological data on the binding of HIV-1 and chemokines. Residues spanning the transmembrane region of CCR5 have been modeled after rhodopsin, and their functional significance examined using the evolutionary trace method. The receptor cavity shares six residues with CC-chemokine receptors CCR1 through CCR4, while seven residues are unique to CCR5. The contribution of these residues to ligand binding and selectivity is tested by molecular docking simulations [9] of TAK779 to CCR1, CCR2, and CCR5. TAK779 binds to CCR5 in the cavity formed by helices 1, 2, 3, and 7 with additional interactions with helices 5 and 6. TAK779 did not dock to either CCR1 or CCR2. The results are consistent with current site-directed mutagenesis data and with the observed selectivity of TAK779 for CCR5 over CCR1 and CCR2. The specific residues responsible for the observed selectivity were identified. The four EC regions of CCR5 have been modeled using constrained simulated annealing simulations [9]. Applied dihedral angle constraints are representative of the secondary structure propensities of these regions. Tertiary interactions, in the form of distance constraints, are generated from available epitope mapping data. Analysis of the 250 simulated structures provides new insights to the design of experiments aimed at determining residue-residue contacts across the EC domains and for mapping CC-chemokines on the surface of the EC domains.

2.1.4. CXCR4

CXCR4 belongs to the peptide receptor and rhodopsin-class of GPCR superfamily. More recent investigations provided molecular support that CXCR4 have a critical role in determining the metastatic destination of breast cancer cells. In a recent paper [10] a computational pipeline, integrating protein structure prediction, molecular dynamics simulations, automated molecular docking, and Brownian dynamics simulations were employed to investigate the dynamic and energetic aspects of CXCR4 associating with SDF-1. The entire simulation revealed the surface distribution feature of electrostatic potentials and conformational "open-close" process of the receptor. The possible binding conformation of CXCR4 was identified, and the CXCR4-SDF-1 binding complex was generated. On the basis of, the modeling and simulation, a binding mechanism hypothesis between CXCR4 and SDF-1 and its relationship to the signal transduction has been proposed.

2.2. Catalytic Receptors

Many cell-surface receptors seem to convert an extracellular signal into an intracellular one by regulating the activity of a G protein, but some seem to signal the cell more directly. These are "catalytic receptor" proteins, and the best experimentally studied examples in animal cells are single transmembrane-spanning proteins, the tyrosine-specific protein kinases with their catalytic domain exposed on the cytoplasmic side of the plasma membrane. When activated by ligand binding, they transfer the terminal phosphate group from ATP to the hydroxyl group on a tyrosine residue of selected proteins in the target cell. Included in this family of receptors are those for insulin, platelet-derived growth factor (PDGF), epidermal growth factor (EGF), which stimulates epidermal cells and a variety of other cells to divide.

2.2.1. Insulin

Molecular dynamics simulation [11] indicates that the dynamical behaviour of the insulin dimer is asymmetric. Atomic level knowledge of the interaction modes and protein conformation in the solvation state identifies dynamical structures, held by hydrogen bonds that stabilize, mainly in one monomer, the interaction between the chains. Dynamic cross-correlation analysis shows that the two insulin monomers behave asymmetrically and are almost independent. Solvation energy, calculated to evaluate the contribute of each interface residue to the dimer association pattern, well compares with the experimental association state found in protein mutants indicating that this parameter is an important factor to explain the association properties of mutated insulin dimers.

2.2.2. Epidermal Growth Factor (EGF)

Molecular dynamics simulations [12] of murine epidermal growth factor (mEGF), human epidermal growth factor (hEGF), and human transforming growth factor (hTGF) structures with explicit water molecules were used to investigate differences in backbone stability, changes in secondary structure, interdomain flexibility, and weakly polar interactions. Backbone root mean square deviations of sections of each peptide show that the most stable regions in mEGF and hEGF are the A-, B-, and C-loops, whereas the most stable regions in hTGF are the A- and B-loops. The secondary structure in the B-loops of mEGF and hEGF differ significantly from the nuclear magnetic resonance (NMR) structures of mEGF and hEGF. The position and type of turns in the B-loop of mEGF and hEGF increase the interstrand distance of the antiparallel β-sheets thereby disrupting their structure.

2.3. Chemotactic Sensors

Bacterial chemoreceptors mediate chemotaxis by recognizing specific chemicals and regulating a noncovalently associated histidine kinase. Ligand binding to the external domain of the membrane-spanning receptor generates a transmembrane signal that modulates kinase activity inside the cell. This transmembrane signaling is being investigated by novel strategies, which have revealed a remarkably subtle conformational signal carried by a signaling helix that spans the entire length of the receptor (Fig. 2a). Multiple, independent lines of evidence [13] indicate that, in the periplasmic and transmembrane domains, conformational signaling is a piston-type sliding of the signaling helix towards the cytoplasm.

2.3.1. Sensory Rhodopsin

Microbial rhodopsin is a seven-helix membrane protein with retinal as a prosthetic group. Rhodopsins are distributed throughout the Bacteria, Archaea, and Eukaryota. These

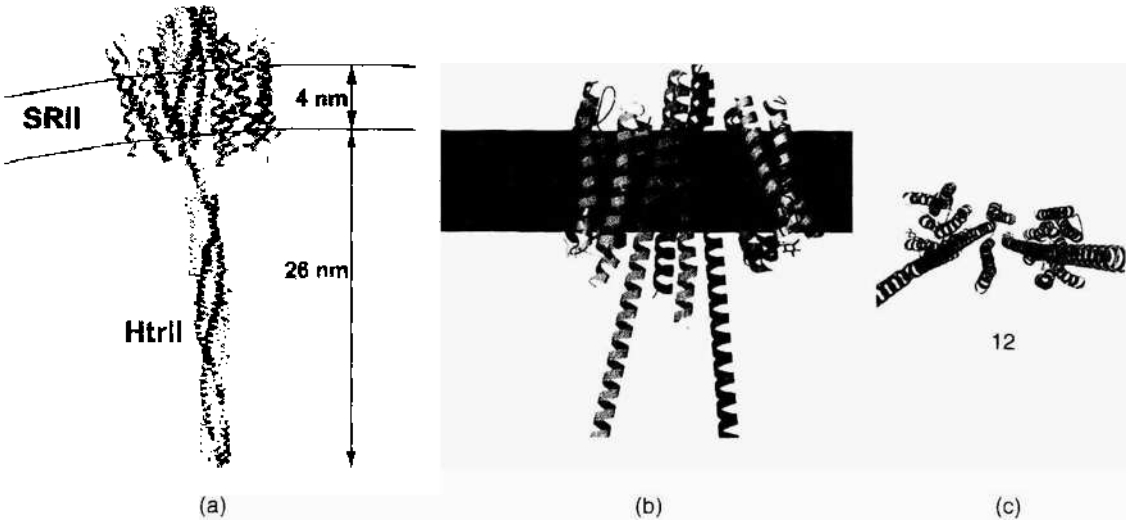


Figure 2. Dimeric structure of sensory rhodopsin II; (a) assumed total configuration; (b) side view, and (c) top view of the crystallographic structure (PDB code 1H2S).

photoactive proteins have common structural features: light-driven ion transport and phototaxis. The sensors activate a signal transduction chain similar to that of the two-component system of eubacterial chemotaxis. The link between the photoreceptor and the following cytoplasmic signal cascade is formed by a transducer molecule (Fig. 2a). In the case of the archaebacterial sensory rhodopsin SRII, the activation of the transducer HtrII by its receptor SRII leads to a signaling cascade. The conformational change of the transducer's transmembrane helices propagates to the tip of the coiled-coil cytoplasmic domain.

The next steps in the signalling cascade involve the kinase CheA, the coupling protein CheW, and the response regulators CheY and CheB. Phosphorylated (P) CheY functions as a switch factor of the flagellar motor. CheB together with CheR are involved in the adaptation processes of the bacteria. The archaebacterial sensory rhodopsin SRII is a green-sensitive photosensor that transduces light signals into attractant and repels migratory responses. Signal transduction by this photoreceptor occurs by means of the membrane-bound transducer protein HTRII, whose activation initiates phosphorylation cascades that modulate the flagellar motors. In addition, analogous to bacteriorhodopsin (BR), SRII can operate as a proton pump when not bound to the transducer, suggesting that key elements of the mechanism of energy transduction in BR have been conserved in SRII. Unlike BR, SRII has a slow photocycle and its signaling state characterizes with a half-life time of hundreds ms, which makes its pumping to be not efficient. The mechanism of SRII proton pumping and signaling is not understood yet. A comprehensive understanding of the mechanism of action of SRII requires a knowledge of its high-resolution structure. Recently several high resolution structures of SRII receptor have been obtained from crystals grown in cubic lipid phase [14].

The first high-resolution structure of NpSRII (PDB code 1JGJ), which, compared with the structure of BR, revealed relatively minor changes near the chromophore with slightly less strain on its conjugated system, but a 1 Å repositioning of the more distant positive charge of Arg72. The cytoplasmic side is significantly more hydrophobic than that of BR, and bilayer-exposed face it was proposed to interact with the cognate transducer, NpHtrII (PDB code 1JGJ). An independent high-resolution NpSRII structure (PDB code 1H68) that agrees well with the structure of 1JGJ with the exception of an additional partially occupied putative chloride ion near Arg72 is also available [15]. More recently the structure of SRII complex with a shot transducer has been reported [16] (PDB code 1H2S). The structures of SRII and the truncated HTRII are depicted in Figures 2a and 2b.

To understand the signal transduction by SRII we also need to know the structure of a linker domain and the water soluble part of HTRII, but their structures so far are unresolved. Molecular dynamics simulations and normal mode analysis of the membrane parts of the SRII-HTRII complex are currently performed (S. Grudinin, private communication).

2.4. Receptor-Ligand Interactions

Since signal transduction by transducer complexes is initiated in many cases by receptor-ligand interactions, it is useful to consider computational studies on the unbinding, unfolding, and docking of proteins. This is summarized briefly at various examples below. Two different molecular simulation methods had been applied: "targeted molecular dynamics" and "steered molecular dynamics" simulations. Both methods are related but different. In steered molecular dynamics studies, pioneered by Grubmüller et al. [17], a force is applied directly to certain atoms in the molecule and the simulation time is many orders of magnitude less than the time of the experiment. This technique [18] is related to the "targeted molecular dynamics" of Schlitter et al. [19], where a holonomic constraint is used to displace a generic function of the atomic coordinates. Many molecular dynamics simulations have been performed using both methods.

2.4.1. Titin

Titin (or connectin) is a gigantic protein molecule that spans half of the sarcomere in vertebrate. The molecule is of the order of a micrometer in length and is thought to provide a passive restoring force as muscles elongate. Atomic force microscopy and laser tweezer experiments have been used to stretch titin molecules to more than ten times their

folded lengths. To explain the observed relation between force and extension, it has been suggested that the immunoglobulin and fibronectin domains unfold one at a time in an all-or-none fashion. Various steered molecular dynamics simulations [17] have been employed to study the forced unfolding of titin [20–22]. The forced unfolding of fibronectin by "targeted molecular dynamics" simulations have been performed by Karplus and coworkers [20]. An external biasing potential on the two ends (N and C terminal) was employed and the protein is treated in the polar hydrogen representation with an implicit solvation model. The latter provides an adiabatic solvent response, which is important for the nanosecond unfolding simulation method. The biasing potential results are confirmed by constant force unfolding simulations. The unfolding forces found in the simulations are of the order of those observed experimentally, even though the speed of the former is more than six orders of magnitude greater than that used in the latter.

Steered molecular dynamics simulation of force-induced titin immunoglobulin domain unfolding [22] led to the discovery of a significant potential energy barrier at an extension of ~14 Å on the unfolding pathway that protects the domain against stretching. Simulations showed that this barrier is due to the concurrent breaking of six interstrand hydrogen bonds (H-bonds) that is preceded by the breaking of two to three hydrogen bonds between different strands, the latter leading to an unfolding intermediate. The simulation results are supported by atomic force microscopy data. The authors [22] performed a structural and energetic analysis of the H-bonds breaking. In nanosecond simulations, water molecules were found to repeatedly interact with the protein backbone atoms, weakening individual interstrand H-bonds until a certain group of H-bonds break simultaneously under the influence of external stretching forces. Only when those bonds were broken the generic unfolding took place, which involves hydrophobic interactions of the protein core and exerts weaker resistance against stretching than the key event.

2.4.2. GroEL

The molecular chaperonin GroEL mediates protein folding in the bacterial cell. GroEL-assisted protein folding is regulated by a cycle of large coordinated domain movements in the 14-subunit double-ring assembly. GroEL can recognize a wide range of protein substrates due to the intrinsic flexibility of the active site. Upon binding of ATP and the co-chaperonin GroES, a large closed cavity is created into which the unfolded peptide is released so that it can refold in a shielded environment, referred to as the "Anfinsen cage." When ATP is hydrolyzed to ADP in the cis ring and ATP is bound in the trans ring, the native (or still misfolded) protein, ADP and GroES are released and the corresponding assembly is formed in the trans ring. The Anfinsen cage is created from the unliganded (closed) structure by large, nearly rigid body movement of the intermediate and apical domains of the cis-ring subunits. These movements are coupled to the binding of ATP and GroES in a process with high positive cooperativity. The behavior of the trans ring is out of phase with the cis ring, due to the negative cooperativity between the rings. The transition path between the closed (unliganded) and the open (liganded) states have been calculated with a targeted molecular dynamics simulation [23]. The simulation results, which indicated that transient interactions along the transition path are essential for GroEL function, provided a detailed structural description of the motions that are involved in the GroEL allosteric cycle.

2.4.3. AN02/DNP

Unbinding of a spin-labeled dinitrophenyl (DNP) hapten from the monoclonal antibody AN02 Fab fragment has been studied by force probe molecular dynamics (FPMD) simulations by Grubmüller and Heymann [24]. In their nanosecond simulations, unbinding was enforced by pulling the hapten molecule out of the binding pocket. Detailed inspection of the FPMD trajectories revealed a large heterogeneity of enforced unbinding pathways and a correspondingly large flexibility of the binding pocket region, which exhibited induced fit motions. Principal component analyses were used to estimate the resulting entropic contribution of about 6 kcal/mol to the AN02/DNP-hapten bond. This large contribution may explain the surprisingly large effect on binding kinetics found for mutation sites that are not directly

involved in binding. The authors proposed that such "entropic control" optimizes the binding kinetics of antibodies. Structural analysis of the FPMD simulations suggests that, in contrast to free energies of unbinding, the effect of mutations on unbinding forces is generally nonadditive.

2.4.4. Scorpion Toxin Lq2

The association of the scorpion toxin Lq2 and a potassium ion channel has been studied using the Brownian dynamics (BD) simulation method [25]. All of the 22 available structures of Lq2 in the Brookhaven Protein Data Bank (PDB) determined by NMR were considered during the simulation, which indicated that the conformation of Lq2 affects the binding between the two proteins significantly. Among the 22 structures of Lq2, only four structures dock in the binding site of the K channel with a high probability and favorable electrostatic interactions. From the four candidates of the Lq2-K channel binding models, the authors identified a good three-dimensional model of Lq2-K channel complex through triplet contact analysis and electrostatic interaction energy estimation by BD simulation. The favorable interactions were identified which stabilize the association between the two proteins. These observations are in good agreement with the experimental results and can explain the binding phenomena between scorpion toxins and potassium channels at the level of molecular structure. The consistency between the BD simulation and the experimental data indicates that the three-dimensional model of Lq2-K channel complex is reasonable and can be used in further biological studies such as rational design of blocking agents of K channels and mutagenesis in both toxins and K channels.

2.4.5. CD2-CD58

The force-induced detachment of the adhesion protein complex CD2-CD58 was studied by steered molecular dynamics simulations [26]. The forced detachment of CD2 and CD58 shows that the system can respond to an external force by two mechanisms, which depend on the loading rate. At the rapid loading rates of 70 and 35 pN/ps (pulling speeds of 1 and 0.5 Å/ps) the two proteins unfold before they separate, whereas at slower loading rates of 7 and 3.5 pN/ps (pulling speeds of 0.1 and 0.05 Å/ps), the proteins separate before the domains can unfold. When subjected to a constant force of 400 pN, the two proteins separated without significant structural distortion. These findings suggest that protein unfolding is not coupled to the adhesive function of CD2 and CD58. The simulations further confirmed that salt bridges primarily determine the tensile strength of the protein-to-protein bond, and that the order of salt bridge rupture depends mainly on the position of the bond, relative to the line of action of the applied force. Salt bridges close to this line break first. The importance of each of the salt bridges for adhesion, determined from the simulations, correlates closely with their role in cell-to-cell adhesion and equilibrium binding determined by site-directed mutagenesis experiments.

3. MEMBRANE CHANNELS

Membrane channels are proteins, which reside in membranes acting as "valves" thereby permitting selective or nonselective flow of molecules along their concentration gradients across the membrane. One may distinguish between pores and ion channels. In contrast to pores exhibiting weak ion selectivity, ion channels are highly specific filters regulating the ion balance of living cells. In contrast to transporters (discussed in the Section 3), channels are passive devices whose opening and closing ("gating") is controlled externally by ligands, voltage, pH value, or mechanical stress. The vast variety of membrane functions (transport of matter and energy, signal processing, and transduction, etc.) is provided by a relatively simple structure of lipid bilayer with incorporated (integral) and/or attached (peripheral) proteins.

A lot of experimental effort has been put into resolving membrane-bound processes, from the structural as well as dynamical point of view, but traditionally with a relatively small input from the theoretical approaches. The primary functional structures responsible for completion of membrane processes are proteins, so they have always called for the central attention. As far as the molecular structure determination of membrane proteins is

concerned, obtaining a crystallized functional form of the membrane protein is still a generally unsolved problem; up to now, only few structures have been resolved up to the atomic details. Consequently, modeling and simulations of membrane processes has been long facing the problem of lack in structural data, not to mention the problem of computational limitations for such complex systems. An excellent review on simulation approaches to ion channels is given by Tieleman et al. [27].

In the past few years, the most rapid development in the field of membrane proteins simulations has been seen specifically in the area of ion channels. The detailed insights into the molecular architecture provided by breakthroughs in determination of crystal structures at the atomic resolution, as well as advances in computer simulations (thanks to the increased speed and power of computational facilities), have been gained. This is, by and large, seen in a recent successful structural determination of the bacterial potassium channel KcsA [28–31], accompanied by the computational simulation of the structure [32–37] that almost completed the detailed picture of structure-function relationship for the potassium ion transport process.

3.1. Pores

MD simulations provide a valuable tool for studying membrane proteins, enabling us to probe their conformational dynamics in both membrane and detergent micelle environments (Bond and Sansom [39]). They are of particular value in enabling us to extrapolate from the essentially static (time- and space-averaged) structure revealed by X-ray diffraction to a more dynamic picture of the behavior of a single proteins in an environment mimicking a small patch of the bacterial membrane. MD simulations have been employed in a number of studies, most notably to probe protein and solvent dynamics in relationship to permeation mechanisms in porins.

3.1.1. Porins

Some types of bacteria are characterized by the construction of their cell wall: the cytoplasmic or inner membrane engulfes the cytoplasm; a second, outer membrane surrounds the cell with few contacts to the inner membrane. In between is the periplasm, where some metabolic activities take place. The exchange of substances between cytoplasm and periplasm is reguated by highly specific transport systems, usually ion channels, whereas the exchange between periplasm and environment occurs via outer membrane proteins (OMPs), which may be non- or weekly selective for groups of substances. Perhaps the best-characterized family of OMPs would be the porins. These include both relatively nonspecific general diffusion pores and also more specific passive pores (e.g., for oligosaccharides) across the outer membrane. Thus, they are an important component of the transport properties of the bacterial membrane. Other transport proteins in outer membranes include those for ferric ions and for vitamin B12, and export pathways for polypeptide toxins and hydrophobic drugs. In addtion to transport proteins, outer membranes include a number of membrane-bound enzymes. Several structures of such enzymes have been determined, including those of a protease OmpT and two OMPs acting on lipid substrates, OMPLA and PagP.

Structure. The construction principle of porins is the same irrespective of their type: a chain of 300–420 amino acids folds to an antiparallel β-barrel of 16 or 18 strands (Fig. 3). The wall of the pore has a thickness of one amino acid only. On the side of the barrel facing the periplasm, the beta strands are connected by short loops or turns. On the other side, the loops directed to the environment are larger and variable. The loop connecting beta strands 5 and 6 is of special importance: it is folded into the barrel and constricts the cross section. At the narrowest point, there are some ionizable amino acids. The filter properties of the pore are defined at this point. Porins are inserted in the outer membrane as trimers. Amino and carboxy termini of the single molecules face the threefold symmetry axis of the complex. As found in other transmembrane proteins there are two belts of aromatic amino acids pointed to the surfaces of the membrane.

To date, the structures of 20 such OMPs have been solved by X-ray diffraction and by nuclear magnetic resonance. Thus, OMPs provide an opportunity for molecular dynamics

Figure 3. Ribbon view of OmpF porin from *Rhodobacter capsulatus* (PDB entry 2omf).

simulation studies to explore the conformational dynamics of a whole family of structurally related membrane proteins, to define both common dynamic properties and functionally important differences between individual species of OMPs [38, 39].

Function. Porins are tubes with a diameter of about 1 nm that are filled with water. Nonspecific porins allow the diffusion of ions and molecules up to a molecular weight of 600. The diffusion speed depends on both the difference of concentration in the periplasm and outside and the molecular weight of the solute. The passing of ions may be regulated electrically. If (*in vitro*) a voltage of + or −100 mV is applied, the channel is closed for ions (voltage gating). This phenomenon is also found in another class of beta-barrel pores, the toxins. From mutation experiments, it was concluded that there is no movement of the loop constricting the inside of the channel. Probably the applied voltage changes the electrostatic properties of the interior wall. A physiological function of voltage gating is not obvious. Because they are well characterized, both structurally and functionally, the porins represent ideal systems for addressing questions about the fundamental principles underlying ion flow in molecular pores at the molecular level using theoretical models. There have been a number of theoretical studies of porins, focusing on different aspects of porin activity.

3.1.2. OmpF

Realistic simulations of porins have been performed with explicit ions, solvent molecules [40] and also phospholipid bilayer membrane [41]. To explore the mechanism of ion conduction, Suenaga et al. simulated OmpF in the presence of an applied transmembrane potential [40]. The translocation of a single Na^+ through the channel was observed in 1.3 ns under the influence of a potential of 500 mV. Tieleman and Berendsen [41] generated a 1-ns MD simulation of an atomic model of OmpF trimer embedded into an explicit phospholipid bilayer membrane with a few explicite counterions (those needed to balance the total charge of OmpF). Since the MD simulations are computationally intensive, and yet, the timescale of ion permeation is significantly longer than can be currently simulated, a number of studies have been used BD to explore the ion conduction mechanism [42–44]. In this approach, the channel and the ions were represented explicitly while the influence of the surrounding solvent molecules was incorporated implicitly via a stochastic random force, a friction coefficient damping the velocity, and some effective potential (generally calculated on the basis of a continuum electrostatic approximation). A good correlation was achieved between calculated single-ion transmission probabilities and experimental ion selectivity. Although approximations such as BD and PB calculations are less computationally intensive than all-atom MD and hence very attractive, their validity, however, is still unknown in the context of molecular pores. For example, it is unclear whether a structureless dielectric continuum is a valid representation. Furthermore, the channel structure is generally kept rigidly fixed during BD

simulations and the coupling to protein fluctuations is ignored. To answer some of these questions, Im and Roux [45] have performed molecular dynamics simulations of OmpF porin embedded in an explicit (DMPC) bilayer bathed by a 1 M [KCl] aqueous salt solution [45] and have explored the microscopic details of the mechanism of ion permeation. The atomic model includes the OmpF trimer, 124 lipids, 13470 water molecules as well as 231 K^+ and 201 Cl^-, for a total of 70,693 atoms. The structural and dynamical results are in excellent agreement with the X-ray data. In the narrowest part of the aqueous pore it was observed that the water molecules are markedly oriented perpendicular to the channel axis because of the strong transversal electrostatic field arising from pore-lining residues. On average the size of the pore was smaller during the simulation than in the X-ray structure, undergoing small fluctuations. Remarkably, it was observed that K^+ and Cl^- follow two well-separated average pathways spanning over nearly 40 Å along the axis of the pore. In the center of the monomer, the two screwlike pathways have a left-handed twist, undergoing a counterclockwise rotation of 180 degrees from the extracellular vestibule to the pore periplasmic side. In the pore, the dynamical diffusion constants of the ions are reduced by about 50% relative to their value in bulk solvent. Analysis of ion solvation across the channel reveals that the contributions from the water and the protein are complementary, keeping the total solvation number of both ions nearly constant. Analysis suggested that ion-ion pairs play an important role. In particular, it is observed that the passage of Cl^- occurs only in the presence of K^+ counterions, and isolated K^+ can move through the channel and permeate on their own. The presence of K^+ in the pore screens the negative electrostatic potential arising from OmpF to help the translocation of Cl^- by formation of ion pairs.

3.1.3. OmpT

Molecular dynamics simulations [46] have been performed on the outer membrane protease OmpT embedded in a lipid bilayer. It was observed that the β-barrel was tilted relative to the bilayer plane. The greatest degree of conformational flexibility was seen in the extracellular loops. A complex network of fluctuating H-bonds was detected between the active site residues, which supports a catalytic mechanism. A configuration yielded by docking calculations of OmpT simulation snapshots and a model substrate peptide Ala-Arg-Arg-Ala was used as the starting point for an extended Hückel calculation on the docked peptide. The trajectories of water molecules reveal exchange of waters between the intracellular face of the membrane and the interior of the barrel but no exchange at the extracellular mouth. This suggests that the porelike region in the center of OmpT may enable access of water to the active site from below. The simulations appeared to reveal the presence of specific lipid interaction sites on the surface of the OmpT barrel, revealing the ability of extended MD simulations to provide meaningful information on protein–lipid interactions.

3.1.4. OmpA

OmpA is a relatively simple bacterial outer membrane protein, for which two X-ray structures are available. The bacterial outer membrane protein OmpA is composed of an N-terminal 171-residue β-barrel domain (OmpA171) that spans the bilayer and a periplasmic, C-terminal domain of unknown structure. OmpA has been suggested to primarily serve a structural role, as no continuous pore through the center of the barrel can be discerned in the crystal structure of OmpA171. However, several groups have recorded ionic conductances for bilayer-reconstituted OmpA171. To resolve this apparent paradox Sansom et al. [47] have used molecular dynamics simulations on OmpA171 to explore the conformational dynamics of the protein, in particular the possibility of transient formation of a central pore. A total of 19 ns of MD simulations of OmpA171 have been run, and the results were analyzed in terms of (1) comparative behavior of OmpA171 in different bilayer and bilayer-mimetic environments, (2) solvation states of OmpA171, and (3) pore characteristics in different MD simulations. Significant mobility was observed for residues and water molecules within the β-barrel. A simulation in which putative gate region side chains of the barrel interior were held in a nonnative conformation led to an open pore, with a predicted conductance similar to experimental measurements. The OmpA171 pore has been shown to be somewhat

more dynamic than suggested by the crystal structure. A gating mechanism was proposed to explain its documented channel properties.

3.1.5. FhuA

One of the more complex members of the OMPs is FhuA. Its primary function is to provide a binding site on the outer membrane surface for siderophores, such as ferrichrome, and subsequently to facilitate their energy-dependent transport across the membrane, presumably powered by the TonB-ExbBD protein complex that resides in the cytoplasmic membrane. Crystal structures of FhuA with and without a bound ferrichrome molecule have provided some clues as to the initial stages of the siderophore transport mechanism. Sansom et al. [48] have performed molecular dynamics simulations of FhuA and of the FhuA-ferrichrome complex, both embedded in a phospholipid bilayer, to probe the short timescale dynamics of this integral membrane protein, and to explore possible mechanistic implications of this dynamic behavior. Analysis of the dynamics of the protein suggests that the extracellular loops move as relatively rigid entities relative to the transmembrane β-barrel. Comparison of the two simulations (with and without bound ferrichrome) revealed some ligand-induced changes in loop mobility. These simulations support the proposal that binding of ferrichrome initiates a signaling mechanism that ultimately leads to the TonB-mediated partial or total removal of the core domain from the β-barrel, thus opening up a permeable pore. These simulations are among the longest that have been performed on a complex membrane protein. However, a simple analysis of sampling reveals that the description of protein motions is far from complete.

3.1.6. Aquaporin (AQP1)

The extraordinary permeation rate of 3 billion water molecules per second per single aquaporin-1 (AQP1) molecule, combined with the strict selectivity for water, have challenged several MD simulations in order to elucidate the relation between structural determinants and selectivity in porins [49–53]. Bacterial porins, for example, from the outer membrane of *E. coli* allow diffusion of hydrophilic molecules with molecular weight up to 600 Da and exhibit modest ionic selectivity. The existence of membrane water channels was predicted in the 1950s [54]. Today the detailed structures of several human [55, 56] and bacterial [57] porins are known.

Selectivity. The nature of the control of water/proton selectivity in biological channels is a problem of fundamental importance. Most studies have proposed that an interference with the orientational requirements of the so-called proton wire is the source of selectivity. The elucidation of the structures of aquaporins, which have evolved to prevent proton transfer (PT), provided a clear benchmark for exploring the selectivity problem. Previous simulations of this system have not examined, however, the actual issue of PT but only considered the much simpler task of the transfer of water molecules. In a paper of Burykin and Warshel [58] aquaporin is taken as a benchmark and the origin of the water/proton selectivity in this and related systems was quantified. This was done by evaluating in a consistent way the free energy profile for transferring a proton along the channel and relating this profile to the relevant PT rate constants. It was found that the water/proton selectivity is controlled by the change in solvation free energy upon moving the charged proton from water to the channel. The reason for the focus on the elegant concept of the proton wire and the related Grotthuss-type mechanism was also considered. It was concluded that these mechanisms are clearly important in cases with flat free energy surfaces (e.g., in bulk water, in gas phase water chains, and in infinitely long channels). However, in cases of biological channels, the actual PT mechanism is much less important than the energetics of transferring the proton charge from water to different regions in the channels.

Permeation. The aqueous pore and the mechanism of ion conduction have been studied by MD simulations with explicit ions and solvent molecules [59] and also with a lipid bilayer membrane [41]. The complete translocation of a single Na^+ ion in the OmpF under a potential of 500 mV was observed in 1.3 ns [59]. The simulation [41] revealed that a strong electric

field-oriented transverse to the pore axis influences the ion transport. The flow of one ion through various bacterial porins (OmpF, PhoE, OmpK36, and mutants thereof) has been studied using BD simulations [42–44] which showed a good agreement between calculated transmission probabilities and experimental ion selectivity.

The MD simulations of the protein tetramer for several nanoseconds [51] reveal that the water molecules are strongly oriented in the channel interior with their dipoles rotating by about 180 degrees during permeation. The dipoles are aligned with the electric field, which originates from the dipoles of two specific helices, HB and HE, of the protein. Hydrogen competition between water molecules and a few polar groups in the pore was found to dominate the permeation process. During the course of a 10 ns simulation, a total of 16 such permeation events are observed, a result that is in good agreement with the very high rate of permeation known from experimental work. Hydrogen bond statistics for water molecules show that there are two major interaction sites inside the channel: the Asn-Pro-Ala (NPA) region and the aromatic-arginine (Ar/R) constriction region. The two highest enthalpic barriers to the passage of water molecules are located directly adjacent to NPA region. This together with the water rotation that also occurs here, suggest that the NPA region is a major selectivity filter. Porins are often designed such that proton conduction is strictly prevented to maintain the proton gradient across the cell membrane that serves as a major energy storage mechanism. In the simulations, frequent simultaneous hydrogen bonding of water molecules to the two NPA asparagines has been observed, thereby weakening interactions among adjacent water molecule in the pore. As contiguous hydrogen bonded water chains are known to be efficient proton conductors, it was suggested that this region is the main proton filter [52].

Water permeation and electrostatic interactions between water and channel have been investigated [53] in the *Escherichia coli* glycerol uptake facilitator GlpF, a member of the aquaporin water channel family, by molecular dynamics simulations. A tetrameric model of the channel embedded in a lipid bilayer membrane was used for the simulations. During the simulations, water molecules pass through the channel in single file. The movement of the single file water molecules through the channel is concerted, and it was shown that it can be described by a continuous-time random-walk model. The integrity of the single file remains intact during the permeation, indicating that a disrupted water chain is unlikely to be the mechanism of proton exclusion in aquaporins. Specific hydrogen bonds between permeating water and protein at the channel center, together with the protein electrostatic fields enforce a bipolar water configuration inside the channel with dipole inversion at the NPA motifs. At the NPA motifs water–protein electrostatic interactions facilitate this inversion. Furthermore, water–water electrostatic interactions are in all regions inside the channel stronger than water–protein interactions, except near a conserved, positively charged Arg residue. Variations of the protein electrostatic field through the channel, owing to preserved structural features, completely explain the bipolar orientation of water. This orientation persists despite water translocation in single file and blocks proton transport. Furthermore, It was found that for permeation of a cation, ion-protein electrostatic interactions are more unfavorable at the conserved NPA motifs than at the conserved Arg, suggesting that the major barrier against proton transport in aquaporins is faced at the NPA motifs.

The difference between osmotic permeability (pf) and diffusion permeability (pd) of single-file water channels has been studied by Schulten et al. [60]. They have demonstrated that the pf/pd ratio corresponds to the number of effective steps a water molecule needs to take to permeate a channel. While pd can be directly obtained from equilibrium molecular dynamics simulations, pf can be best determined from simulations in which a chemical potential difference of water has been established on the two sides of the channel. In light of this, they suggested a method to induce in molecular dynamics simulations a hydrostatic pressure difference across the membrane, from which pf can be measured. Simulations using this method were performed on aquaporin-1 channels in a lipid bilayer. The pf/pd ratio is explained in terms of channel architecture and conduction mechanism.

3.1.7. Glycerolporin Facilitator (GlpF)

Aquaporins constitute a class of membrane proteins with more than 150 members, including a glycerol-conducting subclass (aquaglyceroporins). The *E. coli* glycerol facilitator, GlpF,

selectively conducts glycerol and water, excluding ions and charged solutes. The detailed mechanism of the glycerol conduction and its relationship to the characteristic secondary structure of aquaporins and to the Asn-Pro-Ala (NPA) motifs in the center of the channel are unknown. The X-ray structure of the Escherichia coli glycerol facilitator, GlpF, a homotetramer highly similar to AQP1, was reported at a resolution of 2.2 Å [61]. A complete conduction of glycerol through the GlpF was deduced from molecular dynamics simulations [49], and key residues facilitating the conduction were identified. The nonhelical parts of the two half-membrane-spanning segments expose carbonyl groups towards the channel interior, establishing a curve-linear pathway. The conformational stability of the NPA motifs is important in the conduction and critical for selectivity. Water and glycerol compete in a random manner for hydrogen bonding sites in the protein, and their translocations in single file are correlated. The suggested conduction mechanism should apply to the whole family.

3.1.8. Maltoporin

Gram-negative bacteria are enclosed by an inner and an outer membrane. Whereas transport through the inner membrane is accomplished by active transporters, the passage across the outer membrane occurs by passive diffusion through porin channels. Nonspecific porins (e.g., the matrix porin OmpF) allow the uptake of ions and small molecules. Certain larger solutes are translocated by solute-specific porins such as maltoporin which is permeable to maltose and longer maltodextrin sugars. The binding affinity of maltodextrins to maltoporin increases with increasing chain length, while the permeation rate decreases for longer maltodextrins. There is a high energy cost associated with desolvating a polar sugar and crossing the nonpolar interior of a membrane. Maltoporin provides the glucosyl units with tight binding sites within the pore. This lowers the energy of the sugar halfway across the membrane, thus enhancing the local maltodextrin concentration, and allows for substrate specificity by excluding larger or less elongated compounds. In facilitating the translocation, maltoporin has to reconcile two conflicting objectives: (1) tightly binding the sugar in the pore and (2) facilitating the motion of the sugar chain through the pore. In a recent study [62], it was shown how maltoporin achieves this by the creation of a delocalized binding site, which allows relatively free diffusion of the maltodextrin chain along the pore while maintaining strong interactions. The minimum-energy path of maltohexaose translocation was obtained by the conjugate peak refinement method, which optimizes a continuous string of conformers without applying constraints. This reveals that the protein is passive while the sugar glides screwlike along the aromatic lane. Near instant switching of sugar hydroxyl H bond partners results in two small energy barriers (of 4 kcal/mol each) during register shift by one glucosyl unit, in agreement with a kinetic analysis of experimental dissociation rates for varying sugar chain lengths. Thus, maltoporin functions like an efficient translocation enzyme and the slow rate of the register shift is due to high-collisional friction.

3.1.9. M2 Channel

The transmembrane fragment of the influenza virus M2 protein forms a homotetrameric channel that transports protons. Molecular dynamics simulations have been performed by Sansom et al. [63] and Schweighofer and Pohorille [64]. In the more recent paper [64], simulations of M2 were reported that helped to elucidate the mechanism of channel gating by four histidines that occlude the channel lumen in the closed state. The authors tested two competing hypotheses. In the "shuttle" mechanism, the nitrogen atom on the extracellular side of one histidine is protonated by the incoming proton, and, subsequently, the proton on the nitrogen atom is released on the opposite side. In the "water–wire" mechanism, the gate opens because of electrostatic repulsion between four simultaneously biprotonated histidines. This allows for proton transport along the water wire that penetrates the gate. For each system, composed of the channel embedded in a hydrated phospholipid bilayer, a 1.3-ns trajectory was obtained. It was found that the states involved in the shuttle mechanism, which contain either single-protonated histidines or a mixture of single-protonated histidines plus one biprotonated residue, were stable during the simulations. Furthermore, the orientations and dynamics of water molecules near the gate are conducive to proton transfer. In contrast,

the fully biprotonated state is not stable. Additional simulations showed that if only two histidines are biprotonated, the channel deformed but the gate remained closed. These results support the shuttle mechanism but not the gate-opening mechanism of proton gating in M2.

3.2. Ion Channels

The membrane lipid bilayer is highly impermeable to ions, because the energy barrier for transferring a hydrated ion to the low dielectric environment of acyl chain region is prohibitively large. Nevertheless, every cell has to have a pathway for ionic flow through the membrane, to maintain proper physiological functioning. Ion channels are transmembrane protein structures [65, 66], which ensure the continuous transport pathway for diffusive flow of ions across the membrane. Since the net flow of electric charges gives rise to a rapid change in transmembrane potential, ion channels play a key role in generating and propagating action potentials in nervous system. Moreover, it is well established that improper ion channel functioning leads to a number of pathological conditions, thereby presenting those structures as the primary targets for pharmacological use. Ion channels are functionally and structurally simplest among the membrane transport proteins, since the transport itself does not call for a (major) conformational changes of protein structure. Conformational changes occur in the process of gating that is, are needed for discrimination between opened (permeable to selected ions) and closed (impermeable to all ions present) channel states. The tranformation is energy dependent. In contrast to pores, which exhibit weak ion selectivity, ion channels have highly specific filters, which regulate the ion balance of living cells.

From a general point of view with respect to nanotechnology, voltage-gated ion channels are of particular interest. They control electrical activity in nerve, muscle, and many other cell types. The crystal structure of a bacterial voltage-gated channel reveals the astonishingly simple design of its voltage sensor. The voltage-gated ion channels, residing in membranes of living cells, from bacteria to humans, behave like field-effect transistors. In transistors, the flow of electrons through a semiconductor "channel" is governed by the voltage applied to a "gate" electrode. As voltage-sensing devices, these channels can perform much better than their electronic counterparts (Fig. 4). Their high sensitivity to voltage is important because cellular voltage changes are small.

With the protein equivalents, voltage-gated ion channels, an appropriate voltage, imposed across the cell membrane, causes the channels to open and allows a current of ions to cross the membrane. The molecular structures within ion channels that sense the membrane voltage have remained obscure for the 50 years since Hodgkin and Huxley first described [67] their function. The voltage sensors have at last been made visible, in the X-ray structure of a potassium ion channel by MacKinnon and coworkers [68, 69] and a hypothesis for voltage-sensor motion has been suggested.

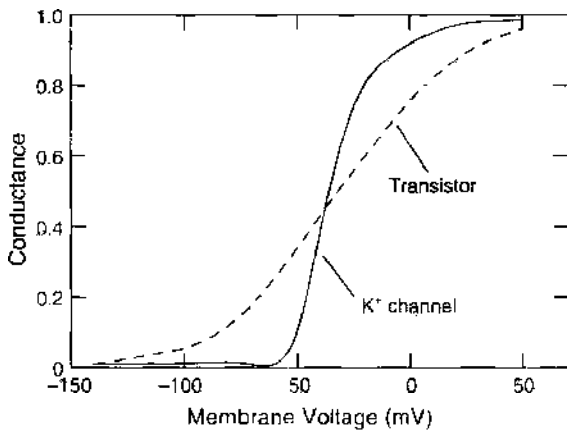


Figure 4. Voltage sensing in a potassium ion channel. The control of ion flow through voltage-gated channels is very sensitive to the voltage across the cell membrane. By comparison, an electronic device such as a transistor is much less sensitive to applied voltage.

One of the most famous examples of an ion channel is the potassium specific channel from *Streptomyces lividans* KcsA, which is regulated by pH. Because of sequence homologies and similar physiologic properties it may be regarded as a prototype for an ion channel. Another class of ion channels, which reside mostly in the cytoplasmic membranes of bacteria, exhibit a significant response to mechanical stress. Upon changes of pressure these mechanosensitive channels alter the diameter of the pores without discrimination of the kind of ions passing through. One outstanding example is MesL which is discussed below.

To assure the precisely controlled transmembrane ionic flow, ion channels have to fulfill three specific points:

1. High selectivity towards specific ion type (or have to highly discriminate one ion type over all the other ions present).
2. Permeation at the diffusion rate (or to provide an energetically barrierless pathway, which ensures the rapid transport of selected ions).
3. Well defined gating control mechanism, in which the conformation of the open state has to be acomplished upon specific stimulus; the known mechanisms include change in transmembrane potential (voltage-gated ion channels), binding of another molecule (ligand-gated or receptor-gated channels) and mechanical stress (mechanosensitive channels or stretch-gated channels).

At present, no single computational technique can describe all the functional properties of an ion channel structure. The choice of the level for the theoretical approach and the use of the computational technique to describe processes of selectivity, permeation, and gating of an ion channel, depend essentially on the timescale of the process itself. While the timescale of the permeation process for the typical channel is in the order of few tenths of nanoseconds, the gating takes the time in the order of milliseconds. The present state of computational speed does not permit the completion of neither of the processes using MD techniques, calling for more coarse-grained modelling. However, BD and continuum theories do not distinguish between monovalent ions, so the modelling of channel selectivity appeals for the MD approach. Although BD simulations have provided usefull data on the total permeation process of ions through channels [70], their reliability concerning details is limited due to the coarse-graining of the channel structure and applications of mean-field approximations of water and lipids. More details, including fluctuations of all constituents of the channel and its environment (all atom model of channel, water and lipids) have been elucidated by use of molecular dynamics (MD) simulations [27, 71] albeit quite limited in time to the few nanoseconds time regime.

Structure. To support the reliability to the results of MD simulations of a full all-atom model of an ion channel (including the channel, the water and the lipid bilayer), it is important to carefully prepare the starting configuration. In particular, the embedding of the crystal structure into a preequilibrated membrane-water system is a nontrivial task, since the insertion of the structure into a prepared cavity in the membrane-water system leaves many local vacancies between the two systems. A too rapid subsequent equilibration renders the channel structure to unfold. The typical rms deviation between the original and this type of equilibrated structure can be of the order of few Å. Obviously, this type of modified structure may not provide correct insights into permeation and selectivity mechanism of the channel. Therefore, a comparison between the crystal structure and the simulation-mediated equilibrated structure is important.

Permeation. The goal of modeling the permeation process is to understand the physical processes underlying the permeation of ions through the channel, with the ability to reproduce present and predict future experimental observations. The timescale of permeation is too long to enable reproducing the experimental data on the process using MD calculations, so the modeling which attempts to reproduce experimental data usually employs more coarse-grained methods like continuum electrostatics and/or BD simulations. Nevertheless, among the computational methodologies used on describing the permeation process, only molecular dynamics has the advantage of providing the data on the molecular kinetic details

at the atomic resolution. However, none of the simulation studies with explicit lipid environment was able to complete the permeation process for the K^+ ion within KcsA channel, because of the current computational limitations for simulations of such a large systems. The permeation process deals with an open state conformation of the protein: the challenge for the theoretical approach is to explain the channel's ability to conduct the ion movement at very high rate and relate it to the detailed molecular structure of the system. In developing the realistic models of ion permeation, one is faced with remarkable complexities that need to be considered: (1) structure determinants of channel-mediated ion transport (i.e., the geometry of the protein-water interface along the permeation pathway) and (2) the distribution of charges in the protein wall, which determine ion–protein interactions along the pathway. Although the precise geometry of ion–transport pathway varies among different channels, there are two characteristic regions that appear to be a general feature of all biological channels: a wider cavity that accomodates hydrated ion and short and narrow selectivity filter. Those two regions seem to be evolutionary conserved, at least in specific channel type (e.g., potassium channels). The mechanism of ion permeation through the channel is generally determined by interactions of permeating ion with the channel wall and ion–ion interaction within the channel pore. Additionally, the interactions are modified with the physical determinants of water molecules within the narrow channel structure, that can be very much different from the bulk water outside the protein structure.

Selectivity. For every experimentally identified ion channel, there is a well-defined ion selection sequence, according to which channels are usually named as potassium, sodim, chlorine or calcium channels. It appears that for the monovalent-selective channels the main selection criterion is the size of the ion, whereas for the calcium channel it is the magnitude of ion charge (calcium channel highly discriminates Na^+ over Ca^{2+}, although the radii of Na^+ and Ca^{2+} do not differ much—0.95 versus 0.99 Å, respectively). Thus, there are dominantely two different selection mechanisms at play: the one operates on the basis of ion size, and the other on the basis of ion charge. The deciding factor on selectivity in channels is the free energy of permeation, namely the variation in free energy of the system as different ion species pass through the channel. In the simple case, one can observe the differences in potential energies at various points along the pore, but for a more quantitative description, a free-energy calculations are needed.

Gating. Ion channels regulate the selective transfer of ions across the membrane in response to different types of stimuli, as for example, changes of pH, transmembrane potential, mechanical stress, or ligand binding. A channel can generally assume two stable conformations, the open and the closed one. The structural part of the channel responsible for the gating controls the accessability of ions to a centrally located water-filled pore. The opening and closing of the gate is accompanied by conformational changes in the protein during gating. The structural and dynamical details of the gating mechanism are the least known properties of ion channels, mostly because of the fact that an opened state of the channel is a transient one, thus not easily fixed to be isolated by crystallization. Consequently, for the majority of known structures no direct comparison of X-ray structures is available and one has to use other experimental techniques that reveal the structural determinants of the gating mechanism.

The next paragraphs deal with some results obtained on specific ion channel structures and the simulation techniques applied to resolve their specific features.

3.2.1. Alamethicin

Alamethicin is an antimicrobial peptide that forms stable channels with well-defined conductance levels. Sansom and coworkers [72] have used extended molecular dynamics simulations of alamethicin bundles consisting of four, five, six, seven, and eight helices in a POPC bilayer to evaluate and analyse channel models and to link the models to the experimentaly measured conductance levels. Their results suggest that four helices do not form a stable water-filled channel and might not even form a stable intermediate. The lowest measurable conductance level is likely to correspond to the pentamer. At higher aggregation numbers the bundles become less symmetrical. Water properties inside the different-sized

bundles are similar. The hexamer is the most stable model with a stability comparable with simulations based on crystal structures. The simulation was extended from 4 to 20-ns or several times the mean passage time of an ion. Essential dynamics analyses were used to test the hypothesis that correlated motions of the helical bundles account for high-frequency noise observed in open channel measurements. In a 20 ns simulation of a hexameric alamethicin bundle, the main motions are those of individual helices, not of the bundle as a whole. A detailed comparison of simulations using different methods to treat long-range electrostatic interactions (a twin range cutoff, particle mesh Ewald, and a twin range cutoff combined with a reaction field correction) shows that water orientation inside the alamethicin channels is sensitive to the algorithms used. In all cases, water ordering due to the protein structure is strong, although the exact profile changes somewhat. Adding an extra 4-nm layer of water only changes the water ordering slightly in the case of particle mesh Ewald, suggesting that periodicity artifacts for this system are not serious. Subsequent studies of the same group of scientists have performed MD simulations on alamethicin K18, which is a covalently linked alamethicin dimer in which the glutamine residue at position 18 in each helix has been replaced by a lysine residue. As described in their previous work, channels formed by this peptide show pH-dependent selectivity. The maximum anion selectivity of the putative octameric conducting state is obtained at pH 7 or lower. Inasmuch as no change in selectivity is seen between pH 7 and pH 3, and because protons are expected to be in equilibrium with the open state of the channel during a selectivity measurement, the channel is believed to be fully charged (i.e., all eight lysines protonated) at pH 7. In an effort to understand how such a highly charged channel structure is stable in membranes and why it is not more selective for anions, Sansom and coworkers [73] have performed a number of computer simulations of the system. Molecular dynamics simulations of 10 ns each of the octameric bundle in a lipid bilayer environment were presented, with either zero, four, or eight lysines charged in the absence of salt, and with eight lysines charged in the presence of 0.5 M and 1 M KCl. When no salt is present and all lysines are charged, on average 1.9 Cl^- ions are inside the channel and the channel significantly deforms. With 0.5 M KCl present, 2.9 Cl^- ions are inside the channel. With 1 M KCl present, four Cl^- ions are present and the channel maintains a regular structure. Poisson-Boltzmann calculations on models of the octameric channel also predict an average of 2-4 Cl^- ions near the lysine residues as a function of ionic strength. These counterions lower the apparent charge of the channel, which may underlie the decrease in selectivity observed experimentally with increasing salt concentrations. They suggested that to increase the selectivity of Alm K18 channels, positive charges should be engineered in a narrower part of the channel.

3.2.2. AChR

The nicotinic acetylcholine receptor (nAChR) is one of most intensively studied ligand gated ion channels, and thus provides a paradigm for the molecular mechanism of fast synaptic neurotransmission. Depolarization of the presynaptic membrane causes release of acetylcholine into the synaptic cleft. Acetylcholine diffuses to the postsynaptic membrane where it binds to the extracellular domains of the nAChR. Details of the structure of the nAChR have been unveiled recently [74]. The nAChR is pentameric, the five subunits arranged around a central pore (Figs. 5b, 5c). The transmembrane pore is lined by the second transmembrane helix (M2) of each subunit. Binding of acetylcholine to the extracellular domains is thought to initiate a wave of conformational change that is propagated through the protein into the transmembrane domain and the inner (M2) helices. This results in a "twist to open" [74] mechanism of channel gating whereby the central pore is opened to the permeation of cations. A leucine side chain is highly conserved throughout the nAChRs and other members of the superfamily and plays a central role in the gating mechanism of these receptors. Unwin and coworkers [74] have proposed the following gating mechanism (Fig. 5a). Binding of ACh opens the channel by initiating rotational movements (arrows) of the inner β-sheets of the α subunits in the ligand-binding domain. The rotations destabilize the gate, causing the helices to adopt an alternative configuration which is permeable to the ions. These movements are communicated to the inner (M2) helices lining the pore and break apart the gate—a hydrophobic girdle in the middle of the membrane—so that ions can flow through.

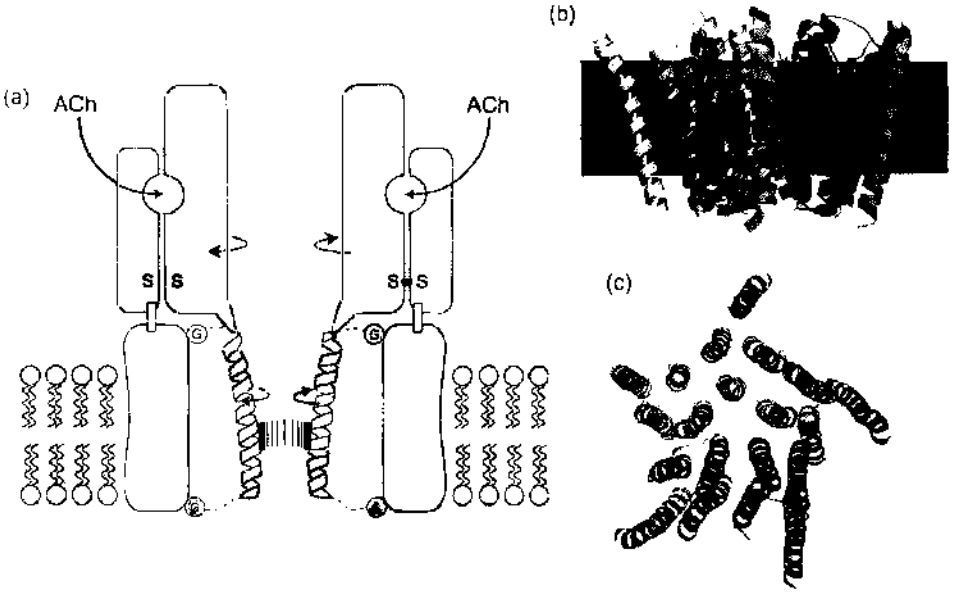


Figure 5. (a) Proposed model for the gating mechanism. The ACh-induced rotations in the α-subunits are transmitted to the gate—a hydrophobic barrier to ion permeation—through the M2 helices. The rotations destabilize the gate, causing the helices to adopt an alternative configuration, which is permeable to the ions. The helices move freely during gating because they are mainly separated from the outer protein wall and connected to it by flexible loops, containing glycine residues (G). S-S is the disulphide-bridge pivot in the ligand-binding domain, which is anchored to the fixed outer shell of the pore. The relevant moving parts are shaded. The height of the membrane-spanning pore is about 40 Å, the β-sheet structure about 60 Å. Reprinted with permission from [74], A. Miyazawa et al., *Nature (London)* 423, 949 (2003), Fig. 6. © 2003, Macmillan Publishers Ltd. (b) Side view and (c) top view of the pore, as seen from the synaptic cleft, with subunits shown in different colors (PDB code 1OED).

A tryptophan side chain in the ligand-binding domain identifies the ACh-binding region; a valine-side chain links the inner sheet to the inner helix; leucine and valine-side chains on the inner helices make the gate; the locations of the membrane surfaces are indicated by broken lines. The helices move freely during gating because they are mainly separated from the outer protein wall and connected to it by flexible loops, containing glycine residues (G). S-S is the disulphide-bridge pivot in the ligand-binding domain, which is anchored to the fixed outer shell of the pore.

The self-assembly of M2 peptides to form a pentameric bundle of transmembrane helices is a model of the pore-lining region of the nicotinic acetylcholine receptor. This has been studied [75] by molecular dynamics simulations of a model of the M2 bundle in a POPC bilayer in order to explore the conformational dynamics of the channel assembly. On the timescale of the simulation, the bundle remains relatively stable, with the polar pore-lining side chains remaining exposed to the lumen of the channel. Fluctuations at the helix termini, and in the helix curvature, result in closing/opening transitions at both mouths of the channel, on a timescale of 10 ns. On average, water within the pore lumen diffuses 4x more slowly than water outside the channel. Examination of pore water trajectories reveals both single-file and path-crossing regimes to occur at different times within the simulation.

The analysis of short MD simulations by the McCammon group [76] shows a pattern of motions which indicate how ligand binding may correlate with larger-scale subunit motions that would connect with the transmembrane region that controls the passage of ions.

3.2.3. Chloride Channels

The ClC family of chloride channels is present in virtually all tissues and organisms, and is widely expressed in most mammalian cells [77]. It is now recognized that chloride channels regulate a variety of important physiological and cellular functions. Chloride-conducting proteins are vital for regulating pH, cell volume, electrical impulses, transport of salts across cells, and voltage stabilization of excitable muscle cells [77]. Mutations in ClC channels cause diseases. The original structures [78, 79] permit theoretical study of chloride permeation,

and determination of the bacterial systems' curvilinear ionic pathways. Even though the bacterial systems are transporters, not channels, the prokaryotes share signature sequence identities with their eukaryotic relatives. Consequently such atomic level analyses might help in understanding ion transport in eukaryotes and ClC channel function generally. Despite a wealth of experimental data, major issues need to be resolved at the atomic level. What is the exact chloride pathway? Which amino acids coordinate the translocating ion and what are their roles? What is the effect of charged amino acids located within the pore mouths on the anion translocation? Identification of the role of pore-lining amino acids is important for understanding both conduction and gating, as mutations of charged residues believed to be chloride-coordinating affect both these properties simultaneously.

Miloshevsky and Jordan [80] have determined the lowest-energy curvilinear pathway, identified anion-coordinating amino acids, and calculated the electrostatic potential energy profiles. They have studied the effect of mutating the charge of some strongly conserved pore-lining amino acids on the electrostatic potential energy profiles. When a certain residue (E148) is neutralized, it creates an electrostatic trap, binding the ion near midmembrane. This suggests a possible electrostatic mechanism for controlling anion flow: neutralize E148, displace the side chain of E148 from the pore pathway to relieve the steric barrier, then trap the anion at midmembrane, and finally either deprotonate E148 and block the pore (pore closure) or bring a second Cl ion into the pore to promote anion flow (pore conductance). The authors have developed a new computational approach, the Monte Carlo ion channel proteins (MCICP) method, to simulating water–protein, ion–protein, and protein–protein interactions, for application to the study of permeation and gating in ion and water channels. It exploits the Metropolis Monte Carlo method and kinetic Monte Carlo techniques. The implementation of the computational model for ClC Cl^- crystal structure is based on various assumptions. Simulations are based on the four crystallographic X-ray structures of bacterial ClC Cl^- assemblies: StClC (Protein Data Bank (PDB) entry 1KPL), EcClC (PDB entry 1OTS), and two EcClC mutants (PDB entries 1OTT).

Cohen and Schulten [81] have investigate Cl^- conduction by means of an all-atom molecular dynamics simulation of the ClC channel in a membrane environment. On the basis of their simulation results, they proposed a king-of-the-hill mechanism for permeation, in which a lone ion bound to the center of the ClC pore is pushed out by a second ion that enters the pore and takes its place. Although the energy required to extract the single central ion from the pore is enormous, by resorting to this two-ion process, the largest free-energy barrier for conduction is reduced. At the narrowest part of the pore, residues Tyr-445 and Ser-107 stabilize the central ion. There, the bound ion blocks the pore, disrupting the formation of a continuous water file that could leak protons, possibly preventing the passage of uncharged solutes.

3.2.4. Gramicidin

Structure. Gramicidin A (gA), an antibiotic from *Bacillus brevis*, has remained for almost 20 years the only channel with a known structure, therefore presenting the favorite object for theoretical modeling. From physiological, as well as structural and functional point of view, it is not considered a biological channel. It is a 15-residue peptide (Fig. 6), which forms a head-to-head dimer in the membrane, assuming a form of a 25 Å long cylinder with a central pore of 2 Å radius. The structure is permeant to monovalent cations (H^+, Li^+, Na^+, K^+, Rb^+, Cs^+), binds divalent cations and rejects all anions. The featureless pore of gA is lined only by peptide backbone atoms, with symmetrical binding sites of moderate affinity near the entry and the exit. Alkali metal cation transport is considered, in simplest terms, to be carried out in three steps: two primarily extrachannel steps (diffusion through bulk aqueous solution up to the channel combined with entry into the binding site and exit from the exit site followed by aqueous diffusion away from the channel) and one intrachannel step (usually referred to as a cation translocation). The rate of the intrachannel step strongly depends on membrane potential, causing current–voltage relationships (I–Vs) to be superlinear when the intrachannel step is rate limiting. Thus I–Vs are sublinear at low permeant ion concentrations, where the extrachannel entry process is rate limiting, and become more superlinear with higher concentrations of permeating ion [82]. Detailed studies on different

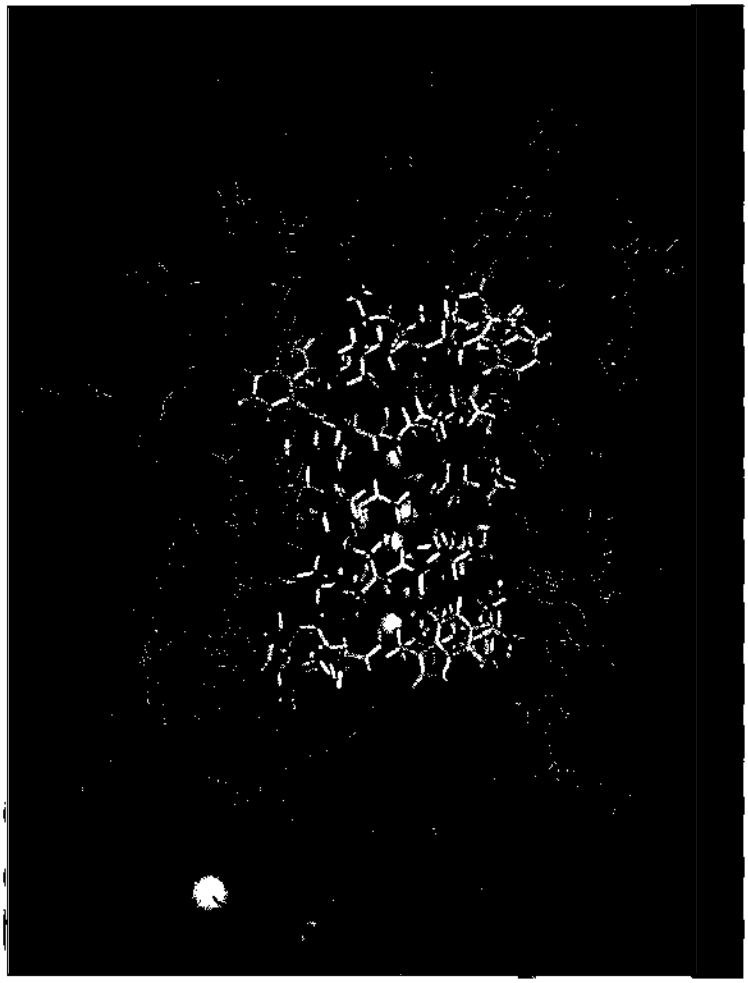


Figure 6. Structure of gramicidin A dimer. The cylindrical pore contains a file of water molecules (H and O atoms in white and red) surrounded by peptide bonds (yellow) (PDB code 1MAG). The lipids are depicted as gray bonds. the K^+ and Cl^- ions as green and gray spheres.

gramicidin structures have been performed using MD simulations recently [83] indicating the great utility of MD studies in the analysis and interpretation of solid-state NMR data. Earlier simulations of a gramicidin channel inserted into a fluid phase DMPC bilayer with 100 lipid molecules with constant surface tension boundary was reported by Jakobsson and co-workers [84, 85]. Among other results, it was found that hydrocarbon chains of lipids adjacent to the channel had higher-order parameters than those farther away. The thickness of the lipid bilayer immediately adjacent to the channel was greater than it was farther away. The thickness of the hydrocarbon region adjacent to the gramicidin was much thicker than what other studies have identified as the "hydrophobic length" of the gramicidin channel.

Permeation. The permeation properties of gA channel seem to be also very dependent on the nature of the lipid environment (which is less likely to be the case for much larger, biological channels). Some interesting results appeared on effects of membrane electrostatics on gA conductance properties. The data provided through experiments [86, 87] were, to a large extent, confirmed by continuum theory studies (3D Poisson-Nernst-Planck calculations) [88]. The calculated large changes in conductance of the channel (e.g., ionic current through gA is six fold larger in charged phosphatidylserine membrane than in a phosphatidylcholine membrane) caused by changed properties of the membrane (including lipid charges and dipoles, as well as membrane thickness) were consistent with experimental results.

By use of the formal correspondence between diffusion and random walk, the permeability for water through the gramicidin channel was calculated [85], and was found to agree closely with the experimental results for permeation of water through gramicidin in a phospholipid membrane. By using fluctuation analysis, components of resistance to permeation were

computed for movement through the channel interior, for the transition step at the channel mouth where the water molecule solvation environment changes, and for the process of diffusion up to the channel mouth. The majority of the resistance to permeation appears to occur in the transition step at the channel mouth. A significant amount is also due to structurally based free energy barriers within the channel. Only small amounts are due to local friction within the channel or to diffusive resistance for approaching the channel mouth. Among other results obtained from simulations of gA, also some important aspects of H^+ transport were revealed. Generally, proton fluxes across membranes constitute one of the most fundamental aspects of cell physiology, since the passive flow of H^+ through H^+ conducting channels drives ATP-synthesis. The nonequilibrium electrochemical potential for protons is achieved by H^+ pumps (e.g., in bacteria, via bacteriorhodopsin, later discussed in the text), which use different forms of energy (from photochemical or redox reactions) to perform the uphill transport. Proton transport is postulated to occur via mechanism similar to one for salt-based conductivity suggested in 1805 by C. J. von Grotthuss [89] and therefore termed *Grotthuss conductance*. Within aqueous channels, proton transport occurs as covalent bonds exchange with hydrogen bonds between hydronium and a neighboring water to produce a charge transport, which is followed by a slower reorientation of the electroneutral water molecule. Ion-free gramicidin channels contain eight or nine water molecules in a single file [82], which are free to interact and rotate. The water molecules thus form an ideal "proton wire," with fully aligned dipole moments as the lowest energy conformation. Channel permeability to protons is 40- to 60-fold higher than permeability to Na^+, and single-channel proton conductance at infinite dilution is 9.1-fold higher than Na^+ conductance. Water flux is not induced by proton-mediated currents, as it is with alkali metal currents thus suggesting that H^+ transport through the gA channel occurs by means of Grotthuss conductance [82]. A theoretical analysis of proton transport in a nine-water wire in vacuo emphasizes the difference between the time scales for proton passage (<ps) and the subsequent water reorientation (>ns) [90]. Positive charge transport first requires that a (hydrated) proton approach a channel containing a water column aligned with water oxygens toward the channel entry. After rapid exchange of hydrogen bonds and covalent bonds, which results in the release of a hydrogen at the channel exit, the water column becomes aligned in the opposite direction and must completely reorient before the next proton transport can occur. Reorientation is expected to be rate limiting and occurs by propagation of a hydrogen-bonding defect through the channel. Furthermore, it was found [90] that the mobility of H^+ in the gramicidin channel is essentially determined by the fine structure and the dynamic fluctuations of the hydrogen-bonded network. The process of H^+ permeation is mediated by thermal fluctuations in the relative positions of oxygen atoms in the wire. When permeating proton is not present, the water chain adopts one of the two polarized configurations, each corresponding to an oriented donor-acceptor H-bond pattern along the channel [91]. As the water molecules are ubiquitous in biological systems and form modulable H-bonded networks, it is expected that similar features in the coordination of these networks apply also to all the membrane proteins that provide efficient pathways for H^+ transport.

Roux and co-workers [92] have developed a model for proton conduction through gramicidin based on molecular dynamics simulations. The transport of a single proton through the gramicidin pore is described by a potential of mean force and diffusion coefficient obtained from molecular dynamics. In addition, the model incorporates the dynamics of a defect in the hydrogen bonding structure of pore waters without an excess proton. Proton entrance and exit were not simulated by molecular dynamics. The single proton conduction model includes a simple representation of these processes that involves three free parameters.

Multinanosecond molecular dynamics simulations of gramicidin A embedded in a dimyristoylphosphatidylcholine bilayer have been performed by Grubmüller et al. [93]. Water permeability was found to be much higher in the double helical conformation, which was explained by lower hydrogen bond-mediated enthalpic barriers at the channel entrance and its larger pore size. Free-energy perturbation calculations indicated that the double helical structure is stabilized by the positive charges at the N termini, whereas the helical dimer is destabilized. Together with the experimental observation that desformyl gramicidin conducts

water hundred-fold better than gramicidin. this suggests that desformyl gramicidin A predominantly occurs in the double helical conformation.

In a recent study Chung and co-workers [94] used the well-known structural and functional properties of the gramicidin A channel to test the appropriateness of force fields commonly usec in molecular dynamics (MD) simulations of ion channels. For this purpose, the high-resoution structure of the gramicidin A dimer was embedded in a POPC bilayer, and the potential of mean force of a K^+ ion was calculated along the channel axis using the umbrella sampling method. Calculations were performed using two of the most common force fields in MD simulations: CHARMM and GROMACS. Both force fields lead to large central barrers for ion permeation. that are substantially higher than those deduced from the physiological data by inverse methods. In long MD simulations lasting more then 60 ns, several ions were observed to enter the binding site but none of them crossed the channel despite the presence of a large driving field. The present results, taken together with many earlier studes, highlighted the shortcomings of the standard force fields used in MD simulations of ion channels and calls for construction of more appropriate force fields for this purpose.

Gating. The dissociation of gA channels into monomers is the simplest example of a chamel gating process. Gating presumably occurs via dissociation and association of the monomers with the closing transition triggered by breaking the dimer's stabilizing hydrogen bonds. The deletion of a single hydrogen bond at the junction between the monomers destabilizes the gA dimer and reduces its conductance. The initial steps in this process have been studied [95] via a computational model that simulates the reaction coordinate for dimer-monomer dissociation. The nonbonded interaction energy between the monomers is deermined, allowing for their free relative translational and rotational motion. Lowest energy pathways and reaction coordinates of the gating process are determined. Partial rupture of the six hydrogen bonds at the dimer junction took place by coupling monomer rotaton and lateral displacement. Coupling rotation with axial separation was far more expensive energetically. The transition state for channel dissociation occurs when monomers are displaced laterally by 4–6 Å, separated by 1.6–2 Å, and rotated by 120 degrees, breaking two hydrogen bonds. In membranes with significant hydrophobic mismatch there is a much greater likelihood of forming four hydrogen bonds (HB) and possibly even two hydrogen bond states. In the 4HB state the pore remains fully open and conductive. However, transtions from the 6HB to 4HB and 4HB to 2HB states take place via intermediates in whict the gA pore is closed and nonconductive. These lateral monomer displacements give rise to transitory pore occlusion at the dimer junction, which provides a rationale for fast closure events (flickers). Local dynamics of gA monomers also leads to lateral and rotational diffusion of the whole gA dimer, giving rise to diffusional rotation of the dimer about the chamel axis.

3.2.5. Glutamate Channels GluR

Fast synaptic transmission between nerve cells in mammals is carried out predominantly by ionotropic glutamate receptors (iGluR). These receptors are a family of ligand-gated ion chamels that open in response to the binding of glutamate. Glutamate is released presynaptically and binds to a post-synaptic receptor gating a cation-selective channel, thus depolarizing the post-synaptic cell. Although glutamate is the natural ligand, the various iGluRs identfied by sequence comparisons may also be classified in terms of their agonist pharmacology. For example. those receptors that show greatest sensitivity to the agonist AMPA are termed *AMPA receptors* (GluR2), those that show greatest sensitivity to kainate are referred to as kainate receptors (GluR5-7). In all of these GluRs the agonist/antagonist binding site is located within the extracellular region of the protein. Recently, Sansom and co-workers [96] lave used multiple molecular dynamics simulations of 2–5 ns duration to explore the structural dynamics of GluR2 in the presence and absence of glutamate and in a complex with kainate. Their studies indicate that not only is the degree of domain closure dependent upon interactions with the ligand, but also that protein/ligand interactions influence the notion of the S2 domain with respect to S1. Differences in domain mobility between the three states (apo-S1S2. glutamate-bound, and kainate-bound) are surprisingly clear-cut.

They discuss how these changes in dynamics may provide an explanation relating the mechanism of transmission of the agonist-binding event to channel opening. They also show how the glutamate may adopt an alternative mode of binding not seen in the X-ray structure, which involves a key threonine (T480) side chain flipping into a new conformation. This new conformation results in an altered pattern of hydrogen bonding at the agonist-binding site.

Excitatory synaptic transmission is mediated by ionotropic glutamate receptors (iGluRs) through the induced transient opening of transmembrane ion channels. The three-dimensional structure of the extracellular ligand-binding core of iGluRs shares the overall features of bacterial periplasmic binding proteins (PBPs). In both families of proteins, the ligand-binding site is arranged in two domains separated by a cleft and connected by two peptide stretches. PBPs undergo a typical hinge motion of the two domains associated with ligand binding that leads to a conformational change from an open to a closed form. The common architecture suggests a similar closing mechanism in the ligand-binding core of iGluRs induced by the binding of specific agonists. Starting from the experimentally determined kainate-bound closed form of the S1S2 GluR2 construct, Mendieta et al. [97] have studied by means of molecular dynamics simulations the opening motion of the ligand-binding core in the presence and in the absence of both glutamate and kainate. Their results suggest that the opening/closing interdomain hinge motions are coupled to conformational changes in the insertion region of the transmembrane segments. These changes are triggered by the interaction of the agonists with the essential Glu 209 residue. A plausible mechanism for the coupling of agonist binding to channel gating has been discussed.

3.2.6. Calcium Channels

Our understanding of ion permeation and selectivity in calcium channels is still poor. In the absence of a molecular sieve mechanism, which selects between ions on the basis of ionic radii, reconciling their high selectivity and high conductance has been a difficult problem. Calcium channels are very selective against Na^+ ions and exhibit a multi-ion Coulomb repulsion mechanism to achieve a high conduction of Ca^{2+} ions. The fact that the two ions have similar raddii but different charges indicates that selectivity must be based on charge. Since the detailed structure of a calcium channel is not available, Chung and co-worker [98] have studied a coarse-grained model using BD simulations. They have used the available information on the structure and conductance properties to construct a model channel consisting of inner and outer vestibules and a selectivity filter. The filter was designed such that two critical elements, its size and charges on its wall, were determined from experimental data. The radius was set to 2.8 Å according to the size of tetramethylammonium, the largest permeable ion. Four negatively glutamate residues were assigned to the filter region. The BD simulations have quite successfully reproduced experimental current-voltage curves, saturation of conductance with concentration, selectivity against Na^+, the anomalous mole fraction effect, attenuation of the calcium current by external sodium ions, and the effect of mutating glutamate residues on blocking of sodium current.

3.2.7. Potassium Channel KcsA

A major breakthrough in our understanding of ion channels at the atomic level had taken place after the recent successful structural determination of the bacterial potassium channel KcsA from *Streptomyces lividans* by MacKinnon and co-workers [28–31]. The structure is depicted in Fig. 7.

Structure. First MD simulations [34, 36] have shown that the simulated atomic structure remained stable and very close to the original structure. The rms deviation of the C_α atoms between the crystal structure and the simulated structure was about 3.7 Å, which is comparable to the crystal structure of 3.2 Å resolution [28] and 2.0 Å resolution [30]. Deviations between crystal structure and simulated structure had been found at the extracellular mouth where a significant widening had been observed. It was speculated that the absence of polarization force fields in the simulation protocol caused this effect. All simulations indicate that the KcsA structure remains closed at the intracellular mouth. It should be noted that the fluctuation profile of the radius of gyration of each residue along the channel shows

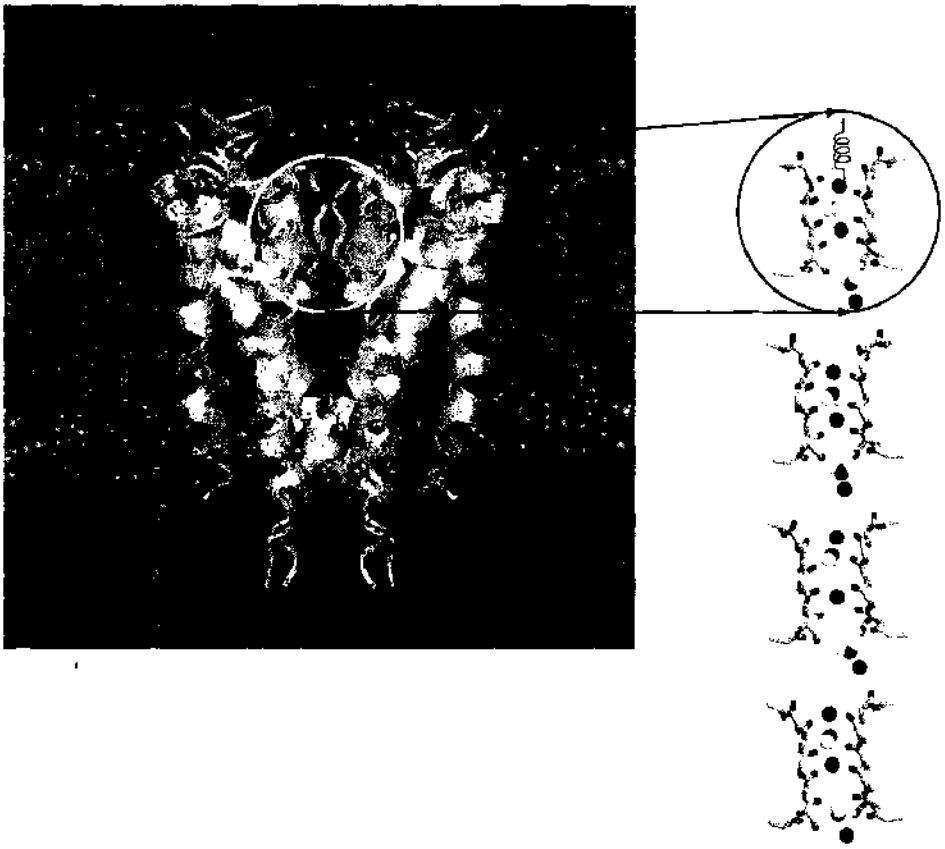


Figure 7. Structure of the KcsA potassium channel embedded in a fully hydrated lipid bilayer membrane. The narrow selectivity filter is the encircled region.

that the residues in the hydrophobic core of the bilayer are relatively more stable than those located close to the head group region and the aqueous phase. This may explain the unstable position of the ion near the exit of the selectivity filter [99]. Since the potassium ions, which are localized in the selectivity filter, represent an integral part of the crystal structure, it was of interest to compare their possible locations as predicted by X-ray analysis [28, 30] to simulation findings. Based on the most recently published structure [30], it was conjectured that the K^+ ions may occupy seven different locations: four inside the selectivity filter, one in the cavity and two ions weakly bound to the extracellular mouth. MD simulations, however, [34, 36, 99] indicate that only one or at most two ions may occupy the filter simultaneously. Some MD results report on only one ion in the filter, where the second ion exits the filter either to the intracellular side [34] or to the extracellular side [99]. However, in the latter both cases, it cannot be excluded that the findings have to be attributed to incorrectly equilibrated structures. In particular, the results of the simulation showed that the potassiun ion K_1, which originally binds to the oxygen atom of the Tyr^{78} carbonyl group, finally dissociates into the aqueous phase. The K_2, which binds to the oxygen atom of the Thr^{75} carbonyl group and K_3, which is located in the cavity, both remain bounded within the channel for the whole 3 ns of simulation. At equilibrium the K_2 ion remains located between Gly^{77} and $Val.^{78}$ Indeed, more elaborate simulations are necessary to clarify the details on ion localization in the selectivity filter. The MD simulations predict an average number of water molecules within the cavity of about 21, which is enough to solvate the cavity ion K_3, but which is in contrast to the conjecture of 50 water molecules based on solving the finite difference Poisson Equation [100]. However, the latter result has to be viewed in the light of the recently reported shortcomings of the Poisson-Boltzmann theory applied to inhomogeneous systems [101].

Permeation. The mechanisms underlying transport of ions across the potassium channel have been examined [102] using electrostatic calculations and three-dimensional Brownian dynamics simulations. To build the open-state configurations of the channel with molecular dynamics simulations, the transmembrane helices were pulled outward until the channel

attains the desired interior radius. To gain insights into ion permeation, the potential energy profiles experienced by an ion traversing the channel in the presence of other resident ions were constructed. These profiles reveal that in the absence of an applied field the channel accommodates three potassium ions in a stable equilibrium, two in the selectivity filter and one in the central cavity. In the presence of a driving potential, this three-ion state becomes unstable, and ion permeation across the channel was observed. These qualitative explanations are confirmed by the results of three-dimensional Brownian dynamics simulations. It was found that the channel conducts when the ionizable residues near the extracellular entrance are fully charged and those near the intracellular side are partially charged. The conductance increased steeply as the radius of the intracellular mouth of the channel was increased from 2 to 5 Å. The simulation results reproduced several experimental observations, including the current–voltage curves, conductance-concentration relationships, and outward rectification of currents.

Our current understanding of ion transport through biological potassium ion channels is based on the concept of the multi-ion transport mechanism: permeating ions line in a queue in the narrow channel pore and move in a single file through the filter. The multi-ion concept had been accepted for many decades, its molecular mechanism, however, remained still elusive. More insights have been obtained recently using molecular dynamics simulations [99]. Steered molecular dynamics simulation permit to monitor the collective motion of ions and water molecules through the narrow selectivity filter, which is the encircled region in Fig. 7. The simulations reveal that the high conductivity is based on the cooperative diffusion of ions and water molecules mediated by the charged flexible carbonyl groups lining the selectivity filter which are shown in Fig. 7 at the right hand side by the red tips at the protein's backbone. The sequence of snapshots represents at subsequent times four successive configurations of ions and water molecules passing through the selectivity filter. A detailed analysis has shown the following microscopic mechanism of ion permeation. At rest, the queue of ions and intercalated water molecules (e.g., K1-W1-K2-W2) are residing at the minima of the periodic pore potential made by the pore lining carbonyl groups, C–O. The pore exiting ion K1 reorients and attracts the neighboring water molecule W1, which thereby induces a local transformation of the neighboring pore potential where the ion K2 is located. This neighboring pore potential is an asymmetric double well potential, bistable, and shelters the K2 in the lower one of its two minima. The W1-induced transformation switches the lower minimum from one to the other well, toward the position of W1. Hence, K2 moves toward this new minimum, thereby following the movements of K1 and W1.

The very remarkable property of this narrow filter is its high conductivity: once the outer ion is pulled out of the filter, all other ions and water molecules follow the same direction, apparently at almost no cost of energy. This becomes more obvious during the transport through very long artificial nanotubes constructed in analogy to the selectivity filter of this ion channel, which is shown in Fig. 8. From top to the bottom, this figure shows at subsequent times, in steps of 50 ps, six successive configurations of ions and water molecules passing through the nanotube. The transport occurs by the hopping of vacancies from the right to the left, as indicated by the arrows. It is fascinating to speculate about a possible construction of such a kind of very efficient conducting device on the nanometer scale under experimental conditions in the laboratory. In fact, the computer simulations of the ion conduction in such a nanotube reveal the same effects as observed for the ion channel. Therefore, since the molecular mechanism of the ion-water transport in ion channels and corresponding nanotubes is based on general physical principles, the efficiency of this type of linear organic ionic conductor has to be compared to known solid state superionic conductors. It would not be a surprise if such biomolecular conductors would be superior to the anorganic one. By following the visions of Feynman and Drexler, it seems conceivable that future inventions of nanotechnological devices may be constructed based on principles copied from biomolecular machines. The discovery of the molecular transport mechanism in ion channels may currently serve only as a paradigm pointing toward a future challenge: the understanding of major biomolecular machines such as rotary ATPases providing the necessary torque for the bacterial flagellar motor, or, translocases which are transmembrane conducting devices for protein transport. In any case, the knowledge of such biomolecular

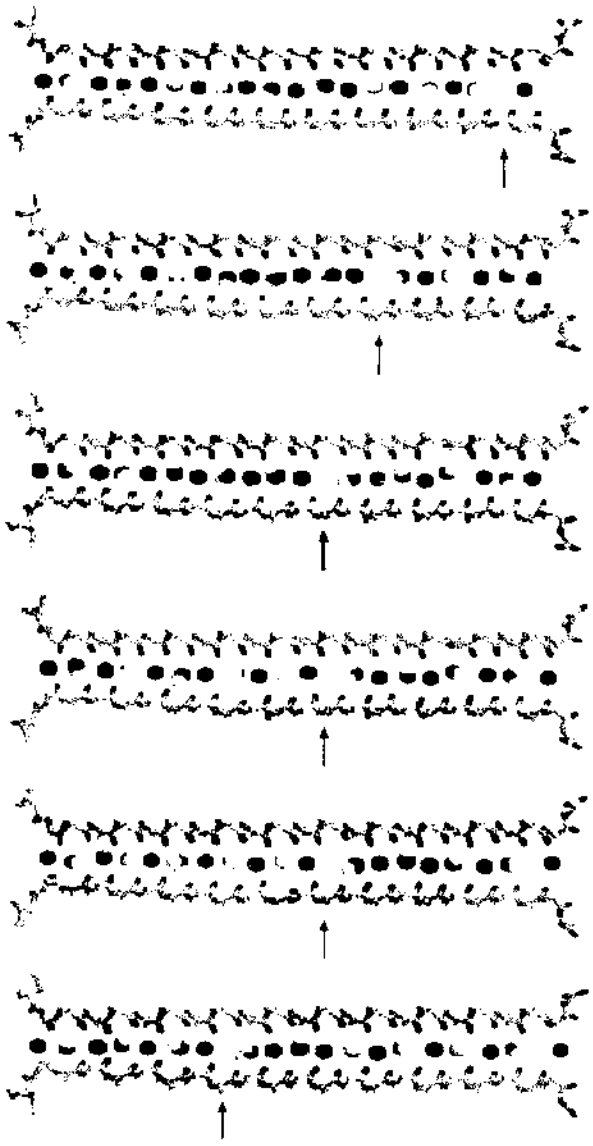


Figure 8. Sequence of snapshots of water molecules and K^+ ions moving in a defectlike motion through a nanotube constructed as a periodically continuated KcsA selectivity filter.

mechanisms, which can be considered to serve as a bridge between nanotechnology and biology, cannot be achieved without the indispensible tool of massive computer simulations.

Selectivity. Another focus of MD simulations of the KcsA potassium channel has been the selectivity filter and the permeation process from the cavity to the filter. The question of selectivity against Na^+ ions has been addressed in several studies through free-energy perturbation calculations, where a K^+ ion in one of the binding sites is alchemically transformed into a Na^+ ion, the calculated free-energy barrier range from 11 kT to 8 kT [103], and 5 kT [104], which are in rough agreement with the experimental value of 9 kT extracted from the K^+/Na^+ selectivity ratio of about 10^4. It has been suggested [103] that hydrophobic residues lining the intrapore and cavity are responsible for the relative high diffusion of ions in those segments. Despite the large suppression of the diffusion coefficient in the filter region, permeation through this segment has been found by BD simulations to be the fastest step in a full conduction cycle thanks to the Coulomb repulsion. Recent MD simulation of Sansom and co-workers [105] indicate that K^+ ions and water molecules within the filter undergo concerted single-file motion in which they translocate between adjacent sites within the filter on a nanosecond timescale. In contrast, Na^+ ions remain bound to sites within the filter and do not exhibit translocation. Furthermore, entry of a K^+ ion into the filter from the extracellular mouth is observed, whereas this does not occur for a Na^+ ion. It is

argued [105] that these differences in interactions in the selectivity filter may contribute to the selectivity of KcsA for K^+ ions in addition to the differences in the dehydration energy between K^+ and Na^+ and the block of KcsA by internal Na^+ ions.

The selective interactions of various cations, Na^+, K^+, Rb^+, and Cs^+, within the selectivity filter of the KcsA potassium channel have been investigated [106] via multiple molecular dynamics simulations (total simulation time, 48 ns) based on the high-resolution structure of KcsA, embedded in a phospholipid bilayer. As in simulations based on a lower-resolution structure of KcsA, concerted motions of ions and water within the filter are seen. Despite the use of a higher resolution structure and the inclusion of four buried water molecules thought to stabilize the filter, this region exhibits a significant degree of flexibility. In particular, pronounced distortion of filter occurs if no ions are present within it. The two most readily permeant ions, K^+ and Rb^+, are similar in their interactions with the selectivity filter. In contrast, Na^+ ions tend to distort the filter by binding to a ring of four carbonyl oxygens. The larger Cs^+ ions result in a small degree of expansion of the filter relative to the X-ray structure. Cs^+ ions also appear to interact differently with the gate region of the channel, showing some tendency to bind within a predominantly hydrophobic pocket. The four water molecules buried between the back of the selectivity filter and the reminder of the protein show comparable mobility to the surrounding protein and do not exchange with water molecules within the filter or the central cavity. A preliminary comparison of the use of particle mesh Ewald versus cutoff protocols for the treatment of long-range electrostatics suggests some difference in the kinetics of ion translocation within the filter.

Gating. The *Streptomyces lividans* potassium channel (KcsA) is pH regulated [107]. A gating mechanism was proposed by Perozo and coworkers [108–110] by using site-directed spin-labeling methods and electron paramagnetic resonance spectroscopy. Results from these experiments indicate that the channel undergoes a "twisted" motion where each of the four TM2 helices tilts away from the permeation pathway, towards the membrane plane, and rotates about its helical axis, supporting a scissoring-type motion with a pivot point near residues [107–108]. These movements result in a large increase of the diameter of the intracellular mouth up to the central water-filled cavity. Although the possible collective motion of the helices can be constructed, mainly based on steric considerations, the origin of the pH-mediated driving force is still unclear. It seems likely that the four long cytoplasmic C-termini of residues 123–160 may play the crucial role, because some of the charged residues may change their protonation state during pH variation and hence provide the necessary variation in their inter-chain Coulomb interaction. Another candidate for pH-mediated conformational transitions are the four N-terminii 1-23 carrying also titrable groups. It is unclear how these chains interact with each other, eventually indirectly via the lipid head groups of the membrane which are in general pH sensitive. Unfortunately, the gating mechanism cannot be studied by standard MD simulation because the time scale of gating is in the order of at least microseconds. The simulation of channel gating is therefore still one of the great challenges in biophysics.

The structure of the bacterial potassium channel, KcsA, corresponds to the channel in a closed state. Two lines of evidence suggest that the channel must widen its intracellular mouth when in an open state: (1) internal block by a series of tetraalkylammonium ions and (2) spin labeling experiments. Thus it is known that the protein moves in this region, but it is unclear by how much and the mechanisms that are involved. To address this issue Sansom and co-workers [111] have applied a novel approach to generate plausible open-state models of KcsA. The approach can be thought of as placing a balloon inside the channel and gradually inflating it. Only the protein sees the balloon, and so water is free to move in and out of the channel. The balloon is a van der Waals sphere whose parameters change by a small amount at each time step, an approach similar to methods used in free energy perturbation calculations. They have shown that positioning of this balloon at various positions along the pore axis generates similar open-state models, thus indicating that there may be a preferred pathway to an open state. They also showed that the resulting structures from this process are conformationally unstable and need to undergo a relaxation process for up to 4 ns. They showed that the channel can relax into a new state that has a larger pore radius at the region

of the intracellular mouth. The resulting models may be useful in exploring models of the channel in the context of ion permeation and blocking agents.

Other studies on blocking agents of K^+ channels have been performed for the docking of the scorpion toxin 1.q2 on binding sites of potassium channels [25] using Brownian dynamics simulations.

3.2.8. Voltage-Gated KvAP

The functional unit of a voltage-gated channel is an assembly of four proteins, or subunits; in each, the polypeptide chain snakes back and forth across the membrane six times [68, 69] (Fig. 9). This "six-transmembrane" structure is seen in the voltage-gated potassium, sodium, and calcium channel families, and also in other channel types.

The resulting X-ray structure [68] of KvAP shows the expected potassium channel (Fig. 9) core, consisting of transmembrane segments S5 to S6, surrounded by S1 through to part of S3. What was unexpected is that the S4 helix, along with the second part of S3, forms an α-helical hairpin—a "paddle" that extends out from the channel core into the membrane's fluid interior. The paddle has a flexible connection to the rest of the channel, as MacKinnon and co-workers [68] showed by comparison with another crystal structure, of segments S1 to S4 alone. This flexibility explains the difficulty that the authors encountered in crystallizing the protein; it also suggests a mechanism for voltage sensing. The paddle is a hydrophobic, charged particle that can move in the membrane interior, transporting its four positive charges from one membrane surface to the other. It is the location of S4—not embedded in the protein core, but loose in the membrane—that is the big surprise here. It explains an old puzzle, that small lipid-soluble molecules somehow have ready access to ion-channel voltage sensors. Such molecules include local anaesthetics, the alkaloid nerve toxins and the well-known insecticides allethrin and DDT. It is now easy to imagine them diffusing up to the voltage-sensor paddle from within the lipid membrane interior. An X-ray crystal structure is like a posed photograph; in the KvAP crystal, for instance, the voltage-sensor paddle is held firmly in place by an antibody scaffold.

Still a few questions have been left to be answered. The actual conformation of the channel in the membrane will need to be clarified, because in the crystal the membrane is replaced by a blanket of detergent molecules. Questions also remain about the disposition of the amino-terminal end of the protein (thought to be intracellular) and of the loop between the S3 and S4 segments in related channels (in the well-studied Shaker potassium channel, this loop is always accessible from the outside surface). Moreover, details of the motions of the voltage sensor—in some channels the charge movement occurs in several discrete steps—remain to be worked out, as does the energetic issue of moving the quadruply charged paddle through

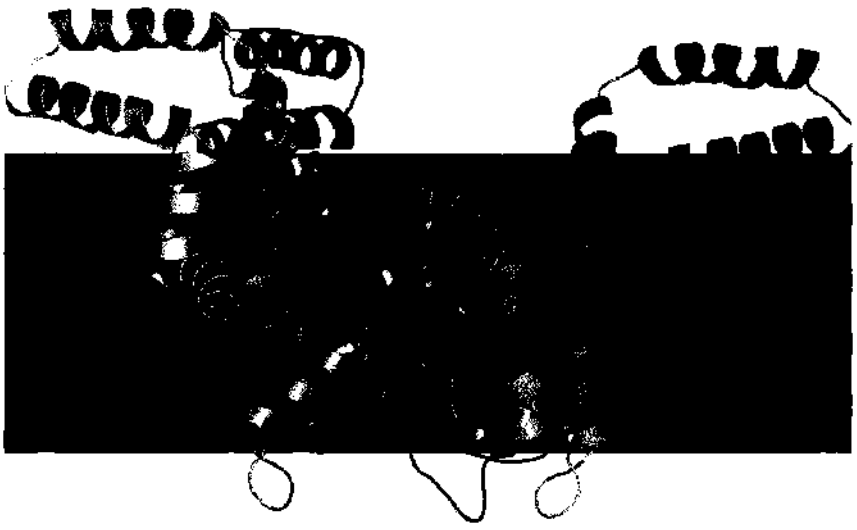


Figure 9. Structure of the voltage-gated ion channel KvAP (PDB code 1ORQ). The front subunit is omitted to unveil the three ions (green balls).

the membrane interior. Still MD simulations have to be performed on this voltage-gated channel in order to elucidate some of the unsolved problems.

3.2.9. Potassium Channel KirBac1.1

Inward rectifier (Kir) channels have two main physiological roles: they regulate cell excitability by stabilizing the membrane potential close to the potassium equilibrium potential, and they are involved in potassium transport across membranes [65]. For example, Kir3.1/Kir3.4 channels modulate cardiac electrical activity, and Kir6.2 is involved in insulin release from pancreatic cells. Two recent structures, of the intracellular domain of mammalian Kir [112] and of a complete bacterial Kir homolog [113], open up the prospect of a detailed understanding of structure/function relationships in this important family of K channels. KirBac is the first bacterial Kir whose structure has been solved at 3.65-Å resolution [113] (Fig. 10). The overall transmembrane topology is similar to that of the simple bacterial channel KcsA, with the addition of a "slide" helix N-terminal to the M1 helix. However, unlike KcsA, the crystal structure of KirBac reveals an intracellular domain consisting mostly of β-sheets. The channel is assumed to be in a closed nonconducting conformation because the intracellular pore mouth is hydrophobic and very narrow (only 0.05 nm in radius, compared with the 0.133-nm radius of a K^+ ion). The selectivity filter of KirBac is quite similar to that of KcsA and is supposed to contain a succession of five ion binding sites formed by cages of eight oxygen atoms. There is considerable interest in the conformational dynamics of the filter of Kir channels. In particular, it has been suggested that, in addition to the intracellular gate, there may also some gating at the selectivity filter. This is not restricted to Kir channels: this phenomenon of inactivation of voltage-gated (Kv) channels is also thought to involve changes in conformation of the selectivity filter. It is therefore of some interest to characterize the conformational dynamics of the filter region of KirBac and to compare with the experimental and computational behavior of other K channels. Molecular dynamics simulations of KirBac have been performed by Sansom and coworkers [114, 115] on a 10-ns timescale, which is comparable to the timescale of ion permeation. The results of five simulations (total simulation time 50 ns) based on three different initial ion configurations and two different model membranes were reported. These simulation data provide evidence for limited (<0.1 nm) filter flexibility during the concerted motion of ions and water molecules within the filter, such local changes in conformation occurring on an 1-ns timescale. In the absence of K^+ ions, the KirBac selectivity filter undergoes more substantial distortions.

Figure 10. Structure of the potassium ion channel KirBac1.1 (PDB code 1P7B).

These resemble those seen in comparable simulations of other channels (e.g., KcsA and KcsA-based homology models) and are likely to lead to functional closure of the channel. The suggested filter distortions may provide a mechanism of K-channel gating in addition to changes in the hydrophobic gate formed at the intracellular crossing point of the M2 helices. The simulation data also provide evidence for interactions of the "slide" helix of KirBac with phospholipid headgroups.

3.2.10. Mechanosensitive Channels

Mechanical forces act on living organisms from all directions throughout the biosphere, making mechanosensory transduction one of the fundamental sensory transduction processes in the biological world. As molecular switches, mechanosensitive (MS) ion channels convert mechanical forces exerted on cellular membranes into electrical or biochemical signals in physiological processes ranging from cellular turgor control in bacteria to touch and hearing in mammals. Mechanosensitive channels are a class of ubiquitous membrane proteins gated by mechanical strain in the cellular membrane. To adapt to osmotic stress, several types of MS channels in the bacterial cell membrane, including the channel of small conductance MscS, can open when osmotic forces stretch the membrane. Without this response, the bacteria lyse. In recent years, the molecular details of how MscS and other bacterial MS channels are gated by mechanical forces have been a subject of intensive research.

Some toxic peptides are selective inhibitor of MS channels. The mechanism of inhibition remains unknown, but it is known [116] that they modify the gating, thus violating a trademark of the traditional lock-and-key model of ligand-protein interactions. Suspecting a bilayer-dependent mechanism, the effect of toxins on gA channel gating have been examined experimentally [116]. The inhibition increases with the degree of hydrophobic mismatch between bilayer thickness and channel length, meaning that the toxic peptide decreases the energy required to deform the boundary lipids adjacent to the channel. These results suggest that modulation of membrane proteins by amphipathic peptides (mechanopharmacology) invoves not only the protein itself but also the surrounding lipids. This has important therapeutic implications.

Structure and Gating of MscL. MscL, a bacterial mechanosensitive channel of large conductance, is the first structurally characterized mechanosensor protein [117–121]. The protein is a pentamer (Fig. 11), approximately 50 Å wide in the plane of the membrane and 85 Å tall. Each 151-residue subunit consists of two transmembrane helices, labeled TM1 and TM2, and a cytoplasmic helix that extends some 35 Å below the membrane. The TM1 helices

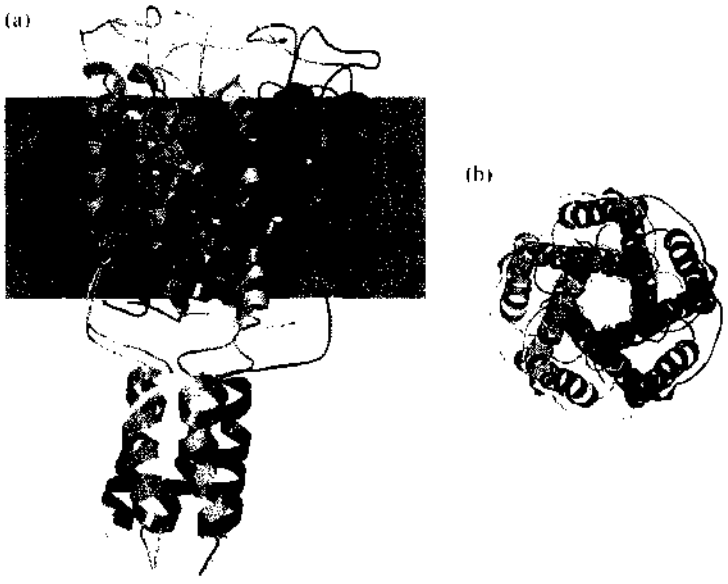


Figure 11. Ribbon representations of the structures of MscL. The top part (a) depicts the view from the plane of the membrane (gray area), whereas the transmembrane region viewed down the membrane normal is illustrated at the bottom (b). Individual subunits are represented in different colors (PDB code 1MSL).

are arranged so as to block diffusion through the channel at their N-terminal ends. Excision of the cytoplasmic domains has been found to have little effect on the gating properties of the channel. In prokaryotes, the channel plays a crucial role in exocytosis and in response to large osmotic pressure changes. In general, it is believed that the gating of mechanosensitive channels is induced by changes in the intrabilayer pressure profiles, which originate from bilayer deformation. To change the membrane tension it has been suggested [119] that different hydrophobic mismatches at the protein-lipid interface induced by different types of lipids may cause an asymmetry of tension across the bilayer membrane and hence lead to a spontaneous curvature that controls the open and the closed state. At least, it is clear from experiments with different types of mixtures of lipids [119, 120] that the protein-lipid interaction must play a fundamental role in defining the physical principles that underly MscL gating. Recently, MD simulations [122, 123] have been performed to contribute to the experimental data on gating mechanism. Standard MD simulation in the nanosecond time regime and the analysis of the structural fluctuations have indicated that the least mobile part of the protein could be identified as the gate, on the same location suggested by expermental findings. This part comprises the first five residues of the TM1 helices, which are shown to be pinched together to form a nonleaky occlusion. In addition, steered MD simulation [122] of the bare protein without membrane and without water have been carried out. The effect of the membrane on the protein has been taken into account by applying a constant surface tension on the protein. Under a range of conditions, it has been shown that the transmembrane helices tilted considerably as the pore opened. The protein refolded into an open conformation, where the transmembrane helices flattened as the pore widened, with a minima loss of secondary structure. The rate at which the protein refolded has been nearly inversely proportional to the applied surface tension. The results indicate that membrane thinning and hydrophobic mismatch within the transmembrane helices my indeed drive gating.

More recently [124] steered molecular dynamics simulations have been used to investigate how forces arising from membrane tension induce gating of the channel. A homology model of the closed form of MscL from was subjected to external forces of 35–70 pN applied to residues near the membrane-water interface. The magnitude and location of these forces corresponded to those determined from the lateral pressure profile computed from a lipid bilayer simulation. A fully expanded state was obtained on the 10-ns timescale that revealed the mechanism for transducing membrane forces into channel opening. The expanded state agrees well with proposed models of MscL gating [119–121], in that it entails an irislike expansion of the pore accompanied by tilting of the transmembrane helices. The channel was most easily opened when force was applied predominantly on the cytoplasmic side of MscL. Comparison of simulations in which gating progressed to varying degrees identified residues that pose steric hindrance to channel opening.

The crystal structure of the *Mycobacterium tuberculosis* homolog of the bacterial mechanosensitive channel of large conductance (Tb-MscL) provides a unique opportunity to consider mechanosensitive signal transduction at the atomic level. Molecular dynamics simulations of the Tb-MscL channel embedded in an explicit lipid bilayer and of its C-terminal helical bundle alone in aqueous solvent have been performed by Elmore and Dougherty [123]. C-terminal calculations imply that although the helix bundle structure is relatively unstable at physiological pH, it may have been stabilized under low pH conditions such as those used in the crystallization of the channel. Specific mutations to the C-terminal region, which cause a similar conservation of the crystal structure conformation, have also been identified. Full-channel simulations were performed for the wild-type channel and two experimentally characterized mutants. The wild-type Tb-MscL trajectory gives insight into regions of relative structural stability and instability in the channel structure. Channel mutations led to observable changes in the trajectories, such as an alteration of intersubunit interactions in one of the mutants. In addition, interesting patterns of protein-lipid interactions, such as hydrogen bonding, arose in the simulations. These and other observations from the simulations are relevant to previous and ongoing experimental studies focusing on characterization of the channel.

Structure and Gating of MscS. MscS, the mechanosensitive channel of small conductance, is found in the inner membrane of *Escherichia coli* and its crystallographic structure in

an open form has been recently solved [125, 126]. Much of what we know about the molecular mechanisms of gating in MscS channel is derived from its crystal structure. Only recently have molecular dynamics and experimental studies shed additional light on the structural changes that occur upon MscS gating. MscS was the first bacterial ion channel shown to respond to membrane stretch. Since its cloning, MscS has emerged as a prototype of a diverse family of MS channels encompassing several representatives from bacteria, archaea, fungi, and plants. The crystal structure of *Escherichia coli* MscS solved at a resolution of 3.9 Å reveals that the channel folds as a homoheptamer and has a large cytoplasmic region. Each subunit contains three TM domains (Fig. 12). Of these, the TM3 helices line the channel pore, whereas the TM1 and TM2 helices constitute the sensors for membrane tension and voltage. The precise conformation of MscS in the crystal form is controversial at present, as a recent study using molecular dynamics simulations [127] implied that water and ions cannot pass through the channel pore, suggesting that the crystal structure may reflect an inactive or desensitized state rather than the open state. Nevertheless, the crystal structure strongly suggests that the channel was captured in an open state because the transmembrane pore reveals an opening of 11 Å, which could account for the 1-ns conductance observed for the MscS activity [128]. Booth and colleagues [126] proposed a new structural model of MscS gating that involves rotation and tilt of pore-lining transmembrane helices. They also show how disruption of a conserved pattern of glycine and alanine residues and their interactions along the pore-lining transmembrane helix TM3 affects the gating of this channel. They proposed a model for the closure of MscS based on mutagenesis experiments affecting structural properties of the pore-lining transmembrane TM3 helix (Fig. 13). TM3 is rich in glycine and alanine residues that the authors propose form a structural motif that plays a crucial role for gating of this channel. The position of the glycine-alanine pattern on the helix faces is conserved in the MscS family of proteins, further supporting the notion that this structural motif is important for proper function in these channels. Gating transitions in MscS thus must involve cooperative action of glycines and alanines along the TM3 helix. The current model of MscS closure indicates that all TM helices decrease in tilt, becoming more vertical in the transition from the open to the closed states. During channel closure TM3 domains must slide against each other and rotate, resulting in a change in the packing of glycines and alanines along the helix. This leads to a much more compact channel with a smaller pore in which certain leucines come into close proximity, thus forming a hydrophobic barrier preventing ion flow. The experimental evidence is consistent with the gating model in which opening of MscS is facilitated by slight irislike rotations and tilt of the pore-lining helices (Fig. 13). A smooth surface formed by four glycines should facilitate the sliding of alanine "knobs," which underlies the conformational transition required for channel gating. Rotation of the transmembrane domains in MscS closely resembles a current model for

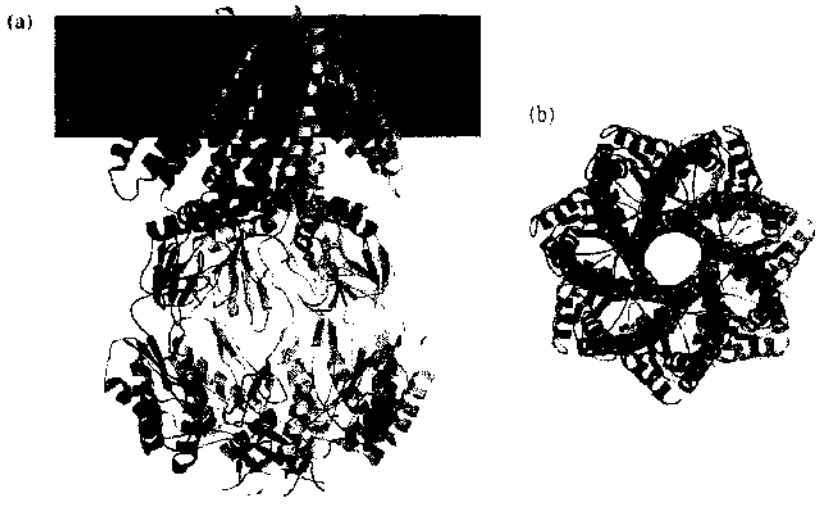


Figure 12. Ribbon representations of the structures of MscS. The top part (a) depicts the view from the plane of the membrane (gray area), whereas the transmembrane region viewed down the membrane normal is illustrated at the bottom (b). Individual subunits are represented in different colors (PDB code 1MXM).

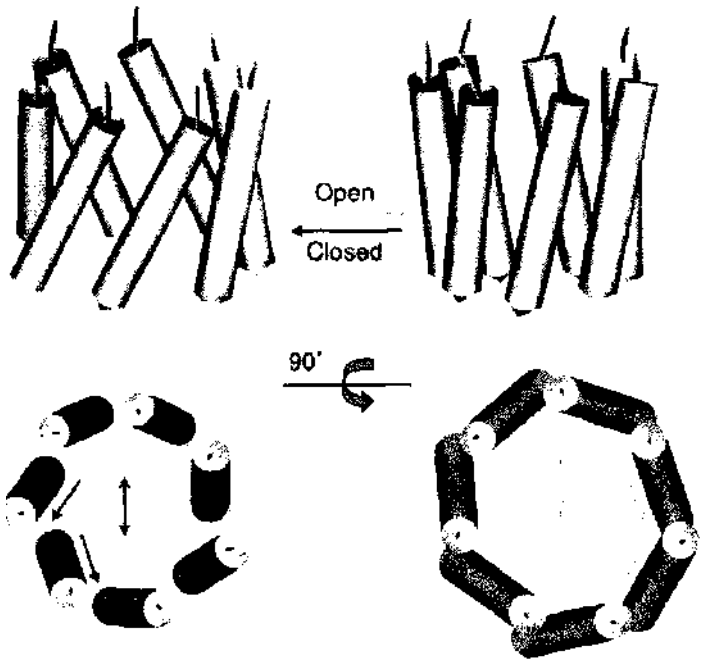


Figure 13. The pore structure of the MscS channel in the open and closed states. The illustration shows how the tilting and rotation of the pore-lining TM3 helices could open and close the channel.

gating of MscL [121], the bacterial MS channel of large conductance. However, consistent with its smaller conductance, the structural changes in MscS are of smaller magnitude compared with that in MscL, whose conductance is three times larger than that of MscS. A search for structurally conserved motifs in other ion channels similar to the conserved pattern of alanines and glycines in MscS channels and examination of the extent of structural plasticity tolerated in other channels may provide insights into how evolution has solved steric hindrance problems to achieve a required protein function. With this model and the structures at hand, we can expect the research on bacterial MS channels to continue to contribute to novel and significant insights into the structure and function of ion channels.

By means of molecular dynamics simulations, Sotomayor and Schulten [127] have studied the stability of the channel conformation of MscS suggested by a previous X-ray crystallographic structure at 3.9 Å resolution [125] in a fully solvated lipid (POPC) bilayer. The combined system encompassing 224,340 atoms. When restraining the backbone of the protein, the channel remained in the open form and the simulation revealed intermittent permeation of water molecules through the channel. Abolishing the restraints under constant pressure conditions led to spontaneous closure of the transmembrane channel, whereas abolishing the restraints when surface tension (20 dyn/cm) was applied led to channel widening. The large balloon-shaped cytoplasmic domain of MscS exhibited spontaneous diffusion of ions through its side openings. Interaction between the transmembrane domain and the cytoplasmic domain of MscS was observed and involved formation of salt bridges between residues Asp62 and Arg128; this interaction may be essential for the gating of MscS. K^+ and Cl^- ions showed distinctively different distributions in and around the channel.

Concluding Remark. There are several computational problems in simulation membrane channels. One problem is related to the common three-dimensional periodic boundary conditions, which seems not to be appropriate for the two-dimensional symmetry of the channel-membrane system. A systematic study on this problem has been made recently [129]. Slab geometric boundary conditions were applied in the molecular dynamics simulation of a simple membrane-channel system. The results of the simulation were compared with those of an analogous system using normal three-dimensional periodic boundary conditions. Analysis of the dynamics and electrostatics of the system showed that slab geometric periodicity eliminates the artificial bulk water orientational polarization that is present while using normal three-dimensional periodicity. Even though the water occupancy and volume of a simple channel is the same when using either method, the electrostatic properties are considerably

different when using slab geometry. In particular, the orientational polarization of water was seen to be different in the interior of the channel. This gave rise to a markedly different electric field within the channel. The implications of slab geometry for the future simulation of this type of system and for the study of channel transport properties have been stressed [129].

4. MEMBRANE TRANSPORTERS

In the following Section, we outline some of the recent advances in modeling and simulating active transport processes across bilayer membranes.

The exchange of substances across cellular membranes may be accomplished by diffusion, facilitated diffusion, or active transport. The result of unaided diffusion is an equilibration of concentrations on both sides of the membrane. In facilitated diffusion, the establishment of equilibrium is sped up by the function of a protein. Active transport is necessary to accumulate a substance against a concentration gradient. Thermodynamics require some kind of energy to perform this, so there has to be another downhill gradient that may be dissipated or some other form of chemical energy. Depending on the source of energy, primary transport is differentiated from secondary transport. Primary transport uses energy directly: light or chemical energy is converted to electrochemical energy as electrochemical potential of the substances to be transported. This category comprises photosynthetic electron transport, light driven ion pumps, redoxenergy dependent respiratory chains, transport ATPases and sodium pumps using decarboxylation energy. In secondary transport, the electrochemical energy originates from the electrochemical potential of another substance that is used up in symport or antiport. Normally, a molecule passes the membrane unchanged. However, in group translocation there is a chemical modification.

The structures of few transport systems are known at atomic resolution. Among these are rhodopsins, some components of the PTS, the proton transporters bacteriorhodopsin and V-ATPase, the ion transporter Ca-ATPase, and the protein transporters as mitochondrial and endoplasmatic pores, and the nuclear pore complex.

4.1. Ion Translocons

4.1.1. Bacteriorhodopsin

Structure and Function. The membrane protein bacteriorhodopsin (bR) serves as a light driven proton pump in bacterial cells. Bacteriorhodopsin is a heptahelical transmembrane protein (Fig. 14) that is found in extremely halophilic archeae like *Halobacterium salinarum*, and in eubacteria. Photoisomerization of the all-trans retinal chromophore, covalently attached to Lys216 through a protonated Schiff base, to the 13-cis, 15-anti configuration initiates ion translocation across the cell membrane, and establishes an electrochemical gradient for ATP synthesis and other energy-requiring membrane processes. The cyclic reaction,

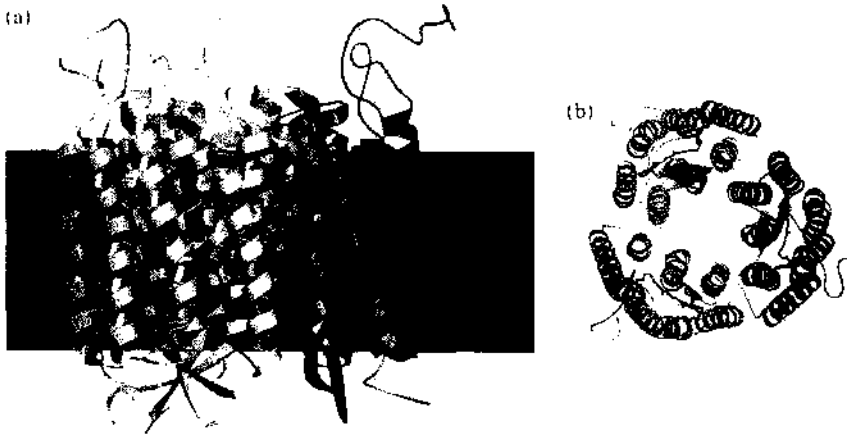


Figure 14. Trimeric structure of bacteriorhodopsin: (a) side view and (b) from the top(PDB code 1CWQ).

the "photocycle," that follows the photoisomerization of the retinal produces distinct, spectroscopically identifiable photointermediates, which have been defined in kinetic and spectroscopic terms as BR_{57}, K_{590}, L_{550}, M_{412}, N_{560}, O_{640}. The photoreaction induces a vectorial transfer of a proton across the membrane, leading to the release of a proton at the extracellular side and an uptake from the cytoplasmic side. Our current knowledge of the structure and the photocycle of bR has been reviewed in detail by several authors (see e.g., Ref. [130]). Three decades of site-directed mutagenesis, static and time-resolved spectroscopy, and low-resolution projection and 3D maps have revealed many details of the photocycle. However, understanding how the energy absorbed by the chromophore leads to unidirectional (vectorial) ion transport requires detailed structural descriptions of each photointermediate at the atomic level. Bacteriorhodopsin was successfully crystallized at very high resolution using lipidic cubic phase methods [131]. But even at highest resolution, X-ray crystallography cannot capture all details, in particular dynamical processes during the coupling of conformational transitions and proton translocation. Therefore, molecular dynamics simulations have proven to provide valuable insights. Among others, molecular dynamics simulations have been used to elucidate conformational fluctuations [132–135] and bR-water mobility [136–139].

Theory and Simulation. Early intermediates of bacteriorhodopsin's photocycle were modeled by means of *ab initio* quantum mechanical/molecular mechanical and molecular dynamics simulations [140]. The photoisomerization of the retinal chromophore and the formation of photoproducts corresponding to the early intermediates were simulated by molecular dynamics simulations. By means of the quantum mechanical/molecular mechanical method, the resulting structures were refined and the respective excitation energies were calculated. Two sequential intermediates were found with absorption maxima that exhibit red shifts from the resting state. The intermediates were therefore assigned to the K and L states. In K, the conformation of the retinal chromophore is strongly deformed, and the NH bond of the Schiff base points almost perpendicular to the membrane normal toward Asp-212. The strongly deformed conformation of the chromophore and weakened interaction of the Schiff base with the surrounding polar groups are the means by which the absorbed energy is stored. During the K-to-L transition, the chromophore undergoes further conformational changes that result in the formation of a hydrogen bond between the NH group of the Schiff base and Thr-89 as well as other rearrangements of the hydrogen-bond network in the vicinity of the Schiff base, which are suggested to play a key role in the proton transfer process in the later phase of the photocycle.

The planarity of the polyene chain of the retinal chromophore in bacteriorhodopsin was studied using molecular dynamics simulation techniques and applying different force-field parameters and starting crystal structures [141]. The largest deviations from a planar structure are observed for the C13=C14 and C15=N16 double bonds in the retinal Schiff base structure. The other dihedral angles along the polyene chain of the chromophore, although having lower torsional barriers in some cases, do not significantly deviate from the planar structure. The results of the simulations of different mutants of the pigment show that, among the studied amino acids of the binding pocket, the side chain of Trp-86 has the largest impact on the planarity of retinal, and the mutation of this amino acid to alanine leads to chromophore planarity. Deletion of the methyl C20, removal of a water molecule hydrogen-bonded to H15, or mutation of other amino acids to alanine did not show any significant influence on the distortion of the chromophore. The results from this study suggest the importance of the bulky residue of Trp-86 in the isomerization process, in both ground and excited states of the chromophore, and in fine-tuning of the pKa of the retinal protonated Schiff base in bacteriorhodopsin. The dark adaptation of the pigment and the last step of the bacteriorhodopsin photocycle imply low barriers against the rotation of the double bonds in the Schiff base region. The twisted double bonds found in the present study are consistent with the proposed mechanism of these ground-state isomerization events.

Although the molecular structure of bR in its ground state is now well determined, details of intermediate photocycle states are still controversial [131, 142]. In addition, the amount

of buried ("internal") water molecules in bR which are assumed to play a decisive role in providing proton pathways and to be involved in the molecular mechanism leading to proton translocation, is still unclear. Very recently [138, 139] new details of the amount and the distribution of internal water molecules, and of the related hydrogen-bonded networks in bR that constitute proton pathways has been reported. The simulations reveal a much higher average number of internal water molecules per monomer than observed in crystal structures. The spatial distribution of water molecules (Fig. 15) and their corresponding hydrogen-bonded networks inside bacteriorhodpsin in its ground state (G) and late M intermediate conformations were determined. In addition, these studies [138, 139] provides important information on hydrogen-bonded networks in bR fluctuating on the ps to ns timescale [139] which was not seen in crystallographic studies.

The proton transfer occurring immediately after retinal trans-cis photoisomerization is still highly controversial. The gradual release of stored energy is inherently nonequilibrium: which photocycle intermediates are populated depends not only on their energy but also on their interconversion rates. To understand why the photocycle follows a productive (i.e., pumping), rather than some unproductive, relaxation pathway, it is necessary to know the relative energy barriers of individual steps. To discriminate between the many proposed scenarios of this process, all the possible minimum-energy paths have been computed [143]. The results show that not one, but three very different pathways have energy barriers consistent with experiment. This result reconciles the conflicting views held on the mechanism and suggests a strategy by which the protein renders this essential step resilient. Using quantum mechanical/molecular mechanical minimum-energy reaction-path calculations the first proton transport step following photon absorption in bacteriorhodopsin, from the 13-cis retinal Schiff base to Asp85, have been investigated [144]. The results suggest that the barrier may arise from breaking and forming of hydrogen bonds involving the Schiff base, Asp85, Asp212, and water w402, and from nonbonded interactions involving protein groups that respond to the charge rearrangements in the Schiff base region.

Data for bR indicate that its permanent dipol is not of some interfacial character but is due to a real asymmetry of the charge distribution. Electron-crystallographic refinement of bR [142] provide a dipole moment of ~570 D. It is suggested [145] that the permanent dipole of bR supports proton transport by attraction of protons inside and repulsion of protons outside of the cell.

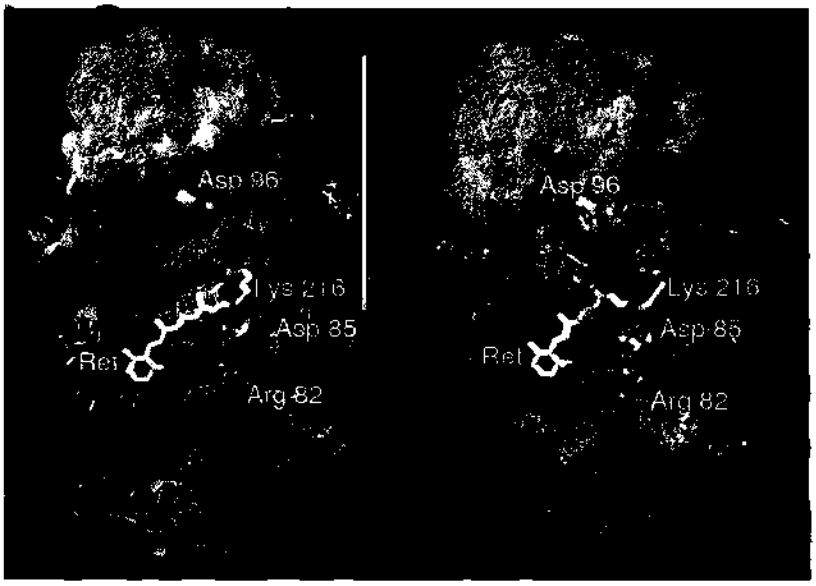


Figure 15. Accessible volumes for internal water molecules of G-state bR (left) and M-state bR (right). The surfaces of the volumes for trapped and diffusive water molecules are represented by yellow and blue triangulated nets, respectively. The red balls represent the positions of trapped water molecules as identified by crystallographic studies. Reprinted with permission from [139], S. Grudinin et al., *Biophys.* 88, 3252 (2005), Fig. 1. © 2005, Biophysical Society.

4.1.2. Halorhodopsin

Anion transport through biological membranes is less well understood than proton transport. Whereas several anion exchangers have been discovered and functionally characterized in eukaryotic and bacterial membranes, only a few anion pumps have been reported. To date, halorhodopsin (HR) is the only known anion pump that is energized by light. HR occurs in halophilic archaea and mediates primarily chloride, but also bromide, iodide, and nitrate import into the cell against their electrochemical gradients. Halorhodopsin belongs to a subfamily of heptahelical membrane proteins, the archaeal rhodopsins, as shown by sequence homology to bacteriorhodopsin (BR) and sensory rhodopsins. The structure of HR halorhodopsin was determined at 1.8 Å [146], a very high resolution for a membrane protein. This study took advantage of the fact that a chloride ion is much easier to determine in the crystal structure than a proton. A single chloride ion was found in the transport site between Lys242 and the retinal chromophore. The protein structure, in combination with energetic calculations, explained why chloride and proton translocation modes are mechanistically equivalent in archaeal rhodopsins. So far, HR has not been investigated in detail using molecular simulations.

4.2. Ion Translocases

4.2.1. V-ATPases

Structure and Function. Intracellular compartments exhibit a significant difference between their lumenal pH and that of the bulk cytoplasm. This pH difference is maintained mostly by the vacular ATPases. The vacuolar H^+-ATPases (or V-ATPases) are a family of ATP-dependent proton pumps responsible for acidification of intracellular compartments and, in certain cases, proton transport across the plasma membrane of eukaryotic cells. Vacuolar acidification is essential for a variety of cellular processes [147] including ligand-receptor dissociation and receptor recycling following receptor-mediated endocytosis, intracellular targeting of newly synthesized lysosomal enzymes, protein processing, and degradation, and coupled transport of small molecules, such as neurotransmitters. Acidification of endosomal compartments is also necessary for budding of transport vesicles and infection by certain envelope viruses, such as influenza virus. V-ATPases have also been identified in the plasma membrane of certain cells where they function in such processes as renal acidification, pH homeostasis, bone resorption, and coupled potassium transport. They have also been implicated in a variety of disease processes, including renal tubular acidosis, osteopetrosis, and tumor metastasis.

V-ATPases are multisubunit complexes (Fig. 16) composed of a peripheral domain (V_1) responsible for ATP hydrolysis and an integral domain (V_0) responsible for proton translocation. Based upon their structural similarity to the F_1F_0-ATP synthases [148], the V-ATPases are thought to operate by a rotary mechanism in which ATP hydrolysis in V_1 drives rotation of a ring of proteolipid subunits in V_0. The peripheral V_1 domain contains eight different subunits (subunits A-H). The integral V_0 domain, responsible for proton translocation, is composed of five subunits (a, b, c, c', c''). A model for the arrangement of subunits in the V-ATPase complex is shown in Fig. 16. The V-ATPases are structurally related to the F-ATPases (discussed in Section 4) that normally function in ATP synthesis in mitochondria, chloroplasts, and bacteria. This relationship between the V- and F-ATPases is evident both in the overall structure of the complexes and in the sequence of certain subunits. Electron microscopy has revealed a similar overall shape for the two proteins, with the peripheral domain of each complex connected to the corresponding integral domain by both central and peripheral stalks. These stalks are believed to play an important role in the mechanism by which these enzymes catalyze ATP-dependent proton transport. However, there are also significant structural differences [148] between the F- and V-ATPases revealed by electron microscopy. For example, images of the V-ATPase reveal a complex array of projections emanating from the V_1 domain, and a cufflike cytoplasmic region on the V_0 domain. Some studies suggest that V_1 and V_0 may be connected by as many as three stalks. Functionally, although the V-ATPases have been shown to be reversible, they normally operate, unlike the F-ATPases, in the direction of ATP-driven proton transport. In terms of primary sequence,

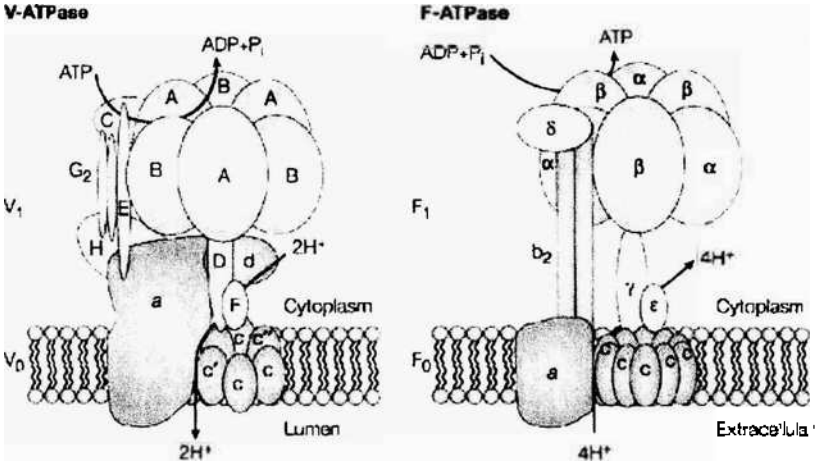


Figure 16. Structural model of the V-ATPase and comparison with the F-ATPase. The V-ATPases are composed of a peripheral V_1 domain responsible for ATP hydrolysis and an integral V_0 domain responsible for proton translocation. The corresponding domains for the F-ATPases are F_1 and F_0. The nucleotide binding sites are located on the A and B subunits of V_1, with the catalytic sites located primarily on the A subunits. Proton translocation is postulated to occur at the interface of the a subunit and the ring of proteolipid subunits (c, c' and c''). The V_1 and V_0 domains are connected by both a central stalk (composed of the D and F subunits) and a peripheral stalk (composed of the C, E, G, and H subunits). These stalks play a crucial role in the assumed rotary mechanism of ATP-driven proton transport. Reprinted with permission from [147]: T. Nishi and M. Forgac, *Nat. Rev. Mol. Cell Biol.* 3, 94 (2002), Fig. 1. © 2002, Macmillan Magazines Ltd.

three subunits show clear sequence homology, indicating a common evolutionary origin for the V- and F-ATPases. These include the two nucleotide binding subunits (A and B), which are approximately 20–25% identical to the and subunits of the F-ATPases, and the proteolipid subunits (c, c', and c''), which are homologous to the c subunit of F_0.

Theory and Simulation. A mechanochemical model of V-ATPase proton pump has been proposed by Grabe et al. [149] and is depicted in Fig. 17. They proposed two models: the two-channel model and the one-channel model for the a-subunit. In the two-channel model (Fig. 17[A]) two half-channels penetrate the stator to the level of the rotor sites; all other parts of the rotor-stator interface are hydrophobic. A horizontal polar strip connects the channels to allow the passage of an unprotonated site, but protons are blocked from leaking by the stator charge. Protons enter the input (basic) channel and bind to a rotor site, largely neutralizing it. Rotation carries the protonated site one complete revolution (to the left) where the site enters the output (acidic) channel (from the right). The stator charge forces the site to relinquish its proton into the lumen. Note that the sizes of the rotor and stator are such that two rotor sites cannot fit in the rotor-stator interface at once. In the one-channel model (Fig. 17[B]) the rotor sites now lie above the level of the membrane and are in equilibrium with the cytoplasm when not in the stator. A single channel penetrates the stator to the level of the rotor sites; all other parts of the rotor-stator interface are hydrophobic. A horizontal polar strip connects the channel to the cytoplasm to allow passage of an unloaded site. The stator charge blocks passage of protons through the strip. Rotation brings a protonated site into the output channel where the stator charge forces it to release its proton to the lumen. The unloaded site then rotates through the polar strip, past the stator charge, exiting the rotor-stator interface to the left.

In their paper, Grabe et al. [149] used a mathematical formulation of the model where the motion of the rotor, in terms of its rotation angle $\theta(t)$, was computed from a force balance equating the viscous drag on the rotor to the torques that act on the rotor and the Brownian force modeling the rotor's thermal fluctuations by a Langevin-type equation:

$$\zeta \frac{d\theta}{dt} = \tau_Q(\theta, s) + \tau_{el}(\theta, s) + \tau_e(\theta, s) - \tau_D(\theta) + \tau_B(t)$$

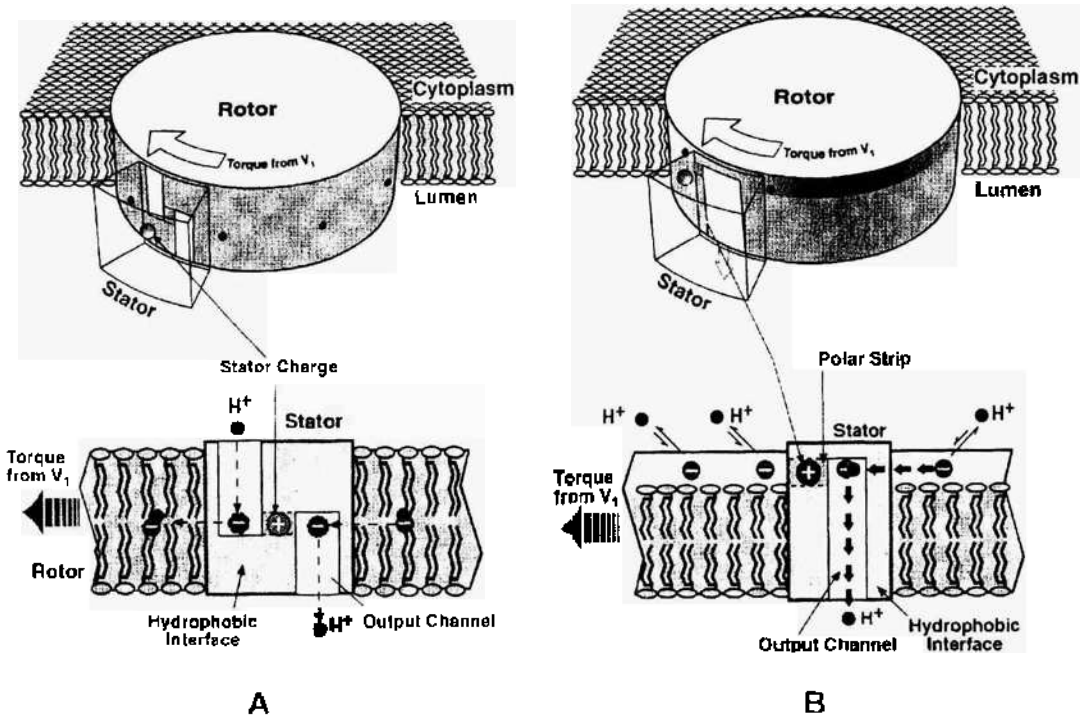


Figure 17. Mechanochemical model of the V-ATPase as proposed by Grabe et al. [149]. Perspective and face-on views of the rotor-stator assembly for the one- and two-channel models showing the paths protons follow in moving from the cytoplasm (top) into the lumen (bottom). Trque supplied by V_1 from ATP hydrolysis turns the rotor to the left (clockwise viewed from the cytoplasm). Reprinted with permission from [149], M. Grabe et al., *Biophys. J.* 78, 2798 (2000), Fig. 2. © (2000), Biophysical Society.

The various terms correspond to the rotator-stator charge interaction, the membrane potential, the dielectric barrier, the driving torque from V_1, and the Brownian torque, respectively. The chemical states of the rotor are described by the binary variable $\mathbf{s}_i$ for full or empty, and determined by a Markov equation $ds/dt = K(\theta)s$, where $K(\theta)$ is the transition matrix. Using appropriate parameters, the model calculations reproduced a variety of experimental measurements of performance of the V-ATPase proton pump.

4.2.2. Ca^{2+}-ATPase

Structure and Function. The most impressive structure in the membrane transporter field is the Ca^{2+}-ATPase solved by Toyoshima et al. [150]. Ca^{2+}-ATPase transports ions across the membrane against a concentration gradient. The structure of the Ca^{2+}-ATPase of skeletal muscle sarcoplasmic reticulum was solved at 2.6 Å in the presence of Ca^{2+} ions, which were found in the transmembrane domain. Comparison with a low-resolution electron density map of the enzyme in the absence of Ca^{2+} suggests that large domain movements, for both the transmembrane and soluble domains, take place during active transport. Ca^{2+}-ATPase uses the free energy derived from the hydrolysis of ATP to pump Ca^{2+} from the cytoplasm to the inside of the sarcoplasmic reticulum, and causes relaxation of the heartbeat, as well as of other muscles. Ca^{2+}-ATPase belongs to the class of P-type ATPases, which differ from the F,V-type ATPases by the fact that ATP is not hydrolyzed directly to ADP and H_2PO4, but that instead two chemical reactions (phosphorylation and dephosphorylation) are involved. There exists experimental evidence that both the phosphorylation and the dephosphorylation of Ca-ATPase are reversible (i.e., they are associated with very small free energy changes). This suggests that Ca-ATPase, like F_1-ATPase, optimizes its use of free energy by coupling the available free energy to protein conformational changes. The similarities between these different types of motor enzymes suggest that there exists a common strategy for the efficient use of chemical energy.

Crystal structures (PDB codes: 1SU4, 1IWO, 1VFP, 1T5S, 1T5T) determined provide atomic models for all four principal states in the reaction cycle. These structures show

that the three cytoplasmic domains rearrange to move 6 out of 10 transmembrane helices (Figs. 18, 19), thereby changing the affinity of the Ca^{2+}-binding sites and the gating of the ion pathway. Release of ADP triggers the opening of the lumenal gate and release of phosphate its closure, effected mainly through movement of the A-domain, the actuator of transmembrane gates.

Theory. Normal mode analysis is a powerful approach for analyzing the structural and dynamical features of macromolecules [151]. Although it is approximate because only the harmonic motion of the system around a single potential minimum is taken into account, qualitative estimates can be made for many properties of macromolecules such as the magnitude of atomic fluctuations, displacement covariance matrix, vibrational entropy [152], and neutron scattering spectra [153]. Several recent studies have demonstrated that low-frequency normal modes are implicated in large-scale conformational transitions in allosteric proteins. The fact that these conformational changes occur typically on the timescale of microseconds or longer and therefore are not accessible to standard molecular dynamics simulations makes the insights available from normal mode analysis very valuable. The low-frequency modes from normal mode analysis have also been used as basis vectors for approximate molecular dynamics simulations or refinement of X-ray or NMR data. Moreover, quantum mechanical effects of nuclear motions on molecular properties can be included straightforwardly in the normal mode framework. The standard normal mode analysis requires the calculation of a mass-weighted second derivative matrix (hessian) followed by the diagonalization of that matrix. Accordingly, there are a number of bottlenecks associated with its application to large systems that contain more than ten thousand atoms. First of all, straightforward calculations of the hessian scale quadratically as the system size, although the scaling can be made linear using a cutoff scheme for nonbonded intermolecular interactions. For long-range interactions such as electrostatics forces, it is possible, in principle, to extend techniques such as fast multipole algorithms to make the evaluation of hessian scale favorably with the system size. The second challenge concerns the storage of the hessian matrix, which formally scales as $(3N)^2/2$, in which N is the number of atoms in the system. Thus, the size of the hessian quickly exceeds the memory of a typical workstation as the number of

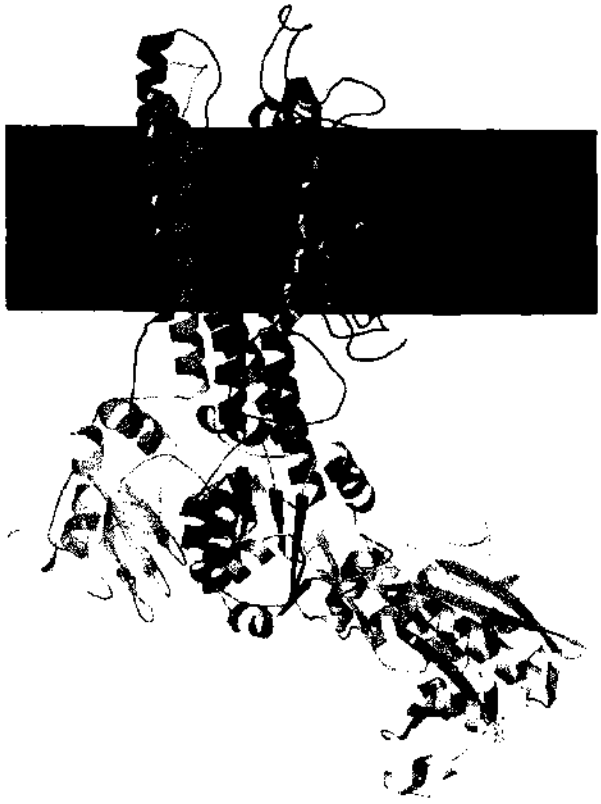


Figure 18. Ribbon view of the Ca^{2+}-ATPase of *Sarcoplasmic reticulum* (PDB code 1EUL).

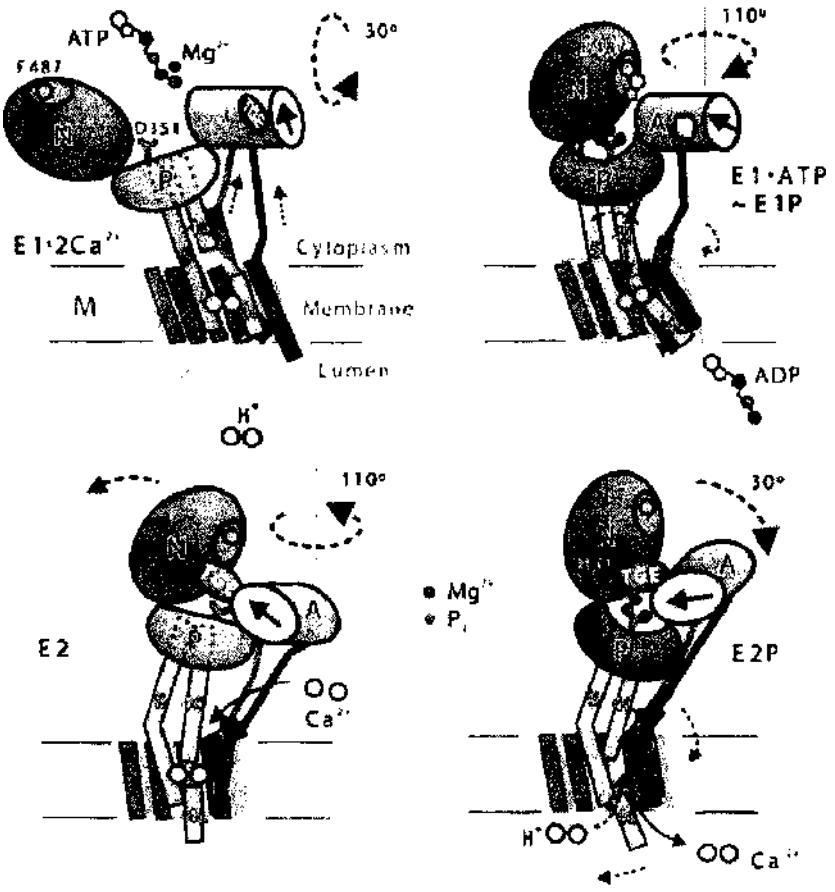


Figure 19. Illustration depicting the structural changes of the Ca^{2+}-ATPase during the reaction cycle, based on the crystal structures in different states. Reprinted with permission from [150], C. Toyoshima et al., *Nature (London)* 432, 361 (2004), Fig. 6, © 2004, Macmillan Publishers Ltd.

atoms increases, for example, for a 10,0000-atom system, storage of the hessian requires more than 4 gigabytes of memory or disk-space. For large systems, however, the hessian is expected to be very sparse, especially when a finite cutoff is used for the intermolecular interactions. The storage problem, therefore, could be resolved by keeping only the nonzero matrix elements and the corresponding indices in the memory. The third challenge is related to the diagonalization of the hessian, which scales as N^3 with standard diagonalization algorithms such as the Jacobi transform. If only a small number of eigenvalues and eigenvectors are of interest, iterative diagonalization approaches with more favorable scaling such as block Lanczos algorithms can be used. An alternative avenue for performing normal mode analysis for large molecules is to use a coarse-grained approach. This is particularly useful when only the low-frequency modes are of interest. This is because these low-frequency motions are typically delocalized throughout the system and involve mainly collective movements of residues. A promising approach along this line is the so-called block normal mode method (BNM, or as rotation translation basis) [154]. In this method, the system is divided into a number of blocks, and each block may contain one or more groups of atoms (such as an amino acid or a secondary structure element). The hessian matrix for the entire system is then projected into a subspace spanned by the translational and rotational motions of these blocks. Only the resulting projected hessian, which is much smaller in dimension than the full atomic hessian, has to be diagonalized. Therefore, the BNM method dramatically reduces the computational cost for the diagonalization step, which can be rate limiting for large systems. As an illustration of the BNM approach for large systems, Li and Cui [155] have applied the BNM to study the low-frequency modes of Ca^{2+}-ATPase, which contains 994 residues. Instead of requiring more than 3 gigabytes of memory, the BNM calculation with one residue per block (for 5964 modes) only used several hundred megabytes of memory and took less than 6 hours on an 1.2-GHz Athlon. The BNM method

was implemented into the simulation program CHARMM. The structural flexibility of the cytoplasmic domains (especially domain N), correlated motions among residues on domain interfaces and displacement patterns for the transmembrane helices observed in the BNM results were discussed in relation to the function of Ca-ATPase. The BNM analysis revealed the structural flexibility of the cytoplasmic domains. Hinge regions were identified for the rotational type of motions of various domains, which are on the interface between domains. Correlated motions were found between two helices in domains P1 and P2 and the loop (1.67) connecting transmembrane helices M6 and M7, as well as among residues on the interface between domain P and transmembrane helices M4 and M5. These correlated motions are expected to be essential for the communication between the calcium-binding site in the transmembrane domain and the phosphorylation site in domain P, which are separated by ~50 Å. The patterns in the motion of transmembrane helices suggest that helices M2, M4, and M8 are likely to be involved in calcium release to lumen, whereas helix M4 seems to be essential to the calcium binding from cytoplasm. The magnitude of their motions, however, is fairly small based on normal mode results at 300 K, which is consistent with the scheme that calcium binding and release are regulated by the binding and hydrolysis of ATP as well as phosphorylation [150]. It is hoped that the direct implementation of BNM as illustrated for Ca^{2+}-ATPase will make normal mode analysis a routine tool for analyzing the structural and dynamical properties of very large molecules. In addition, an implementation of the BNM approach may provide the way for developing efficient sampling algorithms with molecular dynamics or Monte Carlo for studying long-time scale dynamics of macromolecules.

4.2.3. Na^+-K^+-ATPase

One of the best studied transport systems is the Na^+-K^+-ATPase in the plasma membrane of higher eukaryotes. This membrane protein consists of two types of subunits: an α subunit that contains the enzyme's catalytic activity and ion binding sites, and a glycoprotein β subunit of unknown function. Sequence analysis suggests that the α subunit has 8 transmembrane segments and 2 large cytoplasmic domains. The β subunit has a single transmembrane helix a large extracellular domain. The protein may function as an $(\alpha\beta)_2$ tetramer *in vivo*. The Na^+-K^+-ATPase is often called Na^+-K^+-pump because it pumps Na^+ out of and K^+ into the cell. The transport is coupled to hydrolysis of intracellular ATP. The overall reaction is:

$$3Na^+(in) + 2K^+(out) + ATP^{4-} + H_2O \rightarrow 3Na^+(out) + 2K^+(in) + ADP^{3-} + H_2PO_4^-$$

Thus Na^+-K^+-ATPase separates charges across the membrane because three positive charges exit the cell for two charges that enter the cell. The resulting concentration gradient of ions is responsible for the electrical excitability of nerve and muscle cells. More details of the kinetic scheme for the active coupled transport of Na^+ and K^+ can be found in textbooks [156]. So far, no computer simulations on Na^+-K^+-ATPase has been reported.

4.3. Solute Translocases

4.3.1. ABC Transporter

ATP-binding cassette (ABC) transporters utilize the energy of ATP hydrolysis to translocate a wide variety of solutes and lipids across cellular membranes. These proteins share a common architectural organization comprising two hydrophilic nucleotide-binding domains (NBDs) located at the cytoplasmic surface of the membrane and two hydrophobic transmembrane domains (TMDs) that form the translocation pathway. The conformational changes in the NBDs induced by ATP hydrolysis and the means by which they are transmitted to the TMDs to effect substrate translocation remain largely unknown. The ABC transporter is illustrated in Fig. 20(a) by showing the extracellular monomer HisP. Although the crystal structures of ABC transporter NBDs are available, they present a static (time- and space-averaged) snapshots of the structure either in the presence or absence of bound nucleotide. A clearer understanding of conformational change requires elucidation of the dynamic structural changes occurring in response to nucleotide binding and hydrolysis. In this respect molecular dynamics (MD) simulations provide a tool for computational investigations of protein

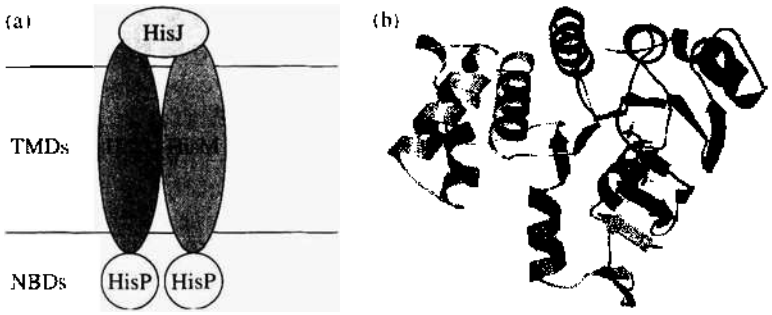


Figure 20. The organization of an ABC transporter: (a) The domain organization of an ABC transporter is illustrated for the $HisQMP_2J$ complex. HisP is the intracellular NBD, HisQ, and HisM are the TMDs, and HisJ is a periplasmic binding protein. (b) The structure of the HisP monomer (PDB code 1B0U) is shown.

conformational changes. Such conformational changes may occur both within a domain (in response to binding of nucleotide, for example) and between domains (allosteric interactions). Evidence has accrued that the initial event in the catalytic/transport cycle, namely the binding of nucleotide, is sufficient to induce a substantial conformational change.

4.3.2. HisP

Two molecular dynamics simulations have been reported for the domain HisP of an ABC transporter (Fig. 20b). Jones and George [157] have performed a molecular dynamics simulation of HisP, the well studied NBD of the bacterial histidine permease, to identify hinges and switches of the NBD conformational transitions and subunit-subunit interfaces. Their analysis revealed that the transmembrane domains (TMDs) regulate ATP hydrolysis by controlling conformational transitions of the NBD helical domains, and identifies the conformational changes and the crucial TMD:NBD interface, by which the energy of ATP hydrolysis is transmitted to the TMDs. They proposed a model for the catalytic cycle of ABC transporters that shows how substrate-binding and transport by the TMDs may be coordinated and coupled with ATP binding and hydrolysis in the NBDs. More recently, Sansom and co-workers [158] have employed molecular dynamics simulations to study the interactions of nucleotide with NBD. The results of extended (20 ns) MD simulations of HisP, the NBD of the histidine transporter $HisQMP_2J$ from *Salmonella typhimurium*, were reported. Analysis of the MD trajectories revealed conformational changes within HisP that are dependent on the presence of ATP in the binding pocket of the protein, and are sensitive to the presence/absence of Mg ions bound to the ATP. These changes are predominantly confined to the helical subdomain of HisP. Specifically there is a rotation of three α-helices within the subdomain, and a movement of the signature sequence toward the bound nucleotide. In addition, they observed considerable conformational flexibility in a conserved glutamine-containing loop, which is situated at the interface between the α-helical subdomain and the F_1-like subdomain. These results supported the mechanism for ATP-induced conformational transitions derived from the crystal structures of other NBDs.

4.3.3. MsbA

Very recently the structure of the transmembrane domain of an ABC lipid transporter was resolved at high resolution by Chang and Roth [159]. They published the structure of *E. coli* MsbA, a lipid flippase that is related to multidrug transporters in eucaryotic systems (Fig. 21). The asymmetric transmembrane distribution of lipids is generated and maintained by a number of phospholipid translocases ("flippases") that vary in lipid specificity, energy requirements, and direction of translocation. Some of the translocase proteins have been isolated, but presently it is not known how translocation works at the molecular level. A process that competes with chloride transport across biological membranes is phospholipid translocation or flip-flop. MsbA is organized as a homodimer, with each subunit containing six transmembrane helices and a nucleotide-binding domain. The asymmetric distribution of charged residues lining a central chamber suggests a general mechanism for the translocation of substrate by MsbA and other ABC transporters. Although the resolution is limited, this structure

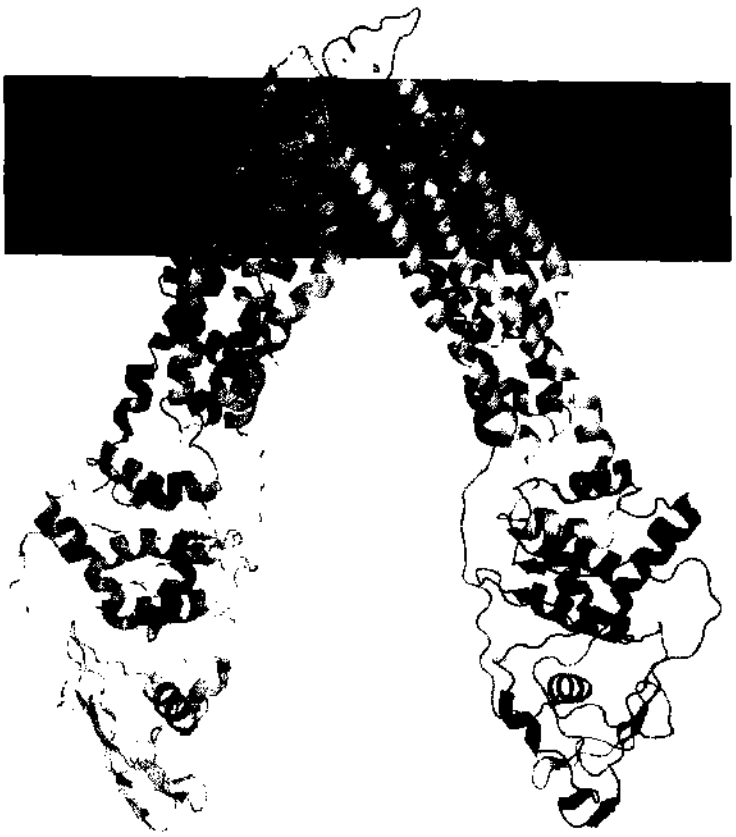


Figure 21. A schematic view of an ABC transporter, MsbA from *E. coli* (PDB code 1PF4).

will be an important foundation in the study of the transportation mechanism of ABC transporters, which form the largest family of membrane transporters. In order to generate a complete model of the prokaryotic ABC transporter MsbA from *E. coli* (PDB code 1JSQ), molecular modeling and simulation approaches have been reported recently [160], starting from the low-resolution structure-based Cα trace (PDB code 1JSQ). The quality of the MsbA model is assessed using a combination of molecular dynamics simulations and static structural analysis. These results suggested that the approach adopted for MsbA may be of general utility for generating all atom models from low-resolution crystal structures of membrane proteins. Molecular dynamics simulations of the MsbA model inserted in a fully solvated octane slab (a membrane mimetic environment) reveal that while the monomer is relatively stable, the dimer is unstable and undergoes significant conformational drift on a nanosecond time scale. This suggests that the MsbA crystal dimer may not correspond to the MsbA dimer *in vivo*. An alternative model of the dimer was suggested in the context of available experimental data.

4.3.4. Glucose Transporter GLUT3

Glucose enters cells via specialized membrane proteins known as glucose transporters (GLUTs) GLUTs are generally comprised of approximately 500 amino acids that are thought to be arranged in 12 transmembrane segments to form a pore. Biochemical studies have shown that both the *N*-terminus and the *C*-terminus of the protein face the intracellular compartment. A molecular model of the three-dimensional (3-D) structure of the glucose transport protein, GLUT3, has been derived by homology modeling [161] based of structural data from the mechanosensitive ion channel MscL and general insights from aquaporin AQP1. The model was subjected to energy minimization and molecular dynamics simulations in the presence of bound substrate (D-glucose). In the proposed structure of GLUT3, the 12 transmembrane helices form a right-hand barrel with a central hydrophilic pore. The pore is shaped like a funnel with dimensions of approximately 5–6 Å by 8 Å at its narrowest point. A network of polar and aromatic amino acids line the pore region and may facilitate

the movement of glucose along the channel. The molecular dynamics studies showed that changes in the tilt and flexibility of key transmembrane helices may modulate the opening of the pore to effect glucose transport.

4.4. Protein Translocases

All cells transport proteins into ("insertion") and across ("translocation") membranes. In doing so, the cells must maintain the membrane's permeability barrier between the cytoplasm and either the extracellular medium or other intracellular compartments. Because proteins are macromolecules, their passage through the membrane creates a nontrivial problem: how can a macromolecule pass through the membrane without creating a hole that destroys the permeability barrier? This is one of the fundamental problems in biology [162]. Relatively small proteins such as toxins and antimicrobial peptides are known to insert spontaneously themselves by physicochemical processes, where the peptides first get adsorbed to the membrane surface where they change their conformation into an α-helix, and then start to insert driven by the hydrophobic nature of the peptide segments [163–166]. Large proteins, however, accomplish their insertion into or translocation accross membranes either assisted by a complex proteinaceous machinery [167] or through fusion of membranes [168]. Protein translocation assisted actively by special membrane proteins ("protein translocases" or "translocons") is the basic mechanism for protein import of various cellular organelles and bacteria, for example, into mitochondria [169, 170], peroxisomes, chloroplasts, from the rough endoplasmic reticulum [171], and across the nuclear envelope [172].

All protein translocases possess several essential features. First, they contain intrinsic signal recognition sites that act as the docking sites for the targeting signals of translocation substrates. The docking sites may act as the primary receptors for polypeptides, or they may act downstream of primary signal receptor systems that target polypeptides from their site of synthesis (cis compartment) to the translocases. Second, translocases form selectively permeable protein-conducting channels that mediate transport from the cis compartment to their destination (trans compartment). Finally, translocons must be coupled to a translocation driving force. In most cases, the association of molecular chaperones with the polypeptide in the trans compartment provides the energy for translocation. Viewed from the perspective of classical membrane transporters, translocases represent remarkably flexible transport complexes. The translocation channels of translocases can accommodate hundreds or thousands of distinct protein substrates while maintaining the permeability barrier of the membrane. In addition to providing a conduit for complete translocation of polypeptides across membranes, many translocases sense stop-transfer signals within integral membrane proteins and gate laterally to allow the diffusion of transmembrane segments into the bilayer. Thus, the channels are not passive players in the translocation process.

Two classes of translocases have evolved in response to these extraordinary demands. The most common class of translocons is envisioned to function as derivatives of gated ion channels [173, 174]. In this class, the nascent or newly synthesized polypeptide is threaded vectorially through a gated protein-conducting channel in a largely unfolded conformation with the aid of molecular chaperones. By maintaining the polypeptide in an unfolded conformation, a single translocase of defined dimensions can accommodate a vast array of substrates. As such, the translocation reactions can be modeled as the transport of polyions through modified channels in a manner analogous to ion or metabolite transport. We refer to these systems as "signal-gated translocases."

The second class of translocases is distinguished by their ability to transport fully folded and/or oligomeric proteins of large dimensions while maintaining the membrane permeability barrier. Stable translocase channels have not been detected in these systems, leading to the proposal that translocases of variable apertures are assembled in response to the size of the translocation substrate. We refer to this class as the "signal-assembled translocons."

General Theoretical Approaches and Concepts. In a pioneering work by Oster and coworkers [175, 176] it was proposed that protein translocases move proteins across membranes using biased thermal motion, the "Brownian ratchet" mechanism, which is based on the chemical asymmetry between the cis and trans sides of the membrane such as the binding

and dissociation of chaperonins to the translocating chain, or chain coiling induced by pH or/and ionic gradients, or glycosylation, or disulfide bond formation. This concept dates back to Feynman [177] who showed that a "Brownian ratchet" can use thermal fluctuations to perform directed work given a temperature gradient.

Motivated by experiments in which a polynucleotide is driven through a proteinaceous pore by an electric field, Lubensky and Nelson [178] have studied theoretically the diffusive motion of a polymer threaded through a narrow channel with which it may have strong interactions. They showed that there is a range of polymer lengths in which the system is approximately translationally invariant, and they developed a coarse-grained description of this regime. From this description, general features of the distribution of times for the polymer to pass through the pore had been deduced. They also introduced a more microscopic model, which provides a physically reasonable scenario in which, as in experiments, the polymer's speed depends sensitively on its chemical composition, and even on its orientation in the channel. They pointed out that the experimental distribution of times for the polymer to pass through the pore is much broader than expected from simple estimates, and speculated on why this might be. Further theoretical studies following similar lines of approaches have been several [179–184]. More general aspects of the dynamics of polynucleotide transport through nanometre-scale pores have been reviewed recently [185].

A theory of channel-facilitated transport of long rodlike macromolecules through thin membranes under the influence of a driving force of arbitrary strength has been developed by Berezhkovskii and Gopich [186]. Analytic expressions were derived for the translocation probability and the Laplace transform of the probability density of time that a macromolecule spends in the channel by translocating and nontranslocating (returning back) macromolecules. These results were used to study how the distribution of the macromolecule lifetime in the channel depends on a polymer chain length and the driving force. It was shown that depending on the values of the parameters, the lifetime probability density may have one or two peaks. This theory is a generalization of the theory developed by Lubensky and Nelson [178].

A quantitative analysis of experimental data for posttranslational translocation into the endoplasmic reticulum was performed by Elston [187]. This analysis reveals that translocation involves a single rate-limiting step, which is postulated to be the release of the signal sequence from the translocation channel. The Brownian ratchet and power stroke models of translocation are compared against the data. The data sets are simultaneously fit using a least-squares criterion, and both models were found to accurately reproduce the experimental results. A likelihood-ratio test reveals that the optimal fit of the Brownian ratchet model, which contains one fewer free parameter, does not differ significantly from that of the power stroke model. Therefore, the data considered there cannot be used to reject this import mechanism. The models were further analyzed using the estimated parameters to make experimentally testable predictions.

Using Brownian molecular dynamics simulations the dynamics of the passage of a stiff chain through a pore into a cell containing particles that bind reversibly to it have been studied [188]. The mean first-passage time as a function of the length of the chain inside for different concentrations of binding particles has been estimated. As a consequence of the interactions with these particles, the chain experiences a net force along its length whose calculated value from the simulations accounts for the velocity at which it enters the cell. This force can in turn be obtained from the solution of a generalized diffusion equation incorporating an effective Langmuir adsorption free energy for the chain plus binding particles. These results suggest a role of binding particles in the translocation process that is in general quite different from that of a Brownian ratchet. Furthermore, nonequilibrium effects contribute significantly to the dynamics; for example, the chain often enters the cell faster than particle binding can be saturated, resulting in a force several times smaller than the equilibrium value.

Monte Carlo simulation studies of the translocation of homopolymers of length N driven through a channel have been performed by Loebl et al. [183]. They found that the translocation time τ depends on temperature in a nontrivial way. For temperatures below some critical temperature T_c, $\tau \sim T^{-4}$. Whereas for $T > T_c$, τ increases with temperature. The low

temperature results are in good agreement with experimental findings as is the dependence of on the driving field strength. The velocity of translocation displays the same characteristics as found in experiment but the N dependence of τ shows the linear relationship observed in experiment only for large values of N. A possible reason for this discrepancy may be attributed to the simple intermonomer potential used and the assumption that the drag is the same at all temperatures.

4.4.1. Mitochondrial Pore

Mitochondria fulfill a large variety of metabolic tasks in eukaryotic cells, such as the generation of ATP, the biosynthesis of haem, and amino-acid metabolism. Similar to proteins that are destined for other organelles, such as chloroplasts or peroxisomes, mitochondrial proteins are synthesized on cytosolic ribosomes and are subsequently transported to mitochondria post-translationally. Pore-forming membrane-embedded multiprotein complexes ("translocases") that are found in the outer and inner mitochondrial membrane mediate the transport of proteins across and into the outer or inner mitochondrial membranes.

The TOM (translocase of the outer mitochondrial membrane) complex (Fig. 22) consists of cytosol-exposed receptors and a pore-forming core, and it mediates the transport of proteins from the cytosol across and into the outer mitochondrial membrane. The SAM protein complex (sorting and assembly machinery), is involved in the biogenesis of β-barrel proteins of the outer membrane. Two translocases of the inner mitochondrial membrane (TIM complexes) mediate protein transport at the inner membrane. The TIM23 complex accomplishes the transport of presequence-containing proteins across and into the inner membrane. The TIM22 complex (a twin-pore translocase) catalyses the insertion of multispanning proteins that have internal targeting signals into the inner membrane. The TIM23 complex requires the PAM complex (presequence-translocase-associated motor complex) and a membrane potential, $\Delta\psi$, for the complete transport of proteins into the mitochondrial matrix. The TIM22 complex mediates the membrane insertion of multispanning inner-membrane proteins that have internal targeting signals, and it uses a $\Delta\psi$ as an external driving force. Membrane insertion by the TIM22 complex is a multistep process, in which the preprotein initially tethers to the translocase in a step that is independent of $\Delta\psi$. Subsequently, in steps that require a $\Delta\psi$, positive charges in the matrix-exposed loops of the precursor protein allow docking in the twin-pore translocase and, eventually, the precursor inserts into the inner membrane.

The protein complex TIM23 is of particular interest (Fig. 23). The transport of proteins across the inner membrane into the matrix requires two energy sources—the membrane potential of the inner membrane and the ATP-dependent activity of the presequence-translocase-associated import-motor (PAM) complex. The so called mitochondrial heat-shock protein-70 (mtHsp70) is the central component of the PAM complex and it binds directly to an incoming, unfolded polypeptide chain. Translocase of the inner mitochondrial membrane, Tim44, serves as a membrane anchor for mtHsp70. The function of mtHsp70 has to be tightly regulated to allow it to cycle between an ADP-bound state and an ATP-bound state, which correspond to states of high and low substrate affinity, respectively. This regulation requires the cooperation of mtHsp70 with cofactors. The regulated cycling of mtHsp70, however, is still unknown.

Although the functional role of mHsp70 could not precisely defined, it was clear that this protein complex is required for post-translational translocation into mitochondria (and the endoplasmic reticulum). Therefore it was assumed that mHsp70 should act as translocation ATPase. Two operating principles have been suggested and examined [175, 176, 189-191]. The "power stroke model" proposes that mHsp70 undergoes a conformational change, which pulls the precursor protein through the translocation pore, whereas, in the "Brownian ratchet model," the role of mHsp70 is simply to block backsliding through the pore.

Theory. Soon after the identification of mHsp70 as the translocation ATPase, it was postulated that protein import is driven by a Brownian ratchet mechanism [175, 176, 189, 190]. According to this model, a precursor chain diffuses within the translocation pore; soluble mHsp70 molecules bind to the chain on the matrix side of the pore, thereby hindering reverse diffusion and causing a net unidirectional movement of the chain into the matrix.

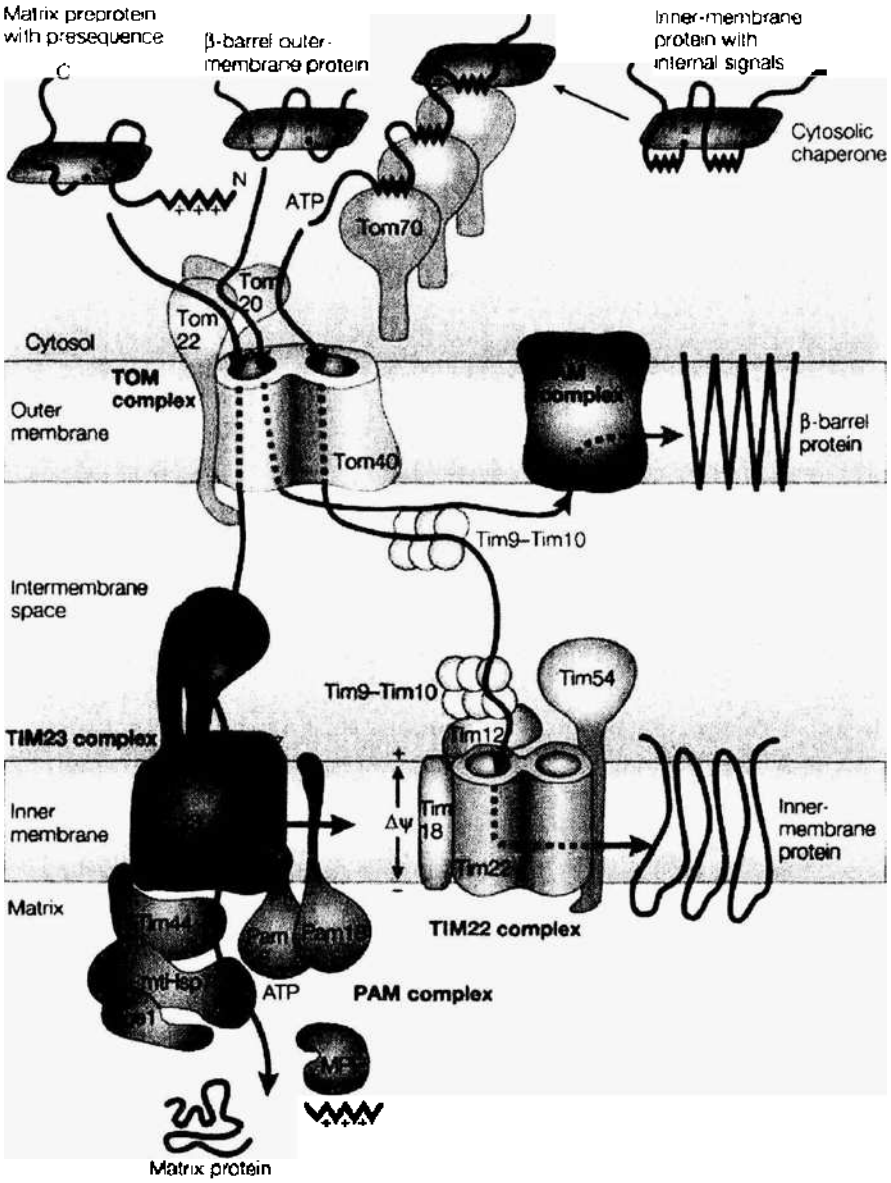


Figure 22. Protein-import pathways for mitochondrial proteins. Precursor proteins (brown) with positively charged amino-terminal presequences, β-barrel outer-membrane proteins (dark green), and multispanning innermembrane proteins (blue) with internal targeting signals are recognized by specific receptors of the translocase of the outer mitochondrial membrane (TOM). The precursor proteins are then translocated through the Tom40 pore. The β-barrel proteins then require the small Tim proteins to guide them through the intermembrane space, and the sorting and assembly machinery (SAM complex) for insertion and assembly into the outer membrane. Reproduced with permission from [169]. P. Rehling et al., *Nat. Rev. Mol. Cell Biol.* 5, 519 (2004), Fig. 1. © (2004), Macmillan Magazines Ltd.

A quantitative study suggested that this mechanism could account for the observed rates of translocation [175, 176]. However, three experimental facts had to be taken into account in the theory:

1. mHsp70 drives translocation in conjunction with Tim44 (Fig. 23), a membrane-associated subunit of the inner membrane import machinery.
2. Some precursors fold before import, and their unfolding can be accelerated by the ATP-dependent action of mHsp70.
3. The diffusion of precursor proteins within the translocation pore is far slower than originally assumed.

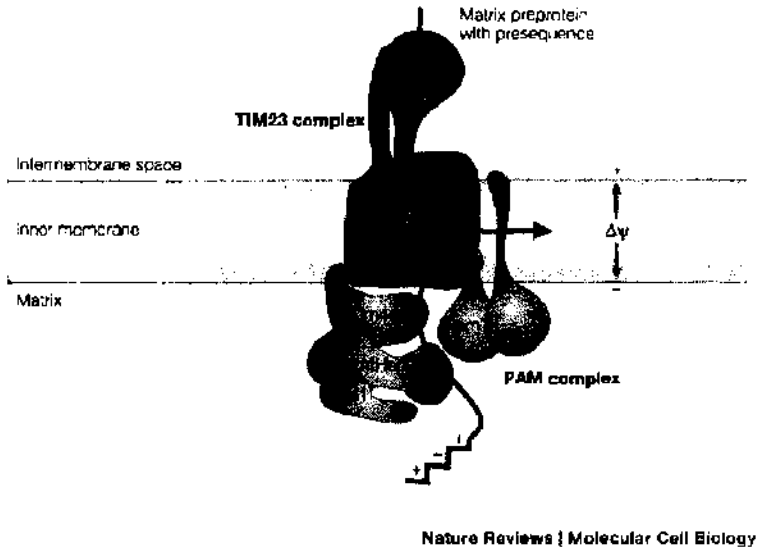


Figure 23. The presequence-translocase-associated import motor PAM. Reprinted with permission from [169], P. Rehling et al., *Nat. Rev. Mol. Cell Biol.* 5, 519 (2004), Fig. 5, © 2004, Macmillan Magazines Ltd.

Attempts have been made to bring the Brownian ratchet model and the power stroke model in accord with the experimental observations. Of course, a theoretical approach will not by itself reveal how proteins are imported into mitochondria. However, this kind of analysis serves two purposes. First, it helps to define possible translocation mechanisms that are consistent both with the experimental data and with the laws of mechanics. Second, it highlights questions that will generate further insights into the import process. However, the major conclusion of the study by Chauwin et al. [190] was that the interactions of precursor proteins with the translocation pore are very significant. On the basis of existing experimental systems, they developed a simple mathematical procedure to calculate diffusion coefficients for precursor chains in the translocation pore. Remarkably, the calculated diffusion coefficients are 106-fold lower than for a freely diffusing chain. This phenomenon has received little attention, but it should be incorporated into any mechanistic description of translocation. Strong drag forces in the pore greatly reduce the predicted efficiency of a Brownian ratchet, and under these conditions a power stroke motor could outperform a Brownian ratchet. Their simulated power stroke motor did not match the translocation rates observed in actual mitochondria, indicating that the current translocation model was too simplistic.

There are two key assumptions underlying this model. The first assumption is that precursor–pore interactions exert a drag force of constant magnitude. In reality, different segments of the precursor chain will interact differently with the pore walls. They performed simulations in which the precursor chain experienced occasional strong resistance instead of a constant, moderate drag force. This modification had virtually no effect on translocation rates in the ratchet model, but it significantly accelerated translocation in the power stroke model. The second important assumption is that the chain segment in the pore is fully extended, capable neither of compression nor of further extension. In fact, the mitochondrial translocation pore is quite flexible, and translocating precursor chains may be able to form helices or other elements of secondary structure. Hence, it would be more accurate to model the precursor as a flexible chain than as a rigid rod. This modification would probably make little difference in the Brownian ratchet model, but in the power stroke model it would be expected to accelerate translocation. The strength of the simulated Brownian ratchet is inversely proportional to the size of the ratcheting interval, which is the distance the chain diffuses between the binding of successive mHsp70 molecules. With a very small ratcheting interval it would be possible to overcome the drag forces caused by strong pore interactions. However, on the basis of the known properties of the mitochondrial translocation machinery, it seems unlikely that such a small ratcheting interval could be achieved in this system. Whether the power stroke model is a plausible representation of mHsp70 function is still a matter of debate. No information is yet available about the putative mHsp70 power stroke.

But even if this power stroke turns out to be weaker than we have assumed, it is conceivable that two mHsp70 molecules can pull simultaneously on the precursor chain. The strength of the mHsp70 motor could be examined directly by adapting optical laser trap technology to study protein translocation into mitochondria. A separate issue concerns the ATPase cycle time of mHsp70. The power stroke motor must turn over between strokes. Therefore, to be compatible with the observed translocation rates, the ATPase cycle time for the complete mHsp70 chaperone system must be smaller than 5 seconds. This number can be measured once all of the components of the mHsp70 chaperone system become available in purified form. Although the calculations favor a power stroke mechanism, they do not allow to distinguish conclusively between the Brownian ratchet and power stroke models.

A recent mathematical analysis of both mechanisms has been performed by Elston [191], which revealed that qualitative differences between the two models, Brownian ratchet and power stroke, occur in the behavior of the mean velocity and effective diffusion coefficient as a function of mHsp70 concentration. However, a final distinction of applicability between ratchet and power stroke will come from other lines of investigation. For example, most studies have focused on measurements of average translocation times, but additional mechanistic information could be obtained by examining the variances of translocation times. A second approach should focus on the details of the mHsp70 reaction cycle. In the Brownian ratchet model, mHsp70 must dissociate from Tim44 while remaining bound to the precursor chain. In the power stroke model, mHsp70 must remain bound to both Tim44 and the precursor chain. One of the most promising ways to examine the mechanism of translocation is to characterize the mHsp70-dependent unfolding of precursor proteins. Indeed, the power stroke notion was inspired by the finding that mHsp70 could unfold a stably folded precursor protein. Although the models investigated so far assume that the precursor chain is already unfolded, they could be extended to incorporate an unfolding event. It seems likely that mHsp70 causes unfolding by pulling on the *N*-terminus of a folded domain This novel reaction could eventually be examined by a combined theoretical and experimental study.

4.4.2. Endoplasmic Pore

Structure and Function. A decisive step in the biosynthesis of many secretory and plasma-membrane proteins is their transport across the endoplasmic reticulum (ER) membrane in eukaryotes or across the cytoplasmic membrane in prokaryotes [192]. These polypeptides are first targeted to the membrane by hydrophobic amino acid sequences, which are either cleavable signal sequences or transmembrane segments (TM) of membrane proteins. Soluble proteins, such as those destined for secretion, are subsequently transported across the membrane through a protein-conducting channel with a hydrophilic interior. In the case of membrane proteins, when a hydrophobic TM arrives in the channel, it is released through an opening in the channel wall into the surrounding lipid phase. The capacity of the channel to open laterally towards the lipid and the wide variety of substrates that it must transport distinguish it from many other channels. An evolutionarily conserved heterotrimeric complex of membrane proteins, called the Sec61 complex in eukaryotes and the SecY complex in eubacteria and archaea, forms the channel [193]. The α-subunit forms the channel pore, and it is the crosslinking partner of polypeptide chains passing through the membrane. Reconstitution experiments have shown that the Sec61/SecY complex is the essential membrane component for protein translocation. The channel itself is a passive conduit for polypeptides and must therefore associate with other components that provide a driving force [194]. In co-translational translocation, the major partner is the ribosome. The elongating polypeptide chain moves directly from the ribosome into the associated membrane channel. The energy for translocation comes from GTP hydrolysis during translation. Many (or perhaps all) cells also have post-translational translocation, in which polypeptides are completed in the cytosol and then transported across the membrane. In yeast (and probably in all eukaryotes), the post-translational translocation partners are another membrane protein complex (the tetrameric Sec62/63p complex) and the lumenal protein BiP, a member of the Hsp70 family of ATPases. BiP may promote translocation by acting as a molecular ratchet or power stroke machine, preventing the polypeptide chain from sliding back into the cytosol. In the eubacterial post-translational pathway, the cytosolic ATPase SecA pushes polypeptides

through the channel. In addition, an electrochemical gradient across the membrane stimulates translocation *in vitro* and is essential *in vivo*.

An understanding of the mechanisms that underlie protein translocation requires detailed structural information. Low-resolution structures have been obtained by single-particle electron microscopy (EM) of either the isolated Sec61/SecY complex or the Sec61 complex bound to the ribosome. Van den Berg and co-workers [193] have reported the structure of the SecY complex from the archae *M. jannaschii*, determined by X-ray diffraction at 3.2 Å resolution, and have proposed based on the X-ray structure the following refined model for the translocation of secretory proteins. Initially, the channel is closed because the plug blocks the pore (Fig. 24, stage 1). Next, a channel partner binds; depending on the mode of translocation and the organism, this can be a ribosome, the Sec62/63p complex or SecA (stage 2). Although part of an oligomer, only one copy of the SecY/Sec61p complex forms the active pore. The closed state of the channel may be destabilized by a conformational change, but binding of the partner alone is insufficient to completely open the channel. In the next step, the substrate inserts as a loop into the channel, with its signal sequence intercalated between TM2b and TM7, and with its mature region in the pore (stage 3). Insertion requires a hinge motion to separate TM2b and TM7, and displacement of the plug to its open-state position close to one of the subunits. The mature region of the polypeptide chain is then transported through the pore, and the signal sequence is cleaved at some point during translocation (stage 4). While the polypeptide chain is moving from an aqueous cytoplasmic cavity to an external one, the pore ring forms a seal around the chain, hindering the permeation of other molecules. Finally, when the polypeptide has passed through, the plug returns to its closed-state position (stage 5).

Models and Theory. To understand the physical basis of the translocation, two mechanistic models for post-translational translocation been proposed and have been referred to as the "Brownian ratchet" and the "translocation motor" [175, 189] or "power stroke motor."

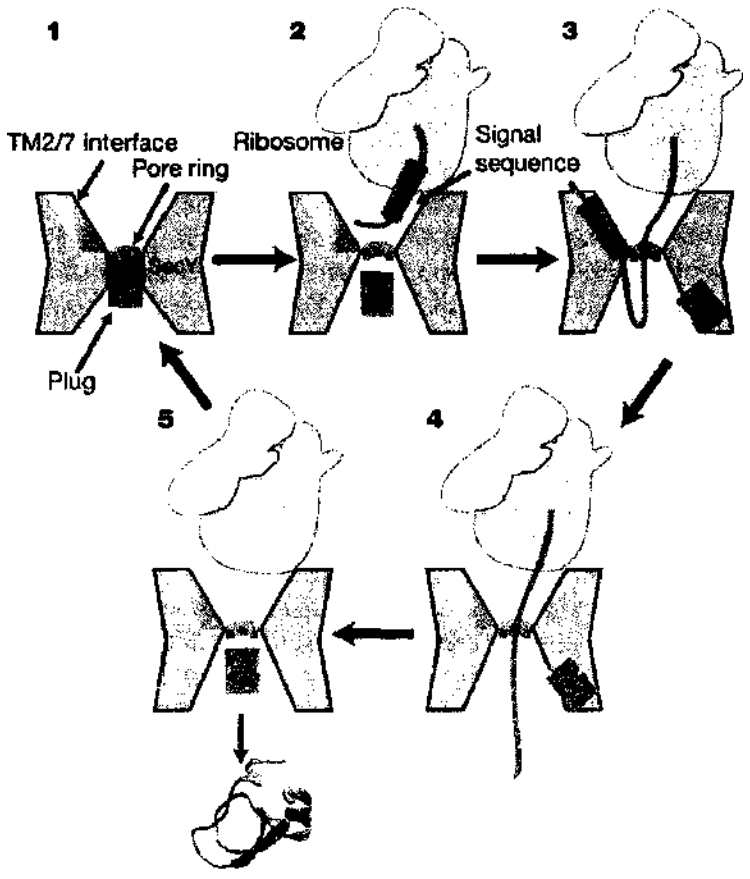


Figure 24. Different stages of translocation of a secretory protein. Reprinted with permission from [193], B. Van den Berg et al., *Nature* (London) 427, 36 (2004), Fig. 11. © 2004, Macmillan Publishers Ltd.

A detailed modeling and mathematical analysis of both models [187] have revealed that the two models make qualitatively different predictions. The differences are observed in the behavior of the mean velocity and effective diffusion coefficient as a function of organellar Hsp-70 concentration. Because direct measurement of these two quantities is not possible, a method has been proposed for determining the mean velocity and effective diffusion coefficient that only requires monitoring the fraction of protein released from the translocation channel as a function of time. This method has applied to one of the most commonly studied post-translational translocation systems in the ER. Similar as for the mitochondrial pore, a signal sequence near the amino terminus targets the precursor protein for import. In the ER, the central channel-forming protein is believed to be Sec-61. On the lumenal side of the membrane, Sec-61 associates with another protein complex (Sec-62/63p). A certain domain of Sec-63p interacts with organellar BiP, which is a member of the Hsp-70 family of ATPases. BiP also binds to the precursor protein and is responsible for providing directionality to the process. The mechanism that drives post-translational translocation is not known, and two different roles for organellar Hsp-70 have been suggested. In the first scenario, Hsp-70 associates with both the membrane bound complex (Sec-63p or Tim-44) and the precursor protein. Hsp-70 then undergoes an ATP-dependent conformational change that pulls the precursor through the pore [189]. In the second scenario, the role of Hsp-70 is simply to prevent backward diffusion of the precursor protein [175], and import relies on biased thermal diffusion. Both models represent "molecular motors" in that chemical free energy is used to produce directed motion. Elston has studied [187] two types of models for ER translocation pore developed on the basis of experimental information. A quantitative analysis of experimental data for post-translational translocation into the endoplasmic reticulum reveals that translocation involves a single rate-limiting step, which is postulated to be the release of the signal sequence from the translocation channel. The Brownian ratchet and power stroke models of translocation are compared against the data. The data sets are simultaneously fit using a least-squares criterion, and both models are found to accurately reproduce the experimental results. A likelihood-ratio test reveals that the optimal fit of the Brownian ratchet model, which contains one fewer free parameter, does not differ significantly from that of the power stroke model.

4.4.3. Nuclear Pore

The eukaryotic nucleus is surrounded by a protective nuclear envelope, which is perforated by trafficking machines termed nuclear pore complexes (NPCs). The NPCs are the sole mediators of exchange between the nucleus and the cytoplasm. Small molecules pass through the NPCs unchallenged. However, large macromolecules are excluded unless chaperoned across by transport factors. Specific carrier molecules, many of which belong to the karyopherin superfamily, mediate transport of cargoes through NPCs and also have an emerging role as regulators of other cellular processes, including mitosis and gene expression.

Structure and Function. Very recently the crystal structure of one carrier molecule, exportin Cse1p complexed with its cargo (Kap60p) and RanGTP, has been resolved [195]. The NPC [196, 197] has a eight-fold rotational symmetry perpendicular to the membrane, and are asymmetric with respect to the plane of the nuclear envelope (Fig. 25). The NPC is consists of a cylindrical core from which numerous peripheral filaments project toward the nucleus and cytoplasm. The complex can be minimally characterized as having three substructures: the cytoplasmic fibrils, a central core, and the nuclear basket. The central core in the NPC contains eight spokes sandwiched between the cytoplasmic and nuclear rings. The spoke structures collectively encircle the central region through which all active transport occurs. One can picture the NPC as a tubular hole in the nuclear envelope, bristling at each entrance with numerous filaments carrying a multitude of binding sites for transport factors. One of the basic questions is how such a simple structure can mediate all the complexities of gated transport. Operationally, NPCs are composed of proteins called nucleoporins (or Nups) forming the stationary phase for nucleocytoplasmic exchange, whereas the mobile phase consists of soluble transport factors and their cargoes. As nucleocytoplasmic transport is driven by a series of specific interactions between components of both phases,

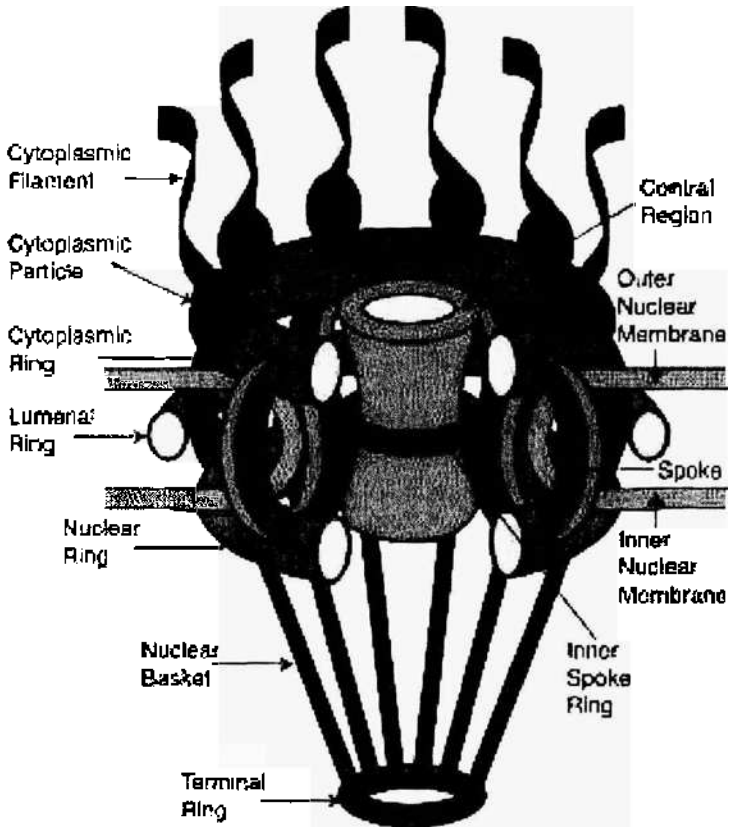


Figure 25. Schematic diagram of the nuclear pore complex (NPC). The structure has an eightfold rotational symmetry perpendicular to the plane of the nuclear membrane. Certain facing portions are removed in this diagram to reveal the architecture of the central region. Labels for all the structures found in vertebrate NPCs are included, with the cytoplasmic face on top. Reprinted with permission from [196]. M. Suntharalingam and S. R. Wente, *Dev. Cell* 4, 775 (2003). Fig. 1. © 2003, Elsevier.

it is frequently difficult to determine which proteins are permanent constituents of the NPC. Nevertheless, to understand how transport occurs, it necessary to characterize the players in both phases and understand how their interplay leads to the coordinated vectorial exchange of macromolecules across the nuclear envelope. Transport involves the recognition of nuclear localization signals on cargos destined for the nucleus and nuclear export signals on cargos destined for the cytoplasm. Most import and export signals are recognized by the β-karyopherin (kap) family of soluble transport receptor proteins (also known as importins and exportins). Once bound to a cargo, a kap negotiates the NPC, releases its cargo in the destination compartment and returns for another round. Because this cycle is directional and can accumulate cargos against a concentration gradient, an energy source and a directional cue are needed. Both are provided by the small GTPase Ran, which is controlled by two regulators—Ran guanosine nucleotide exchange factor (RanGEF) and Ran-specific GTPase activating protein (RanGAP). But how does the NPC select only the karyopherin-cargo complexes and exclude other macromolecules? In other words, how does the NPC act as a sorting machine?

Models and Theory. Several functional models have been proposed. The oily-spaghetti model [198], the selective phase model [199–201], the molecular-latch model [202], and the "virtual gating" model [172]. The virtual gating model has been suggest [172] to explain the mechanism of the rapid and selective macromolecular trafficking through the NPCs. One of the basic assumptions is that the entropic barrier for transporting macromolecules across the NPC plays a pivotal role. But how is the entropic barrier set up? Beyond simple occlusion, other factors could add to the barrier properties of the NPC. The intrinsically disordered nucleoporins (or nups) could act as entropic bristles. Molecules that are large enough to occupy a significant portion of this volume and move on the same timescale as the bristles tend to be excluded from this volume. The disordered filamentous sidearms of

neurofilaments and microtubule-associated proteins act as entropic bristles, whose "push" might help to keep the parallel arrays of their associated filaments regularly spaced. Similarly, the "push" from the nups could keep macromolecules away from the central channel, and the larger the macromolecule, the more it would feel this push. The appeal of this model is that it uses a well-studied polymer phenomenon consistent with the reported structure of nups.

On the basis of the observation that the nuclear pore complex contains a selective permeability barrier. Since experiments on the physical properties of this barrier appear to be in conflict with current physical understanding of the rheology of reversible gels, the paper by Bickel and Bruinsma [201] contains the proposal that the NPC gel is anomalous and characterized by connectivity fluctuations. A simplified model is developed to demonstrate the possibility of enhanced diffusion constants of macromolecules trapped in such a gel.

To reconcile the observed selectivity and the high rate of translocation through nuclear pores, Kustanovich and Rabin proposed [203] that the core of the nuclear pore complex is blocked by a metastable network of phenylalanine and glycine nucleoporins. Although the network arrests the unfacilitated passage of objects larger than its mesh size, the cargo-importin complexes act as catalysts that reduce the free energy barrier between the cross-linked and the dissociated states of the nups, and open the network. Using Brownian dynamics simulations they have calculated the distribution of passage times through the network for inert particles and cargo-importin complexes of different sizes and discuss the implications of their results for experiments on translocation of proteins through the nuclear pore complex.

5. ROTARY MOTORS

There are several membrane proteins that are thought to be rotary motors. These "rotarons" accomplish their primary functions by rotating one group of subunits with respect to the rest. Examples include the F_1F_0-ATPase and the bacterial flagellar motor, which are described in the following Section.

5.1. ATP Synthase

Function. The enzyme F_1F_0-ATPase, also known as ATP synthase, found in virtually every organism, is responsible for most of the ATP synthesis. The nomenclature refers to the two portions of the protein, F_0 and F_1, both of which are rotary motors. Both motors are attached to a common shaft, each attempting to rotate in the opposite direction. The F_1 motor uses the free energy of ATP hydrolysis

$$ATP^{4-} + H_2O \rightarrow ADP^{3-} + H_2PO_4^-$$

to rotate in one direction, while the F_0 motor uses the energy stored in a transmembrane ion or proton gradient to turn in the opposite direction. Which rotation predominates depends on cellular conditions. The normal situation is that F_0 rotates and thereby drives the F_1 motor in reverse whereupon it synthesizes ATP from its constituents, ADP and phosphate. When F_1 dominates, it hydrolyzes ATP and drives the F_0 motor in reverse, turning it into an ion pump that drives ions across the membrane against the electrochemical gradient [204, 205]. A fundamental question is how the various motor proteins harness the free energy derived from this simple chemical reaction to drive the motor in the appropriate direction or, conversely, how F_0F_1-ATP synthase is able to reverse hydrolysis to synthesize ATP. The coupling between ATP hydrolysis and the mechanical steps is the key element in the energy transduction mechanism. This is of interest also because of the essential role of these motors in the regulation of metabolism and other aspects of cellular function.

Structure. The F-ATPase is composed of two protein complexes: the F_0 portion in the cell membrane and the globular non-membrane portion F_1 which catalyzes the synthesis of ATP. Only for F_1-ATPase the structure has been solved at atomic resolution [204–206]. The ion-translocating membrane domain, F_0, is attached via central and peripheral stalks to the catalytic domain, F_1, a spherical globular structure outside the membrane (Figs. 26–28).

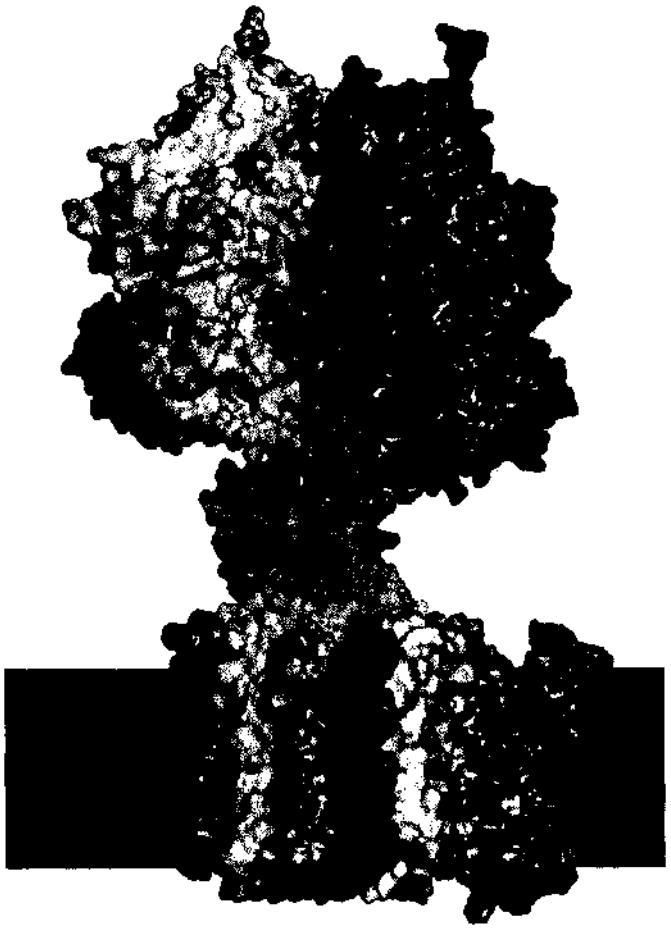


Figure 26. The arrangement of the subunits in the F_1F_0-ATP synthase complex. One α subunit has been removed from the F_1 (upper part) to reveal the γ subunit (orange) within the $\alpha_3\beta_3$ domain. There are three $\alpha\beta$ subunit pairs, shown here in different colors. The $\alpha_3\beta_3$ domain is attached to the a subunit in the F_0 part by a peripheral stalk composed of δ and the two copies of the b subunit. The c subunit ring of F_0 is linked to the γ and ε subunits to form the central rotor.

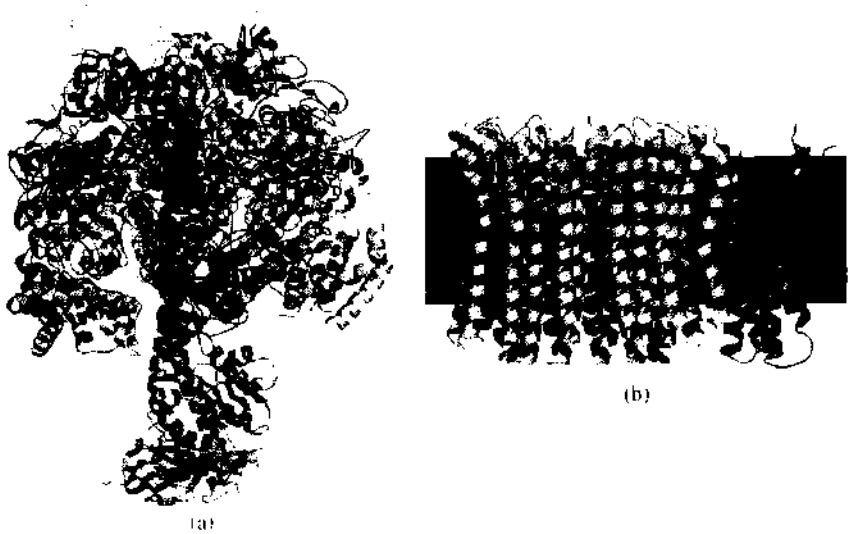


Figure 27. Molecular structures of F_1F_0-ATP synthase. (a) Ribbon structure of F_1-ATPase (PDB code 1E79) [204, 206]. α subunits, red; β subunits, yellow; γ subunit, purple; δ subunit, green; ε subunit, light yellow. (b) Ribbon structure of F_0-ATPase (PDB code 1C17) [205]. F_0-ATPase is composed of a four-helix bundle ($a2$-$a5$, shown in purple color) of subunits a and an oligomer of 10 c-subunits (yellow).

Figure 28. View showing fragment of F_1-ATPase: subunits β (red), the stalk γ (brown), ε (blue), from same structures as in Fig. 27.

Concerning the F_0 portion, it is known that, in mitochondria, it is composed of a ring of 10 identical *c* subunits, whose structure isolated in solution has been determined by NMR, plus several other subunits, at least some of which appear to be involved in the coupling between F_0 and F_1. The F_0 domain also contains one copy of subunit *a*, which contains five transmembrane helices, and two copies of subunit *b*, which contains a single transmembrane helix connected to a polar cytoplasmic loop. Atomic force microscopy shows the *a* and *b* subunits to one side of the *c* subunit ring. The *c* subunit is a highly hydrophobic protein composed of two transmembrane helices connected by a short cytoplasmic loop. In yeast mitochondria, 10 *c* subunits form the ring, and there is evidence for stoichiometries in the range from 10 to 14 in other species. The F_0 and F_1 domains are connected by both a central and a peripheral stalk. The central stalk is composed of the γ and the associated ε and δ subunits, whereas the peripheral stalk contains the *a* subunit and the soluble domains of the *b* subunits. The central stalk and the associated *c* ring are believed to rotate as an ensemble relative to the other parts of the protein. During ATP hydrolysis, the rotation of the central stalk, the γ-subunit, is counterclockwise in 120-degree steps, as viewed from the membrane, whereas for ATP synthesis clockwise. The detailed mechanism, however, is not understood. It has been shown that the average output torque is large (50 pN nm), the variations of the torque are small and that the output free energy of the loaded enzyme is a linear function of the rotational angle. High resolution structures of F_1 reveal a hexameric arrangement of alternating α and β subunits surrounding a central subunit which exists in a highly extended, coiled-coil conformation. The nucleotide binding sites are located at the interface of the subunits. The central subunit is the continuation of the central stalk composed of the γ and ε subunits. The foot of the central stalk makes extensive contacts with a ring of *c* subunits in the membrane domain. The F_1 domain can be separated from F_0, and it still has the ability to hydrolyze ATP.

The α subunits, consisting of α-helical coiled-coils, are asymmetric. The rotation of this asymmetrical structure seems to modulate the binding affinities of the catalytic β subunits for substrate and products. Each of them has three states known as "empty," "loose," and "tight," corresponding to the binding change mechanism of ATP synthesis. An open (or empty) subunit (E state) has a low-catalytic affinity. Corresponding to this mechanism, during the synthesis cycle a rotation of 120-degree converts an empty site, with low affinity, into a loose substrate binding state. A further 120-degree rotation converts the loose state to the tight binding state, and the product ATP is synthesized. Another 120-degree rotation converts the tight state back to the open state, releasing the ATP product.

Presently, there is no detailed structure available for the F_0 sector of the ATP synthase that encompasses the c10 oligomer and the *a*-subunit. The only available crystallographic structure is one for the mitochondrial F_1-c10 complex, at 3.9 Å resolution [205], that does not include subunit *a* critical for ATPase function. Several structural models for components of F_0 (including a model for subunit *a*) have been developed on the basis of disulfide cross-linking data and NMR experiments performed in polar solvents and detergents. It is not clear, however, in how far the protein structure in detergent is similar to that in membranes and to which extent the proposed models are accurate. To model the function of F_0 at atomic resolution, a structural model of F_0 in its native environment needed to be constructed first.

Modeling and Simulations. Although models of how the two rotary motors are coupled have been published [207], the lack of high-resolution structural data on the various F subunits and the F_0F_1 complex has so far prevented a definitive analysis. Recent progress in this area includes an electron microscopy structure at 32 Å of the intact F_0F_1-ATP synthase [208]; atomic models of F_1 and *c*10 were docked into the electron density and used to obtain new information about other portions of the entire enzyme. These results introduce additional constraints that any model must satisfy, but are not at a sufficient level of detail for definitive calculations concerning the mechanism. It has been suggested based on atomic force microscopy, that there is a subunit (perhaps *a*) inside the *c* ring, but the interpretation of the data is not unequivocal. Comparisons of the subunits within a given structure and in different structures have provided important clues concerning the rotational mechanism of F_1-ATPase. However, a direct study of the transition at an atomic level of detail is essential for determining the dynamics of the structural changes and the relative order in which they occur. Therefore simulations of model ATPases are important. However, several assumptions have to be made in the modeling step. For example, it has been assumed in simulations [209, 210] that a clockwise rotation would be required for synthesis.

F_0-ATPase. Schulten and coworkers [211] have performed modeling and molecular dynamics simulations of the F_0 motor. Using structural data of individual components, they have built an all-atom model of F_0 and have investigated the mechanism of torque generation in F_0 (Fig. 27). The mechanism that emerged generates the torque at the interface of the *a*- and *c*-subunits through certain amino acid side chains of the a-subunit and the *c*-subunits. The mechanism couples protonation/deprotonation of two side groups of the *c*-subunits, juxtaposed to the *a*-subunit at any moment in time, to rotations of individual *c*-subunit helices as well as rotation of the entire *c*-subunit. A mathematical model proved the feasibility of torque generation by the stated mechanism against loads typical during ATP synthesis; the essential model characteristics, for example, helix and subunit rotation and associated friction constants, have been tested and furnished by steered molecular dynamics simulations.

A different approach to unveil the mechanochemistry of the F_0 motor has been proposed by Oster and coworkers. This was based on a coarse-grained model and corresponding mathematical fomulations. On the basis of their previous investigations on V- and F-ATPases [149, 212–214] and on ratchet and power stroke models [176, 207], they have developed a model for the F_0 motor of the anaerobic bacterium *Propionigenium modestum* [215]. The structural features of their model are captured in the structural cartoon shown in Fig. 29. They model the rotor as a cylinder with 11 half channels equally distributed on the periphery. Al 11 of the *c* channels are open to the cytoplasm when they are outside the *a*–*c* interface. To maintain the integrity of the stator against leakage, a rotor channel that lines up with the stator channel is occluded. The motion of the rotor, as the slowest degree of freedom, is treated explicitly by the rotation angle, θ, and the chemical states, s_i = (ion bound, empty), of each $i = 1, \ldots, 11$ binding site. All other degrees of freedom settle to their equilibrium values so rapidly that they may be safely assigned their equilibrium values. Therefore the notion of the motor is governed by a set of state-dependent potentials of mean force. Hence the equation of motion for this rotor model can be described by a set of coupled Fokker-Planck equations that describe the evolution of the probability density, $\rho_s(\theta)$, of finding a rotor channel at position θ and chemical state s.

$$\frac{\partial \rho_s(t)}{\partial t} = -\frac{D}{k_B T}\frac{\partial}{\partial \theta}\left[\frac{\partial}{\partial \theta}(V_s(\theta) - \tau_l \times \theta) \times \rho_s(\theta) + k_B T \frac{\partial \rho_s(\theta)}{\partial \theta}\right] + \Sigma K_{s,}(\theta)\rho_s(\theta) \qquad (1)$$

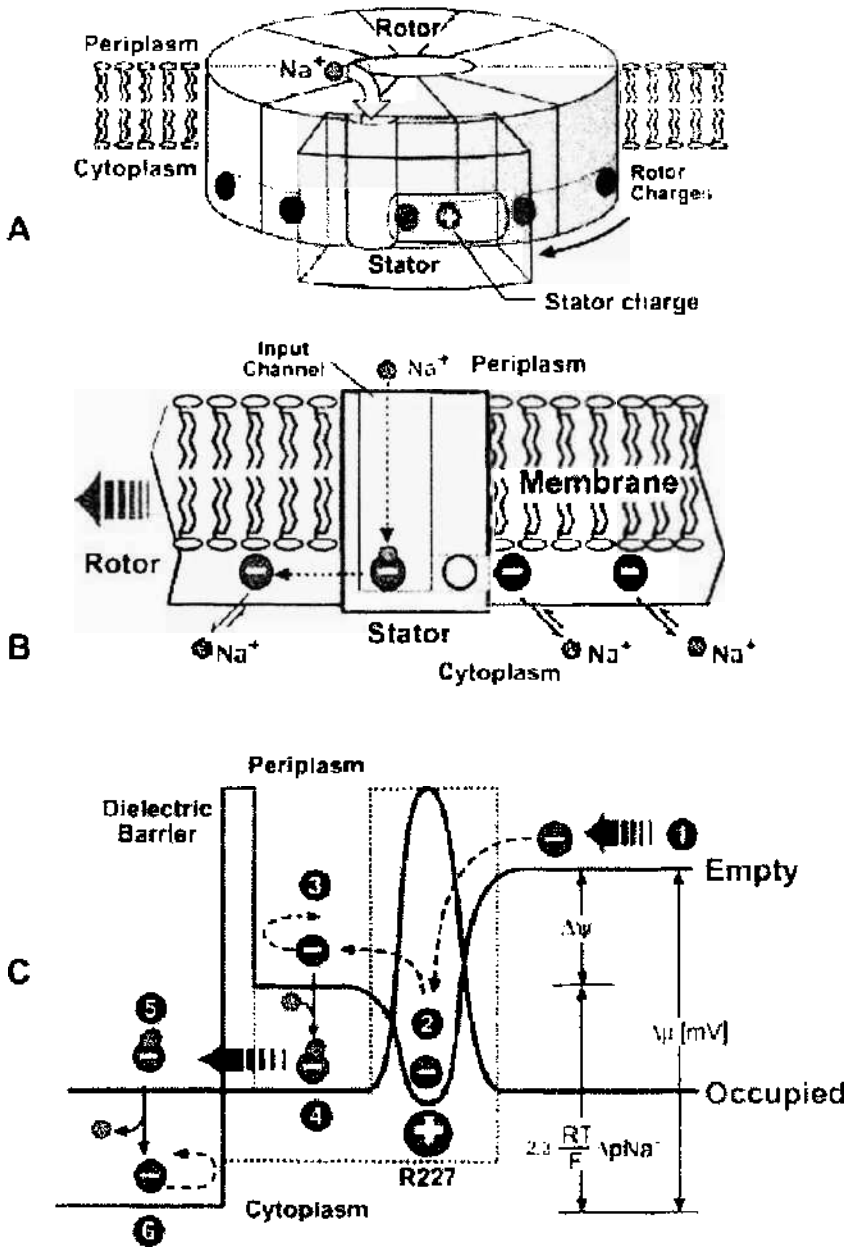


Figure 29. (A) Simplified geometry of the sodium-driven Fo motor. (B) Face-on view of the rotor-stator assembly showing the path of ions through the stator. (C) Free-energy diagram of one rotor site as it passes through the rotor-stator interface. An empty site sees the red potential, while an occupied site sees the blue potential. Binding of an ion switches between the two potentials. Steps 1, 2: the rotor diffuses to the left, bringing the empty (negatively charged) site into the attractive field of the positive stator charge. Steps 2, 3: once the site is captured, the membrane potential biases the thermal escape of the site to the left (by tilting the potential and lowering the left edge). Steps 3, 4: the site quickly picks up an ion from the input channel; this switches the site to the blue potential. Steps 4, 5: an occupied site, being nearly electrically neutral, can pass through the dielectric barrier. If the occupied site diffuses to the right, the ion quickly dissociates back into the input channel as it approaches the stator charge. Steps 5, 6: upon exiting the stator, the site quickly loses its sodium ion. The empty (charged) site once again sees the stator dielectric barrier, which prevents the rotor from diffusing backward. The cycle decreases the free energy of the system by an amount equal to the electromotive force. Reprinted with permission from [216], C. Bustamante et al., *Acc. Chem. Res.* 34, 412 (2001), Fig. 4. © 2001, American Chemical Society.

where N is the total number of chemical states: 16 for normal operation or 81 for ion exchange. D is the relative diffusion constant between the stator and the rotor, $\tau_L(\theta)$ is the load torque from F_1, and $K(\theta)$ is the Markov transition matrix between different chemical states. The forces between the rotor and stator are expressed in terms of the potentials of mean force, $V_s(\theta)$, when the system is at position θ in the chemical state, s. The model reveals the mechanochemical principles underlying the F_0 motor's operation, and

explains existing experimental data on wild-type and mutant F_0 motors. In particular, the model predicts a nonmonotonic dependence of the ATP hydrolysis activity on the sodium concentration, a prediction confirmed by experiments. To explain experimental observations, the positively charged stator residue (R227) must assume different positions in the ATP synthesis and hydrolysis directions. This model calculations demonstrates how a protein can extract energy from an electrochemical gradient and convert it into a mechanical force. It also illustrates how to extract a motor mechanism from dynamical experimental observations in the absence of complete structural information.

F_1-ATPase. The F_1 motor has provided the deepest insights into the mechanochemistry of protein motors. Scientists have compiled the most exhaustive catalog of its properties and behavior. On the basis of atomic structures in several energetic states, it was possible to make predictions of the global conformational transitions accompanying the catalytic cycle. Biochemical studies provided the energy transactions and kinetic rates. Mutation studies isolated the key amino acids. The final necessary information was single-molecule mechanical measurements that measured the forces the motor could generate under different loading and fuel conditions. For example, the rotation of this rotary motor has been visualized and proved to be rotatory using attached actin filaments [217]. Finally, in a synopsis, a comprehensive model could be constructed with all information to create a reasonably complete picture of the operation of this motor protein. It is conceivable that the same procedure must be followed for a similar level of understanding of other molecular motors. The understanding of F_1-ATPase, the smallest biomolecular rotary motor, has made particular progress by the interplay of experimental and theoretical studies; the latter have provided information not available from experiment. In particular simulations made by various groups [209, 210, 218–221] and theoretical studies [212–214, 222] have provided much detailed and global insights. Corresponding extensive reviews exist [207, 223].

One of the first molecular dynamics studies on the rotation mechanism and the concomitant conformational changes in F_1-ATPase had been made by Ma and coworkers [210]. A dynamic analysis of the rotation mechanism for conformational change showed how rotation of the γ subunit induces the observed structural changes in the catalytic β subunits. Both steric and electrostatic interactions contribute. An "ionic track" of Arg and Lys residues on the protruding portion of the γ subunit plays a role in guiding the motions of the β subunits. The trajectory from simulations provides a unified dynamic description of the coupled subunit motions involved in the 120-degree rotation cycles of F_1-ATPase. The simulations reported there were based primarily on targeted molecular dynamics (TMD) method [19], and the biased molecular dynamics (BMD) method [20]. The structural model they used for the simulation was obtained by combining the original X-ray structure [204] and a structure that has resolved most of γ, δ, ε subunits. The simulations correspond to a single 120-degree rotation step in the direction of ATP synthesis, with the γ subunit rotating in a clockwise direction. The α and β subunits do not rotate but undergo conformational changes based on the endstates inferred from the X-ray structure. The TMD simulation applied constraints to the entire system to find a low energy pathway connecting the two endstates. By contrast, in the BMD simulation, a biasing force was applied only to the γ subunit to determine directly the response of the α and β subunits to the rotation of the γ subunit. However, it did not yield the full transition path obtained with the TMD method over the nanosecond timescale of the trajectory. The results demonstrate how the rotation of the γ subunit induces the major changes observed in the structure of the β subunits. The lower part of the γ subunit, including the $\alpha\beta$ domain, which protrudes from the $\alpha_3\beta_3$ complex was shown to play an essential role in the dynamics of the transition. Some α subunit motions occured during the transition, although all three were very similar in the X-ray structure. Certain side chain interactions, such as those involving the ionic track of the γ subunit, were shown to be important in guiding the transition. A more detailed description of the motion in the hinge region than that inferred from the X-ray structure was obtained. The molecular dynamics trajectory reported provided a unified description of the subunit motions of F_1-ATPase in the 120-degree rotation cycle that forms the basis for its function. From a technical point of view several approximations were made to decrease the required computer time by more than two

orders of magnitude. The number of water molecules in the system, relative to full solvation, was decreased. Among others, screening of the charged groups and a distance-dependent dielectric constant were employed. In a subsequent work by Karplus and coworkers, free energy simulations were used [220] to show how to relate the different conformations of the catalytic β-subunits found in the X-ray structure of F_1-ATPase to the binding affinities for ATP measured in solution. For example, it was found that the β_{TP} site is the tight binding site for ATP. Using the results from [220], a free energy diagram was then constructed [221] showing how the ATP and ADP, H_2PO_4 affinities of each of the β-subunit conformations vary with the γ-subunit rotation angle. It was used to develop a structure-based model for the kinetics of F_1-ATPase catalysis in agreement with many experiments.

Recently, MD simulations of the primary mechanical energy transfer steps in F_1-ATP synthase have been reported by Grubmüller and Böckmann [209]. In this paper, steered MD simulations were used to study conformational rearrangements as the central stalk region is forced to undergo rotation. In contrast to previous interpolations between the three known intermediates, the simulations provided a causal picture of a sequence of structural motions, with one triggering the next. Already within the simulation time span of 7 ns, the structural changes proceeded sufficiently far enough to transfer a significant amount of energy to the β_{TP}-binding site, as estimated from its decreased ATP-binding affinity. This decrease was affected gradually by sequential and well-concerted retraction of three charged Arg residues from the bound ATP. Two main conformational energy transfer routes have been identified that couple γ-rotation to this retraction. Unexpectedly, a fast spontaneous closure of the empty β-subunit was seen during γ-rotation. Some of the expected structural changes were not seen in the simulations and, therefore, probably occured at slower timescales than those accessible to simulations. However, despite the fact that rotation was made to occur at a rate several orders of magnitude faster than it happens in reality, a number of features of the simulations appeared realistic. More importantly, the authors suggested a plausible explanation for the release of newly synthesized ATP molecules. More recently, the same authors have studied the conformational dynamics of the F_1-ATPase β-subunit using MD simulations [218]. They had addressed the question whether there exists a spontaneous and nucleotide-independent closure of the open β-subunit upon rotation of the γ-subunit. Therefore they have performed multinanosecond molecular dynamics simulations of the isolated β-subunit with different start conformations and different nucleotide occupancies. Consistent with the available X-ray structures they have observed a fast, spontaneous closure motion of the open β_E-subunit. The observed motions provide a plausible explanation why ADP and P are required for the release of bound ATP and why γ-depleted $\alpha\beta_3$ has a drastically reduced hydrolysis rate.

Schulten et al. [219] have reported an *ab initio* QM/MM study on the mechanism of ATP hydrolysis in the β_{TP} catalytic site of F_1. Their simulations provided an atomic level description of the reaction path, its energetics, and the interaction of the nucleotide with the protein environment during catalysis. The simulations suggested that the reaction path with the lowest potential energy barrier proceeds via nucleophilic attack on the phosphate involving two water molecules. Furthermore, the ATP hydrolysis reaction in β_{TP} was found to be endothermic, demonstrating that the catalytic site is able to support the synthesis of ATP and does not promote ATP hydrolysis in the particular conformation studied.

Using molecular dynamics, a group, lead by Chandler and Oster [222], studied the unbinding of ATP in F_1-ATPase from its tight binding state to its weak binding state. The calculations were made feasible through use of interpolated atomic structures from Wang and Oster [213]. These structures were applied to atoms distant from the catalytic site. The forces from these distant atoms gradually drive a large primary region through a series of sixteen equilibrated steps that trace the hinge bending conformational change in the β-subunit, that drives rotation of γ-subunit. As the rotation progresses, they found a sequential weakening and breaking of the hydrogen bonds between the ATP molecule and the α- and β-subunits of the ATPase. This finding agrees with the "binding-zipper" model of Oster and Wang [214]. In this model, the progressive formation of the hydrogen bonds is the energy source driving the rotation of the shaft during hydrolysis. Conversely, the corresponding sequential breaking of these bonds is driven by rotation of the shaft during ATP synthesis. Their results for

the energetics during rotation suggest that the nucleotide's coordination with Mg^{2+} during binding and release is necessary to account for the observed high efficiency of the motor.

Miki et al. [224] examined the dependence of the physical quantities of the rotatory molecular motor, such as the rotation velocity and the proton translocation rate, on the chemical reaction rate using the model based only on diffusion process. A peculiar behavior of proton translocation was found and the energy transduction efficiency of the motor protein was shown to be enhanced by this behavior. They give a natural explanation that this behavior is universal when certain inequalities between chemical reaction rates hold, which may give a clue to examine whether the motion of the molecular motor is dominated by diffusion process or not.

Since most of the molecular dynamics simulations cannot capture the longtime dynamics of full rotation and concomitant conformational transitions in F_1F_0-ATPase, but on the other hand coarse-grained models as proposed by Oster and Wang [207], neglecting molecular details, but capturing the full cycle of synthesis or hydrolysis, one may speculate whether simulations of, still to be developed, coarse-grained model could conducted using some sort of rigid-body simulation dynamics, similarly as proposed recently by Schlick and Beard [225]. They had introduced an unbiased protocol for performing rotational moves in rigid-body dynamics simulations. This approach-based on the analytic solution for the rotational equations of motion for an orthogonal coordinate system at constant angular velocity removes deficiencies that have been largely ignored in Brownian dynamics simulations, namely errors for finite rotations that result from applying the noncommuting rotational matrices in an arbitrary order. Their algorithm should thus replace standard approaches to rotate local coordinate frames in Langevin and Brownian dynamics simulations.

5.2. Bacterial Flagellar Motor

The bacterial flagellar motor (BFM) can generate torques of about 4000 pN nm near stall. The radius of the rotor is 20 nm, so the motor generates at its rim a force of 200 pN, far greater than any other molecular motor. Similar to the F_0-ATPase, the BFM receives its energy from a transmembrane ion-gradient induced force, being driven by either protons or sodium ions. For each revolution of the motor, approximately 1200 protons pass through it, each contributing about 6 k_BT. The proton-driven motor rotates at 300 Hz, whereas the sodium-driven motor rotates at up to 1700 Hz. Despite intensive investigation, the energy transduction mechanism of the BFM remains elusive; for current review, see [226] and for recent structural experiments [227, 228].

The Motor. Like the F_0 motor, the BFM has a transmembrane rotor and approximately eight peripheral stators (Fig. 30). However, the operating principle might not be the same as F_0 because scaling up the F_0 motor does not produce enough torque to account for the BFM's amazing performance. There have been many proposals and several sophisticated quantitative models [212, 229–231] but none has managed to capture all of the important experimental properties. The only models that have been investigated in quantitative detail use the proton turbine mechanism [229, 230, 232]. This is an electrostatic mechanism based on the idea that there are inclined helical arrays of proton-binding sites on the rotor that interact with vertical rows of stator charges along which protons hop from periplasm to cytoplasm. Although these models can reproduce many of the measurements, recent work on the structure and charge distributions on the stator and rotor are hard to reconcile with this mechanism (Fig. 30, inset). Although the rotor-stator interaction appears to be electrostatic, current thinking centers around some sort of "power-stroke" mechanism with 50 steps per revolution [228]. Just how this works awaits more detailed structures of the rotor and stator elements.

A different model of the flagellar motor was proposed by Schmitt [231] that is based on the established dynamics of the KcsA potassium ion channel and on known genetic, biochemical, and biophysical facts, which accounts for the mechanics of torque generation, force transmission, and reversals of motor rotation. It predicts that proton (or in some species sodium ion) flow generates short, reversible helix rotations of the MotA-MotB channel complex (the stator) that are transmitted by Coulomb forces to the FliG segments

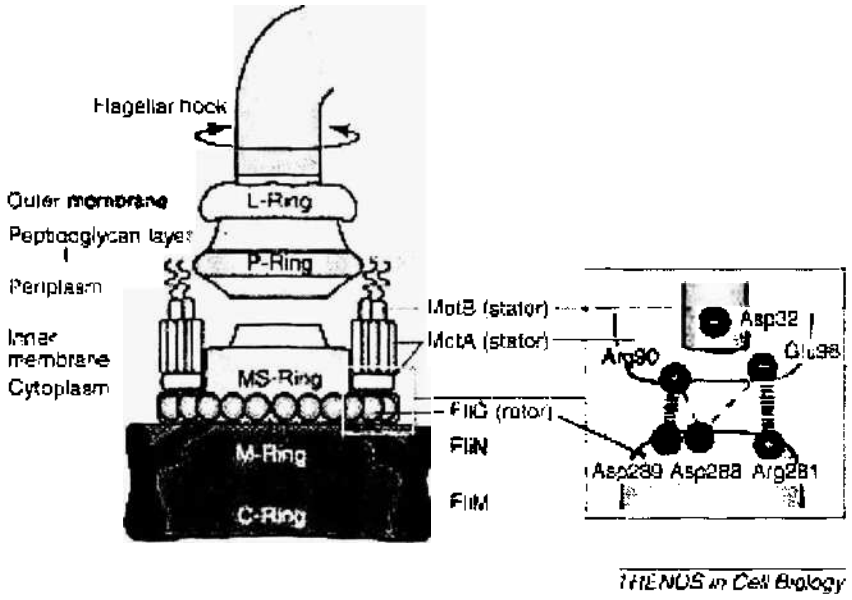


Figure 30. Structure of the bacterial flagellar motor. Ions (protons or sodium) flow through eight stators, each of which consists of four MotA and two MotB subunits. Each MotA comprises four transmembrane α-helices with a large cytoplasmic domain, whereas each MotB is a single transmembrane helix that is anchored to the cell body via the peptidoglycan layer. The torque-generating surfaces are the cytoplasmic domains of the MotA stator units and the FliG proteins on the M ring. Mutation studies have determined the conserved amino acids that participate in torque generation, which are shown in the inset. Of these, only the negative charge on MotB (Asp32) is essential. Reprinted with permission from [207], G. Oster and H. Wang, *Trends Cell Biol.* 13, 114 (2003), Fig. 3, © 2003, Elsevier.

at the rotor surface. Channels are arranged as symmetric pairs, S and T, that swing back and forth in synchrony. S and T alternate in attaching to the rotor, so that force transmission proceeds in steps. The sense of motor rotation can be readily reversed by conformationally switching the position of charged groups on the rotor so that they interact with the stator during the reverse rather than forward strokes. An elastic device accounts for the observed smoothness of rotation and a prolonged attachment of the torque generators to the rotor, that is a high-duty ratio of each torque-generating unit.

A mechanism coupling the transmembrane flow of protons to the rotation of the bacterial flagellum has been studied by Walz and Caplan [230]. The coupling is accomplished by means of an array of tilted rows of positive and negative charges around the circumference of the rotor, which interacts with a linear array of proton binding sites in channels. The authors present a rigorous treatment of the electrostatic interactions using minimal assumptions. Interactions with the transition states are included, as well as proton–proton interactions in and between channels. In assigning values to the parameters of the model, experimentally determined structural characteristics of the motor have been used. According to the model, switching and pausing occur as a consequence of modest conformational changes in the rotor. In contrast to similar approaches developed earlier, this model closely reproduces a large number of experimental findings from different laboratories, including the nonlinear behavior of the torque-frequency relation in *Escherichia coli*, the stoichiometry of the system in *Streptococcus*, and the pH-dependence of swimming speed in *Bacillus subtilis*.

The Switch. A ring of proteins that probably acts as a switch is found in the rotary motors by which bacteria drive their flagella. Each of the six or so motors on the surface of an *E. coli* cell is a roughly cylindrical structure about 50 nm in diameter (Fig. 30). The motors rotate at more than 100 revolutions per second and can abruptly reverse their direction of rotation on the scale of milliseconds. Each motor is built from about 40 different kinds of protein, which form rings around the cylindrical structure. One of these, designated C for cytoplasmic, contains about 34 copies of the protein FliM and is thought to control switching of the motor between clockwise and counterclockwise rotation. Switching is controlled by the

cytosolic concentration of the small diffusible protein CheYp, which binds in a 1 : 1 fashion to FliM. A cylindrical rod forms the central part passing through two hollow rings of protein embedded in the outer membrane. The hollow rings probably work like bearings and are not essential for force production, because they are absent in some species of bacteria. The lower part of the central rod is associated with a circular complex of proteins, the C ring. The outer face of the C-ring is believed to interact with a studlike ring of proteins called MotA and MotB, embedded in the plasma membrane, to contribute directly to torque production. The inner face is built from about 34 copies of the protein FliM, each of which can interact with the cytoplasmic molecule CheYp. High concentrations of CheYp cause the motor to switch to a clockwise direction. In a bacterium without stimulation, the cytosolic concentration of CheYp is such that the flagellar motors show stochastic reversals of rotation every second or so. Even a small change in CheYp concentration is sufficient to disrupt this pattern and produce a marked shift toward either clockwise or counterclockwise rotation. The molecular mechanism of the switch is unknown and the subject of some speculation [229]. However, it is widely believed that binding of CheYp molecules to individual FliM molecules of the C ring causes a change in their conformation, and that this in some way provokes the change in direction of rotation. The motor switching is interpreted by models in which binding of CheYp to the motor shifts the difference in free energy between the cw and ccw states (see also Ref. [226]).

Recently [233], Monte Carlo studies have been performed to investigate the ring of FliM molecules in the BFM. The model assumption was made that this ring operates in isolation from the other part of the motor structure to control motor switching. Using realistic values of concentrations, binding constants and Gibbs free energies to the model, and the response to changes in CheYp molecules was examined. It was found that the simulated ring underwent concerted changes in response to a precisely defined range of concentrations of the cytoplasmic signaling molecule, CheYp, and with a sensitivity coefficient comparable to that determined experimentally. An encouraging feature of this simulation was that the ring switched its state (equivalent to the motor changing its direction of rotation) stochastically at intervals of the order of 1 per second, even though the characteristic flipping time assumed for individual protomers was very much faster (104/s). The conclusion was that similar ability to control the rapid motions of individual proteins into larger transitions on a physiologically relevant timescale might be found widely in living cells.

6. MOTOR PROTEINS

Many essential functions of biological cells are performed by nanoscale motor proteins. The high degree of spatial/temporal organization of molecules and organelles within cells is made possible by these molecular machines that transport components to various destinations within the cytoplasm. Their complexity has given rise to intense investigations to understand their mechanisms. Major progress has been made in the past years by a combination of biophysical experimental techniques and theoretical analysis. Studies of single proteins have played a pivotal role for a variety of motor proteins as listed in the Introduction (Table 1). Motor proteins use the cell's chemical energy repositories of phosphate bonds in nucleotides, generally ATP or GTP. These biomolecular motors make use of the ATP hydrolysis as the free energy source. The motor proteins have a wide range of functions, including organelle transport, muscle contraction, and translocation along DNA/RNA. They play an important role in cell division and cellular motion. A recent example illustrating the complexity of the function of these motors is the experimental demonstration [234] that, when purified, kinesin proteins are attached to unilamellar vesicles and the vesicles are brought into contact with immobilized microtubules, membrane tubes and tubular networks are formed. The study also suggests that multiple kinesin molecules are required to pull the tubes. The malfunction of certain motor proteins can be a source of disease, such as structural myosin mutations in the head and neck region of the molecule. An understanding of the mechanisms of these motor proteins may provide not only a guide for therapy, but may also for developing prototypes of hybrid dynamic nanodevices that operate on micro- or nanofabricated structures with outstanding efficiency and diversity as biological motor proteins.

6.1. Track Motors

Molecular motors such as kinesin superfamily proteins, dynein superfamily proteins, and myosin superfamily proteins have diverse and fundamental roles in many cellular processes, including neuronal development and the pathogenesis of neuronal diseases. The most widely used mechanism for intracellular transport involves molecular motor proteins that carry cargo directionally along a cytoskeletal track:

- myosins along actin filaments, anterograde
- kinesins along microtubules, anterograde
- dyneins along microtubules, retrograde

Therefore, they are also called "linear molecular motors," or "track motors." Track motors are ubiquitous nanomachines that are essential to many bioprocesses at organism level (e.g., muscular contraction that occurs due to the sliding of actin filaments towards the centre of bipolar myosin filaments); cellular level (e.g., in both muscular and nonmuscular cells, the actin-myosin system is responsible for various cellular movement and shape-changing functions); and biomolecular level (e.g., spindle mitosis). Apart from being very important biologically speaking, track motors are complex nanomachines with remarkable efficiency, which makes them very relevant to physics studies. A recent book by Howard [235] reviewed in detail many aspects of the biophysics of linear molecular motors.

Structure. The structures of kinesins, dyneins, and myosins have so-called motor domains that move along the tracks, microtubules, or actin filaments in a specific direction by using the energy derived from ATP hydrolysis. These motor domains are characteristic of and conserved in each superfamily (Fig. 31) [236], but regions outside the motor domains can be divergent and unique to individual motors, enabling various cargoes to be bound. For example, although the interactions between kinesins and cargoes have not been fully elucidated, recent data indicate that many kinesins bind to cargoes via adaptor or scaffolding proteins.

Myosins (class V) were first identified biochemically in vertebrate brain as a myosin-like, calmodulin binding protein and later shown to have motor activity. The principal structural/sequence feature that characterizes myosin Vs is a long lever arm helix that is stabilized by binding one essential light chain and five calmodulins (Fig. 31). Myosin Vs have a conserved $\approx$100 residue C-terminal domain.

Members of the kinesin superfamily share homology with about 340 amino acid motor domain of conventional kinesin. 45 distinct members of this superfamily, known as KIFs, have been identified in the human genome. Although many of the KIFs are microtubule motor proteins that move cargo towards microtubule plus ends (anterograde transport), other KIFs have been shown to move toward microtubule minus ends. Furthermore, some kinesin superfamily members have been shown to destabilize microtubules, affecting cytoskeletal dynamics rather than transport. Conventional kinesin contains an N-terminal motor domain, a long coiled-coil stalk interrupted by a central hinge, and a globular tail domain (Fig. 31). Conserved sequences in the tail domain may be involved in generating a folded, autoinhibited conformation and/or connecting kinesin to its cargo. Kinesin I is a heterodimer composed of two "heavy" chains and two "light" chains. The motor domain that binds to and hydrolyzes ATP, as well as binding in a nucleotide-dependent manner to microtubules, is localized to the N terminus of the kinesin heavy chain. The C terminus of the heavy chain as well as the kinesin light chains mediate binding of the motor to its vesicular cargo.

Dynein was originally identified as a force-generating ATPase, and a cytoplasmic dynein was later discovered to power minus-end-directed motion in nonciliated cells. The dynein heavy chain (DHC) contains a large motor domain that is composed of 6 ATPase-like domains and possibly a seventh domain arranged in a ring (Fig. 31). The first four domains are thought to bind nucleotide (the first being the main ATP hydrolytic site), and the last two domains do not bind nucleotide and may serve a structural role. A coiled-coil extends from the motor domain, and a small globular domain at its tip mediates attachment to microtubules. The nonmotor region of the dynein heavy chaincontains coiled-coil sequences for dimerization as well as binding sites for dynein light chains.

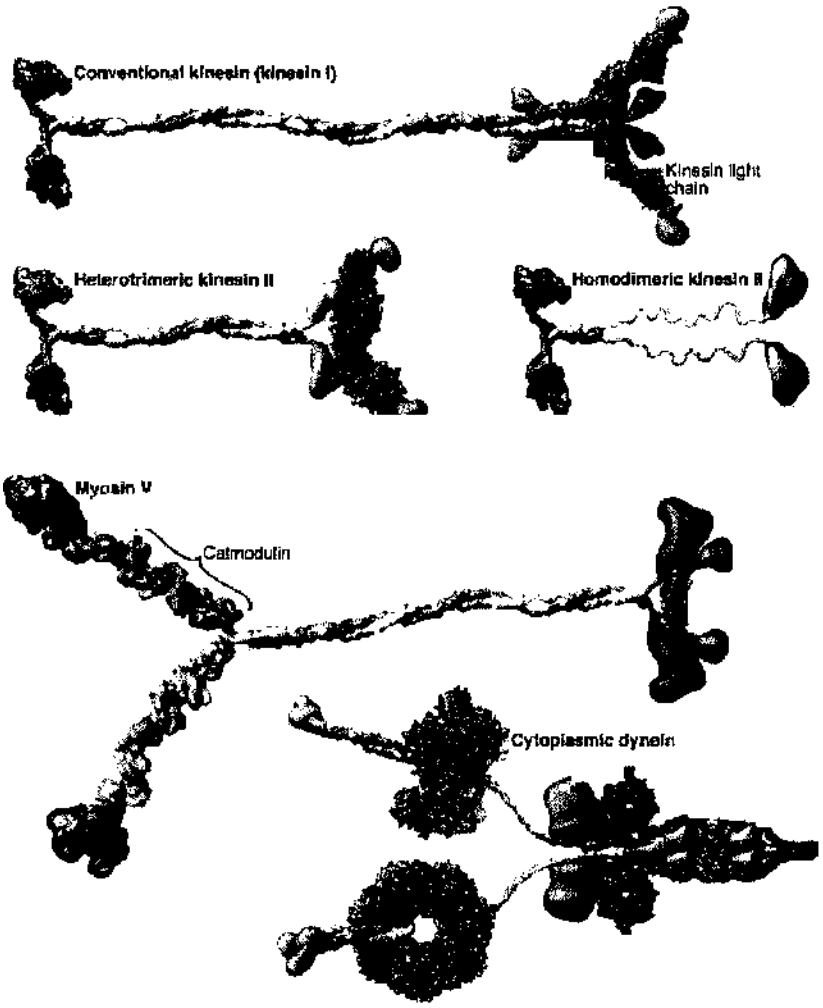


Figure 31. Conformations of various motor proteins based on atomic resolution structures when available. Smooth images for domains of unknown structure. The motor catalytic domains are displayed in blue, mechanical amplifiers in light blue, and tail domains implicated in cargo attachment are shown in purple. Tightly associated motor subunits (light chains) are shown in green. Adapted with permission from [236], R. D. Vale, *Cell* 112, 467 (2003), Fig. 1. © 2003, Elsevier.

Function. Kinesin and dynein proteins are microtubule-dependent motors that slide along microtubules, long hollow cylinders of 25 nm in diameter built by tubulin polymerization, whereas myosin proteins are actin-dependent motors that slide along thin actin filaments or microfilaments of 6 nm in diameter. Kinesin uses the energy of ATP hydrolysis to move along microtubules. Conventional kinesin, also known as kinesin I, moves unidirectionally toward the plus end of the microtubule. Actin filaments are abundant near the plasma membranes and especially in neuronal growth cones. Most kinesins move to the plus end of the microtubules, whereas dynein proteins move to the minus end (Fig. 32). Most members of the myosin proteins move to the plus end of the actin filaments. Some of the plus-end-directed kinesins such as transport cargoes to the dendrites, but most of them transport cargoes to the axons. These motors use the binding energy of the fuel molecule ATP to the enzyme to produce mechanical force, and they use the hydrolysis reaction to weaken the binding between the enzyme and products so that they can be released and the cycle can repeat.

Molecular motor proteins are continuously shuttling vesicles and organelles along the microtubules and actin filaments that make up the cellular cytoskeleton. Nowhere is this transport more essential than in the neuron. These are highly polarized cells with axonal extensions that can reach up to a meter in length. Although axons can represent 99% of the volume of a cell, protein, and lipid synthesis occur almost exclusively in the cell body, and active transport is, therefore, required to supply the axon with newly synthesized material. The microtubule motor kinesin, and additional kinesin-related motor proteins, drive this

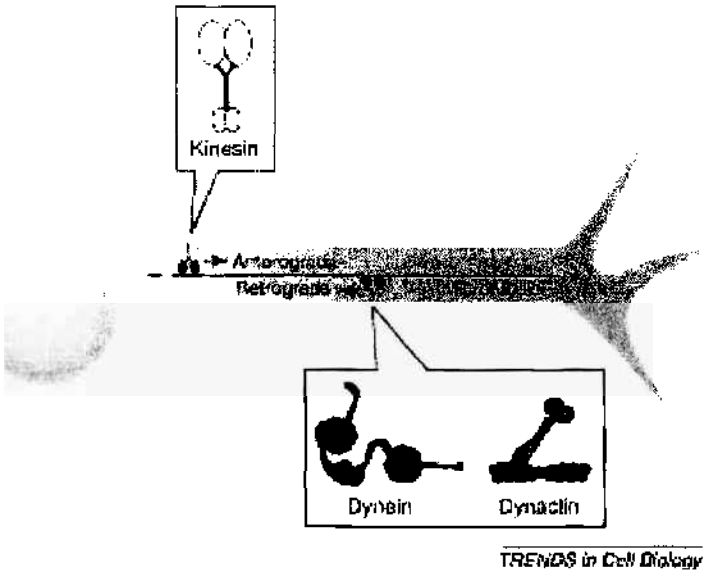


Figure 32. Microtubule motors (kinesin, dynein, dynactin) drive axonal transport in neurons. Reprinted with permission from [237], E. L. F. Holzbauer, *Trends Cell Biol.* 14, 233 (2004), Fig. 1. © 2004, Elsevier.

anterograde transport. In return, material targeted for degradation is actively returned to the cell body by the microtubule motor cytoplasmic dynein. Figure 32 shows how microtubule motors drive axonal transport in neurons. Neurons rely on anterograde transport, powered by kinesin to supply newly synthesized material to the axon and presynapatic terminal. Retrograde axonal transport, driven by cytoplasmic dynein and dynactin, returns material targeted for degradation back to the cell body. Neurotrophic factors and their receptors are also actively transported from synapse to cell body by dynein and dynactin.

General Theoretical Approaches. The pioneering works of Oster et al. [238] and Leibler and Huse [239] have provided one of the first mathematical formulations of phenomenological models of motor proteins. The first paper contains a model for single-motor proteins (myosin, dynein, kinesin) that is powered either by thermal fluctuations or by conformational change. The second paper presents a general phenomenological theory for chemical to mechanical work transduction by motor proteins based on the classical "tight-coupling" mechanism included in a minimal stochastic model. Many theoretical studies are summarized in recent reviews [216, 240], and we refer the reader to these excellent reviews for earlier approaches. In this Section, the more recent developments are summarized.

A theory of molecular motors was presented by Keller and Bustamente [241] that explains how the energy released in single chemical reactions can generate mechanical motion and force. In the simplest case, the fluctuating movements of a motor protein are well described by a diffusion process on a two-dimensional potential energy surface, where one dimension is a chemical reaction coordinate and the other is the spatial displacement of the motor. The coupling between chemistry and motion results from the shape of the surface, and motor velocities and forces result from diffusion currents on this surface. This microscopic description is shown to possess an equivalent kinetic mechanism in which the rate constants depend on externally applied forces. By using this equivalence, the authors explored the characteristic properties of several classes of motor mechanisms and gave general expressions for motor velocity versus load force for any member of each class. They showed that in some cases simple plots of 1/velocity versus 1/concentration can distinguish between classes of motor mechanisms and may be used to determine the step at which movement occurs.

The structural flexibilities of myosin and have been analyzed with normal mode analysis by Li and Cui [155, 242]. The normal mode analysis with physical intermolecular interactions was made possible by an improved implementation of the block normal mode (BNM) approach. The BNM results clearly illustrated that the large-scale conformational transitions implicated in the functional cycles of the motor can be largely captured with a small number of low-frequency normal modes. Therefore, the results support the idea that structural

flexibility is an essential part of the construction principle of molecular motors through evolution. Such a feature is expected to be more prevalent in motor proteins than in simpler systems (e.g., signal transduction proteins) because in the former, large-scale conformational transitions often have to occur before the chemical event of ATP hydrolysis in myosin. This highlights the importance of Brownian motions associated with the protein domains that are involved in the functional transitions. In this sense, Brownian molecular machines is an appropriate description of molecular motors, although the normal mode results do not address the origin of the ratchet effect. The results seem to suggest that it might be more appropriate to describe functional transitions in some molecular motors as *intrinsic elastic motions modulating local structural changes in the active site, which in turn gets stabilized by the subsequent chemical events*, in contrast with the conventional idea of local changes somehow getting amplified into larger-scale motions. In the case of myosin, for example, the authors favor the idea that Brownian motions associated with the flexible converter propagates to the Switch I/II region, where the salt-bridge formation gets stabilized by ATP hydrolysis, in contrast with the traditional concept that *ATP hydrolysis drives the converter motion*. Another useful aspect of the BNM results is that selected low-frequency normal modes have been identified to form a set of collective coordinates that can be used to characterize the progress of a significant fraction of large-scale conformational transitions. Therefore, it is claimed that the BNM analysis had provided a stepping-stone toward more elaborate microscopic simulations for addressing critical issues in free energy conversions in molecular machines, such as the coupling and the causal relationship between collective motions and essential local changes at the catalytic active site where ATP hydrolysis occurs.

6.1.1. Myosin-V

Myosin-V is a motor protein responsible for organelle and vesicle transport in cells. Several mechanisms [243, 244] and theoretical models [245] have been proposed to explain the observed data.

Unlike the single-headed myosins, myosin-V is a kinesin-like processive molecular motor with two heads (Fig. 33). The overall structure of myosin-V can be divided into three functionally distinct regions [246] (Fig. 33):

1. An N-terminal motor domain that contains the actin binding and the active sites, and generates force.
2. A central light-chain binding domain (LCBD) that contains six calmodulin (CaM) light chains and might function as mechanical lever.
3. A C-terminal tail (or stalk) domain that possesses the coiled-coil segments responsible for heavy-chain dimerization and a cargo-binding domain at its extreme C-terminus.

Single-molecule experiments have shown that it is an efficient processive motor where ATP hydrolysis energy fuels the translational movement of myosin-V toward the plus end of actin. Single-molecule mechanical assays have shown that myosin-V moves processively along actin filaments by taking large, ~37 nm steps. The large size of these steps corresponds to the linear repeat of binding sites that appear along the helical actin filament. Electron microscopy along with kinetic analyses of the myosin-V-actin interaction, have led several groups (e.g., Refs. [247, 248]) to propose that myosin-V moves in a hand-over-hand fashion [model (a)], walking along the filament, with the two motor domains taking turns in the lead. The Yanagida group [244, 249] proposed that the motor domain (and not the LCBD) determines how far myosin-V advances per step. Accordingly, they have proposed an alternative model [model (b)] suggesting that the myosin-V motor domain slides along an actin filament down a series of "hotspots," essentially a potential energy slope (the green wedge along filament) induced by the binding of myosin to the filament. A third model (c) is similar to that recently proposed for kinesin [250], suggesting that one head always remains in the lead during stepping, creating a movement reminiscent of a crawling inchworm. Each of these three models shows a single ~37 nm advance of the head-stalk junction (denoted by vertical broken lines), with individual LCBDs color-coded red and blue. The top and bottom rows show the long-lived intermediates before and after the step, respectively, whereas the middle

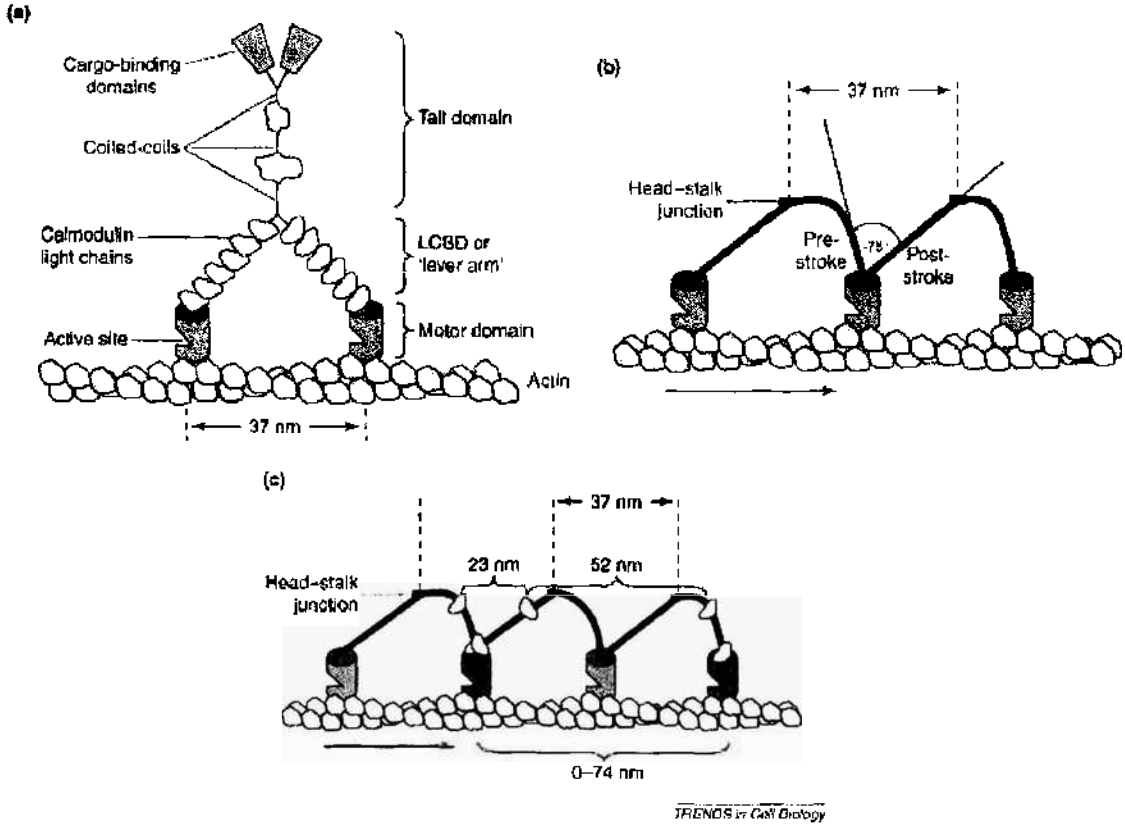


Figure 33. Structure and motility mechanism of myosin-V. (a) The structure. (b) Mechanism by hand-over-hand step. (c) The positions of myosin-V in three consecutive dwells are shown superimposed on a single actin filament. Reprinted with permission from [246], M. J. Tyska and M. S. Mooseker, *Trends Cell Biol.* 13, 447 (2003), Fig. 1. © 2003, Elsevier.

row shows the transition between the two. The brackets underneath each model show the motor domain step size expected in each case.

A theoretical study of myosin-V motility was presented very recently by Kolomeisky and Fisher [245] by following an approach used successfully to analyze the dynamics of conventional kinesin but also taking some account of step-size variations (Fig. 34). Much of the present experimental data for myosin-V can be well described by a two-state chemical kinetic model with three load-dependent rates. The analysis predicted the variation of the

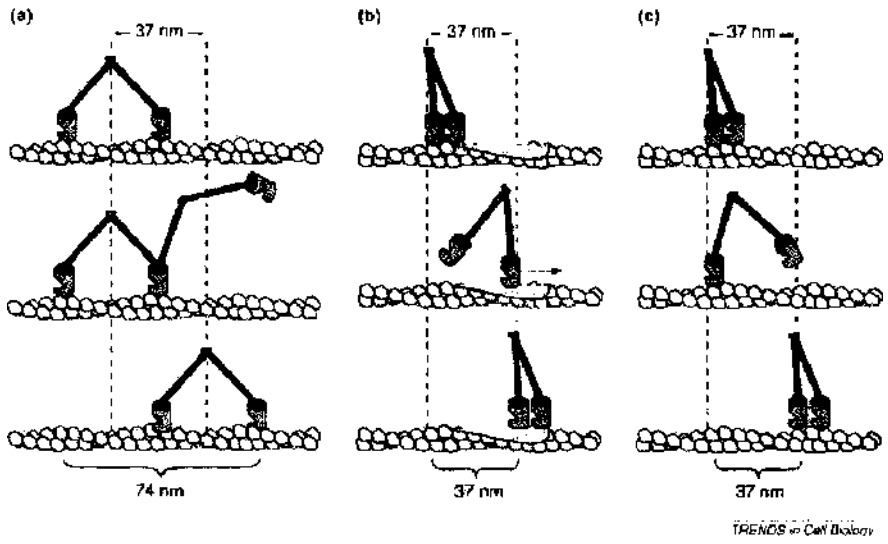


Figure 34. Processivity of myosin-V on actin filament: motility models. Reprinted with permission from [246], M. J. Tyska and M. S. Mooseker, *Trends Cell Biol.* 13, 477 (2003), Fig 2. © 2003, Elsevier.

mean velocity and of the randomness (a quantitative measure of the stochastic deviations from uniform, constant-speed motion) with ATP concentration under both resisting and assisting loads, and indicates a substep of size 13–14 nm (from the ATP-binding state) that appears to be in accord with independent observations.

In a recent simulation work Nicolau and Nicolau [251] developed a computer simulation package for the motility of filaments on the surface functionalised with molecular motors, usually known as motility assays, that is a situation where many myosin molecules are attached to a surface and the surface is covered by motile actin filaments. The simulation of motility assays allows the user to input several fundamental parameters (force exerted by the motors, flexibility of filaments, etc.); collect the predicted statistics and compare the simulated statistics with actual ones, being a valuable tool for a less demanding hypothesis testing of molecular motors functions. The program (source in C and executable) and future updates can be downloaded from www.bionanoeng.com. The myosin molecules are represented by square patches of length l_{myo}. Each of the model myosins has two "arms" that act independently of each other. At any given time, an arm may be "awake" or "asleep." If the arm is "awake," it is free to interact with an actin filament, that is, push it along; after this push, the arm "sleeps" for a time t_{sleep}, during which it cannot interact with nearby filaments. The actin filaments were modeled as chains of "elements"(representing monomers) connected together. The model works with the assumption that the direction of the filaments' motion is controlled by the head alone. This assumption produces simulations that bear a striking resemblance (visually) to in At each time step, all elements of a filament that are "on top of" a myosin "dot" interact with one of the myosin arms (except if both arms are "asleep"). As a result of this interaction, the actin filament experiences a force in the current direction of motion (and thus an acceleration). After pushing the filament along, the arm in question "goes to sleep." If no arm is "awake," no interaction occurs. If none of the filament's elements is "on top of" a model myosin molecule with at least one arm awake, the filament stops. If it is the head element that is interacting, a change in direction will accompany the acceleration. Also, at each time step, all elements experience a retarding force (and hence a deceleration) opposing the motion, proportional in magnitude to the filament's velocity. The combination of the forces due to myosin and the velocity-proportional drag produces motion at a roughly constant speed. The simulations showed that the velocity varies linearly with surface coverage until the latter reaches 100% coverage, after which it is constant. The results also predicted that the average velocity is the result of a tug-of-war between the force applied by the myosin and the drag on the filament, as would be expected.

By combining the mechanical structure of the motor with the ATP hydrolysis kinetics, Lan and Sun [252] have constructed a dynamical model that accounts for the stepwise processivity. They concluded from available data that phosphate release during hydrolysis leads to a powerstroke and a swinging of the lever arm. In contrast with F_1-ATPase where ATP binding is the powerstroke step [207], myosins use ATP binding to release from actin. Since the powerstroke motion is a change in the average conformation of motor domain, and due to thermal motion, there is always some fluctuation in the conformation of myosin-V. The authors postulated that after the phosphate release during hydrolysis, there is a thermodynamic driving force (powerstroke) that rotates the light-chain domain with respect to the motor domain. The exact sequence of events that translate the movements in the binding pocket through the converter region to the lever arm, however, was left to speculation. Their model has four variables: two for the chemical states of the myosin heads 1 and 2, representing the various occupancies of the nucleotide and actin binding sites, and the two for the angles for the geometric angle formed by the projected vector and the tangent vector of the light-chain filament at the joints. Each motor domain can be in any 1 of 10 possible states and the total number of states for the motor complex is 10^2. The dynamics of these variables are given by a system of Langevin equations. The molecular properties of the protein chains connecting the myosin heads are important. A simple elastic model demonstrates that the stress transmitted from the leading head to the trailing head leads to net forward motion. The step-sizes are nonuniform. They also predicted there are several substeps. The translational speed and step-size distributions are computed for several different conditions. The computed force-versus-velocity curve shows that under an

external load, myosin-V slows down. The model made several predictions. It was shown that the unidirectional stepping motion of myosin-V is due to the conformation-dependent ADP release.

6.1.2. Kinesin

Conventional kinesin is a dimer composed of two heavy chains that travels long distances on microtubules by taking 8-nm steps from one tubulin subunit to the next. Each kinesin (Fig. 31) monomer possesses a globular motor domain that has a binding site for the track and a binding site for the fuel, ATP. The monomers are linked to one another though coiled-coil regions in the carboxy-terminal stalk. Binding and unbinding to the track and the fuel, respectively, are delicately interwoven so that a single motor can move for long distances without "falling off" the track, a property referred to as "processivity." To accomplish this, the motor has long been thought to use a mechanism where one head remains attached to the track while the other swings forward, and the roles of the two heads are swapped at each step. Much of the biochemical and structural evidence accumulated over the past decade supports this hand-over-hand model, similar as in myosin-V (see Fig. 34). In fact, it made so much intuitive sense that it has found entry into all major cell biology and biochemistry textbooks. However, as usual, the problem may be in the detail. Unlike humans, kinesin does not have a left foot (head) and a right foot (head). The molecule is composed of two identical subunits that should behave exactly the same. This implies that the lagging head should swing by the leading head always in the same fashion, which in turn would rotate the motor by 180 degree at each step, a "symmetric hand-over-hand" model [250]:

1. In the symmetric hand-over-hand model, the trailing head always passes the leading head on the same side (red arrows in front of the coiled-coil neck).
2. In the asymmetric inchworm model, only the leading head hydrolyses ATP, while the trailing head is pulled up passively. Here, "asymmetric" refers to ATP hydrolysis occurring only in one head.
3. In a symmetric inchworm model, both heads would hydrolyse ATP, and hydrolysis in the trailing head would push the leading head forward.
4. In an asymmetric hand-over-hand model, torsion generated during a step would be accommodated by the flexible hinge domain above the neck during one step and relieved by uncoiling in one of the next steps.

One of the first mathematical formulation for force generation in kinesin was developed by Peskin and Oster [253] in 1995. They had developed a model for force generation in kinesin in which ATP hydrolysis reactions are coordinated with the relative positions of the two heads. The model explains the experimental data and permits to study the relative roles of Brownian motion and elastic deformation in the motor mechanism of kinesin.

A two-dimensional stochastic model for the dynamics of microtubules in gliding-assay experiments was reported by Gibbons et al. [254]. Their model includes the viscous drag acting on the moving fiber and the interaction with the kinesins. For this purpose, they modeled kinesin as a spring, and explicitly used parameter values to characterize the model from experimental data. They numerically computed the mean attachment lifetimes of all motors, the total force exerted on the microtubules at all times, the effects of a distribution in the motor speeds, and also the mean velocity of a microtubule in a gliding assay. They find quantitative agreement with experimental results. They performed additional numerical analysis of the individual motors, and showed how cancellation of the forces exerted by the many motors creates a resultant longitudinal force much smaller than the maximum force that could be exerted by a single motor. They also examined the effects of inhomogeneities in the motor speeds. They presented a simple theoretical model for microtubules dynamics in gliding assays. We show that the model can be analytically solved in the limit of few motors attached to the microtubule and in the opposite limit of high motor density. They found that the speed of the microtubule goes like the mean speed of the motors in good quantitative agreement with the experimental and numerical results.

Vilfan et al. [255] had reported a theoretical and experimental analysis of the interaction between microtubules and kinesin, with special emphasis on the stoichiometry of the interaction, cooperative effects, and their consequences for the interpretation of biochemical and image reconstruction results. Monomeric motor proteins can bind equivalently to microtubules without interference, at a stoichiometry of one motor head per tubulin subunit. By contrast, dimeric motors, as kinsein, can interact with stoichiometries ranging between one and two heads per tubulin subunit, depending on binding constants of the first head and the subsequent binding of the second head, and the concentration of dimers in solution. They showed that an attractive interaction between the bound motor molecules can explain the higher periodicities observed in decorated microtubules (e.g., 16 nm periodicity), and the nonuniform decoration of a population of microtubules and give an estimate of the strength of this interaction. They reported on computer modeling of microtubule decoration by kinesin and simulations performed on a rectangular lattice of binding sites. They calculated the mean dwell time in a certain state and studied the effects of nucleation and annealing of the kinesin lattice on the microtubule surface.

The motility assay of Visscher et al. [256] in which the movement of a bead powered by a single kinesin motor can be measured is a very useful tool in characterizing the force-dependent steps of the mechanochemical cycle of kinesin motors, because in this assay the external force applied to the bead can be controlled (clamped) arbitrarily. However, because the bead is elastically attached to the motor and the response of the clamp is not fast enough to compensate the Brownian motion of the bead, interpretation or analysis of the data obtained from the assay is not trivial. In a recent paper Chen et al. [257] showed how to evaluate the mean velocity of the bead and the motor in the motility assay for a given mechanochemical cycle, and the related fluctuation or the randomness of the velocity using a Monte Carlo simulation method. They found that the randomness of the velocity of the motor is also influenced by the parameters that affect the dynamic behavior of the bead, such as the viscosity of the medium, the size of the bead, the stiffness of the elastic element connecting the bead and the motor. The method presented in this paper should be useful in modeling the kinetic mechanism of any processive motor (such as conventional kinesin and myosin V) based on measured force-clamp motility data.

6.1.3. Dyneins

Dynein was first identified as the motor protein responsible for the motility of eukaryotic cilia and flagella, for example tracheal cilia and sperm-tail flagella. The cytoplasmic form of dynein is similar in structure to the flagellar enzyme and also uses ATP hydrolysis to generate force toward microtubule minus ends. Both the cytoplasmic and flagellar enzymes are large multisubunit complexes: cytoplasmic dynein is composed of two heavy chains, each 500 kDa, as well as intermediate, light intermediate, and light chains. Cytoplasmic dynein is the major motor for retrograde axonal transport in neurons but also is required for a wide range of cellular functions, including endoplasmic-reticulum-to-Golgi vesicular transport, neurofilament transport, mRNA localization, and mitotic spindle assembly.

6.1.4. Dynamin

A recent addition to the class of microtubule-activated nucleotidases is dynamin [258, 259]. Dynamin contains a GTP-binding domain, and exhibits motility characteristics distinct from those of kinesin and dynein. Kinesin and dynein each promote ATP-dependent microtubule motility in a polarized fashion, with dynein mediating force production towards the minus ends of microtubules and kinesin mediating force production towards the plus ends. In contrast, dynamin do not promote gliding of microtubules, but is involved in microtubule-microtubule sliding, a mechanism proposed, for example, for elongation of the mitotic spindle during anaphase.

6.1.5. Problems

The microscopic observations of cargo transport in axons and flagella raise a number of questions. How do the opposite polarity motors, kinesin, and dynein, coordinate their activities?

What kind of machinery processes the incoming cargo and switches motor direction at the "turnaround" zones? Molecular answers to these questions are beginning to emerge but are far from complete. A communication mechanism between motors is suggested by findings in several systems showing that inhibition of either kinesin or dynein function impairs both directions of travel [260]. The molecular basis of this cross-talk between the opposite polarity motors is unknown. However, a recent study revealed that dynactin (traditionally thought to only regulate dynein) is required for both anterograde and retrograde movement. Little is known about the molecules that signal to motors at the "turnaround" zones, although protein kinases and G proteins represent candidates.

6.2. Nucleotidases

The DNA and RNA molecules in living cells work intimately with a myriad of proteins to maintain, replicate, transcribe, and translate the genetic code. The study of the detailed interactions between proteins and nucleic acids responsible for these processes has evolved over the past decade to the point at which the roles of individual atoms in large macromolecular assemblies can now be visualized using the tools of modern structural biology. Recent years have seen considerable progress in simulations of nucleic acids [261, 262] and protein-nucleic acid complexes [263, 264]. Improvements in force fields, simulation techniques [261, 265] and protocols, and increasing computer power have all contributed to making nanosecond-scale simulations of both DNA and RNA commonplace. The results are already helping to explain how nucleic acids respond to their environment and to their base sequence and to reveal the factors underlying recognition processes by probing biologically important nucleic acid-protein interactions and medically important nucleic acid-drug complexation. Molecular dynamics simulations on DNA and RNA that include solvent are now being performed under realistic environmental conditions of water activity and salt [262]. Improvements to force-fields and treatments of long-range interactions have significantly increased the reliability of simulations. New studies of sequence effects, axis bending, solvation, and conformational transitions have appeared.

DNA-Ligand Complexes. Studying ligand binding with MD simulations is apparently proving to be a reliable way of understanding the static and dynamic aspects of these complex conformations in solution, avoiding difficulties that can occur in a crystalline environment, such as modified backbone conformations or hindered curvature. Free-energy calculations also allow reasonable calculations of binding strengths and can expose unexpected cancellations in the factors contributing to binding [266]. Simulations remain an integral part of structure determination by NMR spectroscopy. One can note a considerable increase in the quality of such calculations, which now frequently include explicit solvent, reliable treatment of long-range electrostatics, and time-averaged constraints. Ligand-DNA complexes can still hold a number of surprises, especially as concerns dynamics. Future possibilities are illustrated by an *ab initio* dynamics study using density functional theory-based molecular dynamics [267].

RNA Complexes. Simulations involving RNA are accumulating rapidly, although the size, charge, and structural variety of these molecules continue to pose significant challenges. Several studies have addressed the dynamics of the basic motifs found in folded RNA, including mispairs such as flexible water-bridged UC pairs, or coaxial stacking, where the absence of backbone continuity causes kinking and untwisting but does not enhance flexibility. Studies of stem loops show that nanosecond dynamics maintain stable conformations but cannot sample extensively enough to identify conformational substates [268]. Similar problems occur with bulged bases, where energy barriers prevent sampling of stacked and expelled base conformations during single trajectories. Correctly predicted RNA tetraloop conformations and, in the case of PB calculations, hairpin/duplex equilibria have been reported [269]. Despite sampling problems, it is impressive to see the size of the systems which are now being treated, the most striking being a stable simulation of tRNA with more than 8000 waters on the nanosecond time scale in low salt concentrations [270, 271]. Recognition processes depend on initial electrostatic interactions but also involve the deformation

and stiffening of the macromolecular partners and specific hydrogen bonding with bases. studies have again been studied using MD simulations and free energy analyses to estimate binding free energies and the consequences of site mutations [272]. RNA-drug complexes are also attracting much attention. The readers are referred to a recent review [269].

DNA-Protein Complexes. Both the rapid accumulation of X-ray and NMR data and increasing computer power have contributed to an increasing number of simulations of protein-DNA complexes. Depending on whether the simulation package CHARMM or AMBER is employed, the systems modeled generally solvate the complex within a water sphere and either use shift/switch technology for long-range interactions or use periodic boundary conditions and particle-mesh Ewald electrostatics. Examples of recent studies are Refs. [273–278]. Systems solvated with water spheres typically involve of the order of 20,000 atoms, while this number can rise to 40,000 when periodic boundary conditions are used on large systems [279]. Simulations have reached the nanosecond range only in recent years, and we can assume, as for smaller systems, that longer simulations will reveal new categories of dynamic movement. With this caveat, present simulations nevertheless result in stable protein-DNA interfaces, reproduce most, or all, of the specific interactions observed experimentally, and, often, emphasize the role played by water-mediated hydrogen bonding between the two partners [274]. Computer simulations are also powerful tools for analyzing the components of recognition processes. This can involve looking at structural adaptation upon binding [274], or the consequences of point mutations [273]. Another interesting approach used a simulation of the largely hydrophobic protein-DNA complex to interpret hydroxyl radical cleavage data in terms of both sugar accessibility and fluctuations of the protein side chains contacting the DNA backbone [276]. A challenge which still lies largely in the future is the development of reliable protein-DNA docking procedures [275].

The cooperative binding of proteins to DNA plays a significant role in the regulation of gene expression. In particular, it is well known that transcription factor proteins [5] exhibit a significant level of cooperativity. The structural basis of the cooperativity is not fully understood, but it is known that long-range cooperativity is possible through loops, formed as the result of association between two DNA-binding proteins. Looping is also believed to play an important role in gene access control. Cooperativity at shorter distances may be related to specific protein–protein interactions or to a generic cooperativity resulting from structural distortions induced by the binding of a protein to DNA. For example, the binding of transcription regulation proteins such as the important TATA-box promoters (TPB) involves amino acid intercalation into the stack of basepairs. The result is that kinks are produced in the form of sharp local bending angles in the DNA strand. This deformation may permit a better fit for other DNA-associating proteins, such as the polymerases. Disruptions of the base pair stacking sequence have no effect beyond about half a turn of the double helix, so it is expected that this form of cooperativity is restricted to the immediate neighborhood of the primary binding protein. Protein-induced deformations of the DNA strand are not restricted to transcription factors. DNA may wrap itself once or more around a protein, as happens in the case of complexation of DNA with nucleosomes or bacterial RNA. Recently, a simple mechanism for cooperativity, in the form of a tension-mediated interaction between proteins bound to DNA at two different locations, was proposed by Rudnick and Bruinsma [280]. The proteins were not in direct physical contact. DNA segments intercalating bound proteins were modeled as a worm-like chain, which is free to deform in two dimensions. The tension-controlled protein–protein interaction is the consequence of two effects produced by the protein binding. The first is the introduction of a bend in the host DNA and the second is the modification of the bending modulus of the DNA in the immediate vicinity of the bound protein. The interaction between two bound proteins may be either attractive or repulsive, depending on their relative orientation on the DNA. Applied tension controls both the strength and the range of protein–protein interactions in this model. Properties of the cooperative interaction were discussed, along with experimental implications.

Because simulations of DNA-protein complexes are too large for longtime MD simulations, there have been attempts to provide coarse-grained model studied by Monte Carlo methods. A Monte Carlo model for the structure of chromatin was presented by Wedemann

and Langowski [281]. On the basis of their previous work on superhelical DNA and polynucleosomes, their approach reintegrates aspects of the "solenoid" and the "zig-zag" models. The DNA is modeled as a flexible elastic polymer chain, consisting of segments connected by elastic bending, torsional, and stretching springs. The electrostatic interaction between the DNA segments is described by the Debye-Hückel approximation. Nucleosome core particles are represented by oblate ellipsoids. The interaction potential has been parameterized by a comparison with data from liquid crystals of nucleosome solutions. DNA and chromatosomes are linked either at the surface of the chromatosome or through a rigid nucleosome stem. Equilibrium ensembles of 100-nucleosome chains at physiological ionic strength were generated by a metropolis–Monte Carlo algorithm. For a DNA linked at the nucleosome stem and a nucleosome repeat of 200 bp, the simulated fiber diameter of 32 nm and the mass density of 6.1 nucleosomes per 11 nm fiber length are in excellent agreement with experimental values from the literature. The experimental value of the inclination of DNA and nucleosomes to the fiber axis could also be reproduced. Whereas the linker DNA connects chromatosomes on opposite sides of the fiber, the overall packing of the nucleosomes leads to a helical aspect of the structure. The persistence length of the simulated fibers was 265 nm.

6.2.1. DNA Polymerases

DNA polymerases are responsible for replication of normal and damaged DNA, and faithful DNA replication is crucial for genomic stability. The actual motor mechanism, which is the translocation of DNA polymerase along the DNA, is not well understood, but new high-resolution structures of the incorporation of nucleotides in a polymerase crystal make it likely that more detailed descriptions will become available to complement the phenomenological models. It is also of interest to mention the more complex type III polymerases, which have a high degree of processivity due, at least in part, to the involvement of a ring (clamp) that fits around the DNA. An ATP-dependent motor protein called a clamp loader functions to open the ring and close it around the DNA. Both structural studies and molecular dynamics simulations have given insights concerning the function. A useful review on the complex motor system that loads a polymerase on DNA is Ref. [282]. For certain DNA polymerases, single-molecule experiments have shown that the application of pico-Newton forces to the template DNA can decrease the rate of polymerization and, for forces greater than 35 pN, the enzyme actually depolymerizes the double-stranded DNA. Interpretation of this behavior has been given by phenomenological models [283]. Also, simplified normal mode calculations [284] have been applied to polymerases.

In a recent study Kuriyan et al. [285] have combined X-ray crystallographic studies with MD simulations to gain insight into the process of loading the ring-shaped clamp protein necessary for promoting processive replication by DNA polymerase. The clamp consists of a highly stable dimer of crescent-shaped monomers and placing it around DNA requires that this arrangement is broken open through disruption of one of the monomer-monomer interfaces. Such a process, known as clamp loading. The crystal structure of the essential δ subunit of γ complexed with a single β monomer (Figs. 35 and 36) indicated that the curvature of the crescent-shaped β monomer was significantly decreased from that observed in the β dimer. In fact, this loss of curvature was such that a model of the β dimer constructed from two such monomer structures would be sufficiently open at one end that the passage of at least a single strand of DNA into the inside of the ring would now be possible. This suggested to the authors that the conformation of the monomers observed in the dimeric structure might be intrinsically strained. To test this hypothesis, they conducted two separate MD simulations, one of the β dimer and one of a single β monomer (initially in the same conformation as found in the dimer). Amazingly, the MD simulations confirmed their ideas; although stable in the dimer, the conformation of the isolated monomer spontaneously relaxed to a conformation almost identical to that observed in the $\beta - \delta$ complex. As the authors rightly point out, because such large-scale transitions are rarely observed in MD simulations conducted on a nanosecond timescale, this provides strong evidence for the presence of considerable conformational strain in the monomer. As such, this study clearly demonstrates that fundamental insight into the operation of a large macromolecular "machine" can be obtained from simulations, especially when guided by hypotheses based on experimental results.

Figure 35. DNA polymerase: δ subunit (PDB code 1A5T).

More general theoretical considerations about the mechanism of DNA polymerization are based on the Brownian ratchet concept [216]. In a Brownian ratchet, the role of the chemistry is to select forward fluctuations (or prevent backward fluctuations) of the load, rather than to apply a mechanical force directly to the load. That is, the load force is driven by its own

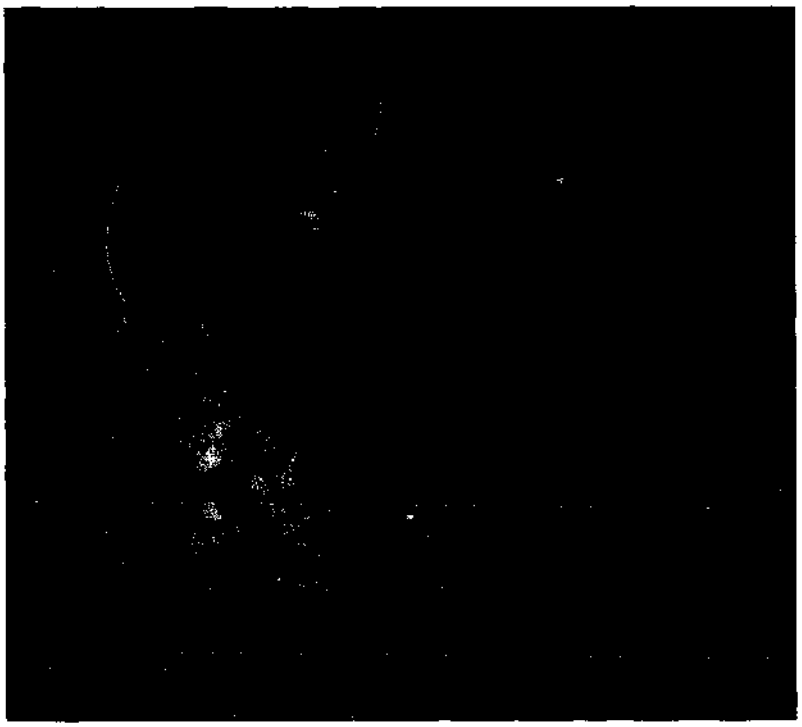


Figure 36. DNA polymerase: β-δ complex (PDB code 2POL).

Brownian fluctuations, and the chemistry provides the energy to rectify the diffusive motion of the load. Though this is a slightly more subtle mechanism than the power stroke, even very simple mechanisms can act as Brownian ratchets. For example, consider the simple model for a DNA polymerase shown in Fig. 37. A DNA polymerase enzyme is bound at the junction between double-stranded DNA and single-stranded DNA. It can slide rapidly between a "closed" state and an "open" state one step forward by opening a space for a new nucleotide (NTP), but no farther. While in the open state, a new nucleotide can bind and be incorporated into the growing strand, thus locking the polymerase one step ahead. Both open and closed states have equal free energy, so, though it can step back and forth under the influence of thermal fluctuations, neither is preferred statistically, and no net mechanical forces push the molecule either forward or backward. Net forward motion occurs because binding of a new nucleotide in the open state prevents the backward step. This physical model can be described by a two-step kinetic scheme, where the single chemical step includes both binding and incorporation of the nucleotide into the chain. In this example, the motor moves forward because the chemical step is irreversible, that is its free energy of reaction is large and negative. This example clearly illustrates that the chemical energy is expended to preferentially select forward steps (or prevent backward steps) and hence to favor forward motion, rather than doing mechanical work on the motor directly. It is also worth noting that a Brownian ratchet can be tightly coupled and efficient in the sense that each fuel molecule burned results in exactly one step forward. A full description of this molecular motor can be given either in a continuum model by the Smoluchowski equation, or in a discrete model by kinetic equations, or in a "mixed model" in which the chemistry is discrete but the mechanics are continuous [216].

6.2.2. RNA Polymerase

RNA polymerase (RNAP) plays a key role in the transcription of the genetic information encoded in DNA by controlling the synthesis of RNA chains. RNA polymerase polymerizes an RNA transcript via a sequence of reactions taking place on its surface. First, the incoming DNA chain is separated into single strands, one of which will serve as the template for the RNA. The two strands reanneal before they exit the posterior end of the enzyme; in-between they form a "transcription bubble," approxiately 15 bp (1 bp = 0.34 nm) long. The entire enzyme is about 30 bp long. The growing RNA chain is synthesized at a catalytic site within the transcription bubble by the addition of nucleotides complementary to the sequence of the template strand. During the elongation of an RNA chain, after the binding of the polymerase to a promoter sequence on the DNA, the polymerase slides along the DNA while maintaining a high level of stability against dissociation. Viewed as a molecular machine, RNA polymerase processes along the DNA by converting the free energy

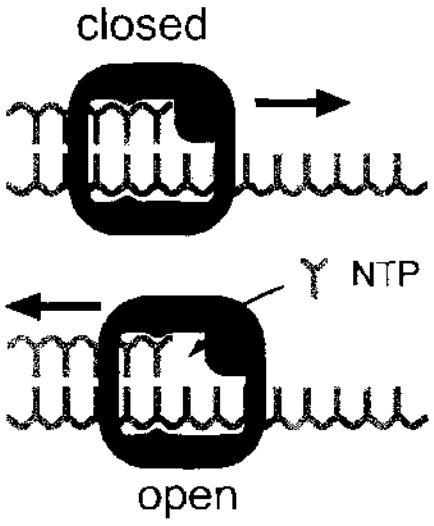


Figure 37. Brownian ratchet model for DNA polymerase. Reprinted with permission from [216], C. Bustamante et al., *Acc. Chem. Res.* 34, 412 (2001). Fig 3. © 2001, American Chemical Society.

of nucleotide binding and hydrolysis into a force directed along the DNA axis. During sliding, the polymerase performs work to overcome various sources of energy dissipation such as viscous drag due to the relative motion of polymerase and DNA (sliding and rotation), and work required to open up the transcription bubble of the double-stranded DNA. The free energy required for this work is provided by the polymerization reaction of the nascent RNA chain when a ribonucleoside triphosphate (NTP) is added to the RNA chain under release of phosphate. This polymerization occurs at a catalytic site C of the enzyme, which furthermore has binding sites for DNA and the nascent RNA strand. Motion of RNAP is influenced by signals in the DNA sequence that is transcribed to RNA. Several types of such signals exist where motion slows down (pause) or stops reversibly for an extended time (arrest), or where the enzyme detaches from the DNA (termination). However, the molecular mechanism by which these chemical bond energies are transduced into the force that drives the RNA polymerase along the DNA strand is not well understood.

Micromechanical experiments permit quantitative studies of forces and motion generated by the enzyme. On the basis of experimental facts Jülicher and Bruinsma [286] have presented a chemical kinetics description of RNA polymerase motion. The model is based on a classical chemical kinetics description of polymerization reactions driven by a free energy gain that depends on forces applied externally at the catalytic site. The RNA polymerase controlled activation barrier of the reaction is assumed to be strongly dependent on inhibitory internal strains of the RNA polymerase molecule. The sequence sensitivity of RNA polymerase is described by a linear coupling between the height of the activation barrier and the local DNA sequence. Their model can simulate optical trap experiments and allows to study the dynamics of chemically halted complexes that are important for footprinting studies. They found that the effective stall force is a sequence-dependent, statistical quantity, whose distribution depends on the observation time. The results were consistent with the experimental observations to date.

In a subsequent paper Oster et al. [287] have shown that RNA polymerase is a processive molecular motor capable of generating forces of 25–30 pN, far in excess of any other known ATPase, where this force derives from the hydrolysis free energy of nucleotides as they are incorporated into the growing RNA chain. The velocity of procession is limited by the rate of pyrophosphate release. They showed how nucleotide triphosphate binding free energy can rectify the diffusion of RNA polymerase, and showed that this is sufficient to account for the quantitative features of the measured load-velocity curve. Predictions were made for the effect of changing pyrophosphate and nucleotide concentrations and for the statistical behavior of the system.

6.2.3. DNA Helicase

Nucleic acid helicases are defined as enzymes that translocate directionally through double stranded nucleic acid substrates to catalyse the separation of the complementary strands. They facilitate various biological processes such as DNA replication, recombination and repair, RNA transcription, editing, and splicing [288]. There are several structural varieties of helicases such as monomeric, dimeric, trimeric (e.g., RecBCD), tetrameric (e.g., RNA polymerase) or closed hexameric, but all use the hydrolysis of ATP to ADP as the preferred source of energy. For example, the bacterium *E. coli* contains at least 17 different helicases, though the need and the function of each of these are not yet clear. The helicase activity involves a motor action fed by NTPs (nucleotide triphosphate) and eventual opening of DNA by dissociating successive base pairs along the chain at certain rates of base pairs per second. Attempts to categorize these varieties of helicases as per their common features have led to various classification schemes. These are (1) active versus passive depending on the direct requirement of NTP for the opening; (2) families and superfamilies (SF) based on the conserved motifs of the primary sequence; (3) monomeric, dimeric, hexameric, oligomeric, depending on the number of units required for activity; and (4) mode of translocation: whether it translocates on the single strands of DNA or on double-stranded DNA. For example, dnaB, the main helicase involved in the replication of DNA in *E. coli*, is a hexameric, passive helicase belonging to the dnaB-like family translocating on a single-strand DNA. PcrA is an active, SF1, monomeric helicase translocating on single-strand DNA,

while recG is an active, SF2, monomeric, double-stranded-DNA translocating helicase. Apart from these gross classifications, very little information is available on the detailed mechanism of the helicase activity. Bulk behavior in solutions such as average unwinding rates, step size, average number of base pairs opened per helicase, are known for a few helicases. The use of single-molecule, micromanipulation tools allowed for monitoring in detail the translocation, unwinding, and processivity (rate of dissociation) of individual RecBCD helicase on DNA. Such experiments elucidated several new aspects of helicase behaviour and showed that many properties could be more related to general principles than on specific chemical details. In particular, it has been observed that (1) RecBCD unwinds DNA at a uniform rate, over a wide range of ATP concentrations, as it moves on one strand, (2) the nuclease activity does not affect unzipping, and (3) the helicase can work in the presence of DNA gaps up to certain lengths. Despite these varieties of experimental findings no clear mechanism coupling the motor action and the helicase activity is known yet. Many biochemical and structural studies of helicases have been reported and our basic understanding of the helicase mechanism based on thermodynamic and kinetic properties have been summarized very recently [288].

Theoretical Concepts. However, our theoretical understanding is not yet complete. Oster and coworkers [289] have developed a "flashing field" model suitable for hexameric ring helicases. Chen et al. [290] desribed the motion a helicase as a biased random walk, and considered how the density of histones affects the random walk.

Bhattacharjee and Seno [291] proposed a "phase coexistence" mechanism of single- and double-stranded DNA in the presence of a helicase. Since the helicase-DNA complex constitutes a fixed-stretch ensemble, it was proposed that this entails a coexistence of domains of zipped and unzipped regions of DNA, separated by a domain wall. The action of the motor protein "helicase" leads to a change in the position of the fixed contraint thereby shifting the domain wall on the double-stranded DNA. The authors suggested that this off-equilibrium domain wall motion is related to the unzipping action of the helicase. This work presents a simple, but powerful argument, based on the principle of phase coexistence, that provides a clear and robust explanation for the gross observed features. The proposed model gives a thermodynamic basis to a model called the "wedge model" [292] according to which the motion of the helicase provides enough force to enable the helicase to destabilize the base pairs at the junction by a process resembling the action of a wedge. In the proposed mechanism, energy (from ATP) is required for translocation activity or the motor action of the helicase and not directly for base pair breaking and therefore, according to the classification scheme, this corresponds to passive helicases. Additional features required for active helicases are ignored in this study. The proposed mechanism is supported by computer simulation of an exactly tractable model. Helicase and the DNA are modeled on the square lattice (fork Y-model). The two strands of DNA are represented by two directed and mutually avoiding walks formed by N bases each (Fig. 38). In the Y-model unzipping can occur only processively, for example, bubbles are suppressed along the chain: the only conformations considered have the first N–m monomers bounded, whereas the remaining m are separated

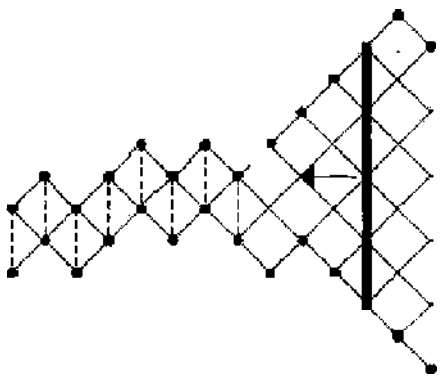


Figure 38. DNA helicase Y-model [291]: thick rod denotes the helicase, nucleotide base denoted by dots, paired bases by dotted lines. The arrow indicates the motion of the helicase.

in a Y-like conformation. The fact that the Y-model does not allow rejoining of the unzipped portion of the double-stranded DNA is similar to the geometry observed in the experiment. A typical configuration of the simulated DNA with the helicase (thick rod) as modeled on a square lattice (thin lines). Bases are represented by dots and paired bases are shown by dotted lines. The position of the fork coincides with the last paired base. The arrow indicates the motion of the helicase. Except for the rigid hard-core constraint, there is no other direct interaction between the DNA and the helicase. Schematic phase diagram in the separation (stretch) versus temperature plane. At fixed temperature, a finite endpoint separation leads to a coexistence of the two phases indicated by the filled circle. In this coarse-grained model helicity is ignored and restricted to two dimensions. It is argued that this is more for simplification of the calculation but would not lead to artificial results. The spirit behind their approach is that the key element that can influence universal behaviour of helicase translocation is the competition between a Y-fork conformation which can be unzipped by paying energy and the movement of an opening molecular motor. In a subsequent paper [293] a propagating-front analysis in terms of a local probability of unzipping was developed. It was shown that the local drive created by the helicase leads to a traveling wave solution with a selected velocity that depends on the nature of the drive.

Recently, Jülicher and Betterton [294] have proposed a physical description of kinetic model for helicase unwinding of nucleic acids (NA), which contains both active and passive opening as different cases in a general framework. In active opening, the helicase directly destabalizes the double-stranded nucleic acid (dsNA) to promote opening. Passive opening implies that the helicase binds single-stranded NA (ssNA) which can be available if a thermal fluctuation partially opens the dsNA. They formulated a discrete ("hopping") model for helicase motion where the helicase binds to a single DNA strand and move towards the double-stranded region by a discrete hopping process. An interaction potential causes the helicase to unwind the duplex provided the helicase is near the junction of ssNA and dsNA. It was shown that different choices of the potential correspond to the cases of active and passive opening. An optimal choice of the interaction potential leads to a helicase which can unwind NA as rapidly as it translocates on a single strand.

Molecular Models. Crystallographic data available for a few helicases have been used to model mechanisms for specific helicases. Though crystal structures cannot give a dynamic view, such proposals are the only ones available so far. Two general models, the inchworm and active rolling models, have been proposed for the mechanism of helicase-catalyzed unwinding of double-stranded nucleic acid (reviewed by [295]. According to these proposals, a hexameric helicase like dnaB opens up dsDNA like a wedge by virtue of its motor action

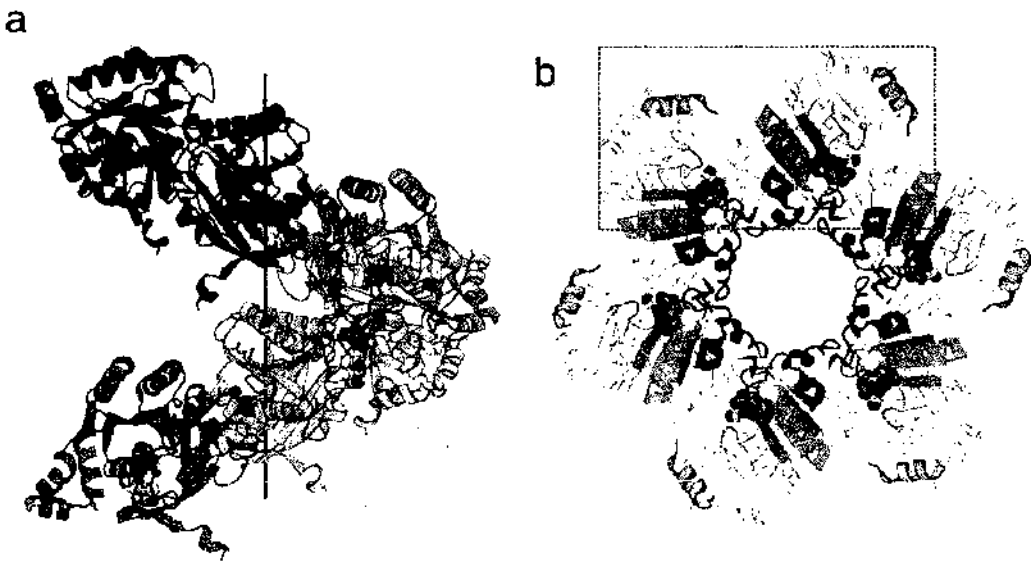


Figure 39. Crystal structure of the T7 helicase domain (PDB code 1CR0). Reprinted with permission from [292], S. S. Patel and K. M. Picha, *Annu. Rev. Biochem.* 69, 651 (2000). Fig. 11. © 2000, Annual Reviews. www.annualreviews.org.

on ssDNA [292]. A rolling mechanism has been proposed for dimeric helicases [295]. In case of PcrA, the helicase activity and the motor action can be decoupled. Crystal structure, supplemented by biochemical evidences on mutants, favours a mechanism where the helicase moves forward on ss-DNA and during its sojourn a different domain of the helicase pulls a few bases of a strand of the ds-DNA beyond the Y-fork, the junction between ss and ds-DNA [296, 297]. RecG is more complicated because it moves in opposite direction from zipped to unzipped, a fork reversal process forming a Holliday junction of four DNA strands [298]. A modified form of the active rolling model proposed in analogy to F_1-ATPase is also plausible [299]. Since helicases are members of the AAA family proteins, a large class of ATPases, many of which are (like F_1-ATPase) hexameric and whose catalytic sites are homologous to that of F_1, it was thought (but have not yet been proved) the helicase to be a rotary motor [289].

Further insights into helicase mechanism await a crystal structure of the hexameric enzyme complexed to a DNA substrate.

The cooperative interactions of replication proteins have been studied extensively in the bacteriophage T7 replication system where only four proteins are required at the replication fork, the DNA helicase, a single-stranded DNA-binding protein, the DNA polymerase catalytic subunit, and a processivity factor thioredoxin.

Figure 39 shows the subunit interactions of the T7 helicase domain.

1. The crystal packing arrangement of the T7 helicase domain, viewed perpendicular to the screw axis (black line), shows the helical filament formed by six molecules (colored differently) in the crystallographic unit cell.

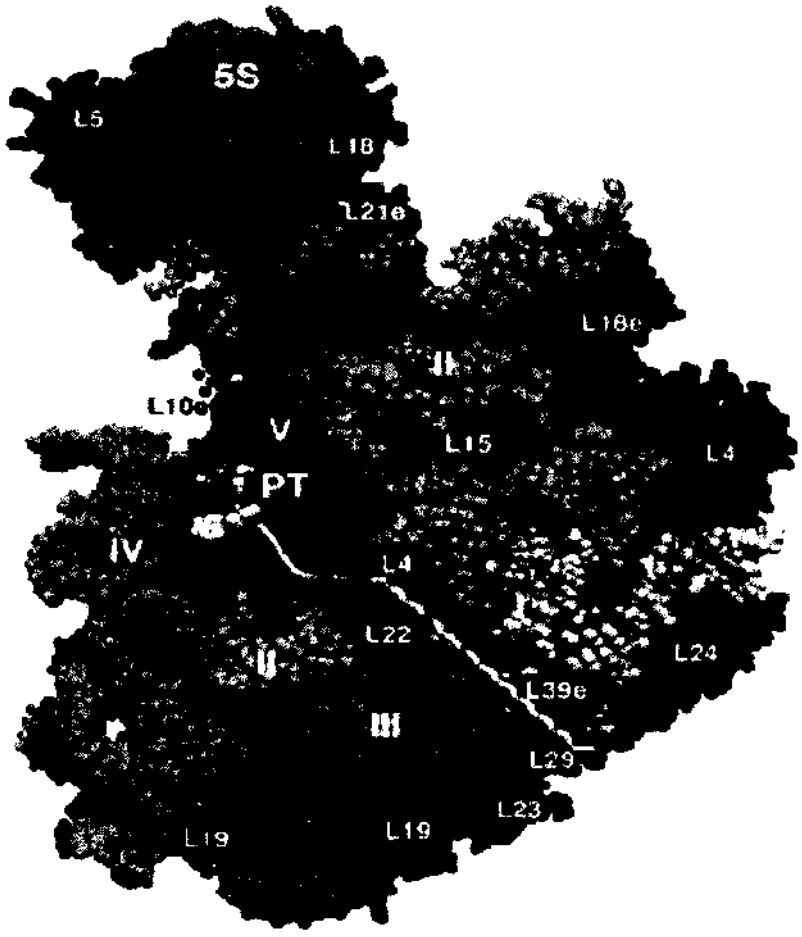


Figure 40. The bacterial ribosome. A slice through the 50S subunit of *Haloarcula marismortui* showing the locations of the modules of rRNA (roman numerals) and the proteins (blue and numbered). A white line represents the model of a nascent protein chain and the pathway it might take through the ribosomal exit tunnel. Reprinted with permission from S. A. Etchells and F. U. Hartl, *Nature Struct. Mol. Biol.* 11, 391 (2004). © 2004, Macmillan Magazines Ltd.

2. The view parallel to screw axis (black line)(rotated 90 degree away from the view shown in [a]) reveals the ringlike appearance of the 6 molecules in the crystallographic unit cell. The filament has an outer diameter of approximately 120 Å with a central hole that is about 35 Å in diameter [292].

7. CONCLUDING REMARKS

Many of the complexes described in the previous chapters and listed in Table 1 of the Introduction are classified as "molecular machines," "molecular motors," or "molecular devices," depending on their sizes, complexity, and tasks. As mentioned in the Introduction, this designation captures many of the aspects characterizing these biological complexes: modularity, complexity, cyclic function, and, in most cases, the consumption of energy. For example, as molecular motors, a rotary or linear movement is used for motility, nucleic acid processing, folding or unfolding, or as transducers, the transduction of light or chemical energy is performed. Most of the work as presented in this review deal with structures at an atomic level of resolution, but it is increasingly common to see studies in which the focal point involves interactions between modules, as in large "cellular machines" or "modular machines" [300] such as the ribosome (Fig. 40), the replisome, the transcriptional machinery, and the spliceosome.

In general, a cell, like a social community, must organize its bustling community of macromolecules and motors. Setting meeting points and establishing the timing of transactions are of fundamental importance for cell behavior. The high degree of spatial/temporal organization of molecules and organelles within cells requires interacting signaling pathways by which the various activities of the protein machines and motors are orchestrated, for example, the correct transport of cell components to various destinations within the cytoplasm. Therefore, in the near future we will witness novel investigations on the next higher level of description, in which patterns of interactions in such "networks" are themselves the focus of the analysis. Theory and simulation play an important role in characterizing these networks, and many of the key concepts are still being developed. A recent review [301] on "Biological Networks" discusses possible decompositions of networks into modules and motifs, and draw distinctions between "scale-free" and "scale-rich" sets of interactions. As the size of the database of protein structures (both observed and inferred) grows, it should increasingly be possible to both predict and understand the connections among proteins that guide cellular metabolism. The field now called "proteomics" will need to address the organizational characteristics of interaction molecular and/or modular networks, which will be the great challenge in this emerging field [301].

ACKNOWLEDGMENTS

I would like to acknowledge the hospitality of the Isaac Newton Institute for Mathematical Sciences at the University of Cambridge (UK) during 2004, where large parts of the review had been accomplished. I am very grateful to Sergei Grudinin for providing many of the figures included in the review.

REFERENCES

1. D. S. Goodsell, "Bionanotechnology: Lessons from Nature," 1st Edn. Wiley-Liss, New York, 2004.
2. R. P. Feynman, "Miniaturisation" (H. D. Gilber Ed.), p. 282, Reinhold, New York, 1961.
3. K. E. Drexler, "Engines of Creation," Anchor Press, New York, 1986.
4. K. E. Drexler, *Annu. Rev. Biophys. Biomol. Struct.* 23, 377 (1994).
5. B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, "Molecular Biology of the Cell," Garland Science, New York, 2002.
6. K. Palczewski, T. Kumasaka, T. Hori, C. Behnke, H. Motoshima, B. Fox, I. Trong, D. Teller, T. Okada, R. Stenkamp, M. Yamamoto, and M. Miyano, *Science* 289, 739 (2000).
7. J. Saam, E. Tajkhorshid, S. Hayashi, and K. Schulten, *Biophys. J.* 83, 3097 (2002).
8. P. J. Greasley, E. Fanelli, O. Rossier, L. Abouin, and S. Cotecchia, *Mol. Pharmacol.* 61, 1025 (2002).
9. M. G. Paterlini, *Biophys. J.* 83, 3012 (2002).
10. X. Huang, J. Shen, M. Cui, L. Shen, X. Luo, K. Ling, G. Pei, H. Jiang, and K. Chen, *Biophys. J.* 84, 171 (2003).

11. M. Falconi, M. Cambria, A. Cambria, and A. Desideri, *J. Biomol. Struct. Dyn.* 18, 761 (2001).
12. C. R. Watts, S. Lovas, and R. F. Murphy, *Proteins Struct. Funct. Genet.* 33, 396 (1999).
13. J. J. Falke and G. L. Hazelbauer, *Trends Biochem. Sci.* 26, 257 (2001).
14. H. Luecke, B. Schobert, J. K. Lanyi, and E. N. S. J. L. Spudich, *Science* 293 (2001).
15. A. Royant, P. Nollert, K. Edman, R. Neutze, E. M. Landau, E. Pebay-Peyroula, and J. Navarro, *Proc. Natl. Acad. Sci. USA* 98 (2001).
16. V. I. Gordeliy, J. Labahn, R. Moukhametzianov, R. Efremov, J. Granzin, R. Schlesinger, G. Büldt, T. Savopol, A. J. Scheidig, J. P. Klare, and M. Engelhard, *Nature (London)* 419 (2002).
17. H. Grubmüller, B. Heymann, and P. Tavan, *Science* 271, 997 (2001).
18. S. Izrailev, S. Stepaniants, M. Balsera, Y. Oono, and K. Schulten, *Biophys. J.* 72, 1568 (1997).
19. J. Schlitter, M. Engels, P. Kruger, E. Jacoby, and A. Wollmer, *Mol. Simul.* 10, 291 (1993).
20. E. Paci and M. Karplus, *J. Mol. Biol.* 288, 441 (1999).
21. P. E. Marszalek, H. Lu, H. Li, M. Carrion-Vazquez, A. F. Oberhauser, K. Schulten, and J. M. Fernandez, *Nature (London)* 402, 100 (1999).
22. H. Lu and K. Schulten, *Biophys. J.* 79, 51 (2000).
23. J. Ma, P. B. Sigler, Z. Xu, and M. Karplus, *J. Mol. Biol.* 302, 303 (2000).
24. B. Heymann and H. Grubmüller, *Biophys. J.* 81, 1295 (2001).
25. M. Cui, J. Shen, J. M. Briggs, X. Luo, X. Tan, H. Jiang, K. Chen, and R. Ji, *Biophys. J.* 80, 1659 (2001).
26. M. V. Bayas, K. Schulten, and D. Leckband, *Biophys. J.* 84, 2223 (2003).
27. D. P. Tieleman, P. C. Biggin, G. R. Smith, and M. S. P. Sansom, *Q. Rev. Biophys.* 43, 473 (2001).
28. D. A. Doyle, J. Morais-Cabral, R. A. Pfuetzner, A. Kuo, J. M. Gulbis, S. L. Cohen, B. T. Chait, and R. MacKinnon, *Science* 280, 69 (1998).
29. J. H. Morais-Cabral, Y. F. Zhou, and R. MacKinnon, *Nature (London)* 414, 106 (2001).
30. Y. F. Zhou, J. H. Morais-Cabral, A. Kaufman, and R. MacKinnon, *Nature* 414, 43 (2001).
31. Y. Jiang, A. Lee, J. Chen, M. Cadene, B. T. Chait, and R. MacKinnon, *Nature* 417, 523 (2002).
32. T. W. Allen, S. Kuyucak, and S.-H. Chung, *Biophys. J.* 77, 2502 (1999).
33. L. Guidoni, V. Torre, and P. Carloni, *Biochemistry* 38, 8599 (1999).
34. I. H. Shrivastava and M. S. P. Sansom, *Biophys. J.* 78, 557 (2000).
35. L. Guidoni, V. Torre, and P. Carloni, *FEBS Lett.* 477, 37 (2000).
36. S. Berneche and B. Roux, *Biophys. J.* 78, 290 (2000).
37. S. Berneche and B. Roux, *Nature (London)* 414, 73 (2001).
38. C. Domene, P. J. Bond, and M. S. P. Sansom, *Adv. Protein Chem.* 66, 159 (2003).
39. P. J. Bond and M. S. P. Sansom, *Mol. Membrane Biol.* 21, 151 (2004).
40. A. Suenaga, A. Komeiji, M. Uebayasi, T. Meguro, M. Saito, and I. Yamato, *Biosci. Rep.* 18, 39 (1998).
41. D. Tieleman and H. J. C. Berendsen, *Biophys. J.* 74, 2788 (1998).
42. T. Schirmer and P. Phale, *J. Mol. Biol.* 294, 1159 (1999).
43. P. Phale, A. Phillipsen, C. Widmer, V. P. Phale, J. P. Rosenbusch, and T. Schirmer, *Biochemistry* 40, 6319 (2001).
44. W. Im and B. Roux, *Biophys. J.* 79, 788 (2000).
45. W. Im and B. Roux, *J. Mol. Biol.* 319, 1177 (2002).
46. M. Bazden and M. S. P. Sansom, *Biophys. J.* 87, 2942 (2004).
47. P. J. Bond, J. D. Faraldo-Gomez, and M. S. P. Sansom, *Biophys. J.* 83, 763 (2002).
48. J. D. Faraldo-Gomez, G. R. Smith, and M. S. P. Sansom, *Biophys. J.* 85, 1406 (2003).
49. M. Jensen, E. Tajkhorshid, and K. Schulten, *Structure* 9, 1083 (2001).
50. F. Zhu, E. Tajkhorshid, and K. Schulten, *FEBS Lett.* 504, 212 (2001).
51. B. L. de Groot and H. Grubmüller, *Science* 294, 2353 (2001).
52. E. Tajkhorshid, P. Nollert, M. Jensen, L. Miercke, J. O'Connell, R. Stroud, and K. Schulten, *Science* 296, 525 (2002).
53. M. O. Jensen, E. Tajkhorshid, and K. Schulten, *Biophys. J.* 85, 2884 (2003).
54. V. W. Sidel and A. K. Solomon, *J. Gen. Physiol.* 41, 243 (1957).
55. K. Murata, K. Mitsuoka, T. Hirai, T. Walz, P. Agre, J. Heymann, A. Engel, and Y. Fujiyoshi, *Nature (London)* 407, 599 (2000).
56. Y. Fujiyoshia, K. Mitsuoka, B. L. de Groot, A. Philippsen, H. Grubmüller, P. Agre, and A. Engel, *Curr. Opin. Struct. Biol.* 12, 509 (2002).
57. T. Schirmer, *J. Struct. Biol.* 121, 101 (1998).
58. A. Burykin and A. Warshel, *Biophys. J.* 85, 3696 (2003).
59. A. Suenaga, A. Komeiji, M. Uebayasi, T. Meguro, M. Saito, and I. Yamato, *Biosci. Rep.* 18, 39 (1998).
60. F. Zhu, E. Tajkhorshid, and K. Schulten, *Biophys. J.* 86, 50 (2004).
61. D. Fu, A. Libson, L. J. W. Miercke, C. Weitzman, P. Nollert, J. Krucinski, and R. M. Stroud, *Science* 290, 481 (2000).
62. R. Dutzler, T. Schirmer, M. Karplus, and S. Fischer, *Structure* 10, 1273 (2002).
63. L. R. Forrest, D. P. Tieleman, and M. S. P. Sansom, *Biophys. J.* 76, 1886 (1999).
64. K. J. Schweighofer and A. Pohorille, *Biophys. J.* 78, 150 (2000).
65. B. Hille, "Ion Channels of Excitable Membranes," 3rd Edn., Sinauer Sunderland, MA, 2001.
66. D. J. Aidley and P. Stanfield, "Ion Channels: Molecules in Action," Cambridge University Press, Cambridge, 1996.
67. A. L. Hodgkin and A. F. Huxley, *J. Physiol.* 117, 500 (1952).
68. Y. Jiang, V. Ruta, J. Chen, A. Lee, and R. MacKinnon, *Nature (London)* 423, 42 (2003).

69. Q.-X. Jiang, D. Wang, and R. MacKinnon, *Nature (London)* 430, 806 (2004).
70. S. Kuyucak, O. S. Andersen, and S.-H. Chung, *Rep. Prog. Phys.* 64, 1427 (2001).
71. B. Roux, *Curr. Opin. Struct. Biol.* 12, 182 (2002).
72. D. P. Tieleman, B. Hess, and M. S. P. Sansom, *Biophys. J.* 83, 2393 (2002).
73. D. P. Tieleman, V. Borisenko, M. S. P. Sansom, and G. A. Woolley, *Biophys. J.* 84, 1464 (2003).
74. A. Miyazawa, Y. Fujiyoshi, and N. Unwin, *Nature (London)* 423, 949 (2003).
75. R. J. Law, D. P. Tieleman, and M. S. P. Sansom, *Biophys. J.* 84, 14 (2003).
76. R. H. Henchman, H. L. Wang, S. M. Sine, P. Taylor, and J. A. McCammon, *Biophys. J.* 85, 3007 (2003).
77. T. J. Jentsch, V. Stein, F. Weinreich, and A. A. Zdebik, *Physiol. Rev.* 82, 503 (2002).
78. R. Dutzler, E. B. Campbell, M. Cadene, B. T. Chait, and R. MacKinnon, *Nature (London)* 415, 287 (2002).
79. R. Dutzler, E. B. Campbell, and R. MacKinnon, *Science* 308, 108 (2003).
80. G. V. Miloshevsky and P. C. Jordan, *Biophys. J.* 86, 825 (2004).
81. J. Cohen and K. Schulten *Biophys. J.* 86, 836 (2004).
82. B. Roux, *Acc. Chem. Res.* 6, 366 (2002).
83. T. W. Allen, O. S. Andersen, and B. Roux, *J. Am. Chem. Soc.* 125, 9868 (2003).
84. S. W. Chiu, S. Subramaniam, and E. Jakobsson, *Biophys. J.* 76, 1929 (1999).
85. S. W. Chiu, S. Subramaniam, and E. Jakobsson, *Biophys. J.* 76, 1939 (1999).
86. T. K. Rostovtseva, V. M. Aguilella, I. Vodyanoy, S. M. Bezrukov, and V. A. Parsegian *Biophys. J.* 76, 1783 (1998).
87. D. D. Busath, C. D. Thulin, R. W. Hendershot, L. R. Phillips, P. Maughan, C. D. Cole, N. C. Bingham, S. Morrison, L. C. Baird, R. J. Hendershot, M. Cotten, and T. A. Cross, *Biophys. J.* 75, 2830 (1998).
88. A. E. Cardenas, R. D. Coalson, and M. G. Kurnikova, *Biophys. J.* 79, 80 (2000).
89. M. Agmon, *Chem. Phys. Lett.* 244, 456 (1995).
90. R. Pomes and B. Roux, *Biophys. J.* 75, 33 (1998).
91. R. Pomes and B. Roux, *Biophys. J.* 82, 2304 (2002).
92. M. F. Schumaker, R. Pomes, and B. Roux, *Biophys. J.* 79, 2840 (2000).
93. B. L. de Groot, D. P. Tieleman, P. Pohl, and H. Grubmüller, *Biophys. J.* 82, 2934 (2002).
94. T. W. Allen, T. Bastug, S. Kuyucak, and S. H. Chung, *Biophys. J.* 84, 2159 (2003).
95. G. V. Miloshevsky and P. C. Jordan, *Biophys. J.* 86, 92 (2004).
96. Y. Arinaminpathy, M. S. P. Sansom, and P. C. Biggin, *Biophys. J.* 82, 676 (2002).
97. J. Mendieta, G. Ramírez, and F. Gago, *Proteins Struct. Funct. Genet.* 44, 460 (2001)
98. B. Corry, W. Allen, S. Kuyucak, and S.-H. Chung, *Biophys. J.* 80, 195 (2001).
99. J. F. Gwan, Molecular Dynamics Studies on Ion Channels, Ph.D. Thesis, Research Center Jülich, Jülich, 2005.
100. B. Roux and R. MacKinnon, *Science* 285, 100 (1999).
101. G. Moy, B. Corry, S. Kuyucak, and S.-H. Chung, *Biophys. J.* 78, 2349 (2000).
102. S.-H. Chung, T. W. Allen, and S. Kuyucak, *Biophys. J.* 82, 628 (2002).
103. T. W. Allen, A. Bliznyuk, A. P. Rendell, S. Kuyucak, and S.-H. Chung, *J. Chem. Phys.* 112, 8191 (2000).
104. J. Åqvist and V. Luzhkov, *Nature (London)* 404, 881 (2000).
105. I. H. Shrivastava, D. P. Tieleman, P. C. Biggin, and M. S. P. Sansom, *Biophys. J.* 83, 633 (2002).
106. C. Domene and M. S. P. Sansom, *Biophys. J.* 85, 2787 (2003).
107. L. G. Cuello, J. G. Romero, D. M. Cortes, and E. Perozo, *Biochemistry* 37, 3229 (1998).
108. E. Perozo, D. M. Cortes, L. G. Cuello, *Science* 285, 73 (1999).
109. D. M. Cortes, L. G. Cuello, and E. Perozo, *J. Gen. Physiol.* 117, 165 (2001).
110. Y. S. Liu, P. Sompornpisut, and E. Perozo, *Nature Struct. Biol.* 8, 883 (2001).
111. P. C. Biggin and M. S. P. Sansom, *Biophys. J.* 83, 1867 (2002).
112. M. Nishida and R. MacKinnon, *Cell* 111, 957 (2002).
113. A. Kuo, J. M. Gulbis, J. F. Antcliff, T. Rahman, E. D. Lowe, J. Zimmer, J. Cuthbertson, F. M. Ashcroft, T. Ezaki, and D. A. Doyle, *Science* 300, 1922 (2003).
114. C. Domene, A. Grottesi, and M. S. P. Sansom, *Biophys. J.* 87, 256 (2004).
115. S. Haider, A. Grottesi, B. A. Hall, F. M. Ashcroft, and M. S. P. Sansom, *Biophys. J.* 88, 3310 (2005).
116. T. M. Suchyna, S. E. Tape, R. E. Koeppe, O. S. Andersen, F. Sachs, and P. A. Gottlieb, *Nature (London)* 430, 235 (2004).
117. S. Sukharev, M. Betanzos, C. S. Chiang, and H. R. Guy, *Nature (London)* 409, 720 (2001).
118. M. Betanzos, C. S. Chiang, H. R. Guy, and S. Sukharev, *Nature Struct. Biol.* 9, 704 (2002).
119. E. Perozo, A. Kloda, D. Cortes, and B. Martinac, *Nature Struct. Biol.* 9, 696 (2002).
120. E. Perozo, D. Cortes, P. Sompornpisut, A. Kloda, and B. Martina, *Nature (London)* 418, 940 (2002).
121. E. Perozo and D. C. Rees, *Curr. Opin. Struct. Biol.* 13, 432 (2003).
122. J. Gullingsrud, D. Kosztin, and K. Schulten, *Biophys. J.* 80, 2074 (2001).
123. D. E. Elmore and D. A. Dougherty, *Biophys. J.* 81, 1345 (2001).
124. J. Gullingsrud and K. Schulten, *Biophys. J.* 85, 2087 (2003).
125. R. B. Bass, P. Strop, M. Barclay, and D. C. Rees, *Science* 298, 1582 (2002).
126. M. D. Edwards, Y. Li, S. Kim, S. Miller, W. Bartlett, S. Black, S. Dennison, I. Iscla, P. Blount, J. U. Bowie, and I. R. Booth, *Nature Struct. Mol. Biol.* 12, 113 (2005).
127. M. Sotomayor and K. Schulten, *Biophys. J.* 87, 3050 (2004).
128. S. Sukharev, *Biophys. J.* 83, 290 (2002).
129. D. Bostick and M. L. Berkowitz, *Biophys. J.* 85, 97 (2003).
130. J. Heberle, *Biophys. Biochim. Acta* 1458, 135 (2000).

131. E. Sass, G. Büldt, R. Gessenich, D. Hehn, D. Neff, R. Schlesinger, J. Berendzen, and P. Ormos, *Nature (London)* 406, 649 (2000).
132. C. Edholm, O. Berger, and F. Jähnig, *J. Mol. Biol.* 250, 94 (1995).
133. D. Xu, M. Sheves, and K. Schulten, *Biophys. J.* 69, 2715 (1995).
134. I. Logunov, W. Humphrey, K. Schulten, and M. Sheves, *Biophys. J.* 68, 1270 (1995).
135. E. Steinhoff, M. Müller, C. Beier, and M. Pfeiffer, *J. Mol. Liq.* 84, 17 (2000).
136. B. Roux, M. Nina, R. Pomes, and J. C. Smith, *Biophys. J.* 71, 670 (1998).
137. J. Baudry, E. Tajkhorshid, F. Molnar, J. Phillips, and K. Schulten, *J. Phys. Chem. B* 105, 905 (2001).
138. C. Kandt, J. Schlitter, and K. Gerwert, *Biophys. J.* 86, 705 (2004).
139. S. Grudinin, G. Büldt, and A. Baumgaertner, *Biophys. J.* 88, 3252 (2005).
140. S. Hayashi, E. Tajkhorshid, and K. Schulten, *Biophys. J.* 83, 1281 (2002).
141. E. Tajkhorshid, J. Baudry, K. Schulten, and S. Suhai, *Biophys. J.* 78, 683 (2000).
142. S. Subramaniam and R. Henderson, *Nature (London)* 406, 653 (2000).
143. A.-N. Bondar, M. Elstner, S. Suhai, J. C. Smith, and S. Fischer, *Structure* 12, 1281 (2004).
144. A.-N. Bondar, S. Fischer, J. M. Elstner, and S. Suhai, *J. Am. Chem. Soc.* 126, 14668 (2004).
145. D. Porschke, *Biophys. J.* 71, 3381 (1996).
146. M. Kolbe, H. Besir, L. O. Essen, and D. Oesterhelt, *Science* 288, 1390 (2000).
147. T. Nishi and M. Forgac, *Nat. Rev. Mol. Cell Biol.* 3, 94 (2002).
148. N. Perzov, V. Padler-Karavani, H. Nelson, and N. Nelson, *FEBS Lett.* 504, 223 (2001).
149. M. Grabe, H. Wang, and G. Oster, *Biophys. J.* 78, 2798 (2000).
150. C. Toyoshima, H. Nomura, and T. Tsuda, *Nature (London)* 432, 361 (2004).
151. D. A. Case, *Curr. Opin. Struct. Biol.* 4, 285 (1994).
152. B. Tidor and M. Karplus, *J. Mol. Biol.* 238, 405 (1994).
153. J. Smith, K. Kuczera, and M. Karplus, *Proc. Natl. Acad. Sci. USA* 87, 1601 (1990).
154. F. Tama, F. X. Gadea, O. Marques, and Y. Sanejouand, *Proteins Struct. Funct. Genet.* 41, 1 (2000).
155. G. Li and Q. Cui, *Biophys. J.* 86, 743 (2004).
156. D. Voet, J. G. Voet, and C. W. Pratt, "Fundamentals of Biochemistry," Wiley, New York, 1999.
157. P. M. Jones and A. M. George, *Proc. Natl. Acad. Sci. USA* 99, 12639 (2002).
158. J. D. Campbell, S. S. Deol, F. M. Ashcroft, I. D. Kerr, and M. S. P. Sansom, *Biophys. J.* 87, 3703 (2004).
159. G. Chang and C. B. Roth, *Science* 293, 1793 (2001).
160. J. D. Campbell, P. C. Biggin, M. Baaden, and M. S. P. Sansom, *Biochemistry* 42, 3666 (2003).
161. D. S. Dwyer, *Proteins Struct. Funct. Genet.* 42, 531 (2001).
162. R. E. Dalbey and G. von Heijne, Eds., "Protein Targeting, Transport and Translocation," Academic Press, London, 2002.
163. S. H. White and W. C. Wimley, *Curr. Opin. Struct. Biol.* 4, 79 (1994).
164. J. L. Popot and D. M. Engelman, *Annu. Rev. Biochem.* 69, 881 (2000).
165. A. Baumgaertner and J. Skolnick, *Phys. Rev. Lett.* 74, 2142 (1995).
166. A. Baumgaertner, *Biophys. J.* 71, 1248 (1996).
167. D. J. Schnell and D. N. Hebert, *Cell* 112, 491 (2003).
168. L. K. Tamm, J. Crane, and V. Kiessling, *Curr. Opin. Struct. Biol.* 13, 453 (2003).
169. P. Rehling, K. Brandner, and N. Pfanner, *Nat. Rev. Mol. Cell Biol.* 5, 519 (2004)
170. W. Neupert and M. Brunner, *Nat. Rev. Mol. Cell Biol.* 3, 555 (2002).
171. A. E. Johnson and M. A. Waes, *Ann. Rev. Cell Dev. Biol.* 15, 799 (1999).
172. M. P. Rout, J. D. Aitchison, M. O. Magnasco, and B. T. Chait, *Trends Cell Biol.* 13, 622 (2003).
173. G. Blobel and B. Dobberstein, *J. Cell Biol.* 67, 835 (1975).
174. G. Blobel, *Proc. Natl. Acad. Sci. USA* 77, 1496 (1980).
175. S. M. Simon, C. Peskin, and G. Oster, *Proc. Natl. Acad. Sci. USA* 89, 3770 (1992).
176. C. Peskin, G. Odell, and G. Oster, *Biophys. J.* 65, 316 (1993).
177. R. Feynman, R. Leighton, and M. Sands, "The Feynman Lectures on Physics," Addison Wesley, Reading, MA, 1963, Vol. 1.
178. T. Lubensky and D. Nelson, *Biophys. J.* 77, 1824 (1999).
179. M. Muthukumar, *J. Chem. Phys.* 111, 10371 (1999).
180. M. Muthukumar, *Phys. Rev. Lett.* 86, 3188 (2001).
181. S. S. Chern, A. E. Cardenas, and R. D. Coalson, *J. Chem. Phys.* 115, 7772 (2001).
182. J. Chuang, Y. Kantor, and M. Kardar, *Phys. Rev. E* 65, 011802 (2002).
183. H. C. Loebl, R. Randel, S. Goodwin, and C. C. Matthai, *Phys. Rev. E* 67, 041913 (2003).
184. Y. Kantor and M. Kardar, *Phys. Rev. E* 65, 021806 (2004).
185. A. Meller, *J. Phys. C* 15, R581 (2003).
186. A. M. Berezhkovskii and I. V. Gopich, *Biophys. J.* 84, 787 (2003).
187. T. C. Elston, *Biophys. J.* 82, 1239 (2002).
188. R. Zandi, D. Reguera, J. Rudnick, and W. M. Gelbart, *Proc. Natl. Acad. Sci. USA* 100, 8649 (2003).
189. B. S. Glick, *Cell* 80, 11 (1995).
190. J.-F. Chauwin, G. Oster, and B. S. Glick, *Biophys. J.* 74, 1732 (1998).
191. T. C. Elston, *Biophys. J.* 79, 2235 (2000).
192. S. Simon and G. Blobel, *Cell* 65, 371 (1991).
193. B. van den Berg, W. M. Clemens, I. Collinson, Y. Modis, E. Hartmann, S. C. Harrison, and T. A. Rapoport, *Nature (London)* 427, 36 (2004).

194. K. E. Matlack, W. Mothes, and T. A. Rapoport, *Cell* 92, 381 (1998).
195. Y. Matsuura and M. Stewart, *Nature (London)* 432, 872 (2004).
196. M. Suntharalingam and S. R. Wente, *Dev. Cell* 4, 775 (2003).
197. M. P. Rout and J. D. Aitchison, *J. Biol. Chem.* 276, 16593 (2001).
198. I. Macara, *Microbiol. Mol. Biol. Rev.* 65, 570 (2001).
199. K. Ribbeck and D. Gorlich, *EMBO J.* 20, 1320 (2001).
200. K. Ribbeck and D. Gorlich, *EMBO J.* 21, 2664 (2001).
201. T. Bickel and R. Bruinsma, *Biophys J.* 83, 3079 (2002).
202. N. Shulga and D. S. Goldfarb, *Mol. Cell. Biol.* 23, 534 (2003).
203. T. Kustanovich and Y. Rabin, *Biophys. J.* 86, 2008 (2004).
204. J. P. Abrahams, A. Leslie, R. Lutter, and J. E. Walker, *Nature (London)* 370, 621 (1994).
205. D. Stock, A. G. W. Leslie, and J. E. Walker, *Science* 286, 1700 (1999).
206. C. Gibbons, M. G. Montgomery, A. Leslie, and J. E. Walker, *Nat. Struct. Biol.* 7, 1055 (2000).
207. G. Oster and H. Wang, *Trends Cell Biol.* 13, 114 (2003).
208. J. L. Rubinstein, J. E. Walker, and R. Henderson, *EMBO J.* 22, 6182 (2003).
209. R. Böckmann and H. Grubmüller, *Nat. Struct. Biol.* 9, 198 (2002).
210. J. Ma, T. C. Flynn, Q. Cui, A. G. W. Leslie, J. E. Walker, and M. Karplus, *Structure* 10, 921 (2002).
211. A. Aksimentiev, I. A. Balabin, R. H. Fillingame, and K. Schulten, *Biophys. J.* 86, 1332 (2004).
212. T. Elston, H. Wang, and G. Oster, *Nature (London)* 391, 510 (1998).
213. H. Wang and G. Oster, *Nature (London)* 396, 279 (1998).
214. G. Oster and H. Wang, *Biochim. Biophys. Acta* 1458, 482 (2000).
215. J. Xing, H. Wang, C. von Ballmoos, P. Dimroth, and G. Oster, *Biophys. J.* 87, 2148 (2004).
216. C. Bustamante, D. Keller, and G. Oster, *Acc. Chem. Res.* 34, 412 (2001).
217. H. Noji, R. Yasuda, M. Yoshida, K. Kinosit, and H. Itoh, *Nature (London)* 386, 299 (1997).
218. R. Böckmann and H. Grubmüller, *Biophys. J.* 85, 1482 (2003).
219. M. Dittrich, S. Hayashi, and K. Schulten, *Biophys. J.* 85, 2253 (2003).
220. W. Yang, Y. Gao, Q. Cui, J. Ma, and M. Karplus, *Proc. Natl. Acad. Sci. USA* 100, 874 (2003).
221. Y. Gao, W. Yang, R. A. Marcus, and M. Karplus, *Proc. Natl. Acad. Sci. USA* 100, 11339 (2003).
222. I. Antes, D. Chandler, H. Wang, and G. Oster, *Biophys. J.* 85, 695 (2003).
223. M. Karplus and Y. Q. Gao, *Curr. Opin. Struct. Biol.* 14, 250 (2004).
224. H. Miki, M. Sato, and M. Kohmoto, *Phys. Rev. E* 68, 061906 (2003).
225. D. A. Beard and T. Schlick, *Biophys. J.* 85, 2973 (2003).
226. H. C. Berg, *Annu. Rev. Biochem.* 72, 19 (2003).
227. P. Brown, C. P. Hill, and D. F. Blair, *EMBO J.* 21, 3225 (2002).
228. S. Kojima and D. Blair, *Biochemistry* 40, 13041 (2001).
229. T. Elston and G. Oster, *Biophys. J.* 73, 703 (1997).
230. D. Walz and S. R. Caplan, *Biophys. J.* 78, 626 (2000).
231. R. Schmitt, *Biophys. J.* 85, 843 (2003).
232. R. Berry, L. Turner, and H. C. Berg, *Biophys. J.* 69, 280 (1995).
233. T. A. Duke, N. LeNovore, and D. Bray, *J. Mol. Biol.* 308, 541 (2001).
234. G. Koster, M. Van Duijn, B. Hofs, and M. Dogterom, *Proc. Natl. Acad. Sci. USA* 100, 15583 (2003).
235. J. Howard, "Mechanics of Motor Proteins and the Cytoskeleton," Sinauer Sunderland, MA, 2001.
236. R. D. Vale, *Cell* 112, 467 (2003).
237. E. L. F. Holzbauer, *Trends in Cell Biol.* 14, 233 (2004).
238. N. J. Cordova, B. Ermentrout, and G. Oster, *Proc. Natl. Acad. Sci.* 89, 339 (1992).
239. S. Leibler and D. A. Huse, *J. Cell Biol.* 121, 1357 (1993).
240. F. Jülicher, A. Adjari, and J. Prost, *Rev. Mod. Phys.* 69, 1269 (1997).
241. D. Keller and C. Bustamante, *Biophys. J.* 78, 541 (2000).
242. G. Li and Q. Cui, *J. Phys. Chem. B* 108, 3342–3357 (2004).
243. M. Walker et al., *Nature (London)* 405, 804 (2000).
244. Y. Ishii and T. Yanagida, *Cell. Mol. Life Sci.* 59, 1767 (2002).
245. A. B. Kolomeisky and M. E. Fisher, *Biophys. J.* 84, 1642 (2003).
246. M. J. Tyska and M. S. Mooseker, *Trends Cell Biol.* 13, 447 (2003).
247. E. M. De La Cruz, A. L. Wells, S. S. Rosenfeld, E. M. Ostap, and H. L. Sweeney, *Proc. Natl. Acad. Sci. USA* 96, 13726 (1999).
248. M. Rief, R. Rock, A. Mehta, M. Mooseker, R. Cheney, and J. Spudich, *Proc. Natl. Acad. Sci. USA* 97, 9482 (2000).
249. H. Tanaka, K. Homma, A. Iwane, E. Katayama, R. Ikebe, J. Saito, T. Yanagida, and M. Ikebe, *Nature (London)* 415, 192 (2002).
250. W. Hua, J. Chung, and J. Gelles, *Science* 295, 844 (2002).
251. D. V. Nicolau and D. V. Nicolau, *Curr. Appl. Phys.* 36, 122 (2004).
252. G. Lan and S. X. Sun, *Biophys. J.* 88, 999 (2005).
253. C. S. Peskin and G. Oster, *Biophys. J.* 68, 202 (1995).
254. F. Gibbons, J. F. Chauwin, M. Desposito, and J. V. Joso, *Biophys. J.* 80, 2515 (2001).
255. A. Vilfan, E. Frey, F. Schwabl, M. Thormaehlen, Y.-H. Song, and E. Mandelkow, *J. Mol. Biol.* 312, 1011 (2001).
256. K. Visscher, M. J. Schnitzer, and S. M. Block, *Nature (London)* 400, 184 (1999).
257. Y. Chen, B. Yan, and R. J. Rubin, *Biophys. J.* 83, 2360 (2002).

258. C. A. Collins, *Trends Cell Biol.* 1, 57 (1991).
259. J. F. Eccleston, D. D. Binns, C. T. Davis, J. P. Albanesi, and D. M. Jameson, *Eur. Biophys. J.* 31, 275 (2002).
260. S. W. Deacon, A. Serpinskaya, P. Vaughn, M. Fanarraga, I. Vernos, K. Vaughan, and V. Gelfand, *J. Cell Biol.* 160, 297 (2003).
261. J. Norberg and L. Nilsson, *Acc. Chem. Res.* 35, 465 (2002).
262. D. L. Beveridge and K. J. McConnell, *Curr. Opin. Struct. Biol.* 10, 182 (2000).
263. Y. Tang and L. Nilsson, *Biophys. J.* 77, 1284 (1999).
264. E. Giudice and R. Lavery, *Acc. Chem. Res.* 35, 350 (2002).
265. N. Bruant, D. Flatters, R. Lavery, and D. Genest, *Biophys. J.* 77, 2366 (1999).
266. S. B. Singh and P. A. Kollman, *Am. Chem. Soc.* 121, 3267 (1999).
267. P. Carloni, M. Sprik, and W. Andreoni, *J. Phys. Chem. B* 104, 823 (2000).
268. A. Lahiri and L. Nilsson, *Biophys. J.* 79, 2276 (2000).
269. M. Zacharias, *Curr. Opin. Struct. Biol.* 10, 311 (2000).
270. P. Auffinger, S. Louise-May, and E. Westhof, *Biophys. J.* 76, 50 (1999).
271. T. Hermann and E. Westhof, *Nat. Struct. Biol.* 6, 540 (1999).
272. C. M. Reyes and P. A. Kollman, *J. Mol. Biol.* 297, 1145 (2000).
273. A. A. Eriksson and L. Nilsson, *Protein Eng.* 11, 589 (1998).
274. S. Sen and L. Nilsson, *Biophys. J.* 77, 1782 (1999).
275. A. Balaeff, M. E. Churchill, and K. Schulten, *Proteins Struct. Funct. Genet.* 30, 113 (1998).
276. N. Pastor, H. Weinstein, E. Jamison, and M. Brenowitz, *J. Mol. Biol.* 304, 55 (2000).
277. V. Tsui, I. Radhakrishnan, P. Wright, and D. A. Case, *J. Mol. Biol.* 302, 1101 (2000).
278. Y. Komeiji and M. Uebayasi, *Biophys. J.* 77, 123 (1999).
279. A. Suenaga, C. Yatsu, Y. Komeiji, M. Uebayasi, T. Meguro, and I. Yamato, *J. Mol. Struct.* 526, 209 (2000).
280. J. Rudnick and R. Bruinsma, *Biophys. J.* 76, 1725 (1999).
281. G. Wedemann and J. Langowski, *Biophys. J.* 82, 2847 (2002).
282. M. Podobnik, T. Weitze, M. O'Donnell, and J. Kuriyan, *Structure* 11, 253 (2003).
283. A. Goel, R. Astumian, and D. Herschbach, *Proc. Natl. Acad. Sci.* 100, 9699 (2003).
284. M. Delarue and Y. Sanejouand, *J. Mol. Biol.* 320, 1011 (2002).
285. D. Jeruzalmi, O. Yurieva, Y. Zhao, M. Young, J. Stewart, M. Hingorani, M. O'Donnell, and J. Kuriyan, *Cell* 106, 417 (2001).
286. F. Jülicher and R. Bruinsma, *Biophys. J.* 74, 1169 (1998).
287. H. Wang, T. Elston, A. Mogilner, and G. Oster, *Biophys. J.* 74, 1186 (1998).
288. P. H. von Hippel and E. Delagoutte, *Cell* 104, 177 (2001).
289. C. Doering, B. Ermentrout, and G. Oster, *Biophys. J.* 69, 2256 (1995).
290. Y. Z. Chen, H. S. Song, and X. J. Wang, *Phys. Rev. E* 56, 919 (1997).
291. S. M. Bhattacharjee and F. Seno, *J. Phys. A* 36, L181 (2003).
292. S. S. Patel and K. M. Picha, *Annu. Rev. Biochem.* 69, 651 (2000).
293. S. M. Bhattacharjee, *Europhys. Lett.* 65, 574 (2004).
294. M. D. Betterton and F. Jülicher, *Phys. Rev. Lett.* 91, 258103 (2003).
295. T. M. Lohman and K. Bjornson, *Annu. Rev. Biochem.* 65, 169 (1996).
296. S. S. Velankar, P. Soultanas, S. Dillingham, H. Subramanya, and D. B. Wigley, *Cell* 97, 75 (1999).
297. P. Soultanas, *EMBO J.* 19, 3799 (2000).
298. M. R. Singleton, *Cell* 107, 79 (2001).
299. S. S. Patel, *Proc. Natl. Acad. Sci.* 94, 5012 (1997).
300. B. Alberts, *Cell* 92, 291 (1998).
301. E. Alm and A. P. Arkin, *Curr. Opin. Struct. Biol.* 13, 193 (2003).

CHAPTER 4

Knots, Bubbles, Unwinding, and Breathing: Probing the Topology of DNA and Other Biomolecules

Ralf Metzler

NORDITA–Nordic Institute for Theoretical Physics, DK-2100 Copenhagen, Denmark

Andreas Hanke

Department of Physics and Astronomy, University of Texas at Brownsville, Brownsville, Texas, and NanoTech Institute, University of Texas at Dallas, Richardson, Texas, USA

CONTENTS

ISBN: 1-58883-048-9

Handbook of Theoretical and Computational Nanotechnology
Edited by Michael Rieth and Wolfram Schommers
Volume 6: Pages (163–216)

1. INTRODUCTION

In an unprecedented way, modern science has gained access to the nanometer scale. The possibility to create moving mechanical structures on the sub-micrometer scale has led to a nonabating advancement in engineering, the sciences, and deep into medicine technology. Newly emerged experimental techniques have leveraged the potential to both probe and manipulate the equilibrium and out of equilibrium behavior of single biopolymers of the polynucleotide and polypeptide families. Thus, on the single molecular level one can explore the physicomechanical properties of biopolymers, and their interaction with other biopolymers such as protein-DNA, protein-RNA, or protein-protein interplay, as well as their coupling to thermal activation due to their highly Brownian environment.

As a biopolymer possesses an excluded volume, that is, it cannot cross itself or other biopolymers, the topology of its three-dimensional configuration becomes an important issue in both the equilibrium and dynamical behavior [1–8]. This notion of topology needs some explanation, best illustrated at hands of the different systems we have in mind. Bacterial (prokaryotic) DNA mostly comes in circular form, that is, tracing the DNA along its biochemical backbone is always a closed space-curve. However, this circle is not necessarily simply connected such that when imaginarily pressed flat on a plane it could be turned into a simple ring shape. Instead of such an unknot configuration, naturally occurring circular DNA can contain various knot types [4, 9, 10, 13, 14]. An extreme example for DNA-topology is kDNA that consists of thousands of concatenated (interlocked) DNA maxi- and minicircles, performing a well correlated replication process [15, 16]. Different knot types or concatenated ring plasmid configurations are an example of the topological degrees of freedom one biopolymer, or an array of several, can explore. Different topological states effect different physiological and physical behaviours such as the inhibition of replication of knotted DNA [4, 9, 17], or the mobility in gel electrophoresis experiments [18, 19]. Nature has come up with particular enzymes that control the knottedness[1] (as well as other topological states such as twist-induced supercoiling) of DNA, whose exact ability to locate a knot in a circular DNA is an unresolved question in molecular biology [20]. Figure 1 shows a DNA trefoil knot and a three-dimensional visualization of the enzyme topoisomerase II that actively changes the topology of DNA by cutting the DNA double-strand, passing another DNA segment through the gap, and resealing the gap [21, 22].

Similar to DNA-knots occurring in double-stranded DNA (dsDNA), RNA and single-stranded DNA (ssDNA) form a secondary structure of bubbles and double-stranded zones of local regions of matching bases [24, 26–28]. This secondary structure may obtain additional volume interactions, the so-called pseudoknots. Other types of topological effects concern the fluctuation-induced creation of single-stranded DNA-bubbles inside a double-stranded DNA, as well as the melting of DNA at higher temperatures [29, 30]. Topology is an essential ingredient in the targeting of proteins, that is, the question how a protein by combination of one-dimensional diffusion along the DNA and three-dimensional diffusion through the bulk can efficiently locate its target site on the DNA [31, 32]. It also gives rise to subtle constraints in the passage of a biopolymer through a membrane nanopore [33], and the ordered packaging of viral DNA into nanocontainers [34]. Additional topological features, which we are not going to address in the following, come into play in eukaryotic systems, for instance, in the condensation of DNA around histones and their wrapping into chromosomes [35], or

[1] Corresponding to the "knottiness" in early knot literature. Throughout the text, we use the term *ring polymer* or *ring DNA* for a chain which represents the unknot. *Circular polymer/DNA* will refer to a closed chain of any knottedness, in general.

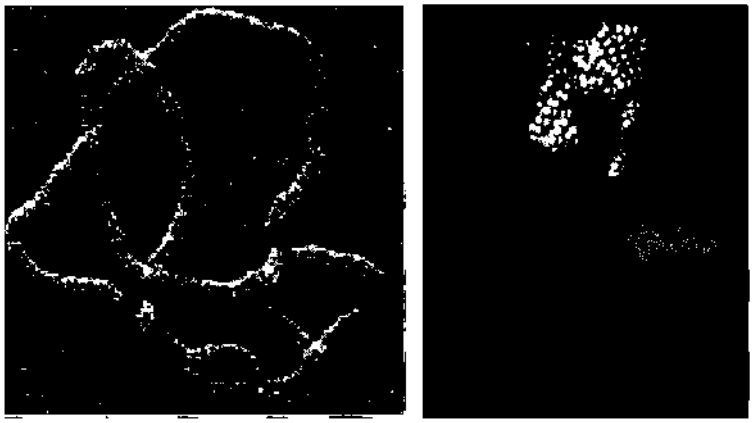


Figure 1. Left: DNA knot. Electron microscope image of a trefoil. Reprinted with permission from [23], S. A. Wassermann and N. R. Cozzarelli, *Science* 229, 171 (1985). © 1985, AAAS. Right: Topoisomerase II. This enzyme can actively change the topology of DNA by cutting the double-strand and passing another segment of double-stranded DNA through the gap before resealing it. The image depicts a short stretch of DNA (horizontally at the bulge of the enzyme, as well as another segment in the lower clamp (perpendicular to the image) after passage through the gap from the upper clamp. This mechanism makes sure that no additional strand passage through the open gap can take place [21, 22]. Figure courtesy James M. Berger, University of California at Berkeley.

the separation of chromosomes during mitosis, powered by molecular motors moving along the microtubules of the mitotic spindle [36].

By probing biopolymers with single molecule techniques, one can study topological effects on single biopolymers and their mutual interaction. This knowledge returns important information to molecular biology and will be helpful in medicinal and biotechnological applications. Being a domain of forefront research, experiment and theoretical modeling are being developed hand in hand. In this chapter, we present some recent results from the biological physics study of topological effects in biopolymeric systems.

2. POLYMERS AND BIOPOLYMERS

Although on a coarse level both man-made polymers and biopolymers essentially behave the same, from a microscopic perspective they differ considerably in respect to their chemical structure, physical properties, as well as means to probe their structure and behavior. In this section, we collect some of the basic properties of both kinds of macromolecules, and point out the importance of biopolymers in nature, science, and applications. In an appendix at the end of this review, we introduce some basic concepts about polymers.

Before proceeding, a few words concerning our nomenclature are in order. We call man-made polymers all those macromolecules that do not occur naturally in biological systems. All molecules consisting of many nucleotides and amino acids occurring naturally or designed in laboratory will be referred to as biopolymers. In a loose sense, we also refer to shorter, oligomeric chains with the same building-blocks, like those often employed in *in vitro* experiments, as biopolymers. We thereby distinguish biopolymers from other higher order biomolecules such as sugars, as detailed below. In our definition, we explicitly disregard low-molecular biological molecules such as simple lipids, as they will not be relevant in the following.

2.1. Man-Made Polymers

The most common man-made polymers: polyethylene, polystyrene, polymethyl acrylate (Plexiglass), polyester, polyamide, and many more, as well as rubbers, have become indispensable in modern life [37–39]. Chemical industry has developed refined synthesis methods, and can achieve degrees of polymerization (chain lengths) of 10^5 to 10^6 monomers through polymerization, polycondensation, or polyaddition. To achieve such lengths, "parasitic" reactions occurring uncontrolled during the synthesis have to be suppressed. Such parasitic reactions lead to branching (i.e., in a polyethylene chain at certain points a Y-shaped branching

occurs). It should be noted that for some purposes branching is actually desired, for example, for stars, comb-like polymers, or polymer brushes, however, at well-defined points. A drawback of chemical polymer synthesis is the relatively large polydispersity, that is, the variation in resulting chain lengths. For purposes requiring a narrow polydispersity, physical methods have to be applied to obtain a narrow distribution of degrees of polymerization [5, 40–43].

Typically, man-made polymers are rather flexible, their persistence length being of the order of 1 nm, which corresponds to some 5 monomeric units, that is, the monomer-monomer distance is of the order of 2 Å. The structure of man-made polymers is generally determined through bulk measurements, mainly by neutron diffraction, small angle X-ray diffraction, NMR, and (inelastic) light scattering. The mechanical response of polymer melts, gels, and rubbers is obtained from macrorheology experiments. Due to the time-temperature superposition used to construct mastercurves, the accessible frequency range of such rheological experiments can cover many orders (often 10 to 15) of magnitude [42, 44–47].

2.2. Biomolecules

Biomolecules can be grouped into unspecific molecules and biopolymers in the stricter sense [1, 5]. *Unspecific* molecules are produced by biological organisms in a large range of molecular weight and structure, such as polysaccharides (cellulose, chitin, starch, etc.), higher fatty acids, actin filaments or microtubules. Also the natural "India rubber" from the *Hevea Brasiliensis* tree, historically important for both industrial purposes and the development of polymer physics [48] belongs to this group.

Biopolymers in the stricter sense comprise the polynucleotides DNA and RNA consisting of the four-letter nucleotide alphabet with A-T and G-C (A-U and G-C for RNA) base pairs, and the polypeptidic proteins consisting of 20 different amino acids, each coded for by 3 bases (codons) in the RNA [1, 2, 5, 50]. Biopolymers are copied and/or created according to the information flow sketched in Fig. 2, the so-called central dogma of molecular biology, a term originally coined by Frances Crick [51]. Accordingly, starting from the genetic code stored in the DNA (in some cases in RNA) DNA is copied by DNA polymerase (replication), and the proteins as the actually task-performing biopolymers are created from DNA via RNA (transcription through RNA polymerase) and further by translation of RNA in ribosomes to proteins. A contemporary view on the process of gene expression, that is, the information flow from DNA (genome) to protein synthesis is depicted in Fig. 3. Note that the transport through the nucleus membrane does not occur in prokaryotic cells, in which the genome is more or less free in the cell. Figure 2 does not show the various proteins involved in the individual steps, such as DNA and RNA polymerases, single-strand binding proteins preventing formation of secondary structure of the exposed ssDNA, and so forth. Most of these steps are represented in Fig. 3 [52]. In all these copying processes, the biochemistry guarantees exact copies of the three types of biopolymers, with extremely low error rates. This exactness essentially is due to the key-lock principle of base-pairing, that is, the exclusive pairing of bases into AT and GC (AU and GC for RNA) [1–3, 50].

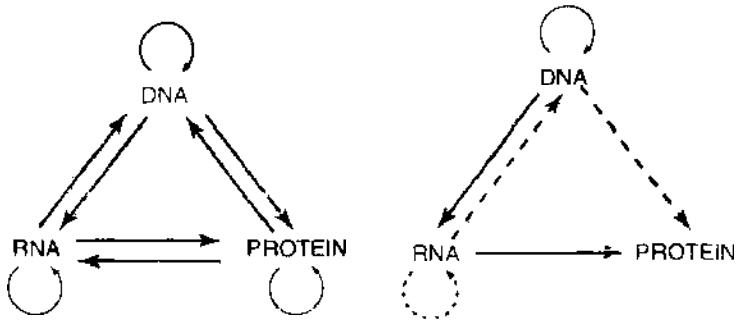


Figure 2. Central dogma of molecular biology according to F. Crick [49]: Potentially, information flow is completely symmetric between the three levels of cellular biopolymers (DNA, RNA, proteins) as shown on the left. However, the recognized pathways are only those included on the right, where solid lines represent probable transfers, and dotted lines possible transfers [49]. Compare to the modern view of gene expression in eukaryotes reproduced in Fig. 3. Reprinted with permission from [49], Crick, *Nature* 227, 561 (1970). © 1970, Nature Publishing Group

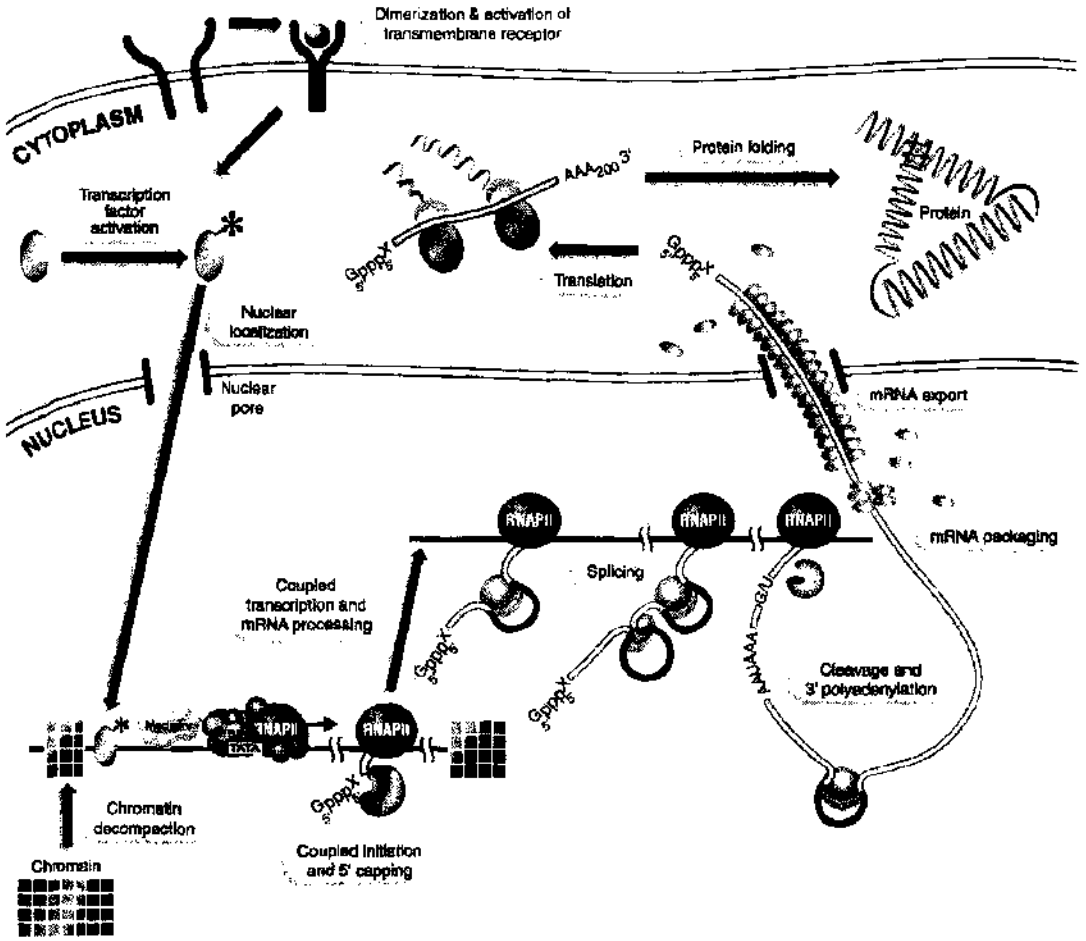


Figure 3. Contemporary view of gene expression, from Ref. [52]. In a continuous process, the information flow from the gene through transcription and translation is regulated by activation processes through regulatory proteins or external signals. Reprinted with permission from [52], Orphanides and Reinberg, *Cell* 108, 439 (2002). © 2002, Elsevier.

DNA can be copied in vast numbers by the polymerase chain reaction (PCR) procedure, as sketched in Fig. 4. PCR uses the melting of dsDNA at elevated temperatures to denature a strand of dsDNA, producing two strands of ssDNA. By lowering the temperature in a solution of invariable primers and single nucleotides, each ssDNA is completed to dsDNA by

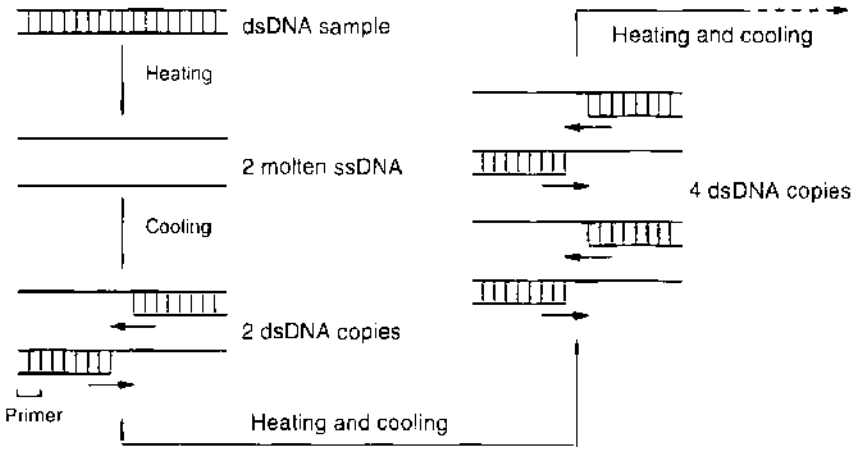


Figure 4. Polymerase chain reaction. By heating and melting of double-stranded DNA into single-strand and successive cooling in a solution of nucleotides and primers, each cycle doubles the amount of double-stranded DNA copies in a highly precise manner.

the key-lock principle of base-pairing [53, 54]. By many such cycles, of the order of 10^9 copies of the original DNA can be produced within the range of hours.[2] Again, the error rate due to the underlying biochemistry can be considered negligible for most purposes. In particular, from the viewpoint of polymer physics/chemistry, the obtained sample is monodisperse and free of parasitic reactions, creating ideal samples for physical studies. This is even more so as any designed sequence of bases can be custom-made in modern molecular biology labs [50].

2.3. The Hierarchical Structure of Biopolymers

All three types of biopolymers exhibit a three-tiered hierarchical structure [1, 5, 55–57]. Given a specific, one-dimensional sequence of nucleotides or amino acids (primary structure), local structural elements come about through the nature of the specific bindings (secondary structure). Volume interactions of chemically remote segments give rise to the global three-dimensional configuration, the tertiary structure. Finally, for proteins one also defines a quarternary structure. We present here some details for the different biopolymers.

2.3.1. The Nucleic Acids: DNA and RNA

DNA is made up of the four bases [1–3, 5, 27, 24]: A(denine), G(uanine), C(ytosine), and T(hymine) that form the DNA double-helix shown in Fig. 5. These building-blocks A, G, C, T base-pair according to the key-lock principle as A-T and G-C, where the AT bond is weaker than the GC bond in terms of stability. Apart from the Watson-Crick base-pairing energy, the stability of dsDNA is effected by the stacking interactions, the specific matching of subsequent base-pairs (bp) along the double-strand, that is, bp-bp interactions. In standard literature, the stacking interactions are listed for pairs of bps (e.g., for AT-GC, AT-AT, AT-TA, etc.) [58]. In Fig. 6, we show the three most abundant DNA configurations A, B, and Z. Based on the AGCT alphabet, the primary structure of DNA can be specified. DNA's six local structural elements twist, tilt, roll, shift, slide, and rise are effected by the stacking interactions between vicinal bps. In Fig. 7, we show a map with the detailed structure elements of the *E. coli* genome. These structural elements also define the three-dimensional shape of DNA within a typical correlation (persistence) length of about 150 base pairs corresponding to 500 Å (i.e., the bp-bp distance measures 3.4 Å, reflecting the rather complex chemical structure of a bp in comparison to the monomer size of man-made polymers) [6, 59–61]. On a larger scale, much longer than the persistence length, DNA becomes flexible. On this level, the tertiary structural elements come into play by so-called looping, that is, the formation of chemical bonds between specific bps which are remote along the DNA backbone. Moreover, dsDNA may locally open into floppy ssDNA bubbles, with a persistence length of about 8 bases or 40 Å. These fluctuation-induced bubbles increase their statistical weight at higher temperatures, until the dsDNA fully denatures (melts). We will come back to DNA denaturation bubbles below. Depending on the external conditions, DNA occurs in several configurations. Under physiological conditions, one is concerned with B-DNA, but there are other states such as A, B′, Z, ps, triplex DNA, quadruplex DNA, cruciform, and H; reviewed, for instance, in [59], compare also [7]. DNA occurs naturally in a large range of length scales. In viruses, DNA is of the order of a few μm long. In bacteria, it already reaches lengths of several mm, and in mammal cells it can reach the order of a few m, roughly 2 m in a human cell and 35 m in a cell of the South American lungfish [2]. Entire DNA in bacteria in vivo, or extracted from bacteria and higher cells for our purposes can therefore be viewed a fully flexible polymer with a persistence length of roughly 50 nm, being governed by generic effects independent of the detailed sequence.

[2] Most proteins denature at temperatures between 40 °C to 60 °C, including polymerases. In early PCR protocols, after each heating step, new polymerase had to be washed into the reaction chamber. Modern protocols make use of heat-resistant polymerases that survive the temperatures necessary in melting. Such heat-resistant proteins occur, for instance, in bacteria dwelling near undersea thermal vents.

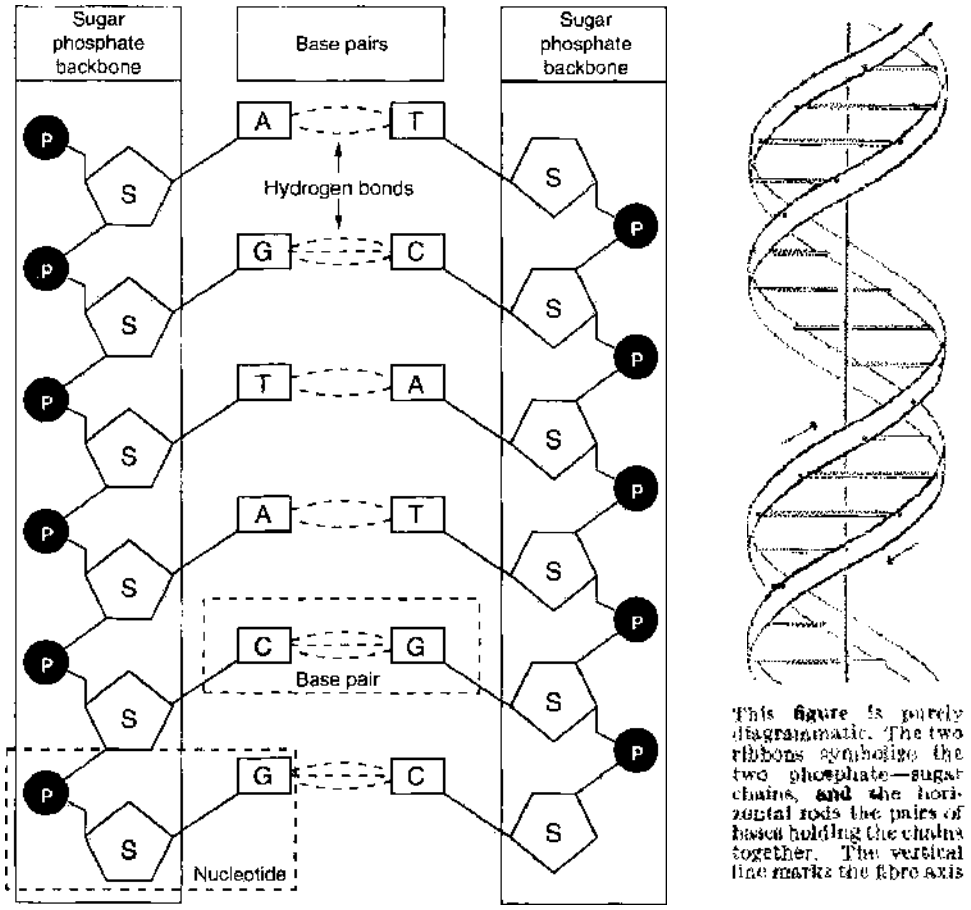


Figure 5. Ladder structure of the DNA formed by its four building-blocks A, G, C, and T, which gives rise to the typical double-helical structure of DNA. A-T base-pairs establish 2 H-bonds, G-C base-pairs 3 H-bonds. Right image from [25] J. D. Watson and F. H. C. Crick, *Nature* 171, 737 (1953). © 1953. Nature Publishing Group.

The stacking interactions govern the local structure of dsDNA. Globally, an additional constraint arises due to the circular nature of the DNA, since it has to satisfy the conservation law [62–64]

$$\mathrm{Lk} = \mathrm{Tw} + \mathrm{Wr} \tag{1}$$

where Lk stands for the linking number, Tw for the twist, and Wr for the writhe of the double helix. Lk is an integer number that is formally given by one-half the number of signed

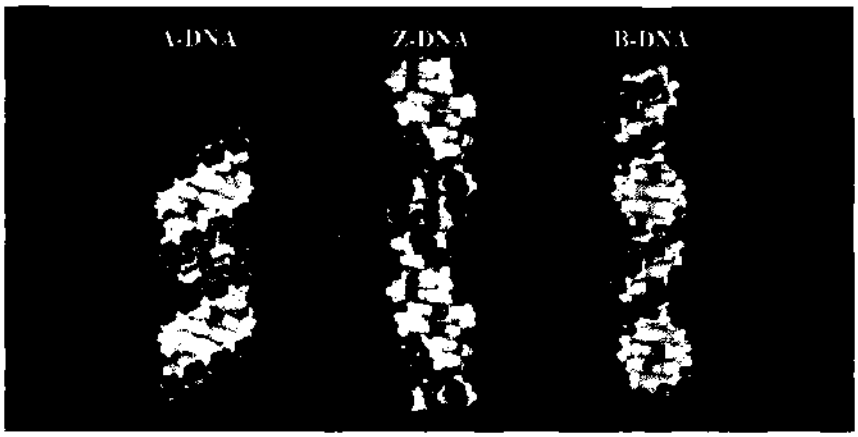


Figure 6. The three DNA configurations labeled A, B, and Z. A-DNA occurs as a dehydrated form of DNA, the A-form is also the preferred conformation of RNA. Z-DNA can occur at stretches of alternating G and C bases and may be important for controlling replication. By far the most important physiological conformation is B-DNA [1, 6–8, 59].

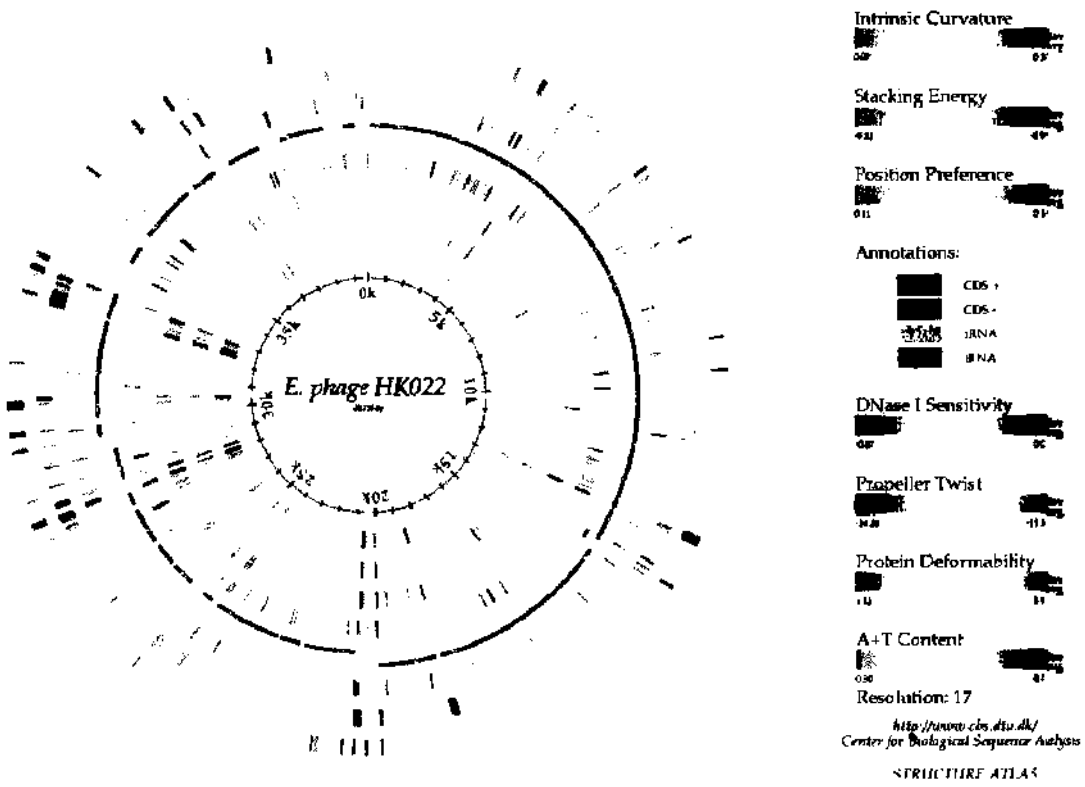


Figure 7. Structure atlas of the complete E.coli genome. Figure courtesy David Ussery, Technical University of Denmark. The structure atlas is available at `www.cbs.dtu.dk/services/GenomeAtlas/`.

crossings of one DNA strand with the other in any regular projection of the molecule. The linking number Lk is a topological property, and no deformation of a closed DNA without breaking and rejoining the DNA strands will alter it. Tw is equal to the number of times that the two strands of DNA wind about the curvilinear central axis of the molecule, and Wr is a number whose absolute value equals roughly the number of times that the DNA axis winds about itself. Whereas Tw is a property of the double-helical structure of DNA, Wr is a property of the DNA axis alone. Tw and Wr do not need to be integers and are not conserved, but coupled through Lk by Eq. (1). A nicked circular DNA, that is, when the twist can fully relax, carries $Lk_0 = N/h$ links, where N is the number of bp and h ($h \simeq 10.5$ in B-DNA) the number of base pairs per turn.

The DNA of virtually all terrestrial organisms is underwound or *negatively supercoiled*, i.e., $\Delta Lk = Lk - Lk_0 < 0$.[1] Often, the superhelix density $\sigma \equiv \Delta Lk/Lk_0$ is used. There are two forms of intracellular supercoiling, the *plectonemic* form, characteristic of plasmid DNA and accessible, nucleosome-free regions of chromatin, and the *toroidal* or *solenoidal* form, where supercoiling is attained by DNA wrapped around histone octamers or prokaryotic non-histone DNA-binding proteins. The former is the active form of supercoiled DNA and is freely accessible to proteins involved in transcription, replication, recombination and DNA repair. The latter is the stored form of supercoiled DNA and is largely responsible for the extraordinary degree of compaction required to condense typical genomes into the cell's nucleus. Negative supercoiling facilitates the local unwinding of DNA by providing a ubiquitous source of free energy that augments the unwinding free energy accompanying the interactions of many proteins with their cognate DNA sequences. The local unwinding of DNA, in turn, is an integral part of many biological processes such as gene regulation and DNA replication (see Section 4). Therefore, understanding the interplay of supercoiling and local helical structure is essential to the understanding of many biological mechanisms [61, 65–68].

[1] An exception are thermophilic organisms living near undersea geothermal vents that have positively supercoiled DNA in order to stabilize the double helix at extreme temperatures.

RNA consists of the same building blocks as DNA, with the exception that T(hymine) is replaced by U(racile). RNA typically occurs in single-stranded form. Therefore, its secondary structure is richer, being characterized by sequences of hairpins: Smaller regions in which chemically remote sequences of bases match, pair and form hairpins which are stiff and energy-dominated, similar to dsDNA. The remaining regions form entropy-dominated floppy loops, analogous to the ssDNA bubbles. In Fig. 8, we show such an RNA secondary structure, along with its connector diagram, in which the nonlocal nature of the bonds becomes obvious. The tertiary structure in RNA comes about by the formation of so-called pseudoknots, chemical bonds established between bases sitting on chemically distant segments of the secondary structure. In RNA-modeling the incorporation of pseudoknots is a non-trivial problem, which currently receives considerable interest; see, for instance [69–71].

2.3.2. The Polypeptides: Proteins

Proteins consist of the 20 amino acids A = Ala(nine), C = Cys(teine), D = Asp(aragine), E = Glu(tamine), F = Phe(nylalanine), G = Gly(cine), H = His(tidine), I = I(so)le(ucine), K = Lys(ine), L = Leu(cine), M = Met(hionine), N = As(paragi)n(e), P = Pro(line), Q = Gl(utami)n(e), R = Arg(inine), S = Ser(ine), T = Thr(eonine), V = Val(ine), W = Tr(y)p(tophan), and Y = Tyr(osine) [55–57]. These amino acids constitute the linear primary structure, that is, the sequence. In contrast to DNA and RNA, which are largely responding to thermal fluctuations and therefore incessantly change shapes, proteins in their native folded state become fairly compact objects. These are, however, distinguished by the following structural elements.

The secondary structure of polymers comprises α-helices and β-sheets, as shown in Fig. 9. Not all amino acids are grouped in α-helices or β-sheets, but function as "hinges" holding different α-helices or β-sheets together.

The compact three-dimensional structure of a protein depends on a complex interaction pattern between secondary structure elements, also between chemically distant parts, building up the tertiary structure. Such protein structures are generally represented either by ribbons or by balls. Several individual amino acid chains can bind together in a specific way, apparently forming one protein. Such agglomerations are often termed quarternary structures. We show these two volume structures in Fig. 10.

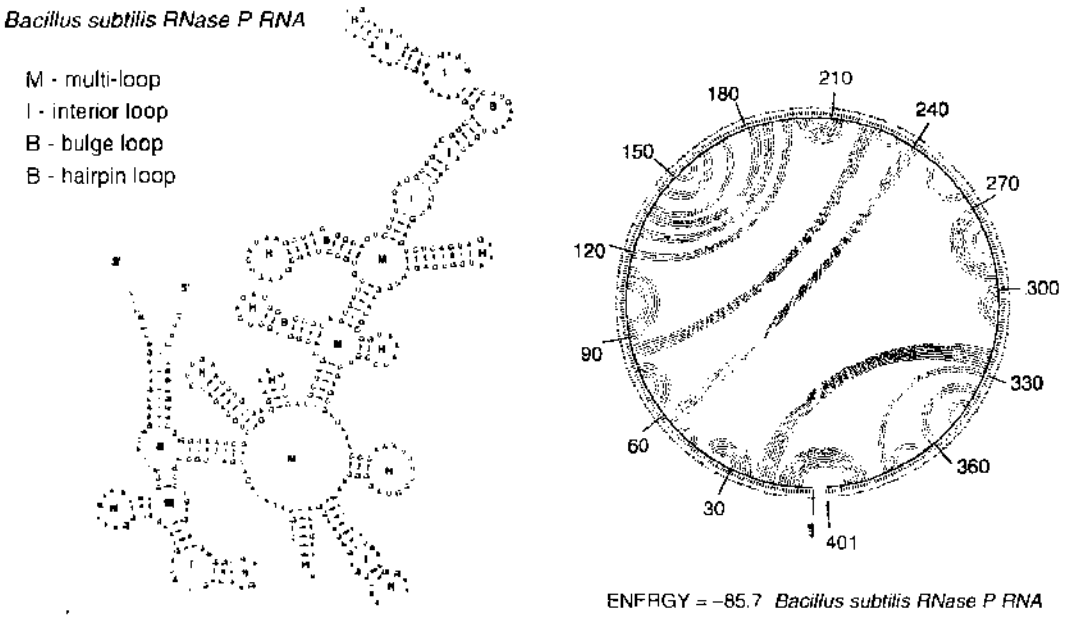


Figure 8. Left: Secondary structure of *Bacillus subtilis* without a pseudoknot. On the right the corresponding connector diagram. A pseudoknot would correspond to a crossing between connectors in this diagram. Figures courtesy Michael Zuker, Rensselaer Polytechnic Institute.

Figure 9. Protein secondary structural elements: α-helix and β-sheet in ribbon representation. Left: Two α-helic structure of Alanine 31 proline mutant of rop protein from E. coli (1B6Q from the protein data bank) drawn with Rasmol. Right: β-sheet structure of the protein concanavalin A (1APN).

2.4. Interacting Biopolymers: From Central Dogma to Breakdown Services

The information flow according to the central dogma sketched in Fig. 2 is propelled by certain classes of proteins involved in accessing, reading and copying/printing this information. During replication of the DNA, additional single-stranded DNA binding (SSB) proteins ensure that the temporarily exposed ssDNA does not form secondary structure, but stays open for new bases to bind during the formation of the DNA duplicate. Additional protein-DNA interactions occur incessantly *in vivo*. Repair proteins and topoisomerases detect structural defects, undesired superstructure (curling up of over/underwound dsDNA), or entangled states, and repair these. Regulatory proteins bind nonspecifically along the entire DNA, or specifically at certain target sites, to activate or inhibit binding of polymerase. Other proteins bind to a specific protein, ratcheting its passage through a nanopore, and so forth [1, 8, 50]. Here, we briefly outline their biochemical background.

2.4.1. Polymerases and Helicases

Polymerases can initiate a DNA chain *de novo*. Respectively, DNA and RNA polymerases are involved in creating DNA and RNA. DNA is always synthesized from the 5′ to the 3′ direction (of the copy). Thus, when replicating DNA after the protein helicase unzips the

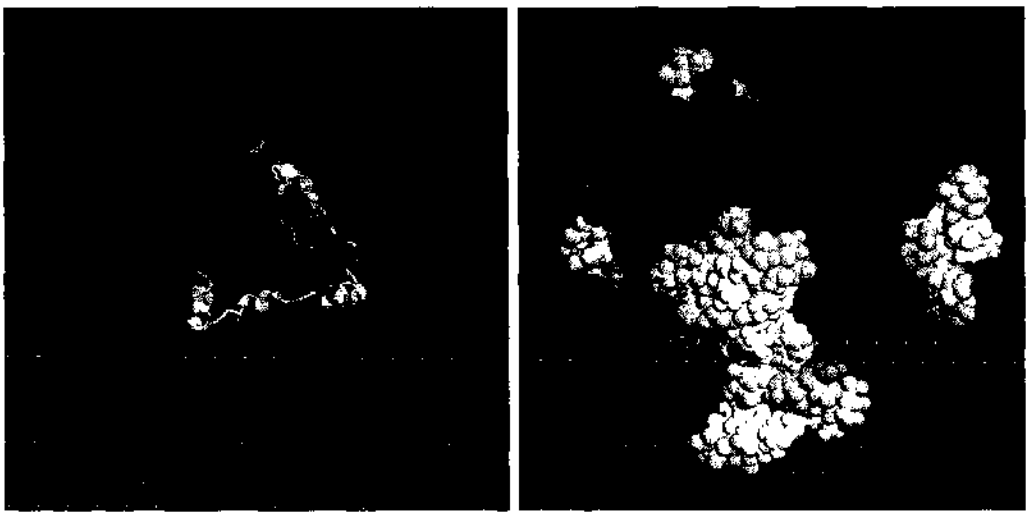


Figure 10. *Left:* Tertiary structure of the single-stranded DNA binding protein gp32 (1GPC) containing both α-helices and β-sheets. *Right:* Quarternary structure showing the individual amino acid chains building up the R6 human insulin hexamer (1A10 from protein data bank) in space-filling representation.

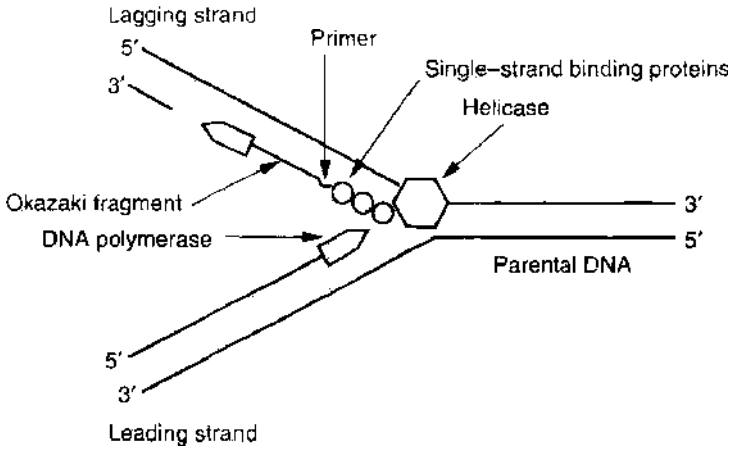


Figure 11. Replication fork. On the lagging strand, polymerase forms Okazaki fragments, starting from a primer catalyzed by primase. SSB proteins prevent the formation of secondary structure in the lagging single-strand.

dsDNA, the leading strand can immediately be complemented by new bases through DNA polymerase, whereas on the lagging strand, patches of bases are initiated by RNA primase and extended to the Okazaki fragments by DNA polymerase. During the entire process, SSB proteins hold the two strands open. In fact, two replication forks move in opposite direction, and create a replication bubble. This quite complex process is sketched in Fig. 11 [1, 8, 50].

Similarly, a transcription bubble is created, which allows access to the genetic code, read out by RNA polymerase. Here, however, only one of the two copies is used to create the single-stranded messenger RNA. RNA polymerase is guided to a promoter site of the gene to be transcribed by regulatory proteins. It starts unwinding the DNA (bubble initiation) at an AT-rich zone, the so-called TATA box [1, 8, 50].

2.4.2. Topoisomerases

These are enzymes that cut and reform DNA. Topoisomerases are grouped according to their tasks: Topoisomerase I cuts, and later reseals, a single strand of a dsDNA. During replication, it relieves rotational strain built up by helicase during separating and unwinding of the helix. (As shown in Fig. 18, under nonphysiological concentrations, it also acts on a nicked dsDNA and introduces knots.) Topoisomerase II relieves the stress when two chromosomes are tangled under attempted separation during mitosis. It cuts both strands of one chromosomal dsDNA, passes the other chromosome through the gap, and then rejoins the cut ends. Remarkably, topoisomerase II can also reduce the knottedness of a single circular dsDNA by passing one segment through another [1, 72, 73]. We show an image of a topoisomerase II in Fig. 1.[4] Apart from DNA topoisomerases, there also exist RNA polymerases that are able to modify the knottedness of RNA.

2.4.3. Other Enzymes

There exist a large number of other enzymes that play important roles in the regular events connected to DNA replication, gene expression and regulation, as well as in the large number of different possible events when something goes wrong in such processes. Thus, repair proteins can correct mismatches during replication or frame-shift mutations, when the reading frame of a gene is changed by insertion or deletion of one or more nucleotides in the coding region; see, for instance, references [1, 50, 52] for details. In general, repair proteins therefore are a necessary ingredient in cancer suppression and prevention of premature aging.

[4] Topoisomerases with higher order number work according to the same principles. Thus, topoisomerase IV occurs in *E. coli* and functions like topoisomerase II [4, 17, 22].

2.5. Single Molecular Manipulation of Biopolymers

In recent years, a number of techniques to manipulate individual molecules and biomolecules have come of age, for example, atomic force microscopes (AFM), optical or magnetic tweezers and traps, hydrodynamic drag, and biomembrane force probes. These devices make it possible to probe the molecular interactions of individual biomolecules by their response to mechanical stress (see [74–77] for reviews). The systems studied by single molecule pulling experiments can be divided in two groups: in *rupture* experiments, receptor and ligand molecules are attached to a substrate and a transducer, respectively, typically via chemical linkers. After allowing receptors and ligands to bind, the transducer, for example, the cantilever of an AFM, is pulled away, which causes the receptor-ligand pairs to rupture. The maximum force the molecule can withstand has been measured in this way for biotin and streptavidin [78, 79], and many other receptor-ligand pairs [80]. Second, *un- and refolding* experiments probe the elastic properties of an individual biomolecule. The molecule is attached between a substrate and a transducer, again via chemical linkers. Force-extension relations are then obtained by measuring the force as a function of the position of the transducer. Investigated systems include DNA [81–83], RNA [84, 85], polysaccharides [86], the muscle protein titin [87], the membrane protein bacteriorhodopsin [88], and many other proteins [89].

In parallel, optical methods based on fluorescence have emerged to study the dynamics of biochemical processes, for example, by energy transfer and evanescent waves. Mechanical and optical methods deliver complementary information and can also be combined in a single experimental setup.

In general, above techniques offer the possibility to monitor single molecular events, and therefore to avoid ensemble averages, which may mask effects due to signal broadening. An additional advantage is the possibility to record the noise around the signal, which allows to extract information about the response of the biomolecule to, for instance, thermal noise. The expense may lie in a somewhat poor data resolution in comparison to bulk techniques. In the course of developing single molecule techniques this disadvantage has been successively reduced, so that presently single molecule techniques for the majority of cases have become the method of choice.

2.5.1. Optical and Magnetic Tweezing, and Atomic Force Microscopy

Optical tweezers use the force exerted by a strongly focused laser beam to trap and thereby move or hold dielectric objects of the size of a few tens of nanometers to some tens of micrometers, due to the high field gradient near the beam focus [77, 90, 91]. Usually, biopolymers to be measured are attached to latex/polystyrene microbeads, which in turn are large enough to be trapped and recorded in an optical tweezer. The beads' elongation can be measured by photodiode detectors, or through a digital camera. Forces transferred to the beads range from one to some 150 pN, and above. Optical tweezers were also used to actually tie knots into actin filaments [92] and DNA [93].

Magnetic tweezers use small magnetic beads instead of the dielectric ones. In such a setup it is possible to exert a torque on the bead, which can then be used, inter alia, to twist an attached DNA double helix [77]. To this end, the DNA molecule is attached at one end to a surface and at the other to a magnetic microbead. The microbead is pulled and rotated by small magnets whose position and rotation can be controlled. In this way one may apply and measure forces over a large range, from 10^{-3} pN to about 100 pN with a relative accuracy of about 10%. This method is absolute and does not require a calibration of the transducer. An advantage of magnetic tweezers is that measurements on DNA at constant force are realized just by maintaining a fixed distance between the magnets and the sample. Moreover, magnetic traps allow simple twisting of the molecule by rotating the magnets. However, because its stiffness is a function of the force, the spatial resolution of magnetic traps is limited to about 10 nm, lower than other manipulation methods at weak forces <1 pN. An interesting example is the study by [94] in which magnetic tweezer beads are rotated, creating supercoiling in the DNA, which can then be monitored to relax stepwise by action of topoisomerase II.

Atomic force microscopes work by optically measuring the elongation of a fine cantilever of a typical size of some 100 μm and spring constant of about 0.1 N/m. It can achieve resolutions of the order of 10 pN. While in probing mode, a cantilever is moved across a surface (or periodically touching down on the surface in tipping-mode) and its response recorded. Apart from monitoring forces, atomic force microscopes may also be used to attach single molecules to (or remove from) a surface [95].

2.5.2. Reconstruction of Molecular Free Energy Landscapes From Nonequilibrium Data

Figure 12 shows a typical setup of single molecule pulling experiments together with a generic free energy profile $G(z)$. The molecule is attached between the substrate surface and the cantilever tip. The position of the cantilever $x(t)$ is moved according to a prescribed experimental protocol. The extension of the molecule is described by a suitable reaction coordinate z given by the position of the cantilever tip. For fixed extension z and time t, the energy of the molecule perturbed by the cantilever spring is given by the time-dependent Hamiltonian

$$H(z,t) = G(z) + V_0(z, x(t)) \equiv G(z) + \frac{k}{2}(x(t) - z)^2 \tag{2}$$

where $G(z)$ is the free energy profile of the unperturbed molecule. The second term describes the external force acting on the molecule in terms of a harmonic potential with effective spring constant k. Because the molecule is coupled to a heat bath at temperature T, the time evolution of z is stochastic.

Traditionally, the cantilever is moved according to a linear ramp,

$$x(t) = x_0 + vt \tag{3}$$

with offset x_0 at $t = 0$ and constant velocity v, and the force $F(t, v)$ acting on the cantilever is recorded. The challenge is to recover from these data the unperturbed molecule's free energy profile $G(z)$ containing the desired information about the molecular interactions. Evans and Ritchie first pointed out that the rupture force of receptor-ligand pairs depends on the loading rate v [96]. Thus, by combining the data $F(t, v)$ for a broad spectrum of loading rates v, referred to as *dynamic force spectroscopy*, important features of $G(z)$ can be determined such as the distance between the minimum and maximum of an energy barrier for rupture [96]. Heymann and Grubmüller refined this technique and obtained the heights and positions of the maxima of a molecular force profile $\partial_z G(z)$ with high spatial resolution [97].

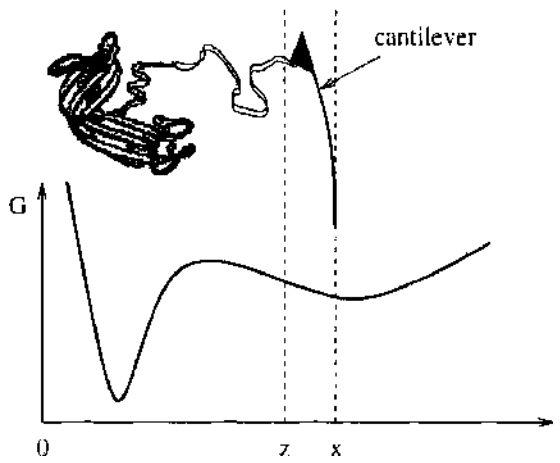


Figure 12. Schematic view of a typical setup of single molecule pulling experiments and a generic free energy potential $G(z)$. The first minimum represents the folded state, whereas the second shallow minimum represents the unfolded state of a biopolymer. The coordinate x denotes the position of the cantilever and z the position of the cantilever tip to which one end of the biopolymer is attached. Reprinted with permission from [104], O. Braun et al., *Phys. Rev. Lett.* 93, 158105 (2004). © 2004, American Physical Society.

However, pulling protocols in real experiments typically generate non-equilibrium data from which one has to recover the equilibrium profile $G(z)$. The difficulty to recover equilibrium properties from nonequilibrium data may be resolved by using a recent advance in nonequilibrium statistical mechanics due to Jarzynski [98], according to which the equilibrium profile $G(z)$ can be inferred by suitably averaging nonequilibrium trajectories $\zeta(t)$ of the reaction coordinate z [99]. Jarzynski's relation, in general, states that the free energy difference ΔG between two equilibrium states can be extracted from averaging the work W required to drive the system from one state to the other according to [98]

$$e^{-\beta\Delta G} = \lim_{N\to\infty} \left\langle e^{-\beta W} \right\rangle_N \tag{4}$$

with $\beta = 1/k_B T$ the inverse temperature. The generalization from two states to a z-resolved free energy profile $G(z)$ perturbed by a harmonic spring (2) reads [99]

$$e^{-\beta[H(z,t)-G_0]} = \lim_{N\to\infty} \left\langle \delta[z-\zeta(t)]\, e^{-\beta W(t)} \right\rangle_N \tag{5}$$

The average $(\ldots)_N$ is over N realizations $\zeta(t')$, $0 < t' < t$, of the stochastic trajectory of the biopolymer's end position, starting in equilibrium at $t' = 0$ and ending at the given position z at $t' = t$ as enforced by the delta function. The external work is a functional of $\zeta(t')$ and given by

$$W(t) = \int_0^t dt'\, \partial_\tau H(\zeta(t'), \tau)|_{\tau=t'} \tag{6}$$

In Eq. (5), the constant

$$G_0 = -k_B T \ln\left(\int_0^\infty dz' \frac{e^{-\beta H(z',0)}}{\lambda_T} \right) \tag{7}$$

is the free energy in the initial state at $t = 0$, where the thermal wavelength λ_T serves for normalisation. Summing up the normalised distributions obtained from Eq. (5) at each time slice with the method of weighted histograms [100] yields the reconstruction formula for the unperturbed free energy profile of the molecule

$$G(z) = -\beta^{-1} \ln \sum_t \frac{\langle \delta(z-\zeta(t)) \exp(-\beta W(t)) \rangle}{\langle \exp(-\beta W(t)) \rangle} \Big/ \sum_t \frac{\exp[-\beta V_0(z, x(t))]}{\langle \exp(-\beta W(t)) \rangle} \tag{8}$$

Recently, Jarzynski's equality (4) has been tested by mechanically stretching a single molecule of RNA reversibly and irreversibly between two conformations in an experiment by Liphardt et al. [101]. The underlying free energy landscape $G(z)$ was recovered from non-equilibrium data to within $k_B T/2$ of its best independent estimate, the mean work of reversible stretching (Figs. 13 and 14). In the experiment by Liphardt et al. (4), individual RNA molecules were attached to microbeads by RNA-DNA hybrid handles and DNA-bead chemical linkers (Fig. 13 B). One bead was held in a force-measuring optical trap and the other one was linked to a piezoelectric actuator through a micropipette tip. Molecules were stretched by moving the tip bead and the force acting on the RNA was measured. From the data the performance of three different estimates for $\Delta G(z)$ were compared, namely the average work $W_A = \langle W \rangle_N$, the fluctuation-dissipation estimate $W_{FD} = \langle W \rangle_N - \beta\sigma^2/2$ where σ is the standard deviation of the work distribution [102], and the estimate obtained from Jarzynski's equality (4), $W_{JE} = -\beta^{-1} \ln\langle \exp(-\beta W) \rangle_N$ (Fig. 14). An RNA molecule was switched N times between the folded and unfolded conformations, and then values for W_A, W_{FD}, and W_{JE} were computed. Seven independent data sets were collected, each with a different RNA molecule and about 40 unfolding-refolding cycles per switching rate. Data for unfolding-refolding rates of 2 to 5 pN/s, 34 pN/s, and 52 pN/s are shown in blue, green, and red, respectively, in Fig. 14. In contrast to W_A and W_{FD}, application of Jarzynski's equality (dashed lines) recovers $G(z)$ to within experimental error ($\pm k_B T/2$) in the whole range from $z = 0$ to 30 nm.

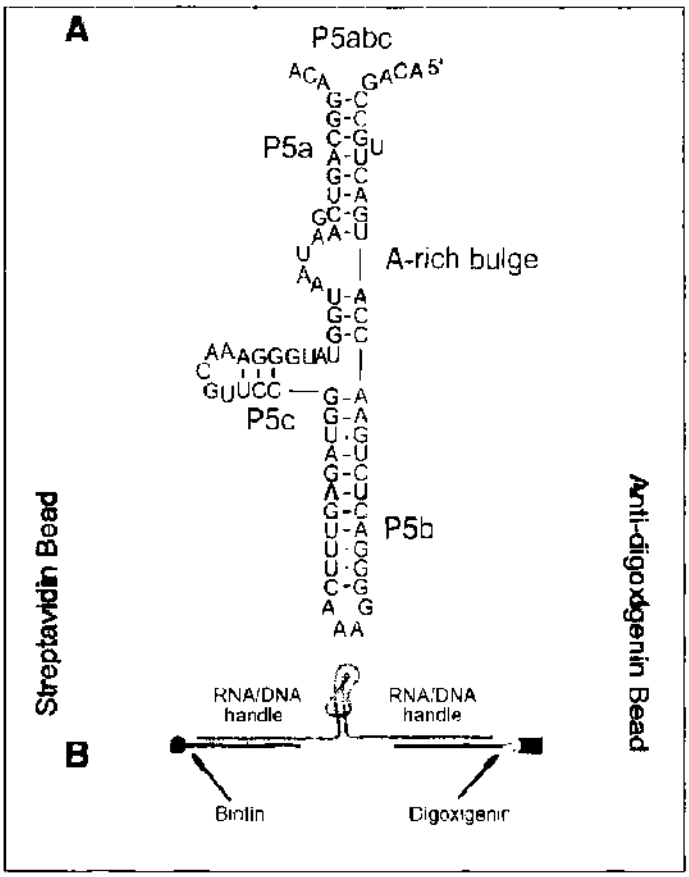


Figure 13. (A) Sequence and secondary structure of the RNA molecule used in the experiment by Liphardt et al. [101]. (B) RNA molecules were attached between two beads with RNA-DNA hybrid handles. Reprinted with permission from [101]. J. Liphardt et al., *Science* 296, 1832 (2002). © 2002, AAAS.

2.5.3. Periodic Loading

Traditional experimental protocols like the linear ramp in Eq. (3) still entail certain drawbacks. First of all, the thermodynamically unstable (concave) barrier regions of $G(z)$, determined by specific molecular interactions and therefore of particular interest, are poorly sampled due to snapping motion and thus hard to determine [97]. In order to improve the quality of data obtained by single molecule pulling experiments, in [104] a new method for obtaining the full free energy profile $G(z)$ has been proposed by introducing a *periodic* ramp, that is,

$$x(t) = x_0 + a\sin(\omega t) \tag{9}$$

with given offset x_0, amplitude a, and frequency ω. Figure 15 shows that periodic loading delivers significantly more accurate data for the sample free energy profile $G(z)$ shown in Fig. 12 than the linear ramp in Eq. (3), under the constraint of equal experimental effort. The improvement of the quality of data in the important barrier region of $G(z)$ around $z \simeq 4$ nm is striking. The better performance of the periodic loading method as compared to linear loading is mainly due to the fact that periodic loading ensures that the barrier region is traversed often and from both sides. The quality of the reconstruction using a periodic ramp moreover depends crucially on the fact that the barrier region is sampled under nonequilibrium conditions taking advantage of Jarzynski's relation to recover the equilibrium profile [98, 99]. Driving the system out of equilibrium is important, as under quasistatic conditions an efficient sampling of the barrier region (where $G(z) \gg k_B T$) is inhibited by the equilibrium Boltzmann factor $\exp[-G(z)/(k_B T)] \ll 1$. For periodic loading, the freedom to choose the frequency ω large enough allows one to probe the same region under nonequilibrium conditions, thus overriding the exponential punishment by the equilibrium Boltzmann factor. The optimal frequency arises from balancing competing effects as quantified in a case

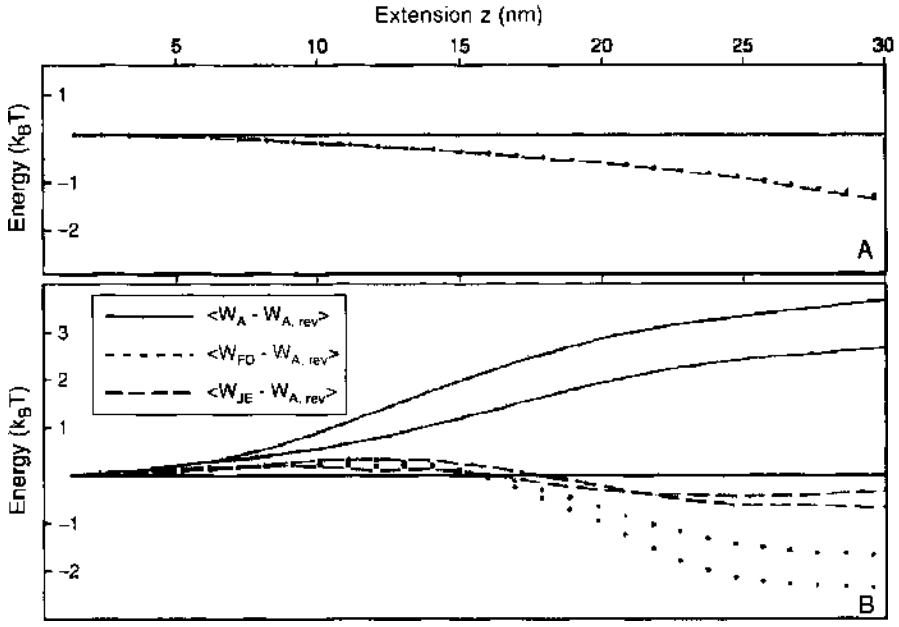


Figure 14. (A) Estimation of free energy profile $\Delta G(z)$ from reversible switching ($r = 2$ to 5 pN/s, blue) by Liphardt et al. [101]. For each of the seven data sets, the mean work W_A, the fluctuation-dissipation estimate W_{FD}, and the estimate W_{JE} from Jarzynski's equality (4) were determined. The mean work of reversible quasistatic switching, $W_{A,rev}$, was subtracted from the above energies and the differences were averaged over data sets, yielding the average dissipated work $W_A - W_{A,rev}$ (solid yellow band, $\pm k_BT/2$), $W_{FD} - W_{A,rev}$ (dotted line), and $W_{JE} - W_{A,rev}$ (dashed line). The estimates W_{FD} and W_{JE} coincide everywhere to within 0.1 k_BT. Both estimates decrease monotonically with extension, and, by $z = 30$ nm, underestimate the average work by about 1.4 k_BT. (B) Estimation of free energy profile from irreversible switching ($r = 34$ pN/s, green, and 52 pN/s, red). Mean energy differences were computed as in (A). Use of the fluctuation-dissipation relation (dotted lines) yields $G(z)$ to within $k_BT/2$ between $z = 0$ and 18 nm. Beyond $z = 18$ nm, however, W_{FD} underestimates $G(z)$ substantially. By contrast, application of Jarzynski's equality (dashed lines) recovers $G(z)$ to within experimental error ($\pm k_BT/2$) in the whole range from $z = 0$ to 30 nm. Reprinted with permission from [101], J. Liphardt et al., *Science* 296, 1832 (2002). © 2002, AAAS.

study below. This frequency should not be too large in order to enable the system to follow the external drive. Moreover, the Jarzynski procedure converges the slower the further one is away from equilibrium [105]. For too small ω, on the other hand, one does not generate enough crossings under the constraint of a finite total measuring time.

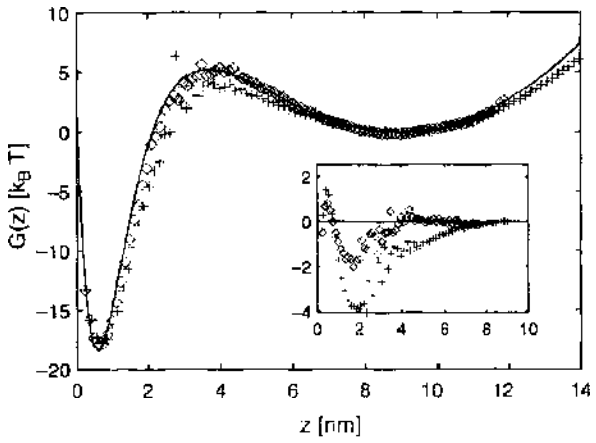


Figure 15. Comparison of reconstructed free energy profiles $G(z)$ by using periodic (◊) *versus* linear loading (+), generated by (8). The solid line is the original free energy profile $G(z)$. For both methods, 10 trajectories of 75 ms length, a spring constant $k = 11.6$ pN/nm, and a diffusion constant $D = 10^{-5}$ cm^2/s were used [see (2) and (10)]. For periodic loading the optimal frequency is $\omega^* = 1.2 \times 10^3$ 1/s (determined in Fig. 16), the preloading offset $z_0 = 6$ nm, and the amplitude $a = 5$ nm in (9). For linear loading the optimal velocity is $v^* = 200$ nm/s (determined in Fig. 16). The inset compares the relative deviation of the reconstructed data from the original profile $G(z)$ corresponding to the zero line. Reprinted with permission from [104], O. Braun et al., *Phys. Rev. Lett.* 93, 158105 (2004). © 2004, American Physical Society.

The periodic loading protocol was tested with simulated experiments and compared with the traditional method of linear loading [104]. To this end a generic free energy profile $G(z)$ was chosen for the unfolding of tertiary structures of biopolymers such as the membrane protein BR [88]. The sample free energy profile has two separated minima, one of which is narrow and deep representing the folded state and one of which is shallow representing the unfolded state, see Fig. 12. For BR, a rich structure of unfolding transitions under force was found [88]. Single force peaks in the unfolding spectra could be allocated to specific changes in molecular configuration. Most of the force peaks scatter between 25 up to 100 pN. With a typical length scale of several nanometres this yields an energy barrier of about 20 k_BT at room temperature. Focusing on one of such transitions, a barrier of 2 nm length and 23 k_BT height was chosen, leading to a typical transition force of about 50 pN in the simulated pulling experiments.

For comparing the periodic with the linear ramp, both kinds of protocols were simulated using typical parameters as given in the caption of Fig. 15. An ensemble of trajectories $\zeta(t)$ of the reaction coordinate z was generated by discretizing the Langevin equation

$$\frac{dz}{dt} = -\gamma^{-1}\frac{\partial H(z,t)}{\partial z} + \sqrt{2D}\,\xi(t) \tag{10}$$

with the Hamiltonian $H(z,t)$ from (2) and a Gaussian random force $\xi(t)$ with zero mean and short-ranged temporal correlations $\langle \xi(t)\xi(t')\rangle = \delta(t-t')$ [104]. The diffusion constant D is related to the friction coefficient γ by the Einstein relation $D = k_BT/\gamma$. The data shown in Fig. 15 were generated by using expression (8).

The overall quality of the data obtained by the periodic ramp is far better than the linear ramp, especially in the barrier region where the data obtained by the linear ramp underestimates the barrier height by several k_BT. In order to ensure an unbiased comparison, the same total number of trajectories and the same total measuring time for both methods was used [102]. In practice, the measuring time of the trajectories, once prepared, is not a limiting factor. The periodic ramp therefore allows measuring as many transitions as necessary to collect the sufficient amount of data. Note in this respect that the periodic protocol requires less equilibration than the previously introduced numerical method of using data both from forward and backward trajectories [106]. In the latter case one has to generate equilibrium states at the beginning of each of these trajectories by waiting sufficiently long at the turning points, whereas with the periodic ramp only the initial state at $t = 0$ has to be equilibrated.

The quality of the reconstructed free energy profile depends crucially on the driving frequency ω for the periodic ramp and the velocity v for the linear ramp, respectively. To quantify this observation the mean square error

$$\sigma^2 \equiv \left\langle \left[\widetilde{G}(z) - G(z)\right]^2 \right\rangle_z \tag{11}$$

of the reconstructed free energy and some characteristic confidence intervals (error bars) for both protocols is shown in Fig. 16. For clarity, $\widetilde{G}(z)$ denotes the reconstructed free energy profile based on (8). For periodic loading, the best results were obtained for the optimal driving frequency $\omega^* \simeq 1.2 \times 10^3$ 1/s, which yields an error of $\sigma^2 \simeq 0.9\ (k_BT)^2$. This frequency is somewhat smaller than the spontaneous transition rate under preloading, which is about 7×10^3 1/s for the model system considered here. Analyzing the work distribution shows that ω^* indeed corresponds to nonequilibrium conditions. For both smaller and larger frequencies than ω^*, the quality of the reconstructed data becomes worse as expected from the reasoning above.

For linear loading, the error increases for increasing driving velocity v, as expected. Since the total measuring time and the total number of trajectories was fixed, the least possible velocity for overcoming the barrier is $v \simeq 100$ nm/s. The smallest error of $\sigma^2 \simeq 1.7\ (k_BT)^2$, however, was observed at a larger, optimal velocity $v^* \simeq 200$ nm/s, with an error bar of 0.03. For the unbiased comparison in Fig. 15, the optimal values ω^* and v^* from the data shown in Fig. 16 were chosen.

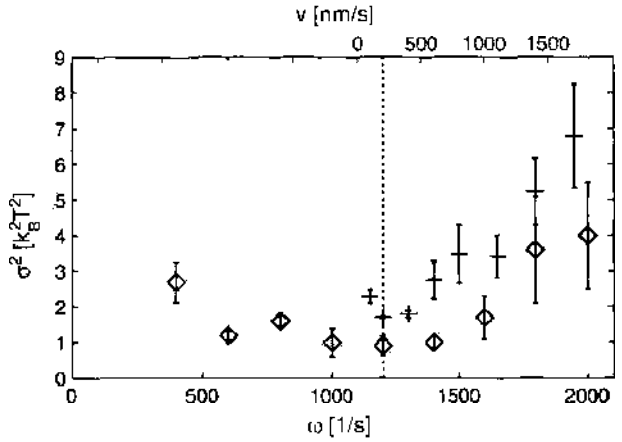


Figure 16. Mean square error σ^2 and error bars for reconstructed free energy profiles by using the linear ramp (3) as a function of v (top scale, +), and the periodic ramp (9) as a function of ω (bottom scale, ◇) (compare Fig. 15). Both methods use 10 runs with 10 trajectories each. The optimal velocity v^* and frequency ω^*, respectively, indicated by the vertical dashed line, correspond to the smallest error. Reprinted with permission from [104], O. Braun et al., *Phys. Rev. Lett.* 93, 158105 (2004). © 2004, American Physical Society.

Thus, the simulated data in Fig. 15 show that the periodic ramp proposed in [104] delivers significantly more accurate data than the traditional linear loading (3), under the constraint of equal experimental effort. An additional advantage of the periodic ramp is the fact that the measuring time may be chosen as long as necessary to collect the sufficient amount of data to recover the barrier regions of the free energy profile, which are hard to determine by previous methods.

2.5.4. Optical Methods Using Fluorescence

Single molecule spectroscopy, usually in a dilute solution of the molecule to measure, allows to record the fluorescence in single events [107–110]. Fluorescence correlation spectroscopy (FCS) records fluctuations in the fluorescence intensity around an equilibrium value. From these data, the autocorrelation function of the fluorescence signal can be determined. Depending on the nature of the process, from this autocorrelation function the relevant timescales of the processes involved (diffusion, quenching kinetics, etc.) may be deduced on the basis of certain models. For example, molecular beacon techniques are based on a pair of fluorophores chemically attached to the molecule. When both fluorophores are close, they are quenched and no fluorescence can be detected. If, due to a molecular configurational transition, the two fluorophores become separated, they fluoresce. This method was used to quantify the bubble fluctuations in dsDNA [111], or the opening dynamics of an RNA hairpin [112].

A second powerful technique in single molecule spectroscopy is fluorescence resonance energy transfer (FRET), based on the nonradiative energy transfer by a resonance dipol-dipol interaction between two fluorophores. The efficiency of this energy transfer depends on the dye-dye distance. By placing such pairs of dyes to the sequence of a biopolymer, valuable information about their distance in a given conformation (e.g., the native or an excited state of a protein) can be obtained [113, 114].

2.5.5. Green Fluorescent Protein and Single Particle Tracking

Green fluorescent protein (GFP) is a protein originally isolated from the luminescent jellyfish *Aequorea victoria*. As a powerful noninvasive labeling tool, it is introduced into the gene of interest (the target gene is "tagged"). Once transcription of the target gene is initiated, also the GFP gene will be transcribed, which allows one to monitor the transcription activation. A second possibility is to attach GFP to a virus or another protein, and then the motion of this object can be followed inside the cell. There also exists a Red Fluorescent Protein isolated from an IndoPacific sea anemone relative, *Discosoma spp.* (disk anemones). It can be clearly distinguished from GFP, and so be used to label different genes. Recent variants also include yellow, cyan, and blue [115].

Single particle tracking is an optical microscopy technique that combines imaging with computational tools to resolve single molecule motions. This is possible for larger objects, which can be distinguished from a background, or for smaller objects such as biomolecules with a bright and stable optical marker. Examples include the monitoring of the motion of microbeads through the cytoskeleton [116], or the study of the lambda receptor in the bacterial outer membrane [117], both *in vivo*.

3. FLUCTUATIONS *VERSUS* FIXED TOPOLOGY: KNOTS AND THEIR SHAPES

Bacterial DNA occurs largely in circular form. Notably, instead of a simply connected ring shape (the unknot), the DNA often exhibits permanently entangled states, such as catenated and knotted DNA. Such configurations have potentially devastating effects on the cell development. In eukaryotic cells additional topological effects occur in the likely entanglement of individual chromosomes. In this section, we concentrate on the simpler prokaryotic case. After a brief introduction to the physiological relevance of DNA-knots we investigate the statistical mechanical behaviour of such permanently entangled chains.

Before proceeding, we note that apart from the biological background, permanently entangled single or many-chain systems are of high interest in traditional polymer physics. Knots are produced in ring cyclization, or two cyclized rings may be concatenated. There have to be invented elaborate techniques to avoid such topological states. Networks like rubber that, by curing via the vulcanization process, are permanently entangled, are another example of systems of primary relevance [42, 43, 48].

3.1. Physiological Background of Knots

The discovery how one can use molecular biological tools to create knotted DNA resolved a long-standing argument against the Watson-Crick double helix picture of DNA [6], namely that the replication of DNA could not work as the opening up of the double helix would produce a superstructure such that the two daughter strands could not be separated. In fact, the topology of both ssDNA and dsDNA is continuously changed *in vivo*, and this can readily be mimicked *in vitro*, although the activity of enzymes *in vivo* is much more restricted than *in vitro* [4, 17]. Different concentrations of enzymes to knotted DNA accessible *in vitro*, that is, makes it possible to probe topology-altering effects by enzymes which *in vivo* do not contribute to such effects. Although it would be likely with a probability of roughly 1/2 that the linear DNA injected by bacteriophage λ into its host *E. coli* would create a knot before cyclization, it turned out to be difficult to detect [6]. First studies therefore concentrated on the fact that under physiological conditions, knots are introduced by enzymes, DNA replication and recombination, DNA repair, and topoisomerization, using these enzymes to prove both knotting and unknotting [9, 23, 118–122]. DNA-knotting is also prone to occur behind a stalled replication fork [10, 123]. Some of the typical topology-altering reactions undergoing in *E.coli* are summarized in Fig. 17. In Fig. 18, DNA-knots created from nicked dsDNA under action of topoisomerase I are charted. The figure also shows the method to detect different topological states of dsDNAs of the same length by electrophoresis motility assays. We note that knots may efficiently be created by using the active packaging of a DNA mutant into phage capsids [124], and then denaturing the capsid proteins.

The existence of DNA-knots has far-reaching effects on physiological processes, and knottedness of DNA has therefore to be eliminated in order to maintain proper functioning of the cell. Among other possible effects, it is immediately clear that the presence of a knot in a circular DNA impedes replication of the DNA, that is, the full separation of the two daughter strands [1, 6]. Moreover, even transcription is impaired [125]. The presence of knots inhibits the assembly of chromatin [126], knotted chromosomes cannot be separated during mitosis [1], and knots in a chromosome may serve as topological barriers between different sections of chromosomes, such that the genomic structural organisation is altered, and certain sections of the chromosomal DNA may no longer interact [127]. Finally, knots

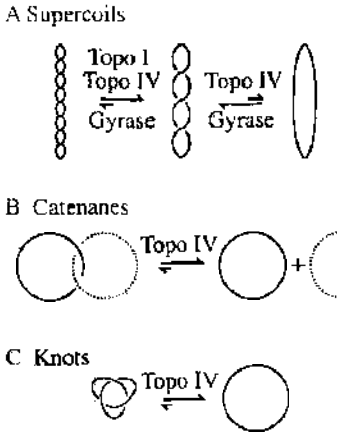


Figure 17. Enzymes changing the topology of dsDNA by cutting and pasting of one or both strands (example for *E. coli*). (A) Torsional stress resulting from the Lk deficit causes the DNA double helix to writhe about itself (negative supercoiling). In *E. coli*, gyrase introduces negative supercoils into DNA and is countered by topoisomerase I (topo I) and topo IV, which relax negative supercoils. (B) Topo IV unlinks catenanes generated by replication or recombination *in vivo*. (C) Topo IV unknots DNA *in vivo*. Reprinted with permission from [17], R. W. Deibler, et al., *Genes & Development* 15, 748 (2001). © 2001, Cold Spring Harbor Laboratory Press.

may lead to double-strand breaks, as they weaken biopolymers considerably due to creation of localised sharp bends [92, 128–130] in analogy to macroscopic lines and ropes [131].[5]

Above we said that knots can be introduced, *inter alia*, by the different enzymes of the topoisomerase family. To remove a knot from a dsDNA, it is necessary to cut both strands, and then pass one segment through the created gap, before resealing the two open ends. *In vivo*, this is usually achieved by topoisomerases II and IV. A reconstruction of topo II is shown in Fig. 1, indicating the upper clamp holding a segment of the DNA, while the bulge-clamp introduces the cut through which the upper segment is passed. In the figure, the segment visible in the pocket of the lower clamp has already been passed through the gap. After resealing, topo II detaches. This process requires energy, provided by ATP. Notably, topo II is extremely efficient, for circular dsDNA of length $\simeq$ 10 kbp it was found that topo reduced the knotted state in between 50- and 100-fold, in comparison to a "dumb" enzyme, which would simply pass segments through at random [132]. We note that the stepwise action of topoisomerase II was recorded in a single molecule setup using magnetic tweezers [94, 133]. Topoisomerases are surveyed in the review of [134]. We should add that the physiological role and the dynamics of knot creation and removal *in vivo* is still not fully explored.

3.2. From Königsberg Bridges and Vortex Atoms to Knot Theory

Knottedness can only be defined on a closed, that is, circular chain. This is intuitively clear as in an open linear chain a knot can always be tied, or an existing knot released. Mathematically, this means that knot invariants are only well defined for a closed space-curve. However, a linear chain whose ends are permanently attached to one, or two walls, or whose ends are extended toward infinity, can be considered as (un)knotted in the proper mathematical sense, that is, their knottedness cannot change.

The classification of knots, or graphs in general, in terms of invariants can essentially be traced back to Euler, recalling his graph theoretical elaboration in connection with the Bridges of Königsberg problem [135], determining a closed path by crossing each Königsberg bridge exactly once. However, the first investigations of topological problems in modern science is most probably due to Kepler, who in earlier work considered stacked ideal bodies as a model to calculate the planetary orbits, and who published elaborately on surface tiling [136]. Further initial steps were due to Leibniz, Vandermonde and Gauss, in whose collection of

[5] The weakness of strings at the site of the knot can be experienced easily by pulling a nylon string in comparison to a knotted one [128].

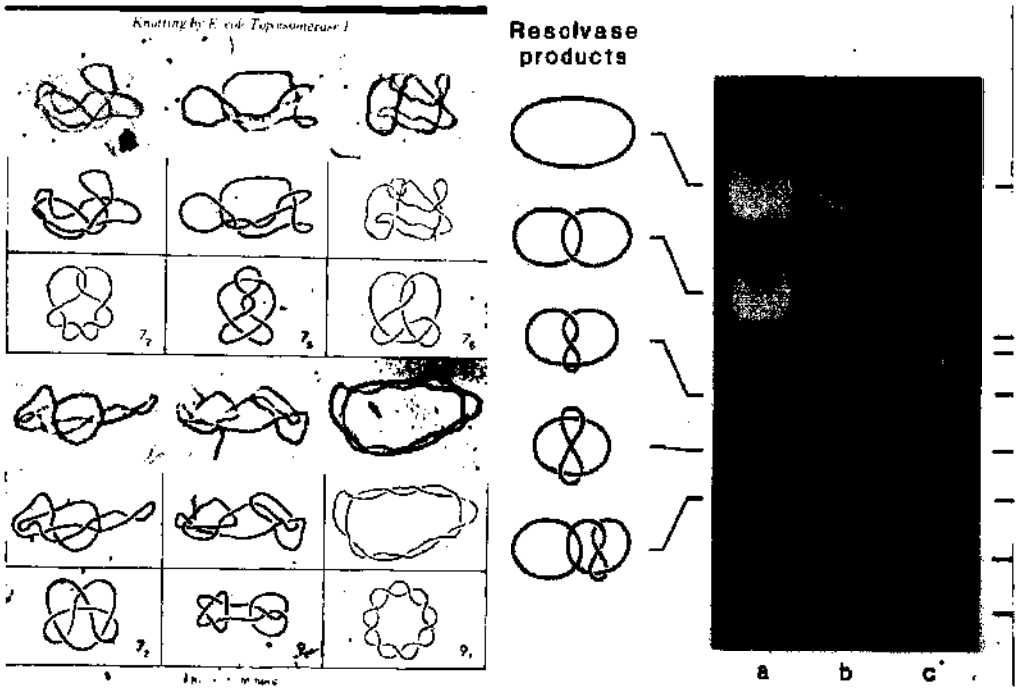


Figure 18. Various DNA knots created by topoisomerase I from a nicked ring DNA of unknot topology. Reprinted with permission from [11]. F. B. Dean et al., *J. Biol. Chem.* 260, 4975 (1985). © 1985. American Society for Biochemistry and Molecular Biology. Right: Different mobility of various DNA knots in gel electrophoresis. The more complex topologies are faster. Reprinted with permission from [23]. S. A. Wassermann et al., *Science* 229, 171 (1985). © 1985. AAAS.

papers drawings of various knots were found[6] whose linking (Umschlingungen = windings) number is indeed a knot invariant [137–139]. Gauss' student, Listing, in fact introduced the term topology, and his work on knots may be viewed as the real starting point of knot theory [140], although his complexions number was proved by Tait not to be an invariant.

Inspired by Helmholtz' theory of an ideal fluid and building on Listing's early contributions to knot theory, Scotsmen and chums Maxwell, Tait and Thomson (Lord Kelvin) started to discuss the possible implications of knottedness in physics and chemistry, ultimately distilled into Thomson's theory of vortex atoms [141, 142]. Out of this endeavor emerged Tait's interest in knots, and he devoted most of his career on the classification of knots. Numerous charts and still unresolved conjectures on knots document his pioneering work [143–146]. The studies were carried on by Kirkman and Little [147–150]. A more detailed historical account of knot theory may be found in the review article by van de Griend [151], and on the St. Andrews history of mathematics webpages.[7]

Planar projections of knots were rendered unique by Listing's introduction of the handedness of a crossing, that is, the orientational information assigned to a point where in the projection two lines intersect. With this information, projections are the standard representation for knot studies. On their basis, the minimum number of crossings ("essential crossings") can be immediately read off as one of the simplest knot invariants. To arrive at the minimum number, one makes use of the Reidemeister moves, three fundamental permitted moves of the lines in a knot projection, as shown in Fig. 19. More complex knot invariants include polynomials of the Alexander, Kauffman and HOMFLY types [137–139].[8] In all what follows, we will only employ the number of essential crossings as classification

[6] Probably copies from an English original.

[7] The MacTutor History of Mathematics archive: http://turnbull.mcs.st-and.ac.uk/history/.

[8] These polynomials all start to be degenerate for higher order knots, that is, above a certain knot complexity, several knots may correspond to one given polynomial [137, 138]. In the case of the simpler knots attained in most DNA configurations and in knot simulations, the Alexander polynomials are unique, in contrast to the Gauss or Edwards invariant; compare, for example, [152].

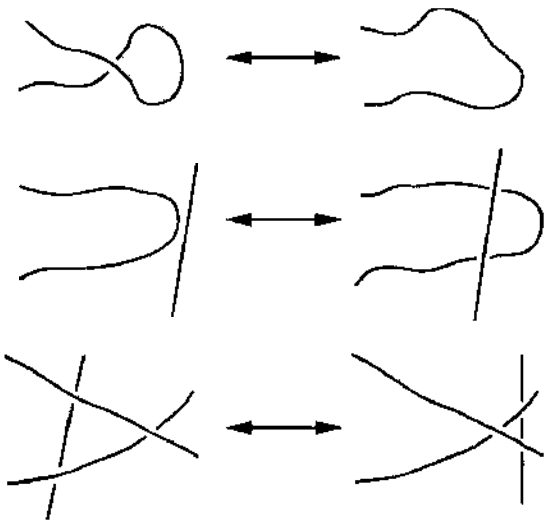


Figure 19. The three Reidemeister moves. All topology-preserving moves of a knot projection can be decomposed into these three fundamental moves.

of knots; in particular, we do not concern ourselves with the question of degeneracy for a given knot invariant.

3.3. Long Chains are Almost Always Entangled

During the polymerization and final cyclization of a polymer grown in a solvent under freely floating conditions, a knot is created with probability 1. This Frisch-Wassermann-Delbrück conjecture [153, 154] could be mathematically proved for a self-avoiding chain (the standard model of a "real" polymer chain) [155, 156], compare also [157]. This is consistent with numerical findings that the probability of unknot formation decreases dramatically with chain length [152, 158]. Indeed, recent simulations results indicate that the probability of finding the unknot in such a cyclisised polymer decays exponentially with chain length [159–162]:

$$P_0(N) \propto \exp\left(-\frac{N}{N_c}\right) \tag{12}$$

Compare Fig. 20.

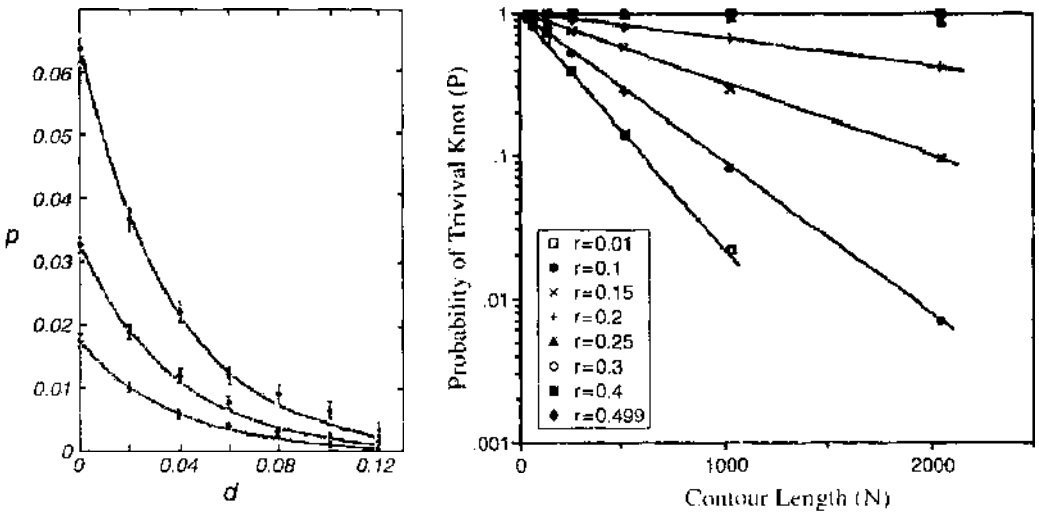


Figure 20. Left: Dependence of the fraction of knotted molecules on the effective diameter d for circular DNA with 14 (lower curve), 20 (middle), and 30 (upper curve) Kuhn segments. d is measured in Kuhn lengths, so to obtain d in nm one multiplies the abscissa by 100; from [59], data from [165]. Reprinted with permission from M. Frank-Kamenetskii, *Phys. Rep.* 288, 13 (1997). © 1997, Elsevier. Right: Semi-log plot of $P_0(N)$ as a function of the polymer length, for various values of the excluded volume (here, thickness) of the chain. Reprinted with permission from [159]: K. Koniaris and M. Muthukumar, *Phys. Rev. Lett.* 66, 2211 (1991). © 1991, American Physical Society.

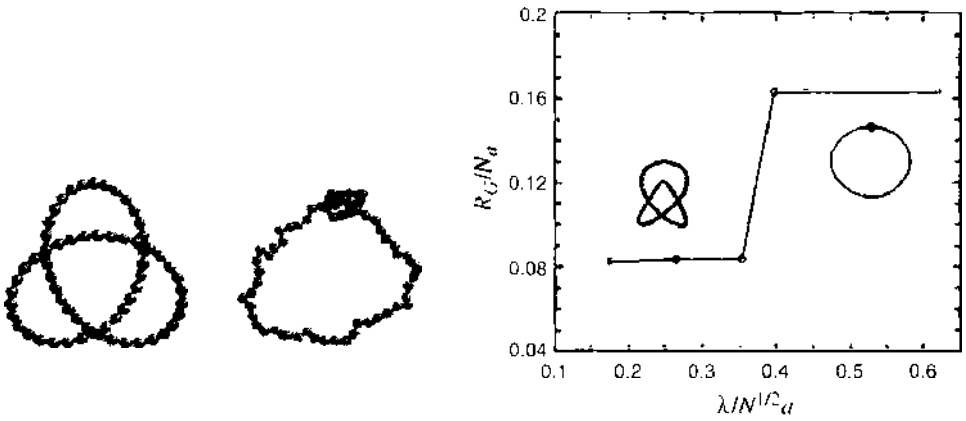


Figure 21. Left panel: Charged trefoil knot with unscreened Coulomb interactions. The left configuration is a completely symmetric harmonic initial condition, to the right we see a typical tight equilibrium configuration. Right panel: Trefoil knot with screened interactions, showing a transition of the gyration radius R_G from a small value for the symmetric (delocalized) configuration toward the tight knot with large gyration radius. Reprinted with permission from [169], P. G. Dommersnes et al., *Phys. Rev. E* 66, 031802 (2002). © 2002, American Physical Society.

However, there exist theoretical arguments and simulations results indicating that the characteristic length N_c occurring in this relation may become surprisingly long [163–166]. The probability to find a given knot type $\mathcal{K}$ on random circular polymer formation has been fitted with the functional form [166–168]

$$P_{\mathcal{K}}(N) = a(N - N_0)^b \exp\left(-\frac{N^c}{d}\right) \tag{13}$$

where a, b, and d are free parameters depending on $\mathcal{K}$, and $c \approx 0.18$. N_0 is the minimal number of segments required to form a knot $\mathcal{K}$, without the closing segment [168]. The tendency toward knotting during polymer cyclization creates problems in industrial and laboratory processes.

3.4. Knot Localization in Charged Knots

Before considering interactions between fixed topology and fluctuations in a knotted polymer chain, we would like to explore briefly the localization phenomenon in charged knots. Consider a simple linear chain or a ring polymer on which each monomer carries an identical charge q. Then the charges will tend to attain maximum distance from each other, leading to a stretched linear chain or a maximally expanded ring.

What happens if we regard a circular chain containing a knot? Compare the two configurations displayed in Fig. 21, depicting a completely delocalized trefoil knot, and the same trefoil when it is tight (note that the tight trefoil is scaled down in size). Disregarding thermal effects (they are in fact sub-leading), we can come to a rough scaling argument. In the delocalised state, the electrostatic energy of the knot has a scaling form $E_1 \simeq Q^2/R_1$ where R_1 is the diameter of the knot scaling with the number of monomers N as $R_1 \sim N$, and the overall charge $Q \sim N$. Thus, $E_1 \simeq N$.[9] The tight counterpart can be split up into two contributions, $E_2 \sim Q^2/R_2$ ($R_2 > R_1$) for the monomers outside the tight knot, and the knot contribution $E_0 \sim q^2/a$ where a is the small diameter of the knot region. Obviously, E_0 is independent of the number of monomers of the entire chain. Thus, the penalty of bringing together the charges within the knot region is a constant for all chain lengths. But as $\Delta E = E_1 - E_2 \simeq N$, the chain length can always be made sufficiently long to overcompensate the energy loss for creating the tight knot. Thus, one would for such longer chains always obtain the tight knot as the stable configuration. Indeed, the careful numerical study in reference [169] could verify this rough scaling result for a collection of prime knots, while

[9] To be more precise, there is also a logarithmic contribution [169].

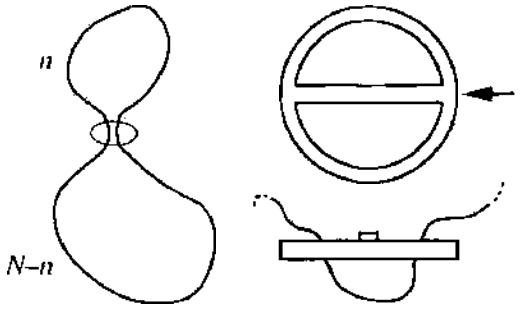


Figure 22. Figure-eight structure, in which a slip-link separates two loops of size n and $N - n$, such that they can freely exchange length among each other, but none of the loops can completely retract from the slip-link. On the right, a schematic drawing of a slip-link, which may be thought of as a small belt buckle.

composite knots split up into their prime components. In those simulations, the initial configuration for the trefoil knot corresponded to the harmonic representation in Fig. 21 on the left, whereas the tight configuration shown on the right corresponds to a typical equilibrium shape.[10] We note that such $1/r^2$ potentials were in fact previously studied by knot theorists with the goal of obtaining a knot invariant, which turned out to be negative for the fact that without excluded volume knots collapse to a point on a continuous curve, providing no way of identifying knots [170, 171].

In all biological systems, the charges are screened by counterions in the liquid environment, which is usually taken into account through a Debye-Hückel form of the electrostatic interaction. The screening length λ entering as an additional length scale will then compete with the radius of the knot: For sufficiently large screening lengths, a tight knot will emerge, as shown in Fig. 21 on the right-hand side for a trefoil knot. This tight knot is, however, stable in the sense that it does not collapse to a point [171].

3.5. Entropic Localization in the Figure-Eight Slip-Link Structure

The simplest single-polymer object with nontrivial (non-unknot) geometry is the figure-eight structure (F8) displayed in Fig. 22. In this compound, a pair contact is enforced by a slip-link, separating off two loops in the circular polymer, such that none of the loops can fully retract, and both loops can freely exchange length among each other. We denote the loop sizes by n and $N - n$, where N is the (conserved) total length of the polymer chain. For such an object, we can actually perform a closed statistical mechanical analysis based on some results from scaling theory of polymers, and compare the result with Monte Carlo simulations of the F8.

The statistical quantities that are of particular interest are the gyration radius, R_g, and the number of degrees of freedom (DOF), ω [43]. R_g, as defined in Eq. (A3) in the appendix at the end of this chapter, measures the root mean squared distance of the monomers along the chain to the gyration center, and is therefore a good measure of its extension. It can, for instance, be measured by light scattering experiment. The DOF ω count all possible different configurations of the chain. For a circular polymer [i.e., a polymer with $\mathbf{r}(0) = \mathbf{r}(N)$], the gyration radius becomes

$$R_g \simeq AN^{\nu} \tag{14}$$

with exponent $\nu = 1/2$ for a Gaussian chain, and $\nu = 0.588$ in the 3D excluded volume case ($\nu = 3/5$ in the Flory model, and $\nu = 3/4$ in 2D). Whereas in 2D this scaling contains truly a ring polymer, in 3D the exponent ν emerges from averaging over all possible topologies, and necessarily includes knots of all types [43, 172, 173]. For a circular chain, the number of DOF contains the number of all possible ways to place an N-step walk on the lattice with connectivity μ (e.g., $\mu = 2d$ on a cubic lattice in d dimensions), μ^N, and the entropy loss for requiring a closed loop, $N^{-d\nu}$, involving the same Flory exponent μ. For the Gaussian case,

[10] In the Monte Carlo simulations we quote, the term *equilibrium* refers to a configuration of the system after several (Rouse) relaxation times.

we recognize in this entropy loss factor the returning probability of the random walk. In the excluded volume case, $N^{-d\nu}$ is an analogous measure [5, 43, 174]. Thus, for a circular chain embedded in d-dimensional space, the number of DOF is

$$\omega \simeq \mu^N N^{-d\nu} \tag{15}$$

Let us evaluate these measures for the F8 from Fig. 22.

As a first approximation, consider the F8 as a Gaussian (phantom) random walk, demonstrating that, like in the charged knot case, entropic effects give rise to long-range interactions. The two loops correspond to returning random walks, that is, the number of DOF for the F8 in the phantom chain case becomes [5, 41, 43][11]

$$\omega_{\mathrm{F8,PC}} \simeq \mu^N n^{-d/2}(N-n)^{-d/2} \tag{16}$$

where d is the embedding dimension. We note that normalization of this expression produces the probability density function (PDF) for finding the F8 with a given loop size $\ell = na$ ($L = Na$).

$$p_{\mathrm{F8,PC}}(\ell) \simeq \mathcal{N}\ell^{-d/2}(L-\ell)^{-d/2} \tag{17}$$

where $\mathcal{N}$ denotes a normalisation factor. The conversion from expressing the chain size in terms of the number of monomers to its actual length is of advantage in what follows, as it allows to more easily keep track of dimensions. Here, we use the length unit a, which may be interpreted as the monomer size (lattice constant), or as the size of a Kuhn statistical segment.

The average loop size $\langle \ell \rangle$ determined through $\langle \ell \rangle = \int_a^{L-a} \ell p(\ell) d\ell$ is trivially $\langle \ell \rangle = L/2$ by symmetry of the structure. Here, we introduce a short-distance cutoff set by the lattice constant a. However, as the PDF is strongly peaked at $\ell = 0$ and $\ell = L$, the two poles caused by the returning probabilities, a *typical* shape consists of one small (*tight*) and one large (*loose*) loop. This can be quantified in terms of the average size of the smaller loop,

$$\langle \ell \rangle_< \equiv 2\int_a^{L/2} \ell p(\ell)\, d\ell \tag{18}$$

In $d = 2$, we obtain

$$\langle \ell \rangle_< \sim \frac{L}{|\log(a/L)|} \tag{19}$$

such that with the logarithmic correction the smaller loop is only marginally smaller than the big one. In contrast, one observes *weak localization*

$$\langle \ell \rangle_< \sim a^{1/2} L^{1/2} \tag{20}$$

in $d = 3$, in the sense that the relative size $\langle \ell \rangle_</L$ tends to zero for large chains. By comparison, for $d > 4$ one encounters $\langle \ell \rangle_< \sim a$, corresponding to *strong localisation*, as the size of the smaller loop does not depend on L but is set by the short-distance cutoff a. Above four dimensions, excluded volume effects become negligible, and therefore both Gaussian and self-avoiding chains are strongly localized in $d \geq 4$.[12]

To include self-avoiding interactions, we make use of results for general polymer networks obtained by Duplantier [172, 175], which are summarized in the Appendix at the end of this review. In terms of such networks, our F8-structure corresponds to the following parameters: $\mathcal{N} = 2$, $J = 2$ physical loops, $n_4 = 1$ vertex of order four, with the identification $s_1 = \ell$ and

[11] Here and in the following we consider two configurations of a polymer chain different if they cannot be matched by translation. In addition, the origin of a given structure is fixed by a vertex point (see below), that is, a point where several legs of the polymer chain are joint. In the F8-structure, this vertex naturally coincides with the slip-link. For a simply connected ring polymer, such a vertex is a two-vertex anywhere along the chain.

[12] Consideration of higher than the physical three space dimensions is often useful in polymer physics.

$s_2 = L - \ell$. By virtue of Eq. (A21), the number of configurations of the F8 with fixed ℓ follows the scaling form

$$\omega_{F8}(\ell) \simeq \mu^L (L-\ell)^{\gamma_{F8}-1} \mathscr{X}_{F8}\left(\frac{\ell}{L-\ell}\right) \tag{21}$$

with the configuration exponent $\gamma_{F8} = 1 - 2d\nu + \sigma_4$. In the limit $\ell \ll L$, the contribution of the large loop in Eq. (21) should not be affected by a small appendix, and therefore should exhibit the regular Flory scaling $\sim (L-\ell)^{-d\nu}$ [176–178]. This fixes the scaling behaviour of the scaling function $\mathscr{X}_{F8}(x) \sim x^{\gamma_{F8}-1+d\nu}$ in this limit ($x \to 0$ with $x = \ell/(L-\ell)$), such that

$$\omega_{F8}(\ell) \simeq \mu^L (L-\ell)^{-d\nu} \ell^{\,c}, \quad \ell \ll L \tag{22}$$

where $c = -(\gamma_{F8} - 1 + d\nu) = d\nu - \sigma_4$. Using $\sigma_4 = -19/16$ and $\nu = 3/4$ in $d = 2$ [175, 172], we obtain

$$c = \frac{43}{16} = 2.6875, \quad d = 2 \tag{23}$$

In $d = 3$, $\sigma_4 \approx -0.48$ [176, 177, 179] and $\nu \approx 0.588$, so that

$$c \approx 2.24 \tag{24}$$

In both cases, the result $c > 2$ enforces that the loop of length ℓ is strongly localized in the sense defined above. This result is self-consistent with the a priori assumption $\ell \ll L$. Note that for self-avoiding chains, in $d = 2$ the localization is even *stronger* than in $d = 3$, in contrast to the corresponding trend for ideal chains.

We performed Monte Carlo (MC) simulations of the 2D figure-eight structure, in which the slip-link was represented by three tethered beads enforcing a sliding pair contact such that the loops cannot fully retract (see Fig. 23). We used a 2D hard core bead-and-tether chain with 512 monomers, starting off from a symmetric initial condition with $\ell = L/2$. Self-crossings were prevented by keeping a maximum bead-to-bead distance of 1.38 times the bead diameter, and a maximum step length of 0.15 times the bead diameter. We estimate the simulation time to be well above the Rouse relaxation time [43] of the chain. As shown in Figure 23, the size distribution for the small loop can be fitted to a power law with exponent $c = 2.68$ in good agreement to the above prediction.

An experimental study of entropic tightening of a macroscopic F8-structure was reported in [181]. There, a granular chain consisting of hollow steel spheres connected by steel rods was once twisted and then put on a vibrating table. From digital imaging, the distribution of loop sizes could be determined and compared to a power-law with index 43/16 as calculated for the 2D excluded volume chain. The agreement was found to be consistent [181].

3.6. Simulations of Entropic Knots in 2D and 3D

Much of our knowledge about the interaction of knots with thermal fluctuations is based on simulations of knotted chains. Before going further into the theoretical modeling of knotted chains, we report some of the results based on simulations studies of both Gaussian and self-avoiding walks.[1] Such simulations either start with a given knot configuration and then perform moves of specific segments, each time making sure that the topology is preserved; or, each new configuration emanates from a new random walk, whose correct topology may be checked by calculating the corresponding knot invariant, usually the Alexander polynomial, and created configurations that do not match the desired topology are discarded. We note that it is of lesser significance that knot invariants such as the Alexander polynomials in fact are no longer unique for more complex knots, because with the highest probability simpler knots are created, for which the invariants are unique. For more details, we refer to the works quoted below.

[1] Although per se a Gaussian chain cannot have a fixed topology due to its phantom character, such simulations introduce a fixed topology by rejecting moves that result in a different knot type.

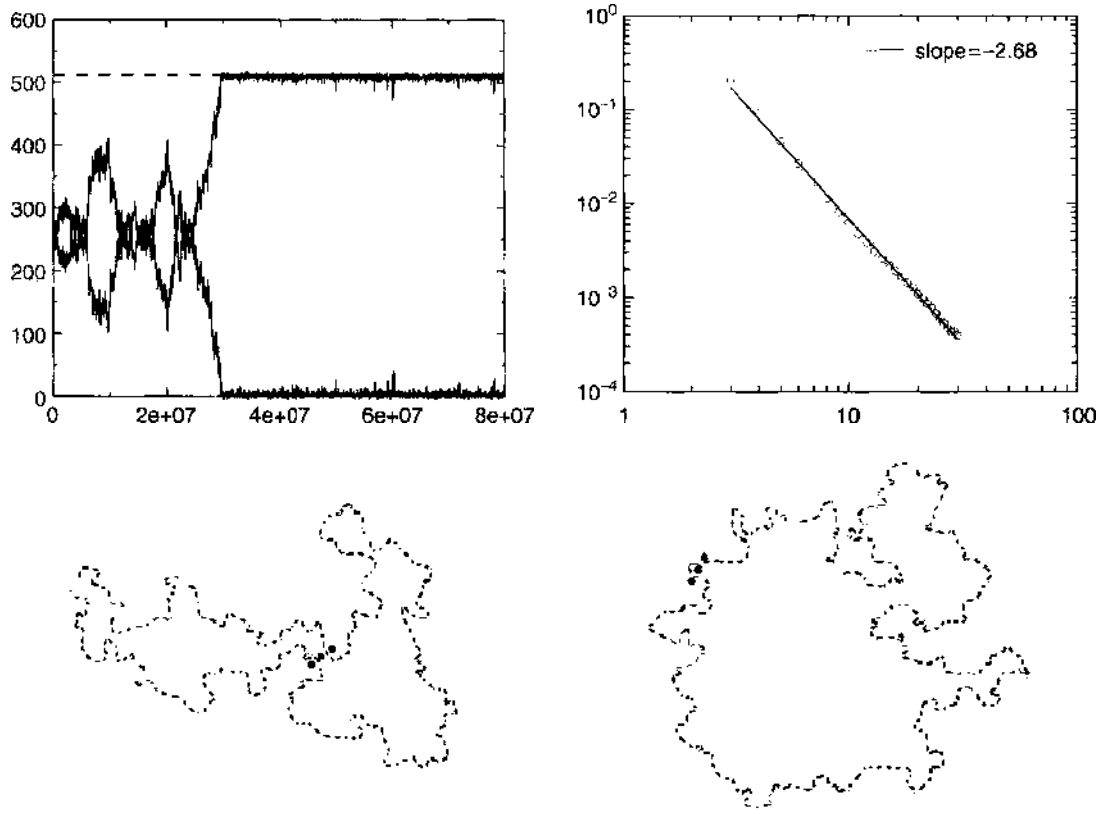


Figure 23. Monte Carlo simulation of an F8-structure in 2D. Top left: Loop sizes ℓ and $L - \ell$ as a function of Monte Carlo steps for a chain with 512 monomers. The symmetry breaking after the symmetric initial condition is distinct. Top right: Power-law fit to the PDF of the smaller loop. The fit produces a slope of -2.68, in beautiful agreement with the calculated value. Bottom left: Bead-and-tether chain used in Monte Carlo simulation, intermediate configuration toward equilibration, starting from a symmetric initial condition. The three black beads are tethered and represent the 2D slip-link. Right: Typical equilibrium configuration. Reprinted with permission from [180], R. Metzler et al., *Phys. Rev. Lett.* 88, 188101 (2002). © 2002, American Physical Society.

In fact, the fixed topology turns out to have a highly nontrivial effect on chains without self-excluded volume. As conjectured in [182], a Gaussian circular chain, whose permitted set of configurations is restricted to a fixed topology, will exhibit self-avoiding behavior. This was proved in a numerical analysis in [173]. The required number of monomers to reach this self-avoiding exponent was estimated to be of the order of 500. Keeping this non-trivial scaling of a Gaussian chain at fixed topology in mind, knot simulations on the basis of phantom Gaussian chains were performed in [183], always making sure that the configurations taken into the statistics fulfil the desired knot topology.

The dependence of the gyration radius R_g on the knot type was investigated for simpler knots in 3D in reference [162]. On the basis of the expression

$$R_g^2 \simeq A_{\mathcal{K}} \left(1 + B_{\mathcal{K}} N^{-\Delta} + C_{\mathcal{K}} N^{-1} + o\left(\frac{1}{N}\right)\right) N^{2\nu_{\mathcal{K}}} \tag{25}$$

including a confluent correction [162, 184, 185] in comparison to the standard expression (14), it was found that the Flory exponent $\nu_{\mathcal{K}}$ is independent of the knot type $\mathcal{K}$ and has the 3D value 0.588. This was interpreted via a *localization* of knots such that the influence of tight knots on R_g is vanishingly small. In fact, Δ is of the order of 0.5 according to the investigations in references [184–187].

Figure 24 shows analogous results from the study in [188]. Extrapolating the curves of $R_g^2/N^{2\nu}$ *versus* $N^{-\Delta}$ for various simple knot types suggests that all approach the same value. Based on longer chains in comparison to reference [162], the study of [188] thus corroborates

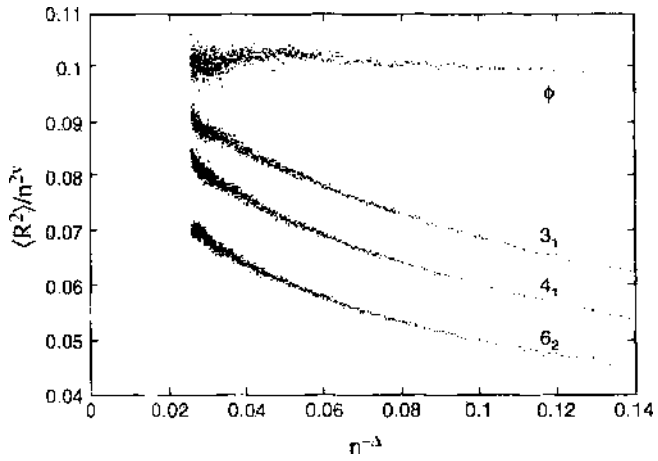


Figure 24. Plot of $R_g^2/N^{2\nu}$ *versus* $N^{-\Delta}$ for unknot 0, and knot types 3_1, 4_1, and 6_2. Extrapolation to $N \to \infty$ suggests that all curves approach the same value, implying that the amplitude of the gyration radius is independent of the knot type. Reprinted with permission from [188], E. Orlandini et al., *J. Phys. A* 31, 5953 (1998). © 1998, IOP Publishing.

the independence of $\nu_{\mathcal{K}} \approx 0.588$ of the knot type $\mathcal{K}$. Moreover, for the number of DOF $\omega_{\mathcal{K}}$, it was found for the form[14]

$$\omega_{\mathcal{K}} \simeq A_{\mathcal{K}} N^{\alpha_{\mathcal{K}}-2} \mu_{\mathcal{K}}^{N} \left(1 + \frac{B_{\mathcal{K}}}{N^{\Delta}} + \cdots\right) \tag{26}$$

that while for the unknot with $\alpha_0 \approx 0.27$ expression (26) is consistent with the standard result (15) ($[0.27-2]/3 \approx -0.58 \approx \nu$), for prime knots $\alpha_{\mathcal{K}} = \alpha_0 + 1$, and for composite knots with N_f prime components,

$$\alpha_{\mathcal{K}} = \alpha_0 + N_f \tag{27}$$

This finding is consistent with the view that each prime component of a knot $\mathcal{K}$ is tightly localized and statistically able to move around one central loop, each prime component counting an additional factor N of DOF. The fact that for a chain of finite thickness the size of the big central loop is in fact diminished by the size of the tight knot is a confluent effect. Not surprisingly, the connectivity factor $\mu_{\mathcal{K}} \approx 4.68$ was found to be independent of $\mathcal{K}$, assuming the standard value for a cubic lattice [189]. Also the amplitude $A_{\mathcal{K}}$ and the exponent $\Delta_{\mathcal{K}}$ of the confluent correction turned out to be $\mathcal{K}$-independent. We note that a similar analysis in (pseudo) 2D[15] also strongly points towards tight localization of the knot [190].

In contrast to the above results, 3D simulations undertaken in [191] (also compare [192]) show the dependence

$$R_g \simeq N^{3/5} C^{-4/15} \tag{28}$$

of the gyration radius on the knot type, characterized by the number C of essential crossings. R_g, that is, decreases as a power-law with C, where the exponent $-4/15 = 1/3 - \nu$ [191]. The functional form (28) was derived from a Flory-type argument for a polymer construct of C interlocked loops of equal length N/C by arguing that each loop occupies a volume $\simeq (N/C)^{3\nu}$, and the volume of the knot is given by $V \simeq C(N/C)^{3\nu}$ (i.e., assuming that due to self-avoiding repulsion the volume of individual loops adds up to the total volume). Equation (28) then follows immediately. This model of equal loop sizes is equivalent to a completely delocalized knot. It may therefore be speculated, albeit rather long chain sizes of up to 400

[14] Note that we changed the exponent by 1 in comparison to the original work, making the counting of nontranslatable configurations consistent with the counting convention specified in footnote 11.

[15] The simulated polymer chain moves in 2D, however, crossings are permitted at which one chain passes underneath another. In that, the simulated polymers are in fact equivalent to knot projections.

were used, whether the numerical algorithm employed for the simulations in [191] causes finite-size effects that, in turn, prevent knot localization. We note that the Flory-type scaling assumed to derive expression (28) is consistent with a modeling brought forward in [193], in which the knot is quantified by the aspect ratio in a configuration corresponding to a maximally inflated tube with the given topology (i.e., a state corresponding to complete delocalisation). In [191], the temporal relaxation behavior of a given knot was also studied. While regular Rouse behavior was found for the case of the unknot, the knotted chains displayed somewhat surprising long time contributions to the relaxation time spectrum [44–48, 194]. This phenomenon was already pointed out by de Gennes within some activation argument to create free volume in a tight knot in order to move along the chain [195]. Note that relatively lose knots in shorter chains do not appear to exhibit such extremely long relaxation time behavior [196].

Simulation of a 3D knot with varying excluded volume showed that if only the excluded volume becomes large enough, the gyration radius of the knot is independent of the knot type [197]. The picture of tight knots is further corroborated in the study by Katritch et al. using a Gaussian chain model with fixed topology to demonstrate that the size distribution of the knot is distinctly peaked at rather small sizes [167].

Apart from determining the statistical quantities R_g and ω_g from simulations, there also exist indirect methods for quantifying the size of the knot region in a knotted polymer. One such method is to confine an open chain containing a knot between two walls, and measuring the finite size corrections of the force-extension curve due to the knot size. This is based on the idea that the gyration radius for a system depending on more than one length scale (i.e., apart from the chain length N) shows confluent corrections, such that [198]

$$R_g = AN^{\nu}\Phi\left(\frac{N_0}{N}, \frac{N_1}{N}, \ldots\right) \simeq AN^{\nu}\left(1 - BN^{-\Delta}\right) \tag{29}$$

when only the largest correction is considered, and in 3D $\Delta \approx 0.5$ is universal [184–187]. If this leading correction is due to the argument N_0/N in the scaling function Φ, the length scale N_0 depends on N through the scaling $N_0 \sim N^t$ with $\Delta = 1 - t$. From Monte Carlo simulations, it could then be inferred that the size of the knot scales like [198]

$$N_0 \sim N^t, \quad t = 0.4 \pm 0.1 \tag{30}$$

This, in turn, enters the force-extension curve $f' = G(R')$ with the dimensionless force $f' = f AN^{\nu}/(k_BT)$ and distance $R' = R/(AN^{\nu})$ of the walls, in the form with confluent correction

$$f' \simeq G(R')\left(1 + g(R')N^{-\Delta}\right) \tag{31}$$

The results of the simulations are reproduced in Fig. 25, and $t = 0.4$ corresponds to the best data collapsing, assuming the validity of the scaling arguments. An argument in favor of this approach is the consistency of the exponent $t = 0.4$ with the inferred $\Delta = 0.6$, which is close to the known value. Note that the force-extension of a chain with a slip-link was discussed in [199] and shown that a loop separated off by a slip-link is confined within a Pincus-de Gennes blob. We also note that results corresponding to delocalisation in force-size relations were reported in [200, 201]. An entropic scale was conceived in [202]: Separating two chains with fixed topology but allowing them to exchange length (e.g., through a small hole in a wall) would enable one to infer the localization behavior of a knot by comparing the equilibrium balance of this knot with a slip-link construct of known DOF until the average length on both sides coincides. The preliminary results in [202] are, however, shadowed by finite-size effects of the accessible system size, as limited by computation power.

Thus, there exists simulations results pointing in both directions, knot localization and delocalization. As the latter may be explained by finite size effects, it seems likely that (at least simple) knots in 2D and 3D localise in the sense that the knot region occupies a portion of the chain that is significantly smaller in comparison to the entire chain. In particular, this

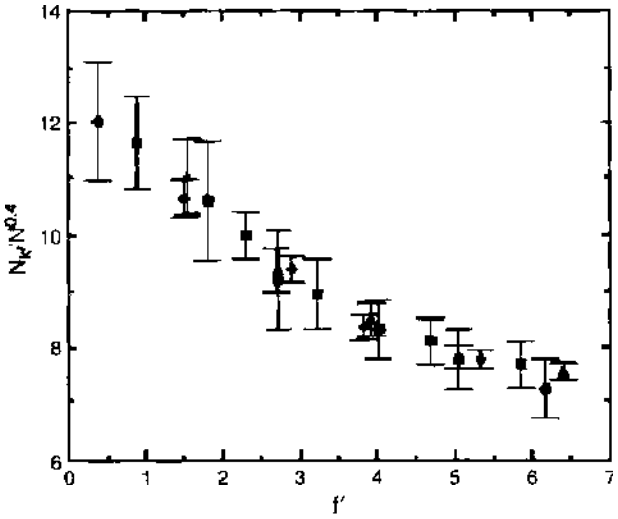


Figure 25. Scaled knot size $N_k/N^{0.4}$ *versus* the dimensionless force f', for chains of lengths $N = 225$ to 750. Reprinted with permission from [198], O. Farago et al., *Europhys. Lett.* 60, 53 (2002). © 2002, EDP Sciences.

would imply that the average size of the knot region $\langle \ell \rangle$ scales with the chain length Na with an exponent less than one, such that

$$\lim_{N\to\infty} \frac{\langle \ell \rangle}{N} = 0 \tag{32}$$

Below, we show from analytical grounds that such a localisation is a natural consequence of interactions of a chain of fixed topology with fluctuations. We note, however, that conclusive results for knot localisation may in fact come from experiments: Manipulation of single chains such as DNA can be performed for rather long chains, making it possible to reach beyond the finite-size corrections inherent in, for example, the force-extension simulations mentioned above.

Before proceeding to these analytical approaches, we note that there have also been performed simulations of knotted chains under nondilute conditions [203, 204]. In (pseudo) 2D, these have found delocalization of the knot, that is, $\lim_{N\to\infty}\langle \ell \rangle/N = const$. We come back to these simulations below in connection with the modeling of dense and Θ-knots.

3.7. Flattened Knots in Dilute and Dense Phases

Analytically, knots are a hard problem to tackle. Statistical mechanical treatments of permanently entangled polymers are so difficult to treat since topological restrictions cannot be formulated as a Hamiltonian problem but appear as hard constraints partitioning the phase space [5, 43, 205, 206].[16] A segment of a 3D knot, in other words, can move without feeling the constraints due to the non-trivial topology of a knotted state, until it actually collides with another segment. The accessible phase space of DOF is therefore characterized by inequalities.[17]

Consequently, only a relatively small range of problems have been treated analytically, starting with the seminal papers by Edwards [207, 208], in which he considers the classification of topological constraints in polymer physics. De Gennes addressed the problem of tight knot motion along a polymer chain using scaling arguments for the activation of free length inside the knot region, producing a double-exponential expression for the corresponding time scale [195], which might explain the extreme long-time contributions in the

[16] For comparison, self-avoidance in 3D is usually treated as a perturbation, that is, as a "soft constraint," in analytical studies [43].

[17] Although a similar statement is true for polymer networks in 3D, the field theoretical results for their critical exponents are in fact obtained as *averages* over *all* topologies. For instance, the exponent ν entering the gyration radius of a 3D polymer ring counts all knotted states [172].

relaxation time spectrum of permanently entangled polymers [42, 44, 45, 48, 194]. Some analytical results were obtained for a pair, or an "Olympic" gel of entangled polymer rings, see, for instance, [209–213]. In a mean field approach based on the Kauffman invariant the entropy of knots was investigated in references [214–216]. Similarly, some statistical properties of random knot diagrams were investigated in [217, 218]. However, some insight can be gained on the basis of phenomenological models, which we will come back to below. Here, we continue with an analytical study of flat knots.

One possibility to treat knotted polymer chains analytically is to confine the DOF of the knot to motion in 2D, only.[18] The knot, that is, is preserved, as at the crossings the chain is allowed to form an over-/underpassing, while the rest of the knot is confined to 2D. Such a confinement can in fact be experimentally realised in various ways. Thus, the chain can be confined between two close-by glass slabs, as demonstrated by [219]. It can be pressed flat on a surface by gravitation or similar forces, for instance in macroscopic systems [181, 220]. Finally, the chain can be adhesively bound to a membrane and still reach configurational equilibrium, as experimentally shown for DNA in references [221, 222]. An example for a knot adsorbed to a polymer-coated mica surface is shown in Fig. 26. From such flat knots as discussed in the remainder of this section, we will be able to infer certain generic features also for 3D knots.

A flat knot therefore corresponds to a polymer network in 2D, but the orientation of the crossings is preserved, such that the network graph actually coincides with a typical knot projection [137–139], as shown in Fig. 27 on the left. This projection of the trefoil, and similar projections for all knots, displays the knot with the essential crossings. A flat knot can, in principle acquire an arbitrary number of crossings by Reidemeister moves; for instance, the bottom left segment of the flat trefoil can slide under the vicinal segment, creating a new pair of vertices, and so on. However, we suppose that such transient additional loops are sufficiently short-lived so that we can neglect them in our analysis. Then, we can apply results from scaling analysis of polymer networks of the most general type shown in Fig. A2, as reviewed in the Appendix. We note that from the Monte Carlo simulations we performed it may be concluded that such additional vertices can in fact be neglected.

3.7.1. Flat Knots in Dilute Phase

We had previously found that for the F8-structure, the PDF for the size of each loop is peaked at $\ell \to 0$ and $\ell \to L$. From the scaling analysis for self-avoiding polymer networks, we concluded strong localization of one subloop. For more complicated structures, the joint probability to find the individual segments with given lengths s_i is expected to peak at the edges of the higher dimensional configuration hyperspace. Some analysis is necessary to find the characteristic shapes. Let us consider here the simplest nontrivial knot, the (flat) trefoil knot 3_1 shown in Fig. 27. Each of the three crossings is replaced with a vertex with four outgoing legs, and the resulting network is assumed to separate into a large loop and a multiply connected region which includes the vertices. Let $\ell = \sum_{i=1}^{5} s_i$ be the total length of all segments contained in the multiply connected knot region. Accordingly, the length of the large loop is $s_6 = L - \ell$.

In the limit $\ell \ll L$, the number of configurations of this network can be derived in a similar way as in the scaling approach followed for the F8. This procedure determines the concrete behaviour of the scaling form

$$\omega_{\mathrm{III}} \simeq \mu^{L}\, \mathscr{W}_{\mathrm{III}}\left(L-\ell, \ell, \frac{s_1}{\ell}, \frac{s_2}{\ell}, \frac{s_3}{\ell}, \frac{s_4}{\ell}\right) \tag{33}$$

including the scaling function $\mathscr{W}$ that depends on altogether six arguments. The index III is chosen according to Fig. 28, where the flat trefoil configuration in the dense phase appears at position III of the scheme (explained below). After some manipulations, the number of DOF yields in the form [180]

$$\omega_{\mathrm{III}}(\ell, L) \sim \mu^{L}(L-\ell)^{-d_F}\ell^{-c} \tag{34}$$

[18] The system in a strict sense is not fully two-dimensional, as we allow crossings.

Figure 26. AFM tapping-mode images of flattened DNA-knot absorbed on a surface. The substrate surface used is AP-mica (freshly cleaved mica reacted with an amino terminal silane to make is positively charged). The DNA knots used are extracted from bacteriophage P4: the DNA is a 11.4 kbp molecule (with a 1.4 kbp deletion resulting in a final length of 10 kbp), which has two cohesive ends. They are not covalently closed, thus no supercoiling is present. On the left, the knot absorbed out of the 3D bulk on to the surface is strongly trapped, that is, the knot is "projected" onto the surface without any equilibration. The knot appears rather delocalised. Right: The knot is localized in a relatively small region of the DNA chain: these kind of images are obtained when the ionic strength in the solution is increased, this reduces the interaction between the DNA and the surface allowing the knots to partially equilibrate on the surface. Courtesy F. Valle and G. Dietler [221].

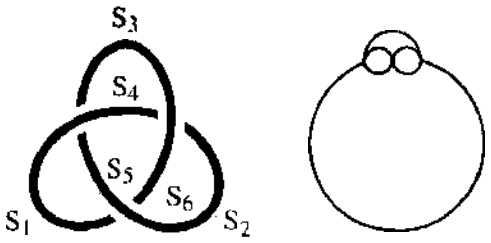


Figure 27. Flat trefoil knot with segment labels. On the right, a schematic representation of a localised flat trefoil with one large segment is shown.

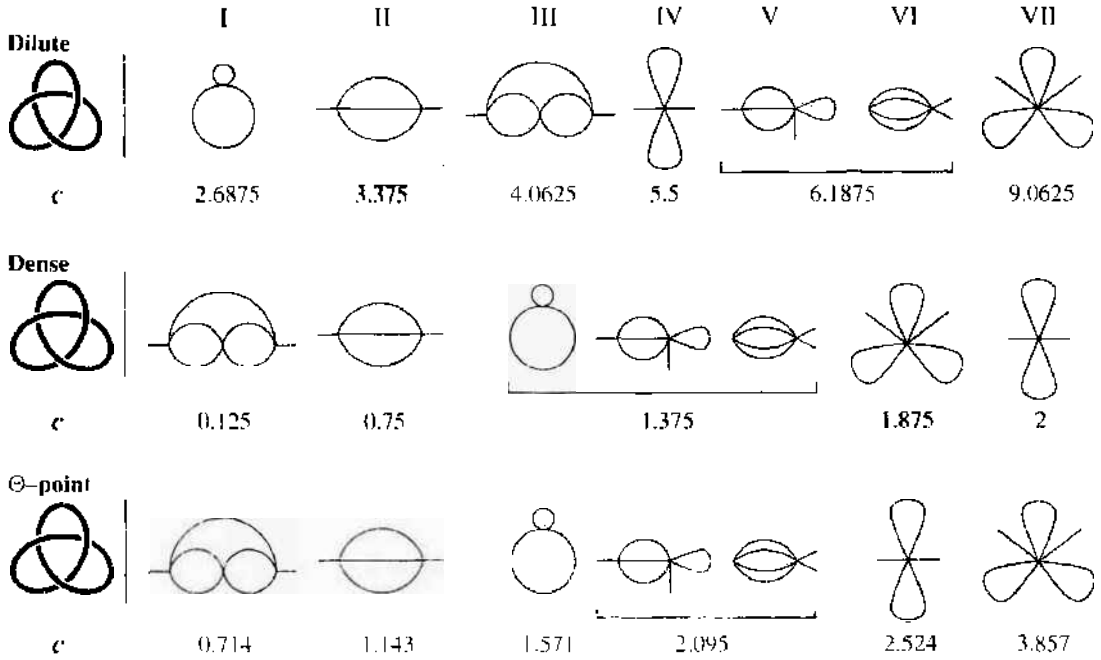


Figure 28. Hierarchy of the flat trefoil knot 3_1. Upper row: dilute phase. Middle row: dense phase. Bottom row: Θ-phase. To the left of each row, the trefoil projection is shown. It splits up into the hierarchies of configurations, with exponents c below each contraction. The small protruding legs represent the big central loop, compare, for instance, Fig. 27 on the right with position III of the top row. See text for details.

with the scaling exponent

$$c = -(\gamma_{\rm III} - 1 + d\nu) - m, \quad m = 4 \tag{35}$$

Here, $m = 4$ corresponds to the number of independent integrations over the segments s_i ($i = 1, \ldots, 4$) of the knot region, as we only retain the cumulative size $\ell = \sum_{i=1}^{4} s_i$ of the knot region. Putting numerical values, we find $c = 65/16$, i.e., *strong localization*.

However, some care is necessary in performing these integrations, as the scaling function $\mathcal{H}_{\rm III}$ may exhibit nonintegrable singularities if one or more of the arguments s_i/ℓ tend to 0. The geometries corresponding to these limits (edges of the configuration hyperspace) represent *contractions* of the original trefoil network $\mathcal{G}_{\rm III}$ in the sense that the length of one or more of the segments s_i is of the order of the short-distance cutoff a. If such a short segment connects different vertices, they cannot be resolved on larger length scales, but appear as a single, new vertex. Thus, each contraction corresponds to a different network $\mathcal{G}_i$, which may contain a vertex with up to eight outgoing legs. For the flat trefoil knot, there exist six different contractions, as grouped in Fig. 28 around the original flat trefoil at position III. As an example, in the top row of Fig. 28 contraction VI follows from the original trefoil III if the uppermost segment becomes very small, and similarly the network VII emanates from contraction VI if one of the four symmetric segments becomes very small. For each of these networks, one can calculate the corresponding exponent c in a similar way as above, leading to the general expression

$$c = 2 + \sum_{N \geq 4} n_N \left\{ \frac{N}{2}(d\nu - 1) + (|\sigma_N| - d\nu) \right\} \tag{36}$$

The σ_N are given in Eq. (A22). In Fig. 28, the various contractions are arranged in increasing exponent c.

Our scaling analysis relies on an expansion in $a/\ell \ll 1$, and the values of c determine a sequence of contractions according to higher orders in a/ℓ: The *smallest* value of c corresponds to the most likely contraction, while the others represent corrections to this leading scaling behavior, and are thus less and less probable (see Fig. 28). To lowest order, the trefoil behaves like a large ring polymer at whose fringe the point-like knot region is located. At the next level of resolution, it appears contracted to the figure-eight shape $\mathcal{G}_{\rm I}$. For more accurate data, the higher order shapes II to VII may be found with decreasing probability. Interestingly, the original uncontracted trefoil configuration ranks third in the hierarchy of shapes.

These predictions were checked by MC simulations with the same conditions as described in Section 3.5, to prevent intersection. The flat trefoil knot was prepared from a symmetric, harmonic 3D representation with 512 monomers, which was collapsed and then kept on a hard wall by the "gravitational" field $V = -k_B T h/h^*$ perpendicular to the wall, where h is the height and h^* was set to 0.3 times the bead diameter. Configurations corresponding to contraction I are then selected by requiring that besides a large loop, they contain only one segment larger than a preset cutoff length (taken to be 5 monomers), and similarly for contraction II. The size distributions for such contractions, as well as for all possible knot shapes are shown in Fig. 29. The tails of the distributions are indeed consistent with the predicted power laws, although the data (especially for contraction II) is too noisy for a definitive statement.

Our scaling results pertain to all flat prime knots. In particular, the dominating contribution for *any* prime knot corresponds to the figure-eight contraction $\mathcal{G}_{\rm I}$, as Eq. (36) predicts a larger value of the scaling exponent c for any network $\mathcal{G}$ other than $\mathcal{G}_{\rm I}$. Accordingly, Fig. 30 demonstrates the tightness of the prime knot 8_{19}. Composite knots, however, can maximize the number of configurations by splitting into their prime factors as indicated in Fig. 30 for $3_1\#3_1$. Each prime factor is tight and located at the fringe of one large loop, and accounts for an additional factor of L for the number of configurations, as compared to a ring of length L without a knot.

Flat knots can experimentally be produced by "projecting" a dilute 3D knot from the bulk onto a mica surface, on which the knot is absorbed. Variation of the ionic strength in the

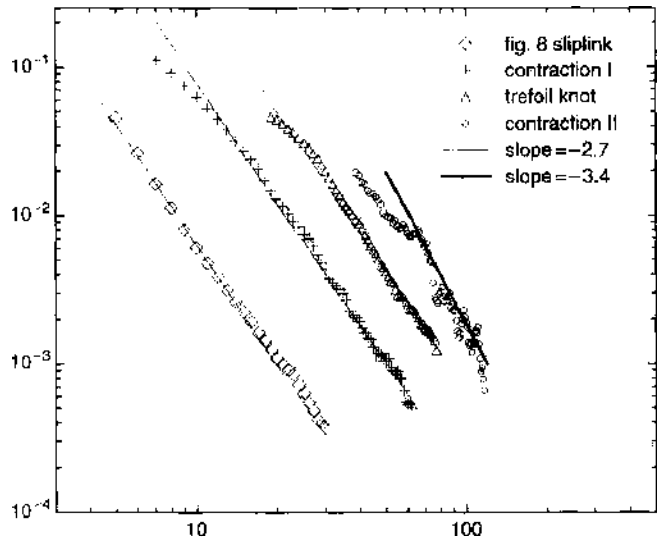


Figure 29. Power law tails in PDFs for the size ℓ of tight segments: As defined in the figure, we show results for the smaller loop in a figure-eight structure, the overall size of the trefoil knot, as well as the two leading contractions of the latter.

solution determines whether the knot is going to be strongly trapped on the surface such that, once captured on the surface, it is completely immobilised (small ionic strength); or whether the adsorption is weaker such that the knot can (slowly) equilibrate while being confined to 2D, that is, equilibrate as a flat knot. In Fig. 26, both situations are shown. Whereas the strongly trapped knot appears rather delocalized, the tendency toward forming a tight knot in the weakly trapped case is distinct [221].

3.7.2. Flat Knots Under Θ and Dense Conditions

In many situations, polymer chains are not dilute. Polymer melts, gels, or rubbers exhibit fairly high densities of chains, and the behavior of an individual chain in such systems is significantly different compared to the dilute phase [42, 43, 194]. Similar considerations apply to biomolecules: in bacteria, the gyration radius of the almost freely floating ring DNA may sometimes be larger than the cell radius itself. Moreover, under certain conditions, there is a non-negligible osmotic pressure due to vicinal layers of protein molecules, which tends to confine the DNA [223–225]. In protein folding studies, globular proteins in their native state are often modeled as compact polymers on a lattice (see [226] for a recent review).

A polymer is considered dense if, on a lattice, the fraction f of occupied sites has a finite value $f > 0$. This can be obtained by considering a single polymer of total length L inside a box of volume V and taking the limit $L \to \infty$, $V \to \infty$ in such a way that $f = L/V$ stays finite [178, 227, 228]. Alternatively, dense polymers can be obtained in an infinite volume through the action of an attractive force between monomers. Then, for temperatures T below the collapse (Theta) temperature Θ, the polymers collapse to a dense phase, with a density $f > 0$, which is a function of T [178, 229–231]. For a dense polymer in d dimensions,

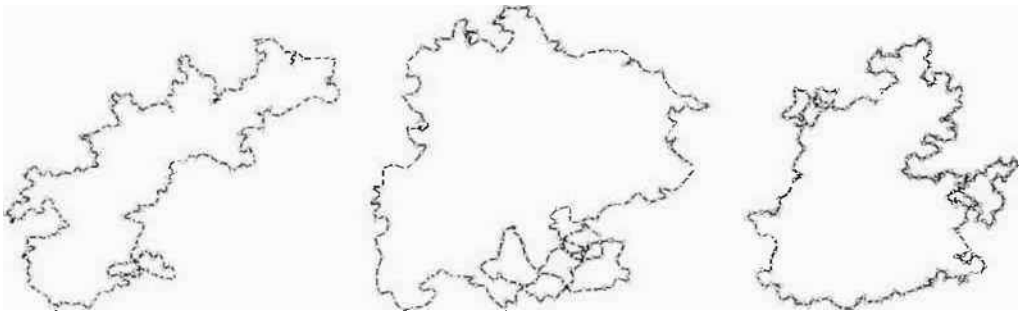


Figure 30. Typical equilibrium configurations of 256-mer chains for the trefoil 3_1, the prime knot 8_{19}, and the composite knot $3_1\#3_1$ consisting of two trefoils, in $d = 2$. The initial conditions were symmetric in all cases.

the exponent ν, defined by the radius of gyration $R_g \sim L^{\nu}$, becomes $\nu = 1/d$. The limit $f = 1$ is realized in Hamiltonian paths, where a random walk visits every site of a given lattice exactly once [232, 233]. Dense polymers may be related to 2D vesicles and lattice animals (branched polymers) [234–237].

As studied in [238], the value of the exponent c for the 2D dense F8 is (compare to the Appendix)

$$c = -\gamma_{F8} = \frac{11}{8} = 1.375 \tag{37}$$

implying that the smaller loop is *weakly localized*. This means that the probability for the size of each loop is peaked at $\ell = 0$ and, by symmetry, at $\ell = L$. An analogous reasoning for the 2D F8 at the Θ point gives

$$c = \frac{11}{7} = 1.571 \tag{38}$$

In both cases the smaller loop is weakly localized in the sense that $\langle l \rangle_c / L \to 0$. Figure 31 shows the symmetric initial and a typical equilibrium configuration for periodic boundary conditions obtained from Monte Carlo (MC) simulations, see [238] for details. In Fig. 31, the lines represent the bonds (tethers) between the monomers (beads, not shown here). The three black dots mark the locations of the tethered beads forming the slip-link in 2D. The initial symmetric configuration soon gives way to a configuration with $\ell \ll L$ on approaching equilibrium. Figure 32 shows the development of this symmetry breaking as a function of the number of MC steps. We note, however, that the fluctuations of the loop sizes in the "stationary" regime appear to be larger in comparison to the dilute case studied in reference [239], compare Fig. 23. We checked that for densities (area coverage) above 40% the scaling behavior becomes independent of the density. (The above simulation results correspond to a density of 55%.) The size distribution data is well fitted to a power law (for over 1.5 decades with 1024 monomers), and the corresponding exponent with 512 and 1024 monomers in Fig. 32 is in good agreement with the predicted value (37).

For our MC analysis, we again used a hard core bead-and-tether chain, in which self-crossings were prevented by keeping a maximum bead-to-bead distance of 1.38 times the bead diameter, and a maximum step length of 0.15 times the bead diameter. To create the dense F8 initial condition, a free F8 is squeezed into a quadratic box with hard walls. This is achieved by starting off from the free F8, surrounding it by a box, and turning on a force directed towards one of the edges. Then, the opposite edge is moved toward the centre of the box, and so on. During these steps, the slip-link is locked, that is, the chain cannot slide through it, and the two loops are of equal length during the entire preparation. Finally, when the envisaged density is reached, the hard walls are replaced by periodic boundary conditions, and the slip-link is unlocked. After each step, the system is allowed to relax for times larger than the localization times occurring at the main stage of the run.

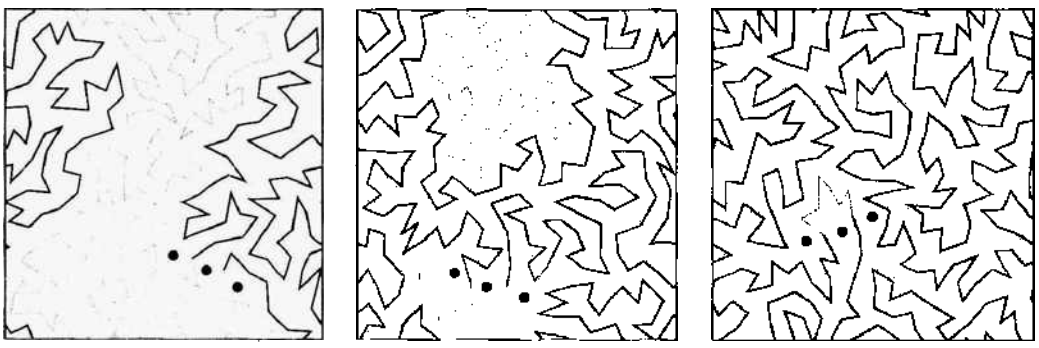


Figure 31. Symmetric ($\ell = L/2 = 128$) initial configuration of a 2D dense F8 (left), intermediate (centre), and equilibrium configuration (right) with periodic boundary conditions. The two different grey values correspond to the two subloops created by the slip-link. The slip-link itself is represented by the three (tethered) black dots.

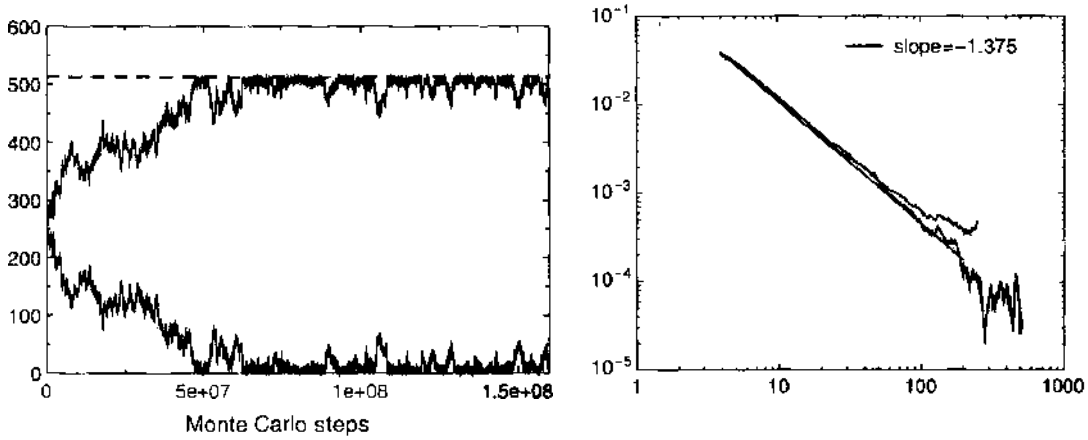


Figure 32. Left: The breaking of the initial symmetry between the two loops of the F8, as a function of MC steps at 55% area coverage. Right: The loop size probability distribution $p(\ell)$ at $\rho = 55\%$ area coverage, for the F8 with 512 (top) and 1024 (bottom) monomers. The power law with the predicted exponent $c = 1.375$ in Eq. (37) is indicated by the dotted line.

A similar analysis as for the dense/Θ-F8 structure and the dilute flat trefoil reveals the number of DOF for the flat dense trefoil in the form [238]

$$\omega_3(\ell, L) \sim \omega_0(L)\ell^{-c} \tag{39}$$

with $c = -\gamma_3 - m$, where $\gamma_3 = -33/8$ from Eq. (A24) in the appendix ($\mathcal{L} = 4$, $n_4 = 3$) and $m = 4$ is the number of independent integrations over chain segments. Thus, $c = 1/8 < 1$ which implies that the dense 2D trefoil is *delocalized*. As above, we have to consider the various possible contractions of the flat knot. For dense polymers, the present scaling results show that both the original trefoil shape ($c = 1/8 < 1$, see above) and position II ($c = 3/4 < 1$) are in fact *delocalized* and represent equally the leading scaling order (cf. middle row of Fig. 28). The F8 is only found at the third position and is weakly localized ($c = 11/8 > 1$). Thus, in an MC simulation of the dense 2D trefoil, we predict that one mainly observes delocalized shapes corresponding to the original trefoil and position II in Fig. 28, and further, with decreasing probability, the weakly localised F8 and the other shapes of the hierarchy in Fig. 28.

These predictions are consistent with the numerical simulations of reference [203], who observe that the mean value of the second largest segment of the simulated 2D dense trefoil configurations grows linearly with L, and conjecture the same behaviour also for the other segments, corresponding to the delocalization of the trefoil obtained above.

An analogous reasoning can be applied to the 2D trefoil in the Θ phase. We find that in this case that the leading shape is again the original (uncontracted) trefoil, with $c = 5/7 < 1$. This implies that the 2D trefoil is *delocalized* also at the Θ point. All other shapes are at least weakly localized, and subdominant to the leading scaling order represented by the original trefoil. The resulting hierarchy of shapes is shown in Fig. 28 (bottom part).

3.8. 3D Knots Defy Complete Analytical Treatment

As already mentioned, 3D knots correspond to a problem involving hard constraints that defy a closed analytical treatment. It may be possible, however, that by a suitable mapping to, for instance, a field theory, an analytical description may be found. This may in fact be connected to the study of knots in diagrammatic solutions in high energy physics [240]. There exists a fundamental relation between knots and gauge theory as knot projections and Feynman graphs share the same basic ingredients corresponding to a Hopf algebra [138]. However, up to now no such mapping has been found, and a theoretical description of 3D knots based on first principles is presently beyond hope. To obtain some insight into the statistical mechanical behavior of knotted chains, one therefore has to resort to simulations studies or experiments. In addition, a few phenomenological models for both the equilibrium

and dynamical behavior of knots have been suggested such as in [164, 191, 193, 196, 200, 201, 241, 242].

When discussing numerical knot studies, we already mentioned the Flory-type model brought forward in [191, 192]. There, a knot is thought of as a polymer construct of C interlocked loops of equal length N/C, where C is the number of essential crossings, a knot invariant. Arguing that the C individual loop-components repel each other and that each occupies a volume $\simeq (N/C)^{3\nu}$, the total volume of the knot is obtained as $V \simeq C(N/C)^{3\nu}$. This produces the gyration radius

$$R_g \simeq N^{3/5} C^{-4/15} \tag{40}$$

as a function of the knot type characterized by C. One may argue that the differences in the knot size for the different knot types corresponding to the same C may be included in the prefactor, that is independent of the chain length N. Obviously, this model of equal loop sizes is equivalent to a completely delocalized knot, as all loops are of identical size. The functional behavior (40) on C could be corroborated by simulations in [191] shown in Fig. 33; however, as discussed above, it seems that the used algorithm in fact is dominated by finite-size effects causing delocalization even of the comparatively long chains used in these simulations.

The latter statement is in fact equivalent to another Flory-type approach to knotted polymers reported in [193]. In this model, the knot is thought of as an inflatable tube: for a very thin tube diameter, the tube is equivalent to the original knot conformation; inflating the tube more and more will increasingly smoothen out the shape until a maximally inflated state is reached. The knot is then characterized by the aspect ratio

$$p = \frac{L}{D} \cong \frac{aN}{aN_D}, \quad \text{therefore} \quad 1 \leq p \leq N \tag{41}$$

between length L and maximum tube diameter D. It appears that p is a (weak) knot invariant, and can be used to characterize the gyration radius of the knot. It is clear that, by construction, the aspect ratio describes a totally delocalized knot, and indeed it turns out that in good solvent, the gyration radius shows the dependence $R_g \simeq AN^{3/5}\tau^{1.5}p^{-4/15}$, where τ is the (dimensionless) deviation from the Θ temperature [193]. Obviously, the aspect ratio appears to be proportional to the number of essential crossings in comparison to expression (40). We note that similar considerations are employed in [242], including a comparison to

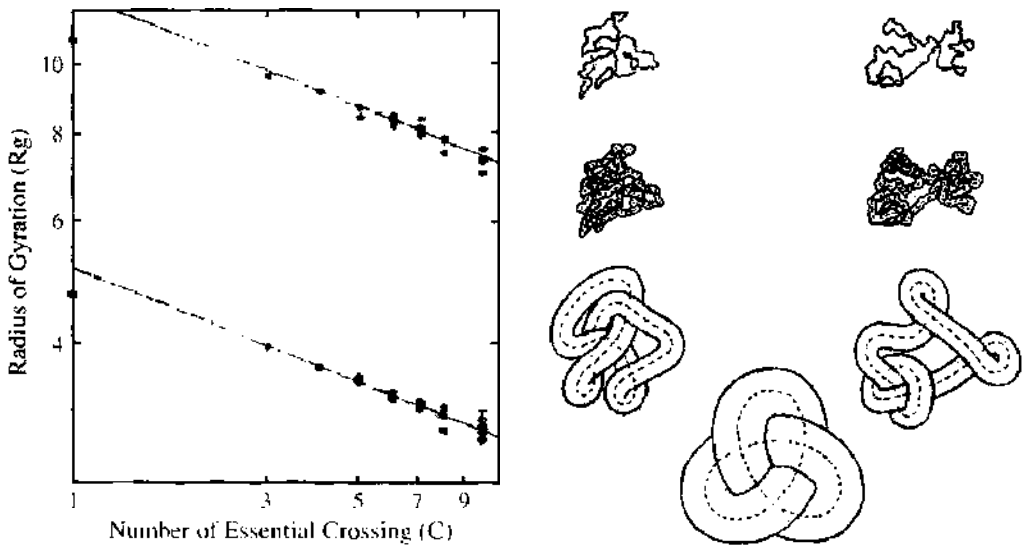


Figure 33. Left: Gyration radius as function of the number of essential crossings from [191]. Reprinted with permission from [191], S. R. Quake, *Phys. Rev. Lett.* 73, 3317 (1994). © 1994, American Physical Society. Right: Illustration of the inflated tube model from [193]. Reprinted with permission from [193], A. Yu. Grosberg et al., *Phys. Rev. E* 54, 6618 (1996). © 1996, American Physical Society.

the entropy of a tight knot, finding comparable entropic likelihood. The modeling based on the aspect ratio p is further refined in [164].

In the case of 3D knots, the knot localization problem is presently not resolved for knots of all complexity. Probably only detailed simulations studies of higher order knots will make it possible to decide for the various models of 3D knots. Major contributions are also expected from single molecule experiments, for instance, from force-extension measurements along the lines of the simulations study in [198], the advantage of experiments being the fact that it should be possible to go toward rather high chain lengths that are inaccessible in simulations.

To overcome similar difficulties in the context of the entropic elasticity for rubber networks, Ball, Doi, Edwards and coworkers replaced permanent entanglements by slip-links [243–246]. Gaussian networks containing slip-links have been successful in the prediction of important physical quantities of rubber networks [194], and they have been used to study a small number of entangled chains [247]. In a similar fashion, one may investigate the statistical behaviour of single polymer chains in which a fixed topology is created by a number of slip-links. Such "paraknots" can be studied analytically using the Duplantier scaling results [239]. Although these findings may be of some use in the discussion of entropic effects in RNA secondary structure formation, it is not clear how to connect paraknots to the behavior of real 3D knots.

4. DNA BREATHING AND DENATURATION

"A most remarkable physical feature of the DNA helix, and one that is crucial to its functions in replication and transcription, is the ease with which its component chains can come apart and rejoin. Many techniques have been used to measure this melting and reannealing behaviour. Nevertheless, important questions remain about the kinetics and thermodynamics of denaturation and renaturation and how these processes are influenced by other molecules in the test tube and cell" [2]. This remarkable 1974 citation in essence encompasses the problem of local and global denaturation of DNA, whose full description is an ongoing problem in both biochemistry and biological physics. DNA denaturation (or "melting") is due to the breakage of the hydrogen bonds in the Watson-Crick base-pairs, and therefore reversible. There are several mechanisms how such "unzipping" of dsDNA occurs. Under physiological conditions, local DNA-breathing occurs spontaneously due to thermal fluctuations, opening up transient bubbles of a few tens of base pairs [248]. These breathing fluctuations may be supported by single-strand binding proteins, thereby lowering the DNA base pair stability [249]. *In vitro*, melting can be induced by increasing the temperature, or by titration with an acid or alkali [2, 3, 29]. Depending on temperature and pH, the melting can be full or partial. Schematically, figure 34 depicts the denaturation behavior of double-stranded DNA.

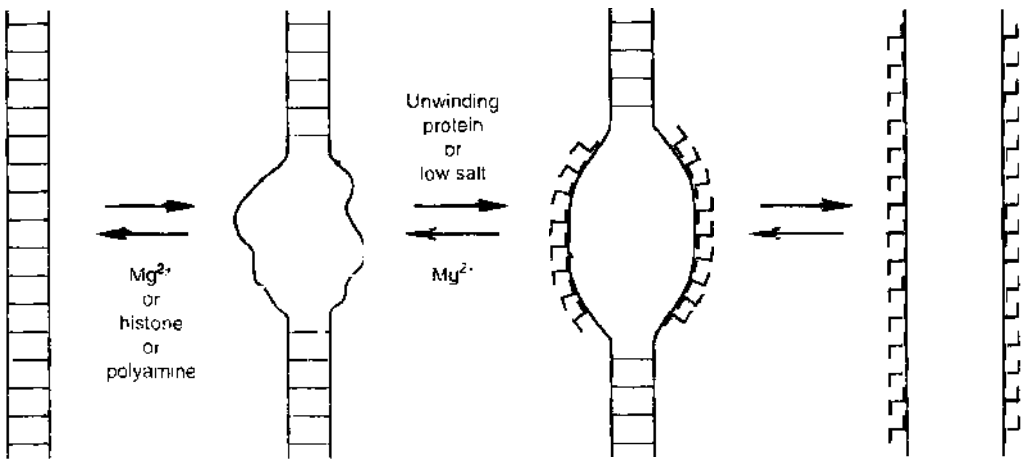


Figure 34. Denaturation of a DNA double-helix. Varying salt conditions and adding single-strand binding proteins ↑ biases the equilibrium distribution of denaturation bubbles. Equivalently, an increase in temperature will foster the tendency toward denaturation.

DNA breathing and the lability of local stretches of the DNA double-helix is essential for numerous physiological processes such as the association of single-stranded DNA binding proteins, and the initiation of replication and transcription. *Vice versa*, denaturation (and thus also renaturation and repair) is strongly biased in the presence of such single-stranded DNA binding proteins [2, 3, 250]. Its quantitative characterization is thus an essential feature in the understanding of numerous genetic processes. Traditional (bulk) methods are mainly based on temperature and titration induced denaturation, which can be monitored in terms of the absorption behavior of UV light [30]. Generally, AT base-pairs melt at lower temperatures than GC; for instance, under physiological salt conditions an AT pair melts at 64°C and a GC at 107°C.[19] Depending on the local and global distribution of ATs in respect to GCs, the distribution of breathing bubbles, or melting regions, along the double-strand as well as the global melting temperature varies. The thermal stability of DNA is due to the Watson-Crick hydrogen bonds between bases of the opposing strands, and of more substantial interactions between closely stacked base-pairs [29, 30]. From melting studies and calorimetry measurement, the thermodynamical data for all local stacking interactions between base-pairs (such as AT/AT, AT/TA, AT/GC, etc.) are known relatively precisely, allowing for sequencing on the basis of the melting data [58]. However, the detailed mechanisms, and the kinetics in particular, of DNA denaturation remain open. We briefly review recent information on the kinetics due to the novel possibility to monitor DNA breathing on the single molecular level by fluorescence methods, and to probe the interaction with single-strand binders in single molecule optical tweezers setups.

The typical behaviour of dsDNA under variation of temperature is shown in Fig. 35, displaying the fraction θ_{ll} of double-helical domains within the DNA as a function of temperature. Figure 36 shows the differential behavior $d\theta_{ll}/dT$. In the latter, relatively sharp peaks occur, prompting speculation whether the melting transition might be of first order.[20] The understanding of the order of the denaturation has received considerable interest in statistical mechanics, and we mention some of the most common approaches.

The DNA-strand undergoing denaturation decomposes into zones of intact double-helix, and single-stranded bubbles, as shown in Fig. 37. In the *Poland-Scheraga model*, energetic bonds in the double-stranded, helical regions of the DNA compete with the entropy gain from the far more flexible, single-stranded loops [29, 30, 252, 253]. The stability of the double helix originates mainly from stacking interactions between adjacent base pairs, aside from the Watson-Crick hydrogen bonds between bases. In addition, the positioning of bases for pairing out of a loop state gives rise to an entropic contribution. The Gibbs free energy $G_{ij} = H_{ij} - TS_{ij}$ for the dissociation of two stacked base pairs i and j is available in terms of the enthalpic and entropic contributions H_{ij} and S_{ij} [58]. In the following, we consider a homopolymer for simplicity, as, for instance, suitable for the AT breathing domain in [111].

Generalization to a heteropolymer can be included in a rather straightforward way, due to the assumption of statistical independence of individual bubbles [29, 177, 254]. For an AT-homopolymer ($i = j = \mathrm{AT}$), the Gibbs free energy per base pair in units of $k_B T$ yields as $\gamma \equiv \beta G_{ij}/2 = 0.6$ at 37°C for standard salt conditions (0.0745 M-Na^+). Similarly, for a GC-homopolymer one finds the higher value of $\gamma = 1.46$ at 37°C. The condition $\gamma = 0$ defines the melting temperature T_m [58], thus $T_m(\mathrm{AT}) = 66.8°\mathrm{C}$ and $T_m(\mathrm{GC}) = 102.5°\mathrm{C}$ (we assume that G_{ij} is linear in T, cf. [255]). Above T_m, γ becomes negative. For given $\gamma = \gamma(T)$, the statistical weight for the dissociation of n base pairs obtains as

$$W(n) = \exp(-n\gamma) \tag{42}$$

Additional contributions arise upon formation of a loop within the dsDNA. Firstly, an initial energy barrier has to be overcome, which we denote as γ_1 in units of $k_B T$. From fitting melting curves to long DNA, $\gamma_1 \approx 10$ was obtained, so that the statistical weight for the initiation of a loop (cooperativity parameter), $\sigma_{ll} = \exp(-\gamma_1)$, is of order 10^{-5} [30, 58]. As the

[19] This is also the reason for the importance of the TATA box for protein binding in transcription; compare Section 2.4.1.

[20] The modeling of DNA denaturation in terms of 1D Ising-like models may suggest that there should not exist a true phase transition [251].

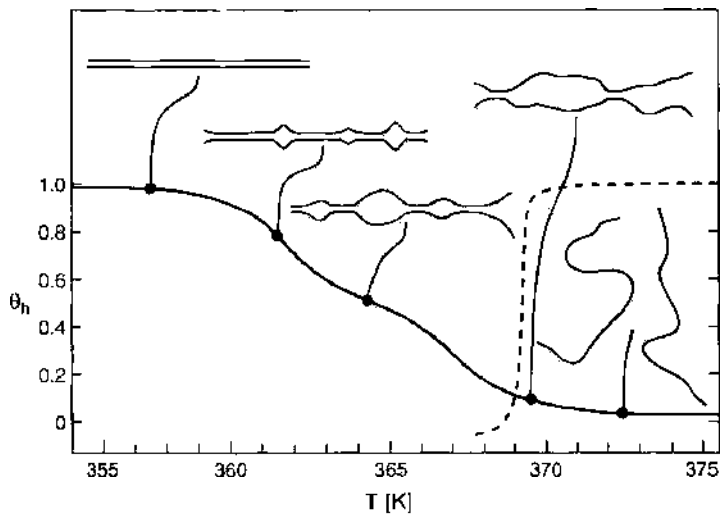


Figure 35. Fraction θ_h of double-helical domains within the DNA as a function of temperature. Schematic representation of $\theta_h(T)$, showing the increased formation of bubbles and unzipping from the ends, until full denaturation has been reached.

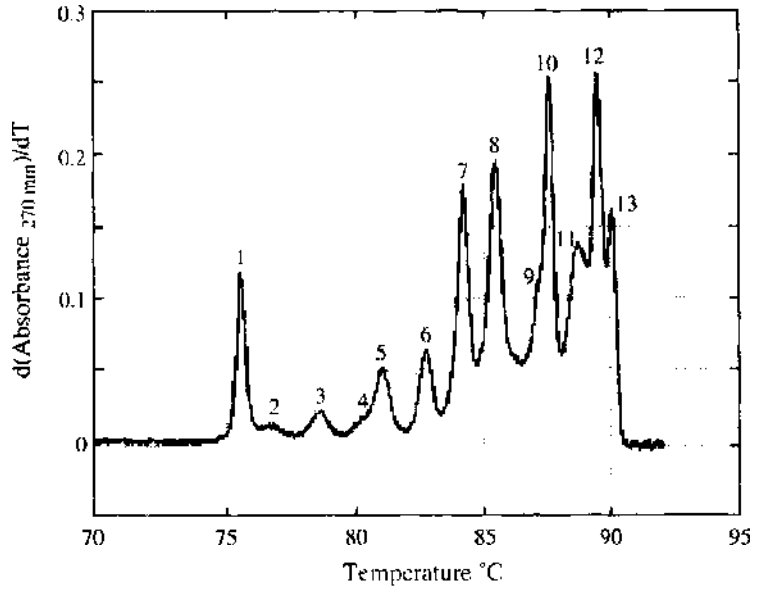


Figure 36. Fraction θ_h of double-helical domains within the DNA as a function of temperature. Differential behavior of the measured absorbance as function of temperature. The sharp spikes are distinct, despite the rather short sequence of some 4.7 kbp. Reprinted with permission from [58], R. D. Blake et al., *Bioinformatics* 15, 370 (1999), © 1999, Oxford University Press.

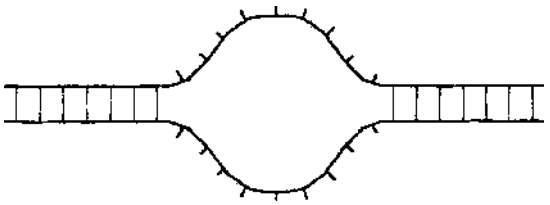


Figure 37. Denaturation bubble embedded in double-stranded helical domains. Due to the heteropolymer character of DNA, bubbles are symmetric and do generally not form secondary structure within or between the single-stranded segments, conforming to the perfect-matching model. In a homopolymer, additional DOF due to imperfect matching may occur [29].

free energy to extend an existing loop by one base pair is smaller than $k_B T$, DNA melts as large cooperative domains [29, 30, 58]. We note that a small value of σ_0 was shown to be essential for suppression of mutations [226], and is fairly independent of the precise value of the critical loop exponent [254, 256]. Closer to, but still below the melting temperature T_m, the bubbles become larger, and long range interactions beyond nearest neighbours become more important. In this case, the probability of bubble formation is larger, and γ_1 ranges between 3 and 5, thus $\sigma_0 \lesssim 0.05$ [254] (also see Section 5 in [30]). Secondly, once a loop of n base pairs has formed, there is a weight $f(n)$ of mainly entropic origin, to be detailed below. The additional weight of a loop of n open base pairs is thus

$$\Omega(n) = \sigma_0 f(n) \tag{43}$$

For the bubble entropy loss one usually assumes the form [30, 58]

$$f(n) \simeq (n + D)^{-c} \tag{44}$$

where D denotes an empirical stiffness parameter [257, 258]. Often, the choice $D = 1$ is made [30, 58]. The loop closure exponent c depends on the specific model used to represent the melting configuration. Most commonly, the value $c = 1.76$ is employed, corresponding to a self-avoiding loop [30, 58, 259]. We come back to the question of the exact value of c below. Here, we note that for smaller bubbles (in the range of 1 to a few tens of base pairs), the appropriate form of $f(n)$ is more involved. In particular, $f(n)$ will depend on the finite persistence length of ssDNA (about eight bases), on the specific base sequence, and possibly on interactions between dissolved but close-by base pairs [30, 111]. Therefore, the knowledge of $f(n)$ provides information on these microscopic interactions. For simplicity, we here adopt the simple form $f(n) = (n + 1)^{-c}$ for all $n > 0$, and consider the loop weight

$$\Omega(n) = \sigma_0 (n + 1)^{-c} \tag{45}$$

We show that at the melting temperature the results for the relaxation times for the bubbles are different for the available values $c = 1.76$ and $c = 2.12$ quoted above, although not significantly.

Focusing on a single bubble in dsDNA and neglecting its interaction with other bubbles, we obtain the total free energy $\mathcal{F}(n)$ of a single bubble with $n > 0$ open base pairs in the form

$$\beta \mathcal{F}(n) = -\log(W(n)\Omega(n)) = n\gamma(T) + \gamma_1 + c \log(n + 1) \tag{46}$$

where the dependence on the temperature T enters only via $\gamma = \gamma(T)$. This approximation is justified as long as the bubbles are not too large since due to $\sigma_0 \ll 1$ the mean distance between bubbles [$\sim 1/\sigma_0$ [29]] is large [254]. It also corresponds to the situation studied in the recent experiment by [111] We show the free energy $\mathcal{F}(n)$ in Fig. 38 for $c = 1.76$ and for the parameters of an AT-homopolymer, for various temperatures. At low temperature ($T = 37°C$), after opening a bubble tends to close; at the melting temperature $T_m = 66°C$, there is only a weak tendency toward closure due to the entropy loss; above the melting temperature ($T = 100°$), after a weak initial activation, the bubble tends to open up further until the DNA is denatured. Obviously, the one-bubble approximation in a general dsDNA breaks down close to and above T_m; however, the approximation may still be meaningful for designed DNA, in which AT-rich zones are clamped by GC-rich ones.

Within this Poland-Scheraga framework, the dynamics of bubbles can be modelled in terms of a Fokker-Planck equation for the probability density function $P(n, t)$ to find a bubble of size n at time t. Such an approach is based on the generally accepted multistate unzipping model, according to which the double strand opens by successively breaking Watson-Crick bonds, like opening a zipper [261, 262]. Thermal fluctuations cause a random walk-like propagation of the zipper locations at both ends of the bubble-helix joints, and this mechanism gives rise to the diffusive nature of $P(n, t)$. In [260], we derived the dynamical equation

$$\frac{\partial P(n,t)}{\partial t} = D\left(\frac{\partial}{\partial n}\left\{\gamma + \frac{c}{n+1}\right\} + \frac{\partial^2}{\partial n^2}\right) P(n,t) \tag{47}$$

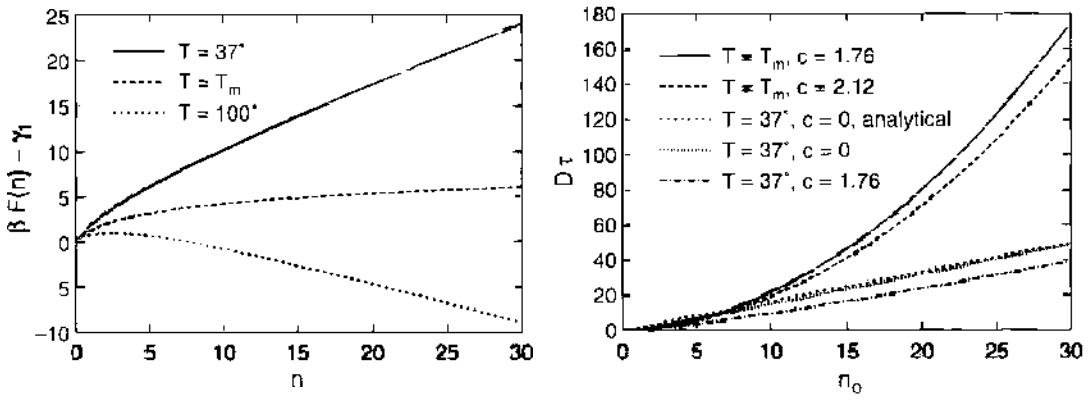


Figure 38. Left: Variable part of the bubble free energy (46). $\beta\bar{\mathcal{F}}(n) - \gamma_1 = n\gamma(T) + c\log(n+1)$, as a function of the bubble size n for $T = 37°C$ ($\gamma \approx 0.6$), $T_m = 66°C$ ($\gamma = 0$), and $T = 100°C$ ($\gamma \approx -0.5$). Right: Characteristic bubble closing times τ as a function of initial bubble size n_0 for an AT-homopolymer, obtained from the Fokker-Planck equation approach. See text and [260] for more details.

Here, the first term on the right-hand side described the drift stemming from the gradient of the free energy, $d\bar{\mathcal{F}}/dn$, and the second term represents the diffusion within n-space (i.e., the relative diffusion of the two zipper locations). Thus, we arrived at a reduced 1D description of the bubble dynamics in a homopolymer, with the bubble size n as the effective "reaction coordinate." For a heteropolymer, also the position of the bubble within the double helix, that is, the index m of the first open base pair, becomes relevant. In this case, the bubble free energy $\bar{\mathcal{F}}$ and thus the PDF depend on both m and n, and on the specific base sequence; the corresponding generalization of Eq. (47) is straightforward. In a random sequence, additional phenomena may occur, such as localization of larger bubbles [254]. Finally, to establish the Fokker-Planck equation (47), we assume that changes of the bubble size n occur slower than other degrees of freedom of the PS free energy within the bubble region (e.g., Rouse-Zimm modes). This assumption seems reasonable due to the long bubble dynamics' relaxation time scales of 20 to 100 μs [111], and the good approximation of bubble independence [254].

The formulation in terms of the Fokker-Planck equation (47) makes it possible to derive the characteristic times for bubble closing and opening in terms of a first passage time problem. That is, for bubble closing, the associated mean closing time follows as the mean first passage time to reach bubble size $n = 0$ after starting from the initial bubble size n_0. In the regime of low temperature (T significantly lower than T_m), the drift term consists of two contributions, the constant drift $D\gamma$ and the loop closure component $Dc/(n+1)$, which decreases with n. For large n, we can therefore approximate the drift by the constant term $D\gamma$, and in this limit the Fokker-Planck equation (47) is equivalent to the continuum limit of the master equation used in [111] to describe the experimental bubble data. In particular, we can identify our two independent parameters D and γ with the rate constants k_+ and k_- to open and close a base pair introduced in [111], respectively: $D \equiv (k_+ + k_-)/2$ and $\gamma \equiv 2(k_- - k_+)/(k_+ + k_-)$. In this approximation, the correlation functions used successfully to fit the experimental results in [111] can be derived from the Fokker-Planck equation (47). Moreover, we can deduce the mean first passage time PDF for a bubble of initial size n_0 to close in the exact analytical form (compare [263])

$$f(0,t) = \frac{n_0}{\sqrt{4\pi Dt^3}} \exp\left\{-\frac{(n_0 - D\gamma t)^2}{4Dt}\right\} \tag{48}$$

which decays exponentially for large n. In particular, from (48) the characteristic (mean) first passage time for bubble closing,

$$\tau = \frac{n_0}{(D\gamma)} \tag{49}$$

follows, which is linear in the initial bubble size n_0. In Fig. 38, we compare this analytical result for the value $\gamma(37°C)$ with the characteristic closing times using the full drift term from

Eq. (47), obtained from numerical integration. It can be seen that the qualitative behavior for both cases with and without the $Dc/(n+1)$ term is very similar, but that in the presence of the loop closure component the characteristic times are reduced. Similarly, characteristic life-times of bubbles can be derived from the Fokker-Planck equation (47) at other temperatures [260]. The free energy landscape changes in the presence of single-strand binding proteins, and there comes about a rich behavior of the bubble dynamics in dependence on the binding rates of these single-stranded DNA binding proteins [264–266]. A statistical mechanical approach to the coupled dynamics of a DNA bubble and single-stranded DNA binding proteins has just been reported, describing the increase of the bubble lifetime, or even full denaturation of the DNA molecule, as induced by the binding proteins [267]. Stochastic simulations techniques using the Gillespie algorithm can be used to produce time series of single bubble events [S. K. Banik, T. Ambjörusson, and R. Metzler, in preparation]. Finally, the complete treatment of heteropolymer bubbles is on the way [T. Ambjörusson and R. Metzler, in preparation].

There exist various other models as well as improvements of the Poland-Scheraga model of DNA denaturation, such as the Peyrard-Bishop model [268–270] or Azbel's work on the elastic properties of the strands [271]. However, there exist a number of review articles [30, 272, 273], and we refer to these for further details. A question of much interest to physicists concerns the actual order of the denaturation transition. It has, for instance, been suggested to take into consideration that the bubble is in fact embedded between two vicinal dsDNA helices. Treating such a bubble as a polymer network with 3-vertices at the zipping location, it was shown that the exponent becomes $c \approx 2.1$, which would suggest a first-order transition [176, 177, 274–277]. However, the relevance to typical systems studied in experiments was questioned [278]. Similarly, a directed polymer approach producing a first-order transition [279] was criticized [280]. Dynamically, no significant change of the denaturation behavior is expected between a critical exponent c that is either slightly above or below the borderline value 2 [254, 260]. Moreover, typical samples are subject to significant finite-size effects whereas a difference between first- and second-order transitions is only relevant for large systems. Thus, the question of the order may be considered of not much concern for all practical purposes; in particular, as the biologically significant impact comes from the bubble *dynamics*. We finally remark that due to the heteropolymer character of DNA, the diffusive motion of DNA-bubbles along the backbone of the DNA is expected to be subdiffusive [254].

5. CONCLUSIONS

Biological systems inspire physics and engineering research. The various interactions of a biopolymer with a fixed topology such as the knotted state while being immersed in a Brownian environment, as well as the interactions with other biomolecules pose numerous questions, the understanding of which will also feed back to biological research itself. With approaches different from biologists' or biochemists' training, physicists and engineers can contribute significant new aspects and techniques such as manipulation and measurement of single molecules. At the same time, these systems inspire new directions in the natural sciences and engineering, in particular, as many biopolymeric processes are prototypical for nanometer systems. It is symptomatic for these new tracks of research that the same laboratory rooms are being shared, for instance, by biologists and physicists the same.

In this chapter, we presented an overview of the current understanding of DNA knots and their statistical behavior. These knots occur naturally in vivo and alter various properties of genetic regulation, such as prevention of replication, the impairment of transcription, reduction of the rupture stability, as well as the modification of the volume interaction of chemically remote domains of the genome. Apart from these vital biological functions, DNA knots may also serve as test tools toward a better understanding of topological constraints in classical polymer physics. Due to the omnipresent fluctuations, statistical mechanical tools are the method of choice to study such phenomena. Investigations in this new field are going hand in hand between experiments and theoretical modeling, such that new modeling studies may provoke experiments, or experiments may be used to gain necessary vital information to build up a theory, such as in the case of 3D knots.

We also briefly touched upon the breathing of DNA, that is, the local denaturing and renaturing of double-stranded DNA. The timescale of this breathing competes with typical binding times of single-strand binding proteins, and the single molecular study of this phenomenon will give new information on the binding dynamics of such proteins, as well as possibilities toward new biotechnologies, for instance, in PCR methods.

There exist numerous other relevant scenarios where topological constraints govern the behaviour of biopolymers. Thus, the passage of biopolymers through narrow membrane nanopores is biologically vital for information and protein exchange. It has been studied extensively in single molecule setups [33, 281, 282] and analyzed through various theoretical models; see, for instance [261, 283–291]. Similarly, the packaging of DNA and RNA into protein-made transport nanocontainers currently receives high interest due to the sheer values involved, for example, high packaging forces exerted by the packaging motor, and pressures inside the capsid of several MPa, reaching almost crystalline packaging of the DNA in the case of the $\phi 29$ phage [34, 292–297]. Finally, related ideas are expected to be useful in designing new types of molecules with entropic elements in the emerging field of topochemistry [298–303].

Dynamically, due to the heterogeneous character of biopolymers and/or the polymer-dynamical properties *per se*, dynamical phenomena involving biopolymers may exhibit anomalous behavior, such as subdiffusion [304, 305] in bubble or protein motion along the dsDNA backbone [254, 306], translocation [307, 308], granule [309], cytoskeleton motion [116, 310].

Topology finally plays an important role in genetic regulation, where the looping of DNA brings chemically remote operator sites close to each other in real space [311–314], or proteins bind nonspecifically along the entire DNA with rather high probability (see, for instance, [315, 316] and references therein), influencing both the targeting dynamics to the specific binding site [31] as well as the regulation dynamics and stability of the involved operators [316–320].

We hope that this chapter will contribute to increase the interest in biological systems and effects of topological constraints, and provide a platform for further investigations.

ACKNOWLEDGMENTS

We acknowledge helpful discussions with Paul Dommersnes, Yacov Kantor, Mehran Kardar, and Mark Williams, and we thank Francesco Valle and Giovanni Dietler for the AFM images of flat DNA knots as well as for discussions. R.M. wishes to thank Tobias Ambjörnsson, Audun Bakk, Richard Karpel, Yossi Klafter, Oleg Krichevsky, Amit Meller, and Lene Oddershede for valuable discussions. A.H. would like to thank Ray Baughman and Stephen Levene for numerous discussions and their kind hospitality at the Dallas NanoTech Institute and the Department of Molecular and Cell Biology at the University of Texas at Dallas.

Appendix: A Polymer Primer

In this section, we introduce some basic concepts from polymer physics. Starting from the random walk model, we define the fundamental measures of a polymer chain before introducing excluded volume. For more details, we refer to the monographs [5 40 42, 43].

The simplest polymer model is due to Orr [321]. It models the polymer chain as a random walk on a periodic lattice with lattice spacing a. Then, each monomer of index i is characterised by a position vector $\mathbf{R}_i$ with $i = 0, 1, \ldots, N$. The distance vector between monomers i and $i+1$ is called $\mathbf{a}_{i+1} = \mathbf{R}_{i+1} - \mathbf{R}_i$. Consequently, the end-to-end vector of the polymer is

$$\mathbf{r} = \sum_{i=1}^{N} \mathbf{a}_i \tag{A1}$$

Different $\mathbf{a}_i$ have completely independent orientations, such that we immediately obtain the average ($\langle\cdot\rangle$ over different configurations) squared end-to-end distance

$$\mathbf{R}_0^2 = \langle \mathbf{r}^2 \rangle = \sum_{i,j} \langle \mathbf{a}_i \cdot \mathbf{a}_j \rangle = \sum_i \langle \mathbf{a}_i^2 \rangle = Na^2 \tag{A2}$$

$R_0 \simeq N^{1/2}a$ is a measure for the size of the random walk. An alternative measure of the size of a polymer chain is provided by its radius of gyration R_g, which may be measured by light scattering experiments. It is defined by

$$R_g^2 = \frac{1}{1+N}\sum_{i=0}^{N}\langle(\mathbf{R}_i - \mathbf{R}_{ci})^2\rangle \tag{A3}$$

and measures the average squared distance to the centre of mass,

$$\mathbf{R}_{ci} = \frac{1}{1+N}\sum_{i=0}^{N}\mathbf{R}_i \tag{A4}$$

Expression (A3) can be rewritten as

$$R_g^2 = (1+N)^{-2}\sum_{i=0}^{N-1}\sum_{j=i+1}^{N}\langle(\mathbf{R}_i - \mathbf{R}_j)^2\rangle \tag{A5}$$

With $\mathbf{R}_j - \mathbf{R}_i = \sum_{n=i+1}^{j}\mathbf{a}_n$, one can easily show that $R_g^2 = a^2N(N+2)/[6(N+1)]$. For large N, that is, $R_g \simeq \frac{a}{\sqrt{6}}N$, and therefore:

$$R_g \sim R_0 \sim aN^{1/2} \tag{A6}$$

On a cubic lattice in d dimensions, each step can go in $2d$ directions, and for a general lattice, each vector $\mathbf{a}_i$ will have μ possible directions. The number of distinct walks with N steps is therefore μ^N. Denote $\mathfrak{N}_N(\mathbf{r})$ the number of distinct walks with end-to-end vector $\mathbf{r}$, the probability density function (PDF) for a given $\mathbf{r}$ is

$$p(\mathbf{r}) = \frac{\mathfrak{N}_N(\mathbf{r})}{\sum_{\mathbf{r}}\mathfrak{N}_N(\mathbf{r})} \tag{A7}$$

For large N, due to the independence of individual $\mathbf{a}_i$, this PDF will acquire a Gaussian shape,

$$p(\mathbf{r}) = \left(\frac{d}{2\pi Na^2}\right)^{d/2}\exp\left(-\frac{dr^2}{2Na^2}\right) \tag{A8}$$

where the normalisation is such that $\langle\mathbf{r}^2\rangle = Na^2$. From this expression, we can deduce that the number of degrees of freedom of a closed random walk chain is proportional to $N^{-d/2}$, the entropy loss suffered by a chain subject to the constraint $\mathbf{r} = 0$. On a general lattice,

$$\omega \simeq \mu^N N^{-d/2} \tag{A9}$$

with the connectivity constant μ, a measure for in how many different directions the next bond vector can point ($\mu = 2d$ in a cubic lattice). At fixed end-to-end distance, the entropy of the random walk becomes $S(\mathbf{r}) = S_0 - k_B dr^2/(2Na^2)$ where S_0 absorbs all constants. For the free energy $\mathcal{F}(\mathbf{r}) = E - TS(\mathbf{r})$ we therefore obtain

$$\mathcal{F}(\mathbf{r}) = \mathcal{F}_0 + \frac{dk_BTr^2}{2R_0^2} \tag{A10}$$

that is, the random walk likes to coil, the restoring force $-\nabla\mathcal{F}(\mathbf{r})$ being linear in $\mathbf{r}$. This is often called the entropic spring character of a Gaussian polymer. Note that the "spring constant" increases with temperature ("entropy elasticity").

In this random walk model of a polymer chain, it is straightforward to define the persistence length of the chain. By this we mean that successive vectors $\mathbf{a}_i$ are not independent, but tend to be parallel. Over long distance, this correlation is lost, and the chain behaves like a random walk. Due to the quantum chemistry of the monomers, an adjacent pair of vectors

$\mathbf{a}_i, \mathbf{a}_{i+1}$ includes preferred angles, for carbon chains leading to the trans/gauche configurations. This feature is captured schematically in the freely rotating chain model as depicted in Fig. A1. Following [42], we can obtain the correlation $\langle \mathbf{a}_n \cdot \mathbf{a}_m \rangle$ as follows. If we fix all vectors $\mathbf{a}_m, \ldots, \mathbf{a}_{n-1}$, then the average $\langle \mathbf{a}_n \rangle_{\mathbf{a}_m, \mathbf{a}_{m+1}, \ldots, \mathbf{a}_{n-1}\text{fixed}} = \mathbf{a}_{n-1} \cos\theta$. Multiplication by $\mathbf{a}_m$ produces

$$\langle \mathbf{a}_m \cdot \mathbf{a}_n \rangle_{\mathbf{a}_m, \ldots, \mathbf{a}_{n-1}\text{fixed}} = \mathbf{a}_m \cdot \mathbf{a}_{n-1} \cos\theta. \tag{A11}$$

Averaging over the $\mathbf{a}_m, \ldots, \mathbf{a}_{n-1}$ leads to the recursion relation $\langle \mathbf{a}_m \cdot \mathbf{a}_n \rangle = \langle \mathbf{a}_m \cdot \mathbf{a}_{n-1} \rangle \cos\theta$. With the initial condition $\langle \mathbf{a}^2 \rangle = a^2$, we find

$$\langle \mathbf{a}_m \cdot \mathbf{a}_n \rangle = a^2 \cos^{n-m}\theta \tag{A12}$$

Thus, if $\theta = 0$, we obtain a rigid rod behavior, while for $\theta \neq 0$, there occurs an exponential decay of the correlation between any two bond vectors $\mathbf{a}_n$ and $\mathbf{a}_m$. This defines a length scale

$$\ell_p \equiv -\frac{a}{\log\cos\theta} \tag{A13}$$

the persistence length of the chain. It diverges for $\theta \to 0$, while it vanishes for $\theta = 90°$, corresponding to the random walk model discussed above ("freely jointed chain"). As

$$\sum_{k=-\infty}^{\infty} \langle \mathbf{a}_{n+k} \cdot \mathbf{a}_n \rangle = a^2 \left(1 + 2 \sum_{k=1}^{\infty} \cos^k\theta \right) = a^2 \frac{1+\cos\theta}{1-\cos\theta} \tag{A14}$$

we find $R_0^2 = a^2 N(1+\cos\theta)/(1-\cos\theta)$, that is, statistically, the freely jointed chain behaves the same as the random walk chain, but with a rescaled monomer length. The statistical unit in a polymer chain is often taken to be the Kuhn length $\ell_K = 2\ell_p$.

The above chain models are often referred to as being phantom, that is, the chain can freely cross itself. A physical polymer possesses an excluded volume and behaves like a so-called self-avoiding chain. Mathematically, this can be modeled by self-avoiding walks. To include the major effects, it is sufficient to follow a simple argument due to Flory. Consider a chain with unknown radius R and internal monomer concentration $c_{\text{int}} \simeq N/R^d$. Assuming that the self-avoiding character is due to monomer-monomer interactions, the repulsive energy is proportional to the squared concentration, that is,

$$\mathcal{F}_{\text{rep}} = \frac{1}{2} T v(T) c^2 \tag{A15}$$

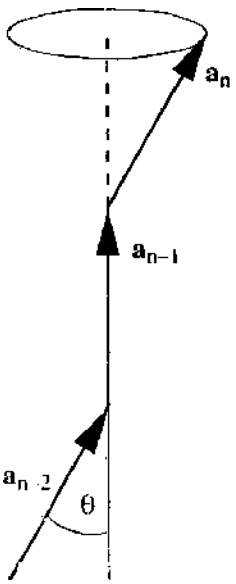


Figure A1. Freely jointed chain, in which successive bond vectors include an angle θ.

with the excluded volume parameter $v(T)$ ($v(T) \equiv (1-2\chi)a^d$ in Flory's notation, where the θ condition $\chi = 1/2$ corresponds to ideal chain behavior). To obtain the total averaged repulsive energy $\bar{\mathcal{F}}_{\text{rep tot}}$, we need to average over c^2. In a mean field approach, we take $\langle c^2 \rangle \longrightarrow \langle c \rangle^2 \sim c_{\text{int}}^2$. We therefore obtain

$$\bar{\mathcal{F}}_{\text{rep tot}} \simeq Tv(T)c_{\text{int}}^2 R^d = Tv(T)\frac{N^2}{R^d} \tag{A16}$$

favoring large values of R. This "swelling" competes with the entropic elasticity contribution $\bar{\mathcal{F}}_{\text{el}} \simeq TR^2/(Na^2)$. The total free energy becomes

$$\frac{\bar{\mathcal{F}}}{T} \simeq v(T)\frac{N^2}{R^d} + \frac{R^2}{Na^2} \tag{A17}$$

with a minimum at $R_F^{d+2} = v(T)a^2N^3$, so that the Flory radius scales like

$$R_f \sim AN^{\nu}, \quad \text{with} \quad \nu = \frac{3}{2+d} \tag{A18}$$

The value of the exponent $\nu(d=2) = 3/4$ coincides with the exact value in $d = 2$, while $\nu(d=3) = 3/5$ is very close to the best known value 0.588 in $d = 3$.

Polymer Networks

A linear excluded volume polymer chain has the size

$$R_g^2 \simeq AN^{2\nu} \tag{A19}$$

with $\nu = 0.75$ in $d = 2$, and $\nu = 0.588$ in $d = 3$. Its number of DOF is given in terms of the configuration exponent γ such that

$$\omega \simeq \mu^N N^{\gamma-1} \tag{A20}$$

where $\gamma = 1.33$ in $d = 2$ and $\gamma = 1.16$ in $d = 3$.

Remarkably, similar critical exponents can be obtained for a general polymer network of the type shown in Fig. A2, as originally derived by Duplantier [172, 175], compare also [179, 322]: In a network $\mathcal{G}$ consisting of $\mathcal{N}$ chain segments of lengths $s_1, \ldots, s_{\mathcal{N}}$ and total length $L = \sum_{i=1}^{\mathcal{N}} s_i$, the number of configurations $\omega_{\mathcal{G}}$ scales as

$$\omega_{\mathcal{G}}(s_1, \ldots, s_{\mathcal{N}}) = \mu^L s_{\mathcal{N}}^{\gamma_{\mathcal{G}}-1} \eta_{\mathcal{G}}\left(\frac{s_1}{s_{\mathcal{N}}}, \ldots, \frac{s_{\mathcal{N}-1}}{s_{\mathcal{N}}}\right) \tag{A21}$$

where $\eta_{\mathcal{G}}$ is a scaling function, and μ is the effective connectivity constant for self-avoiding walks. The exponent $\gamma_{\mathcal{G}}$ is given by $\gamma_{\mathcal{G}} = 1 - d\nu\mathcal{L} + \sum_{N\geq 1} n_N \sigma_N$, where ν is the swelling exponent, $\mathcal{L}$ is the number of independent loops, n_N is the number of vertices with N outgoing legs, and σ_N is an exponent associated with such a vertex. In $d = 2$, this exponent is given by [172, 175]

$$\sigma_N = \frac{(2-N)(9N+2)}{64} \tag{A22}$$

In the dense phase in 2D [178, 227, 229–231, 323], and at the Θ transition [324], analogous results can be obtained.

First, consider the dense phase in 2D. If all segments have equal length s and $L = \mathcal{N}s$, the configuration number $\omega_{\mathcal{G}}$ of such a network scales as [178, 227][21]

$$\omega_{\mathcal{G}}(s) \sim \omega_0(L)s^{\gamma_{\mathcal{G}}} \tag{A23}$$

[21] Note that due to the factor $\omega_0(L)$ the exponent of s is $\gamma_{\mathcal{G}}$ and not $\gamma_{\mathcal{G}} - 1$ like in the expressions used in the dilute phase [175] or at the Θ point, for which $\omega_0(L) \sim L^{-d}$. However, for 2D dense polymers, one has $d\nu = 1$, so that both definitions of $\gamma_{\mathcal{G}}$ are equivalent, cf. section 3 in [178].

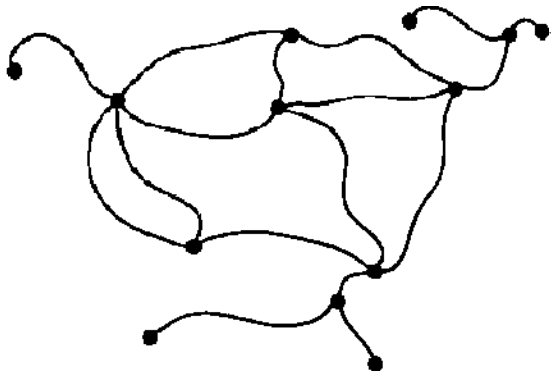


Figure A2. Polymer network $\mathcal{G}$ with vertices (•) of different order N, where N self-avoiding walks are joined ($n_1 = 5$, $n_3 = 4$, $n_4 = 3$, $n_6 = 1$).

where $\omega_0(L)$ is the configuration number of a simple ring of length L. For dense polymers, and in contrast to the dilute phase or at the Θ point, $\omega_0(L)$ (and thus $\omega_{\mathcal{G}}$) depends on the boundary conditions and even on the shape of the system [178, 229, 230, 323]. For example, for periodic boundary conditions corresponding to a 2D torus, one finds $\omega_0(L) \sim \mu^L L^{\Psi-1}$ with a connectivity constant μ and $\Psi = 1$ [178]. However, the network exponent

$$\gamma_{\mathcal{G}} = 1 - \mathcal{L} + \sum_{N \geq 1} n_N \sigma_N \tag{A24}$$

is *universal* and depends only on the topology of the network by the number $\mathcal{L}$ of independent loops, and by the number n_N of vertices of order N with vertex exponents $\sigma_N = (4 - N^2)/32$ [178, 227]. For a linear chain, the corresponding exponent $\gamma_{\text{lin}} = 19/16$ has been verified by numerical simulations [178, 325]. For a network made up of different segment lengths $\{s_i\}$ of total length $L = \sum_{i=1}^{\mathcal{N}} s_i$, Eq. (A23) generalizes to (cf. Section 4 in [178])

$$\omega_{\mathcal{G}}(s_1, \ldots, s_{\mathcal{N}}) \sim \omega_0(L)\, s_1^{\gamma_{\mathcal{G}}-1}\, \mathcal{Y}_{\mathcal{G}}\left(\frac{s_1}{s_{\mathcal{N}}}, \ldots, \frac{s_{\mathcal{N}-1}}{s_{\mathcal{N}}}\right) \tag{A25}$$

which involves the scaling function $\mathcal{Y}_{\mathcal{G}}$.

For polymers in an infinite volume and endowed with an attractive interaction between neighboring monomers, a different scaling behavior emerges if the system is not below but right at the Θ point [178]. In this case the number of configurations of a general network $\mathcal{G}$ is given by

$$\tilde{\omega}_{\mathcal{G}}(s_1, \ldots, s_{\mathcal{N}}) \sim \mu^L\, s_1^{\tilde{\gamma}_{\mathcal{G}}-1}\, \overline{\mathcal{Y}}_{\mathcal{G}}\left(\frac{s_1}{s_{\mathcal{N}}}, \ldots, \frac{s_{\mathcal{N}-1}}{s_{\mathcal{N}}}\right) \tag{A26}$$

with the network exponent

$$\tilde{\gamma}_{\mathcal{G}} = 1 - d\nu\mathcal{L} + \sum_{N \geq 1} n_N \tilde{\sigma}_N \tag{A27}$$

Overlined symbols refer to polymers at the Θ point. In $d = 2$, $\nu = 4/7$ and $\tilde{\sigma}_N = (2 - N)(2N + 1)/42$ [178].

REFERENCES

1. B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, "Molecular Biology of the Cell." Garland, New York, 2002.
2. A. Kornberg, "DNA synthesis." W. H. Freeman, San Francisco, CA, 1974.
3. A. Kornberg and T. A. Baker, "DNA Replication." W. H. Freeman, New York, 1992.
4. N. R. Cozzarelli and J. C. Wang, "DNA topology and its biological effects." Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1990.
5. A. Yu. Grosberg and A. R. Khokhlov, "Statistical Mechanics of Macromolecules." AIP Press, New York, 1994.
6. M. D. Frank-Kamenetskii, "Unraveling DNA: The Most Important Molecule of Life." Perseus, Cambridge, MA, 1997.

7. C. R. Calladine and H. Drew, "Understanding DNA: The Molecule and How It Works," Academic Press, London, 1997.
8. L. Stryer, J. M. Berg, and J. L. Tymoczko, "Biochemistry," W. H. Freeman, San Francisco, 2002.
9. S. A. Wassermann and N. R. Cozzarelli, *Science* 232, 951 (1986).
10. J. Sogo, A. Stasiak, M. L. Martínez-Robles, D. B. Krimer, P. Hernández, and J. B. Schvartzman, *J. Mol. Biol.* 286, 637 (1999).
11. F. B. Dean, A. Stasiak, T. Koller, and N. R. Cozzarelli, *J. Biol. Chem.* 260, 4975 (1985).
12. A. V. Vologodskii, N. J. Crisona, B. Laurie, P. Pieranski, V. Katritch, J. Dubochet, and A. Stasiak, *J. Mol. Biol.* 278, 1 (1998).
13. L. Postow, B. J. Peter, and N. B. Cozzarelli, *Bioessays* 21, 805 (1999).
14. J. B. Schvartzman and A. Stasiak, *EMBO Rep.* 5, 256 (2004).
15. T. A. Shapiro and P. T. Englund, *Annu. Rev. Microbiol.* 49, 117 (1995).
16. L. Simpson and A. da Silva, *J. Mol. Biol.* 56, 443 (1971).
17. R. W. Deibler, S. Rahmati, and E. L. Zechiedrich, *Genes & Development* 15, 748 (2001).
18. A. Stasiak, V. Katritch, J. Bednar, D. Michoud, and J. Dubochet, *Nature* 384, 122 (1996).
19. A. Bucka and A. Stasiak, *Nucl. Acids Res.* 30, e24 (2002).
20. S. Trigueros, J. Salceda, I. Bermudez, X. Fernandez, and J. Roca, *J. Mol. Biol.* 335, 723 (2004).
21. J. M. Berger, S. J. Gamblin, S. C. Harrrison, and J. C. Wang, *Nature* 379, 225 (1996).
22. J. C. Wang, *Quart. Rev. Biophys.* 31, 107 (1998).
23. S. A. Wassermann, J. M. Dungan, and N. R. Cozzarelli, *Science* 229, 171 (1985).
24. R. F. Gesteland and J. F. Atkins, Eds., "The RNA World," Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1993.
25. J. D. Watson and F. H. C. Crick, *Nature* 171, 737 (1953).
26. R. W. Simons and M. Grunberg-Manago, Eds., "RNA Structure and function," Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1998.
27. V. A. Bloomfield, D. M. Crothers, and I. Tinoco, "Physical Chemistry of Nucleic Acids," Harper & Row, New York, 1974.
28. W. Sanger, "Principles of Nucleic Acid Structure," Springer-Verlag, New York, 1984.
29. D. Poland and H. A. Scheraga, "Theory of Helix-Coil Transitions in Biopolymers," Academic Press, New York, 1970.
30. R. M. Wartell and A. S. Benight, *Phys. Rep.* 126, 67 (1985).
31. O. G. Berg, R. B. Winter, and P. H. von Hippel, *Biochem.* 20, 6929 (1981).
32. M. Slutsky and L. A. Mirny, E-print q-bio.BM/0402006 (2004).
33. A. Meller, *J. Phys. C* 15, R581 (2003).
34. W. C. Earnshaw and S. R. Casjens, *Cell* 21, 319 (1980).
35. H. Schiessel, *J. Phys. C* 15, R699 (2003).
36. G. Goshima and R. D. Vale, *J. Cell Biol.* 162, 1003 (2003).
37. J. A. Brydson, "Plastics Materials," Butterworth-Heinemann, Oxford, 1999.
38. J. E. Mark, B. Erman, and F. R. Eirich, "Science and Technology of Rubber," Academic Press, New York, 1994.
39. P. G. de Gennes and J. Badoz, "Fragile Objects," Springer-Verlag, New York, 1996.
40. P. J. Flory, "Principles of Polymer Chemistry," Cornell University Press, Ithaca, New York, 1953.
41. P. J. Flory, "Statistical Mechanics of Chain Molecules," Oxford University Press, Oxford, 1989.
42. M. Doi and S. F. Edwards "The Theory of Polymer Dynamics," Clarendon Press, Oxford, 1986.
43. P. G. de Gennes, "Scaling Concepts in Polymer Physics," Cornell University Press, Ithaca, New York, 1979.
44. J. D. Ferry, "Viscoelastic Properties of Polymers," Wiley, New York, 1970.
45. I. M. Ward and D. Hadley, "An Introduction to the Mechanical Properties of Solid Polymers," John Wiley and Sons Ltd., New York, 1993.
46. N. W. Tschoegl, "The Phenomenological Theory of Linear Viscoelastic Behavior: An Introduction," Springer-Verlag, Berlin, 1989.
47. U. Eisele, "Introduction to Polymer Physics," Springer-Verlag, Berlin, 1990.
48. L. R. G. Treloar, "The Physics of Rubber Elasticity," Clarendon Press, Oxford, 1975.
49. F. Crick, *Nature* 227, 561 (1970).
50. D. P. Snustad and M. J. Simmons, "Principles of Genetics," John Wiley & Sons, New York, 2003.
51. F. C. H. Crick, *Symp. Soc. Exp. Biol.* XII, 138 (1958).
52. G. Orphanides and D. Reinberg, *Cell* 108, 439 (2002).
53. M. J. McPherson and S. G. Møller, "PCR Basics," Springer-Verlag Telos, New York, 2000.
54. C. W. Dieffenbach and G. S. Dveksler, "PCR Primer: a Laboratory Manual," Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1995.
55. T. E. Creighton, "Proteins: Structures and Molecular Properties," W. H. Freeman, New York, 1993.
56. A. Fersht, "Structure and Mechanism in Protein Science: Guide to Enzyme Catalysis and Protein Folding," W. H. Freeman, San Francisco, 1999.
57. G. Petsko and D. Ringe, "Protein Structure and Function," Blackwell Publishing, Oxford, 2003.
58. R. D. Blake, J. W. Bizzaro, J. D. Blake, G. R. Day, S. G. Delcourt, J. Knowles, K. A. Marx, and J. SantaLucia, Jr., *Bioinformatics* 15, 370 (1999).
59. M. D. Frank-Kamenetskii, *Phys. Rep.* 288, 13 (1997).
60. J. F. Marko and E. D. Siggia, *Macromolecules* 28, 8759 (1995).

61. J. F. Marko and E. D. Siggia, *Phys. Rev. E* 52, 2912 (1995).
62. G. Călugăreanu, Czech. *Math. J.* 11, 588 (1961).
63. J. H. White, *Am. J. Math.* 91, 693 (1969).
64. F. B. Fuller, *Proc. Natl. Acad. Sci. U.S.A.* 75, 3557 (1971).
65. C. J. Benham, *Comput. Appl. Biosci.* 12, 375 (1996).
66. S. Goetze, A. Gluch, C. Benham, and J. Bode, *Biochemistry* 42, 154 (2003).
67. A. V. Vologodskii, S. D. Levene, K. V. Klenin, M. Frank-Kamenetskii, and N. R. Cozzarelli, *J. Mol. Biol.* 227, 1224 (1992).
68. H. Tsen and S. D. Levene, *Biophys. J.* (in preparation).
69. M. Pillsbury, H. Orland, and A. Zee, E-print physics/0207110 (2002).
70. H. Orland and A. Zee, *Nucl. Phys. B* 620, 456 (2002).
71. M. Baiesi, E. Orlandini, and A. L. Stella, *Phys. Rev. Lett.* 91, 198102 (2003).
72. J. Yan, M. O. Magnasco, and J. F. Marko, *Nature* 401, 932 (1999).
73. V. V. Rybenkov, C. Ullsperger, A. V. Vologodskii, and N. R. Cozzarelli, *Science* 277, 690 (1997).
74. C. Bustamante, S. B. Smith, J. Liphardt, and D. Smith, *Curr. Opin. Struc. Biol.* 10, 279 (2000).
75. R. Merkel, *Phys. Rep.* 346, 344 (2001).
76. M. Rief and H. Grubmüller, *Chem. Phys. Chem.* 3, 255 (2002).
77. T. R. Strick, M.-N. Dessinges, G. Charvin, N. H. Dekker, J.-F. Allemand, D. Bensimon, and V. Croquette, *Rep. Prog. Phys.* 66, 1 (2003).
78. E. L. Florin, V. T. Moy, and H. E. Gaub, *Science* 264, 415 (1994).
79. R. Merkel, P. Nassoy, A. Leung, K. Ritchie, and E. Evans, *Nature* 397, 50 (1999).
80. J. W. Weisel, H. Shuman, and R. I. Litvinov, *Curr. Op. Struc. Biol.* 13, 227 (2003).
81. J. F. Marko and E. D. Siggia, *Macromolecules* 28, 8759 (1995).
82. T. Strunz, K. Oroszlan, R. Schäfer, and H.-J. Güntherodt, *Proc. Natl. Acad. Sci. U.S.A.* 96, 11277 (1999).
83. M. C. Williams, I. Rouzina, and V. A. Bloomfield, *Acc. Chem. Res.* 35, 159 (2002).
84. J. Liphardt, B. Onoa, S. B. Smith, I. Tinoco, Jr., and C. Bustamante, *Science* 292, 733 (2001).
85. S. Harlepp, T. Marchal, J. Robert, J.-F. Léger, A. Xayaphoummine, H. Isambert, and D. Chatenay, *Euro. Phys. J. E* 12, 605 (2003).
86. M. Rief, F. Oesterhelt, B. Heymann, and H. E. Gaub, *Science* 275, 1295 (1997).
87. M. Rief, M. Gautel, F. Oesterhelt, J. M. Fernandez, and H. E. Gaub, *Science* 276, 1109 (1997).
88. D. J. Müller, M. Kessler, F. Oesterhelt, C. Möller, D. Oesterhelt, and H. Gaub, *Biophys. J.* 83, 3578 (2002).
89. X. Zhuang and M. Rief, *Curr. Op. Struc. Biol.* 13, 88 (2003).
90. M. C. Williams, Biophys. Textbook online: Optical Tweezers: Measuring Piconewton Forces. Available at www.biophysics.org/btol/.
91. C. Bustamante, Z. Bryant, and S. B. Smith, *Nature* 421, 423 (2003).
92. Y. Arai, R. Yasuda, K.-I. Akashi, Y. Harada, H. Miyata, K. Kinosita, Jr., and H. Itoh, *Nature* 399, 446 (1999).
93. X. R. Bao, H. J. Lee, and S. R. Quake, *Phys. Rev. Lett.* 91, 265506 (2003).
94. T. R. Strick, V. Croquette, and D. Bensimon, *Nature* 404, 901 (2000).
95. H. Clausen-Schaumann, M. Seitz, R. Krautbauer, and H. E. Gaub, *Curr. Opin. Chem. Biol.* 4, 524 (2000).
96. E. Evans and K. Ritchie, *Biophys. J.* 72, 1541 (1997).
97. B. Heymann and H. Grubmüller, *Phys. Rev. Lett.* 84, 6126 (2000).
98. C. Jarzynski, *Phys. Rev. Lett.* 78, 2690 (1997); *Phys. Rev. E* 56, 5018 (1997).
99. G. Hummer and A. Szabo, *Proc. Natl. Acad. Sci. U.S.A.* 98, 3658 (2001).
100. A. M. Ferrenberg and R. H. Swendsen, *Phys. Rev. Lett.* 63, 1195 (1989).
101. J. Liphardt, S. Dumont, S. B. Smith, I. Tinoco, Jr., and C. Bustamante, *Science* 296, 1832 (2002).
102. D. A. Hendrix and C. Jarzynski, *J. Chem. Phys.* 114, 5974 (2001).
103. E. G. D. Cohena and D. Mauzerall, E-print cond-mat/0406128 (2004).
104. O. Braun, A. Hanke, and U. Seifert, *Phys. Rev. Lett.* 93, 158105 (2004).
105. D. M. Zuckerman and T. B. Woolf, *Phys. Rev. Lett.* 89, 180602 (2002).
106. M. R. Shirts, E. Bair, G. Hooker, and V. S. Pande, *Phys. Rev. Lett.* 91, 140601 (2003).
107. G. Haran, *J. Phys. C* 15, R1291 (2003).
108. S. Weiss, *Nat. Struct. Biol.* 7, 724 (2000).
109. P. Schwille, *Cell. Biochem. Biophys.* 34, 383 (2001).
110. R. Rigler, M. Orrit, and T. Basché, Eds., "Single Molecule Spectroscopy." Springer-Verlag, Berlin, 2002.
111. G. Altan-Bonnet, A. Libchaber, and O. Krichevsky, *Phys. Rev. Lett.* 90, 138101 (2003).
112. G. Bonnet, O. Krichevsky, and A. Libchaber, *Proc. Natl. Acad. Sci. U.S.A.* 95, 8602 (1998).
113. O. Krichevsky and G. Bonnet, *Rep. Prog. Phys.* 65, 251 (2002).
114. C. Gonzalo, E. Harbron, Y. Zeng, H.-W. Liu, D. O'Connor, C. Eia, K. Musier-Forsyth, and P. F. Barbara, submitted to *Biophys. J.* (2004).
115. J. Lippincott-Schwartz and G. H. Patterson, *Science* 300, 87 (2003).
116. A. Caspi, R. Granek, and M. Elbaum, *Phys. Rev. Lett.* 85, 5655 (2000).
117. L. Oddershede, J. K. Dreyer, S. Grego, S. Brown, and K. Berg-Sorensen, *Biophys. J.* 83, 3152 (2002).
118. L. F. Liu, C.-C. Liu, and B. M. Alberts, *Cell* 19, 697 (1980).
119. K. Mizuuchi, L. M. Fisher, M. H. O. Dea, and M. Gellert, *Proc. Natl. Acad. Sci. U.S.A.* 77, 1847 (1980).
120. T. J. Pollock and H. A. Nash, *J. Mol. Biol.* 170, 1 (1983).
121. S. J. Spengler, A. Stasiak, and N. R. Cozzarelli, *Cell* 42, 325 (1985).
122. S. A. Wassermann and N. R. Cozzarelli, *Proc. Natl. Acad. Sci. U.S.A.* 82, 1079 (1984).

123. E. Viguera, P. Hernandez, D. B. Krimer, A. S. Boistov, R. Lurz, J. C. Alonso, and J. B. Schvartzman, *J. Biol. Chem.* 271, 22414 (1996).
124. J. Arsuaga, M. Vazquez, S. Trigueros, D. Sumners, and J. Roca, *Proc. Natl. Acad. Sci. U.S.A.* 99, 5373 (2002).
125. J. Portugal and A. Rodriguez-Campos, *Nucleic Acids Res.* 24, 4890 (1996).
126. A. Rodriguez-Campos, *J. Biol. Chem.* 271, 14150 (1996).
127. P. Staczek and N. P. Higgins, *Mol. Microbiol.* 29, 1435 (1998).
128. P. Pieranski, S. Kasas, G. Dietler, J. Dubochet, and A. Stasiak, *New J. Phys.* 3, 101 (2001).
129. A. M. Saitta, P. D. Soper, E. Wasserman, and M. L. Klein, *Nature* 399, 46 (1999).
130. A. Stasiak, A. Dobay, J. Dubochet, and G. Dietler, *Science* 286, 11a (1999).
131. T. McNally, "The complete book of fly fishing," Mountain Press, Camden, ME, 1997.
132. V. V. Rybenkov, C. Ullsperger, A. V. Vologodskii, and N. R. Cozzarelli, *Science* 277, 690 (1997).
133. A. Stasiak, *Curr. Biol.* 10, R526 (2000).
134. J. C. Wang, *Annu. Rev. Biochem.* 65, 635 (1996).
135. L. Euler, *Commentationes Academiae Sci. Imp. Petropolitanae* 8, 128 (1736); *Sci. Amer.* 189, 66 (1953).
136. J. Kepler, W. V. Dyck, and M. Caspar, Eds., "Johannes Kepler, Gesammelte Werke," Beck, München, 1937.
137. A. C. C. Adams, "The Knot Book: An Elementary Introduction to the Mathematical Theory of Knots," Freeman, New York, 1994.
138. L. H. Kauffman, "Knots and physics, Series on Knots and Everything" Vol. 1, World Scientific, Singapore, 1993.
139. K. Reidemeister, "Knotentheorie," Springer, Berlin, 1931; "Knot theory," BSC Assocs., Moscow, Idaho, 1983.
140. J. B. Listing, "Vorstudien zur Topologie," Göttinger Studien Vandenhoeck und Ruprecht, Göttingen, 1848.
141. W. Thomson, Lord Kelvin, *Philos. Mag.* 34, 15 (1867).
142. W. Thomson, *Proc. R. Soc. Edinb.* 27, 59 (1875–76).
143. P. G. Tait, *Trans. R. Soc. Edinb.* 28, 145 (1876–77).
144. P. G. Tait, *Trans. R. Soc. Edinb.* 32, 327 (1883–4).
145. P. G. Tait, *Trans. R. Soc. Edinb.* 33, 493 (1884–85).
146. P. G. Tait, "Scientific papers," Cambridge University Press, London, 1898.
147. T. P. Kirkman, *Proc. R. Soc. Edinb.* 13, 120, 363 (1884–5).
148. T. P. Kirkman, *Proc. R. Soc. Edinb.* 32, 483 (1884–85).
149. C. N. Little, *Trans. Connecticut Acad. Sci.* 18, 7, 27 (1885).
150. C. N. Little, *Trans. R. Soc. Edinb.* 35, 771 (1890).
151. P. van de Griend, in "History and Science of Knots" (J. C. Turner and P. van de Griend, Eds.), World Scientific, Singapore, 1996.
152. A. V. Vologodskii, A. V. Lukashin, M. D. Frank-Kamenetskii, and V. V. Anshelvich, *Sov. Phys. JETP* 39, 1059 (1974).
153. H. L. Frisch and E. Wassermann, *J. Am. Chem. Soc.* 83, 3789 (1961).
154. M. Delbrück, in "Mathematical Problems in Biological Sciences" (R. E. Bellman, Eds.), *Proc. Symp. Appl. Math.* 14, 55 (1962).
155. D. W. Sumners and S. G. Whittington, *J. Phys. A* 21, 1689 (1988).
156. N. Pippenger, *Discrete Appl. Math.* 25, 273 (1989).
157. C. Vanderzande, *J. Phys. A* 28, 3681 (1995).
158. M. D. Frank-Kamenetskii, A. V. Lukashin, and A. V. Vologodskii, *Nature* 258, 398 (1975).
159. K. Koniaris and M. Muthukumar, *Phys. Rev. Lett.* 66, 2211 (1991).
160. J. P. J. Michels and F. W. Wiegel, *Proc. R. Soc. (London) A* 403, 269 (1986).
161. J. P. J. Michels and F. W. Wiegel, *Phys. Lett.* 90A, 381 (1982).
162. E. J. Janse van Rensburg and S. G. Whittington, *J. Phys. A* 24, 3935 (1991).
163. M. D. Frank-Kamenetskii, A. V. Lukashin, V. V. Anshelevich, and A. V. Vologodskii, *J. Biomol. Struct. Dyn.* 2, 1005 (1985).
164. A. Yu. Grosberg, E-print cond-mat/0207427 (2002).
165. K. V. Klenin, A. V. Vologodskii, V. V. Anshelevich, A. M. Dykhne, and M. D. Frank-Kamenetskii, *J. Biomol. Struct. Dyn.* 5, 1173 (1988).
166. M. K. Shimamura and T. Deguchi, *Phys. Lett. A* 274, 184 (2000).
167. V. Katritch, W. K. Olson, A. Vologodskii, J. Dubochet, and A. Stasiak, *Phys. Rev. E* 61, 5545 (2000).
168. A. Dobay, P. E. Sottas, J. Dubochet, and A. Stasiak, *Lett. Math. Phys.* 55, 239 (2001).
169. P. G. Dommersnes, Y. Kantor, and M. Kardar, *Phys. Rev. E* 66, 031802 (2002).
170. A. Stasiak, V. Katritch, and L. H. Kauffman, Eds., "Ideal knots, Series on Knots and Everything," Vol. 19, World Scientific, Singapore, 1998.
171. P. Hoidn, R. B. Kusner, and A. Stasiak, *New J. Phys.* 4, 20.1 (2002).
172. B. Duplantier, *J. Stat. Phys.* 54, 581 (1989).
173. J. M. Deutsch, *Phys. Rev. E* 59, R2539 (1999).
174. B. D. Hughes, "Random Walks and Random Environments," Volume 1: Random Walks Oxford University Press, Oxford, 1995.
175. B. Duplantier, *Phys. Rev. Lett.* 57, 941 (1986).
176. Y. Kafri, D. Mukamel, and L. Peliti, *Phys. Rev. Lett.* 85, 4988 (2000).
177. Y. Kafri, D. Mukamel, and L. Peliti, *Eur. Phys. J. B* 27, 135 (2002).
178. B. Duplantier and H. Saleur, *Nucl. Phys. B* 290, 291 (1987).
179. L. Schafer, C. V. Ferber, U. Lehr, and B. Duplantier, *Nucl. Phys. B* 374, 473 (1992).

180. R. Metzler, A. Hanke, P. G. Dommersnes, Y. Kantor, and M. Kardar, *Phys. Rev. Lett.* 88, 188101 (2002).
181. M. B. Hastings, Z. A. Daya, E. Ben-Naim, and R. E. Ecke, *Phys. Rev. E* 66, 025102(R) (2002).
182. J. des Cloizeaux, *J. Phys. (France) Lett.* 42, L433 (1981).
183. A. Dobay, J. Dubochet, K. Millett, P.-E. Sottas, and A. Stasiak, *Proc. Natl. Acad. Sci. U.S.A.* 100, 5611 (2003).
184. J. C. LeGuillou and J. Zinn-Justin, *Phys. Rev. B* 21, 3976 (1989).
185. J. C. LeGuillou and J. Zinn-Justin, *J. Physique (France)* 50, 1365 (1989).
186. E. J. Janse van Rensburg, S. G. Whittington, and N. Madras, *J. Phys. A* 23, 1589 (1990).
187. B. Li, S. N. Madras, and A. D. Sokal, *J. Stat. Phys.* 80, 661 (1995).
188. E. Orlandini, M. C. Tesi, E. J. Janse van Rensburg, and S. G. Whittington, *J. Phys. A* 31, 5953 (1998).
189. A. J. Guttmann, *J. Phys. A* 22, 2807 (1989).
190. E. Guitter and E. Orlandini, *J. Phys. A* 32, 1359 (1999).
191. S. R. Quake, *Phys. Rev. Lett.* 73, 3317 (1994).
192. S. R. Quake, *Phys. Rev. E* 52, 1176 (1986).
193. A. Yu. Grosberg, A. Feigel, and Y. Rabin, *Phys. Rev. E* 54, 6618 (1996).
194. B. Erman and J. E. Mark, "Structures and Properties of Rubberlike Networks." Oxford, New York, 1997.
195. P.-G. de Gennes, *Macromolecules* 17, 703 (1984).
196. P. K. Lai, Y. J. Sheng, and H. K. Tsao, *Phys. Rev. Lett.* 87, 175503 (2001).
197. M. K. Shimamura and T. Deguchi, *Phys. Rev. E* 64, 020801 (2001).
198. O. Farago, Y. Kantor, and M. Kardar, *Europhys. Lett.* 60, 53 (2002).
199. R. Metzler, Y. Kantor, and M. Kardar, *Phys. Rev. E* 66, 022102 (2002).
200. Y. J. Sheng, P. K. Lai, and H. K. Tsao, *Phys. Rev. E* 61, 2895 (2000).
201. P. K. Lai, Y. J. Sheng, and H. K. Tsao, *Macromol. Theor. Simul.* 9, 578 (2000).
202. R. Zandi, Y. Kantor, and M. Kardar, E-print cond-mat/0306587 (2003).
203. E. Orlandini, A. L. Stella, and C. Vanderzande, *Phys. Rev. E* 68, 031804 (2003).
204. E. Orlandini, A. L. Stella, and C. Vanderzande, *J. Stat. Phys.* 115, 681 (2004).
205. T. A. Vilgis, *Phys. Rep.* 336, 167 (2000).
206. A. L. Kholodenko and T. A. Vilgis, *Phys. Rep.* 298, 251 (1998).
207. S. F. Edwards, *Proc. Phys. Soc. (London)* 91, 513 (1967).
208. S. F. Edwards, *J. Phys. A* 1, 15 (1968).
209. M. Otto and T. A. Vilgis, *Phys. Rev. Lett.* 80, 881 (1998).
210. M. Otto and T. A. Vilgis, *J. Phys. A* 29, 3893 (1996).
211. T. A. Vilgis and M. Otto, *Phys. Rev. E* 56, R1314 (1997).
212. F. Ferrari, H. Kleinert, and I. Lazzizzera, *Int. J. Mod. Phys. B* 14, 3881 (2000).
213. F. Ferrari, H. Kleinert, and I. Lazzizzera, *Phys. Lett. A* 276, 31 (2000).
214. A. Yu. Grosberg and S. Nechaev, *Adv. Polym. Sci.* 106, 1 (1993).
215. A. Yu. Grosberg and S. Nechaev, *Europhys. Lett.* 20, 613 (1992).
216. A. Yu. Grosberg and S. Nechaev, *J. Phys. A* 25, 4659 (1992a).
217. S. Nechaev, in "Topological Aspects of Low Dimensional Systems." Lecture notes of Les Houches 1998 summer school; extended version available as E-print cond-mat/9812205.
218. O. Vasilyev and S. Nechaev, *JETP* 93, 1119 (2001).
219. S. W. P. Turner, M. Cabodi, and H. G. Craighead, *Phys. Rev. Lett.* 88, 128103 (2002).
220. E. Ben-Naim, Z. A. Daya, P. Vorobieff, and R. E. Ecke, *Phys. Rev. Lett.* 86, 1414 (2001).
221. F. Valle and G. Dietler, École Polytechnique Fédérale de Lausanne, personal communication (2004).
222. B. Maier and J. O. Rädler, *Phys. Rev. Lett.* 82, 1911 (1999).
223. H. Walter and D. E. Brooks, *FEBS Lett.* 361, 135 (1995).
224. V. V. Vasilevskaya, A. R. Khokhlov, Y. Matsuzawa, and K. Yoshikawa, *J. Chem. Phys.* 102, 6595 (1995).
225. L. Lerman, *Proc. Natl. Acad. Sci. U.S.A.* 68, 1886 (1971).
226. T. Garel, "Remarks on homo- and hetero-polymeric aspects of protein folding," E-print cond-mat/0305053 (2003).
227. B. Duplantier, *J. Phys. A* 19, L1009 (1986).
228. J. L. Jacobsen, N. Read, and H. Saleur, *Phys. Rev. Lett.* 90, 090601 (2003).
229. B. Duplantier, *Phys. Rev. Lett.* 71, 4274 (1993).
230. A. L. Owczarek, T. Prellberg, and R. Brak, *Phys. Rev. Lett.* 70, 951 (1993).
231. A. L. Owczarek, T. Prellberg, and R. Brak, *Phys. Rev. Lett.* 71, 4275 (1993).
232. B. Duplantier and F. David, *J. Stat. Phys.* 51, 327 (1988).
233. J. Kondev and J. L. Jacobsen, *Phys. Rev. Lett.* 81, 2922 (1998).
234. M. E. Fisher, *Physica D* 38, 112 (1989).
235. E. Orlandini, F. Seno, A. L. Stella, and M. C. Tesi, *Phys. Rev. Lett.* 68, 488 (1992).
236. R. Dekeyser, E. Orlandini, A. L. Stella, and M. C. Tesi, *Phys. Rev. E* 52, 5214 (1995).
237. J. Cardy, *J. Phys. A* 34, L665 (2001).
238. A. Hanke, R. Metzler, P. G. Dommersnes, Y. Kantor, and M. Kardar, *Eur. Phys. J. E.* 12, 347 (2003).
239. R. Metzler, A. Hanke, P. G. Dommersnes, Y. Kantor, and M. Kardar, *Phys. Rev. E.* 65, 061103 (2002).
240. R. Gambini and J. Pullin, "Loops, Knots, Gauge Theories and Quantum Gravity." Cambridge University Press, Cambridge, UK, 1996.
241. R. Metzler, *New J. Phys.* 4, 91 (2002).
242. A. Yu. Grosberg, *Phys. Rev. Lett.* 85, 3858 (2000).
243. R. C. Ball, M. Doi, S. F. Edwards, and M. Warner, *Polymer* 22, 1010 (1981).

244. M. Doi and S. F. Edwards, *J. Chem. Soc. Farad. Trans.* 274, 1802 (1978).
245. S. F. Edwards and T. A. Vilgis, *Polymer* 27, 483 (1986).
246. P. G. Higgs and R. C. Ball, *Europhys. Lett.* 8, 357 (1989).
247. J. C. Sommer, *J. Chem. Phys.* 97, 5777 (1992).
248. M. Guéron, M. Kochoyan, and J. L. Leroy, *Nature* 328 89 (1987).
249. A. Revzin, Ed., "The Biology of Non-Specific DNA-Protein Interactions" CRC Press, Boca Raton, FL, 1990.
250. N. Acharya and U. Varshney, *J. Mol. Biol.* 318, 1251 (2002).
251. L. D. Landau and E. M. Lifshitz, "Physical Kinetics," Butterworth-Heinemann, Oxford, UK, 1981.
252. B. H. Zimm and J. K. Bragg, *J. Chem. Phys.* 31, 526 (1959).
253. B. H. Zimm, *J. Chem. Phys.* 33, 1349 (1960).
254. T. Hwa, E. Marinari, K. Sneppen, and L. H. Tang, *Proc. Natl. Acad. Sci. U.S.A.* 100, 4411 (2003).
255. I. Rouzina and V. A. Bloomfield, *Biophys. J.* 77, 3242 (1999).
256. R. Blossey and E. Carlon, *Phys. Rev. E* 68, 061911 (2004).
257. M. Fixman and J. J. Freiere, *Biopol.* 16, 2693 (1977).
258. O. Gotoh and Y. Tagashira, *Biopol.* 20, 1033 (1981).
259. M. E. Fisher, *J. Chem. Phys.* 45, 1464 (1966).
260. A. Hanke and R. Metzler, *J. Phys. A* 36, L473 (2003).
261. D. K. Lubensky and D. R. Nelson, *Phys. Rev. Lett.* 85, 1572 (2000).
262. C. Kittel, *Am. J. Phys.* 37, 917 (1969).
263. S. Redner, "A Guide to First-Passage Processes," Cambridge University Press, Cambridge, UK, 2001.
264. M. Pant, R. L. Karpel, and M. C. Williams, *J. Mol. Biol.* 327, 571 (2003).
265. M. Pant, R. L. Karpel, I. Rouzina, and M. C. Williams, *J. Mol. Biol.* 336, 851 (2004).
266. R. Karpel, *IUBMB Life* 53, 161 (2002).
267. T. Ambjörnsson and R. Metzler, E-print q-bio.BM/0411053.
268. M. Peyrard and A. R. Bishop, *Phys. Rev. Lett.* 62, 2755 (1989).
269. T. Dauxois, M. Peyrard, and A. R. Bishop, *Phys. Rev. E* 47, R44 (1993).
270. T. Dauxois and M. Peyrard, *Phys. Rev. E* 51, 4027 (1995).
271. M. Y. Azbel', *Biopolymers* 19, 61; 19, 81 (1980).
272. C. Richard and A. J. Guttmann, *J. Stat. Phys.* 115, 925 (2004).
273. M. Peyrard, *Nonlin.* 17, R1 (2004).
274. M. Baiesi, E. Carlon, Y. Kafri, D. Mukamel, E. Orlandini, and A. L. Stella, *Phys. Rev. E* 67, 021911 (2003a).
275. M. Baiesi, E. Carlon, and A. L. Stella, *Phys. Rev. E* 66, 021804 (2002).
276. E. Carlon, E. Orlandini, and A. L. Stella, *Phys. Rev. Lett.* 88, 198101 (2002).
277. M. S. Causo, B. Coluzzi, and P. Grassberger, *Phys. Rev. E* 62, 3958 (2000).
278. A. Hanke and R. Metzler, *Phys. Rev. Lett.* 90, 159801 (2003).
279. T. Garel, C. Monthus, and H. Orland, *Europhys. Lett.* 55, 132 (2001).
280. S. M. Bhattacharjee, *Europhys. Lett.* 57, 772 (2002).
281. S. E. Henrikson, M. Misakian, B. Robertson, and J. J. Kasianowicz, *Phys. Rev. Lett.* 85, 3057 (2000).
282. M. Bates, M. Burns, and A. Meller, *Biophys. J.* 84, 2366 (2003).
283. J. Chuang, Y. Kantor, and M. Kardar, *Phys. Rev. E* 65, 011802 (2002).
284. T. Ambjörnsson and R. Metzler, *Phys. Biol.* 1, 77 (2004).
285. M. Muthukumar, *J. Chem. Phys.* 111, 10371 (1999).
286. W. Sung and P. J. Park, *Phys. Rev. Lett.* 77, 783 (1996).
287. R. Zandi, D. Reguera, J. Rudnick, and W. M. Gelbart, *Proc. Natl. Acad. Sci. U.S.A.* 100, 8649 (2003).
288. Y. Kafri, D. K. Lubensky, and D. R. Nelson, *Biophys. J.* 86, 3373 (2004).
289. T. Ambjörnsson, S. P. Apell, Z. Konkoli, E. A. Di Marzio, and J. J. Kasianowicz, *J. Chem. Phys.* 117, 4063 (2002).
290. O. Flomenbom and J. Klafter, *Phys. Rev. E* 68, 041910 (2003).
291. O. Flomenbom and J. Klafter, *Biophys. J.* 86, 3576 (2004).
292. P. K. Purohit, J. Kondev, and R. Phillips, *J. Mech. Phys. Solids* 51, 2239 (2003).
293. R. Metzler and P. G. Dommersnes, *Eur. Biophys. J.* 33, 497 (2004).
294. J. T. Kindt, S. Tzlil, A. Ben-Shaul, and W. Gelbart, *Proc. Natl. Acad. Sci. U.S.A.* 98, 13671 (2001).
295. D. E. Smith, S. J. Tans, S. B. Smith, S. Grimes, D. L. Anderson, and C. Bustamante, *Nature* 413, 748 (2001).
296. T. Odijk, *Biophys. J.* 75, 1223 (1998).
297. S. C. Riemer and V. A. Bloomfield, *Biopol.* 17, 785 (1978).
298. J.-M. Lehn, "Supramolecular Chemistry," VCH, Weinheim, 1995.
299. J.-P. Sauvage and C. Dietrich-Buchecker, Eds., "Catenanes, Rotaxanes and Knots," VCH, Weilheim, 1999.
300. G. Schill, "Catenanes, Rotaxanes and Knots," Academic Press, New York, 1971.
301. A. Hanke and R. Metzler, *Chem. Phys. Lett.* 359, 22 (2002).
302. G. D. Fallon, M. A. P. Lee, S. J. Langford, and P. J. Nichols, *Organic Lett.* 6, 655 (2004).
303. G. Bottari, F. Dehez, D. A. Leigh, P. J. Nash, E. M. Perez, J. K. Y. Wong, and F. Zerbetto, *Angew. Chemie Int. Ed.* 42, 5886 (2003).
304. R. Metzler and J. Klafter, *Phys. Rep.* 339, 1 (2000).
305. R. Metzler and J. Klafter, *J. Phys. A* 37, R (2004).
306. M. Slutsky, M. Kardar, and L. A. Mirny, *Phys. Rev. E* 69, 061903 (2004).
307. R. Metzler and J. Klafter, *Biophys. J.* 85, 2776 (2003).
308. M. Kotulska, S. Koronkiewicz, and S. Kalinowski, *Phys. Rev. E* 69, 031920 (2004).

309. I. M. Tolic-Nørrelykke, E.-L. Munteanu, G. Thon, L. Oddershede, and K. Berg-Sørensen, *Phys. Rev. Lett.* 93, 078102 (2004).
310. M. Weiss and T. Nilsson, *Trends Cell Biol.* 14, 267 (2004).
311. M. Ptashne and A. Gann, "Genes and Signals." Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 2002.
312. C. E. Bell and M. Lewis, *J. Mol. Biol.* 314, 1127 (2001).
313. C. E. Bell, P. Frescura, A. Hochschild, and M. Lewis, *Cell* 101, 801 (2000).
314. A. Hanke and R. Metzler, *Biophys. J.* 85, 167 (2003).
315. A. Bakk and R. Metzler, *FEBS Lett.* 563, 66 (2004).
316. A. Bakk and R. Metzler, *J. Theor. Biol.* 231, 525 (2004).
317. A. Bakk, R. Metzler, and K. Sneppen, *Biophys. J.* 86, 58 (2004).
318. E. Aurell and K. Sneppen, *Phys. Rev. Lett.* 88, 048101 (2002).
319. E. Aurell, S. Brown, J. Johansen, and K. Sneppen, *Phys. Rev. E* 65, 051914 (2002).
320. K. Baek, S. Svenningsen, H. Eisen, K. Sneppen, and S. Brown, *J. Mol. Biol.* 334, 363 (2003).
321. W. J. Orr, *Trans. Faraday Soc.* 43, 12 (1947).
322. K. Ohno and K. Binder, *J. Phys. (Paris)* 49, 1329 (1988).
323. B. Duplantier and F. David, *J. Stat. Phys.* 51, 327 (1988).
324. B. Duplantier and H. Saleur, *Phys. Rev. Lett.* 59, 539 (1987).
325. P. Grassberger and R. Hegger, *Ann. Phys.* 4, 230 (1995).

CHAPTER 5

Receptor Flexibility in Ligand Docking

Claudio N. Cavasotto,[1] Andrew J. W. Orry,[2] Ruben A. Abagyan[1,2]

[1] *Molsoft LLC, La Jolla, California, USA*
[2] *The Scripps Research Institute, La Jolla, California, USA*

CONTENTS

ISBN: 1-58883-048-9

Handbook of Theoretical and Computational Nanotechnology
Edited by Michael Rieth and Wolfram Schommers
Volume 6: Pages (217–257)

1. INTRODUCTION

Corpora non agunt nisi fixata
(substances do not act unless bound)

P. Ehrlich, *Lancet*, 1913, 182, 445

Since this accessory substance is the recipient of stimuli which it transfers to the contractile material, we may speak of it as the receptive substance of the muscle.

J. N. Langley, *Journal of Physiology*, 1905, 33, 374

The Nobel Prize winners physiologist John Newport Langley (1852–1925) and immunologist Paul Ehrlich (1854–1915), working independently of one another, are widely accredited to be the first to introduce the concept of a receptor (see Ref. [1] for an excellent historical review). Ever since these groundbreaking observations, scientists have been attempting to understand how molecules specifically interact with receptors. More recently, it has become clear that receptors are not static but undertake sometimes subtle but critical structural rearrangements on interaction with molecules within the environment of the cell.

The formation of noncovalent complexes between molecules is crucial for the functioning of the biochemical machinery within a cell. Biological processes such as enzyme catalysis, molecular transport, and signal transduction rely on specific recognition between a receptor and the molecule to which it binds, otherwise known and referred to in this chapter as a ligand.

In an attempt to understand the formation of these complexes, structural biologists have been using molecular atomic structure as a means of predicting and detecting the formation of ligand–receptor interactions. Determination of the structures of these interacting molecules on the microscopic level has led to the elucidation of the pharmacological effects of these complexes and resulted in the rational design of drugs.

Recently, many worldwide structural proteomics initiatives and structure-based drug design methods have been having a significant effect on the development of therapeutics for a variety of diseases. Around 1000 new structures are being added to the Protein Data Bank (PDB) each year [2]. These structures reveal some of the secrets of protein–ligand interactions, and by using this information with the development of new computational technology, we can design molecules that can bind to these proteins. Ligand docking into and virtual screening (VS) of a drug target receptor is one method of identifying potential drug compounds based on the structure of the interacting molecules.

The protein–ligand binding process is a complicated equilibrium between a solvated ligand and receptor in isolation and a solvated receptor–ligand complex. In the attempt to calculate the free energy of binding, many energy terms need to be considered in the equilibrium balance, including effects caused by breaking and forming hydrogen bonds, the hydrophobic effect, loss of translational, rotational and conformational entropy of ligand and receptor on binding, and so forth.

The first attempt to predict molecular interactions was made in 1976. Beddell et al. [3] tried to design compounds that could fit into hemoglobin using handmade physical models of the protein structure. Six years later, Kuntz et al. [4] were the first to undertake a computer simulation of protein docking. Since then, many new computational methods have been developed for protein docking, and although the reliability and accuracy of the methods have gotten better, there is still scope for improvement, particularly when a ligand induces some changes in the receptor on binding.

Initially, docking methods were developed based on Fischer's [5] theory of a lock and key in which the three-dimensional (3D) structures of the ligand and receptor complement one another and fit favorably together. However, Koshland [6, 7] proposed a more accurate view of this process whereby the 3D structures of the ligand and receptor adapt to one another on binding, which is known as "induced fit." The structure of the ligand modifies the receptor binding-pocket environment, which gives rise to many low-energy substates before binding. We review this theory and the modern explanation about the ligand binding process in Section 3.1. One of the main problems associated with computational protein docking is how to incorporate flexibility into the ligand and drug-binding pocket and how to appropriately deal with "induced fit" (Fig. 1).

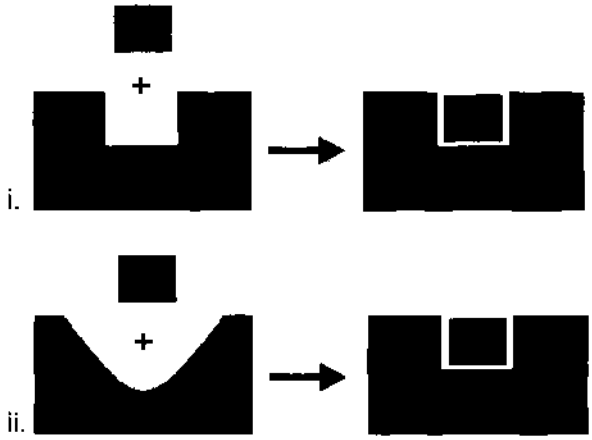


Figure 1. Schematic diagram of the theories relating to ligand (grey) and receptor (black) binding. (i) Fischer's lock and key theory. (ii) Koshand's theory of induced fit on ligand binding.

In general, a crystal structure is used as the basis of the structure-based design project; however, this only presents a rigid picture of the protein in the state in which it was crystallized, and not the many energetically favorable conformational states inherent in most protein receptors. The capability to predict many potential conformational states of a receptor, for example, antagonist in its and agonist forms, enhances the accuracy of structure-based drug design. Many protein-ligand complexes including HIV protease, aldose reductase, and streptavidin support the "induced-fit" model. In Section 5, we will discuss algorithms that incorporate a degree of receptor flexibility into protein docking.

Nowadays most docking algorithms can be used for virtual screening. However, incorporating flexibility into a virtual screening algorithm adds another level of complication to an already challenging problem. Conventional virtual screening processes uses a docking algorithm to screen any number of compounds, but generally in the range of from 200,000 to more than one million compounds, and then assess the docking scores. Different strategies can be used to reduce the pool of best scoring compounds (docking into a refined model, chemical clustering of the compounds, etc.) and the resulting compounds (usually numbering around 50–100) will be tested experimentally for their desired activity. Depending on the outcome of the experimental results, a new, focused virtual library can be developed from the highest-affinity compounds and rescreened, thereby refining the ligand scaffold and eventually increasing the experimental binding affinity (Fig. 2). In Section 6, we will discuss algorithms that attempt to incorporate receptor flexibility into virtual screening process.

In this chapter, we will discuss 1 current ligand docking and scoring methods; and why receptor and ligand flexibility is important for ligand docking; and current methods that incorporate flexibility into protein docking and virtual screening.

2. LIGAND DOCKING

The association of molecules such as a ligand to its receptor relies on a number of different interactions including electrostatics, hydrogen bonds, and van der Waals, but also entropy considerations (loss on binding, hydrophobic effect, etc.). Predicting these interactions and effects is very challenging, and only in recent years have methods begun to solve some of these problems. Prediction accuracy is improving every year; however, most molecules (ligands and receptors) are flexible, which adds an even greater degree of complexity to the prediction of ligand receptor interactions.

Computational docking has three main challenges:

1. Developing an exhaustive search algorithm that will sample all the conformational space for the ligand in a binding pocket.
2. Developing a scoring algorithm that can accurately represent the thermodynamics of interactions between a ligand and the receptor and differentiate between known and unknown binders.

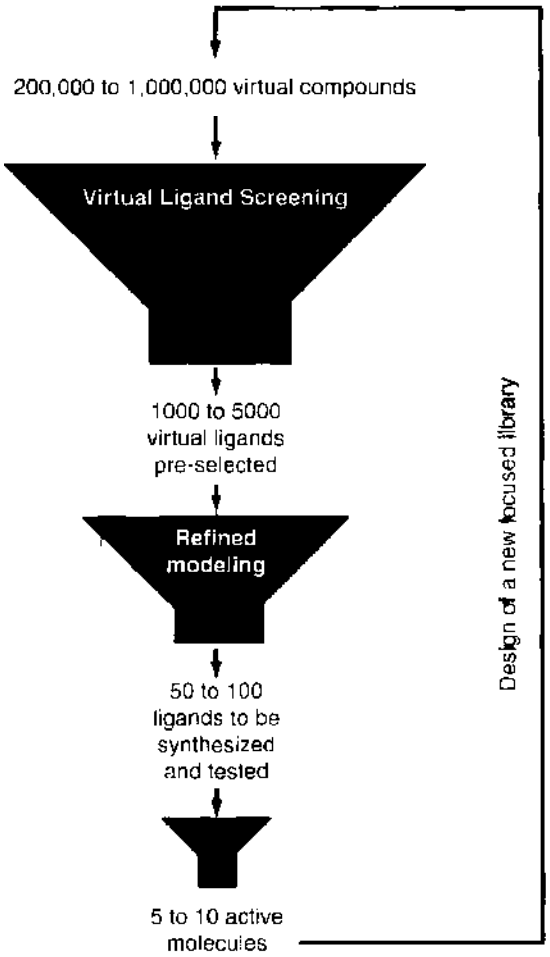


Figure 2. Diagrammatic representation of the conventional virtual screening process. In general, a large number of compounds are screened, scored, and tested experimentally. Good binders identified experimentally are then used as a scaffold and are computationally screened to identify new compounds that may have higher binding affinities or specificities.

3. Developing approximations that will speed up the calculations of challenges 1 and 2, and hence enable large databases of compounds to be screened in a reasonable amount of computational time.

A successful docking algorithm will incorporate all three of the above challenges into the methodology. Here we describe the current algorithms, methods, and scoring functions that are available for protein small-molecule docking. Section 5 will describe how some of these methods have been adapted to incorporate flexibility into the ligand binding pocket. Most docking algorithms can be adapted to include virtual screening (i.e., approximations can be made in the algorithm that speed up the docking process so as to allow large databases to be screened). This chapter is primarily concerned with protein flexibility; however, to understand how the flexible-receptor docking algorithms work, a basic description of current ligand–protein docking methods is required. (A more in-depth review of protein-docking algorithms and scoring methods can be found elsewhere in some excellent review papers [8–10].)

2.1. Ligand-Docking Background

Ligand docking in corporation with virtual screening can be used to screen millions of compounds to a receptor so as to narrow down a subset of potential drug-like compounds for testing in experimental binding assays. A standard experimental method for identifying ligands to receptor drug targets is to rigorously test *in vitro* many hundreds of thousands of potential binding compounds. This is time-consuming and costly and, as a result of the

sheer scale of the task, means that many important potential drug- candidates are missed and never tested. Testing is generally limited to compounds with similar structural characteristics; therefore, new potentially structurally diverse ligands will never be identified. Ligand docking technologies are quickly becoming a vital step in the drug discovery process.

To perform computational ligand docking, a 3D high-resolution structure of the protein of interest is needed. This structure can be obtained in a variety of ways, including searching the PDB [2] to determine whether there is an x-ray crystal or nuclear magnetic resonance structure, or making your own fully validated homology model. If you already know where the active site is within the receptor, then you are ready to perform the docking analysis. However, if the position of the binding site is unknown or unclear, then you may wish to use a ligand-binding-pocket-finder algorithm, along with experimental information such as mutagenesis data, to predict a ligand-binding site.

There are some important considerations to take into account before undertaking a computational docking. These considerations are

1. Symmetry: Is the true ligand-binding pocket formed by chains that are not explicitly present in the PDB entry for your receptor? To solve this problem, you need to find all the molecules/chains/subunits involved in the interaction with the ligand.
2. Occupancies: What is the occupancy value for the receptor pocket? Occupancy is a fraction of atomic density at a given point ranging from 0 to 1. Sometimes electron density is poor in a crystal structures and the occupancy is low (<0.5), and hence this is not a good target for ligand docking.
3. B-Factor: The B-factor is the mean-square displacement of an atom from its position in the structure. If this factor is high (>60), then it is probably not a good structure for ligand docking.
4. Beware of the orientation and protonation states of the following amino acid residues in your receptor binding pocket: histidine, lysine, arginine, glutamine, asparigine, glutamate, and aspartate.
5. Are there any water molecules or metal ions in the ligand binding pocket? Include them in the analysis if they are critical for ligand binding; if not, remove them from the procedure.

In a real-life case, the goal of virtual screening is to select a small number (<2% of a database) of potential binders to the receptor of interest from a large source library. An enrichment factor (EF) is a good benchmark for the performance of a docking algorithm. For a particular focused library built with the top s compounds of the ranked database, the enrichment factor can be calculated as

$$EF(s) = \frac{\mathrm{Hits}_s}{\mathrm{NC}_s} \Big/ \frac{\mathrm{Hits}_{\mathrm{total}}}{\mathrm{NC}_{\mathrm{total}}} \tag{1}$$

where NC is the number of compounds. Defined in this way, the enrichment factor is the ratio of the probability of finding a ligand in the top s compounds of the screened database to the probability of finding it at random.

The first approach to predicting the structure of a protein ligand complex was undertaken with physical models of the receptor and ligand [3]. The first computational approach to attempt to solve the ligand-docking problem was described by Kuntz et al., who based their method on the observation that shape complementarity is evident in complex formation [4]. The DOCK method of Kuntz et al. used geometrical matching surfaces to predict protein–ligand interaction by keeping the receptor and ligand rigid.

Other early docking methods also assumed rigidity of both the ligand and receptor and shape complementarity. In 1986, an extension of Kuntz's method was proposed that used overlapping spheres to represent the ligand-binding site [11, 12]. Each ligand was divided into rigid fragments that were docked separately, rejoined, and then minimized. Katchalski-Katzir et al. [13] used a Fourier transform approach to assess the degree of molecular overlap between the ligand and the receptor. Other methods used least squares to fit the surface patterns [14, 15], distance-matrix matching [16], and "geometric hashing" [17].

Since these early methods, many more ligand docking algorithms have been described. Ligand docking and virtual screening methods (some popular algorithms include: AUTODOCK [18], DOCK [19], PRODOCK [20], EUDOC [21], FlexX [22], FLOG [23], GOLD [24], Internal Coordinate Mechanics (ICM) [25], LUDI [26], Pro_LEADS [27], QXP [28], and SLIDE [29]) have been constantly developed for many years and use a variety of procedures, including Monte Carlo (MC), molecular dynamics (MD), genetic algorithms, and shape complementarity. A good docking algorithm will successfully incorporate a good 'chemical space' search strategy, a reliable scoring system and will be computationally affordable.

2.2. Ligand Docking Methods

2.2.1. Monte Carlo

The MC approach, named after the city where the main attractions are games of chance, randomly generates values for uncertain variables over and over to simulate a model.

MC was first introduced in 1953 by Metropolis et al. [30] and consists of three steps, which are then repeated:

1. One or several system variables are changed randomly.
2. The energy of the system is evaluated.
3. If the energy E_{new} is lower than E_{old}, then the new conformation is always accepted and used in the next iteration. If not, the new conformation is accepted with a probability given by

$$P_{acc} = \exp\left[\frac{-(E_{new} - E_{old})}{kT}\right] \tag{2}$$

where k is Boltzman's constant and T is the effective temperature of the simulation.

This method allows the sampling of many different energy minimums and can cover a wide and rugged energy landscape that is not possible with MD methods (see Section 2.2.2). Methods that incorporate MC include AFFINITY [31], ICM [25], MCDOCK [32], PRODOCK [20], and QXP [28].

In the ICM docking method, the molecular system is described in terms of internal coordinates, which substantially reduces the number of variables defining the conformation of the system, and uses a modified ECEPP/3 [33] force-field with a distance-dependent dielectric constant [25, 34]. The Biased Probability Monte Carlo [35, 36] (BPMC) minimization procedure is used for global energy optimization. The BPMC global energy optimization method consists of the following steps: (1) a random conformation change of the free variables of the system according to a predefined continuous probability distribution [35, 36]; (2) local energy minimization of analytical differentiable terms, as it has also been shown that after each random step full local minimization greatly improves the efficiency of the procedure [37, 38]; (3) calculation of the complete energy including nondifferentiable terms such as entropy and solvation energy; and (4) acceptance or rejection of the total energy based on the Metropolis criterion [30] (see above) and return to step 1.

The ICM method has been successful at identifying agonists and antagonists with a model of the retinoic acid receptor (RAR), which is implicated in a wide range of diseases including breast cancer and diabetes [39, 40]. Two novel agonists and two novel antagonists resulting from the predicted receptor model were active *in vitro*. One of the agonsts displayed novel structural features that may translate into the development of new ligands for cancer therapy. Recently, the ICM docking and virtual screening method has found 14 inhibitors to the thyroid receptor (TR) that have potential to treat hyperthyroidism, a common clinical condition that can have serious manifestations in the elderly [41].

One of the current challenges of structural biology is to solve the atomic structure of a class of membrane proteins called G-protein coupled receptors (GPCRs). GPCRs and membrane proteins in general are very difficult to structurally characterize because of the problems associated with expression and crystallization. At present, only one GPCR has been solved by x-ray crystallography [42] and therefore this class of proteins offers a challenging scenario

for protein modelers and those who wish to use docking as a means of validating, refining protein models, GPCR "deorphanization," or drug design. The ICM docking method has been validated as a GPCR docking tool. The docking method was able to very accurately (~0.15 Å) redock the native ligands of the GPCR bovine rhodopsin and the GPCR-like seven-transmembrane protein bacteriorhodopsin. A high enrichment factor has been shown when using ICM to screen the LIGAND natural products library from the University of Kyoto [43, 44] and ChemBridge library (San Diego, CA) [45] (see Fig. 3). The same scoring was achieved both with and without the crucial second intracellular loop which indicates that it may not be necessary to complicate the GPCR modeling problem by incorporating loop regions.

The ICM method can be adapted to incorporate flexibility within the receptor. This will be discussed more fully in Section 6.3.

PRODOCK also uses a MC approach and after each random move applies a local gradient-based minimization in the AMBER [46] or ECEPP/3 [33] force field, similar to the ICM method. Then, Bezier splines are incorporated into a grid-based technique to evaluate the energy function [20, 47, 48].

MCDOCK, in contrast to the previously described MC methods of ICM and PRODOCK, does not use a gradient optimization method, but the conformations are generated by geometrical-based docking followed by energy-based docking [32]. MCDOCK was successfully tested on 19 different ligand-receptor complexes, with a root mean square deviation

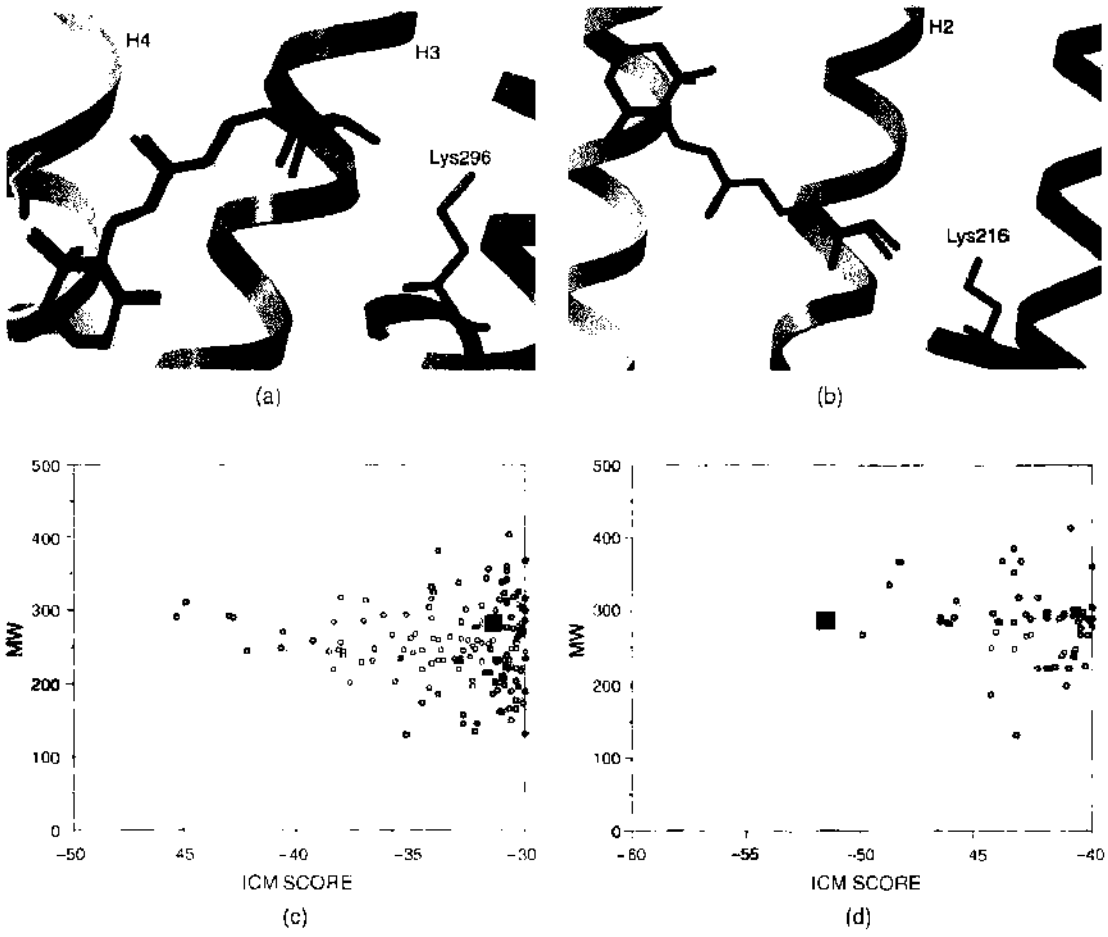


Figure 3. Bound crystal structure retinal (blue) and redocked retinal (red) are shown for bRho (a) and BR (b). Helices (H) VI and VII have been cut away for clarity. (Pictures colored from blue [N terminus] to red [C terminus]). In (c) and (d), virtual screening plots of docking score for the ligands in the LIGAND database docked into the bRho receptor (c) and BR (d) are shown. Retinal, the known binding ligand to both receptors, is highlighted with a square box. Ligands are characterized by their molecular weight (MW). Only best scoring compounds are shown (pictures generated using ICM). Reprinted with permission from [45], C. N. Cavasotto et al., *Proteins-Struct. Funct. Genet.* 51 423 (2003) © 2003, Wiley.

(RMSD) between the experimental and MCDOCK predicted poses of between 0.25 and 1.84 Å [32].

Another MC method called QXP [28] was tested on 12 receptor-ligand pairs, and 11 of them were predicted with less than 0.76 Å RMSD from the original crystal structure. This method uses only random dihedral moves initially, followed by rigid-body translations and rotations using atoms within the binding pocket as guides. Once guided into the pocket, MC is performed on the ligand, using only rigid rotations and translations. Gradient minimization followed by MC is then performed on the ligand torsion angles.

A number of other MC methods have been successful with a variety of real biological docking problems, including a method by Caflisch et al. in 1997 [49], who successfully docked a ligand to the FK506 binding protein (FKBP). This method is described in more detail in Section 5.3.1 because it incorporates some flexibility into the receptor-binding pocket.

Bouzida et al. [50], who docked HIV-1 protease and commercial programs such as GLIDE [51–53] (Schrödinger Inc., San Diego, CA) and AFFINITY [31] (Accelrys Inc., San Diego, CA), all use MC and have been successfully applied to a variety of small-molecule docking problems.

2.2.2. Molecular Dynamics

In a molecular dynamics (MD) simulation, the evolution of the system as a function of time is obtained by integrating Newton's law of motion. Atoms within molecules are considered as point particles with mass and charge, thus obeying the laws of classical mechanics. The energy of the system is described in terms of force-fields such as AMBER (Assisted Model Building with Energy Refinement) [46] and CHARMM (Chemistry at Harvard Macromolecular Mechanics) [54].

MD simulations are a very computationally expensive technique, especially in the field of ligand-protein docking, because traversing a rugged energy landscape may be very problematic. A related problem is that even state-of-the-art MD simulations using parallelization can only simulate periods ranging from 10 to 100 ns, which is not long enough to represent many of the ligand-binding processes.

However, some algorithms have overcome these inherent problems associated with using MD by incorporating approximations that reduce the computational expense inherent with an MD simulation. Di Nola's group has been successful in using MD for docking by separating the center of the mass of the substrate from its internal and rotational motion and then analyzing thermodynamic properties [55, 56]. They used a very high temperature for translational modes and a regular temperature for the internal degrees of freedom, thus allowing the ligand to sample the surface of the receptor much faster. The method was further adapted to incorporate flexibility into the receptor, coupled with a lower-temperature bath that will be discussed in Section 5.3.1.

Other groups have tried a similar approach (see Refs. [31, 57–59]), and so-called Brownian dynamics have also been applied [60, 61]. To reduce the computational time for an MD simulation, Luty et al. [31] restricted the full molecular simulation to the vicinity of the binding site. They correctly predicted the formation of the complex between benzamidine and trypsin, performing 100 independent simulations. On the basis of this method, Wasserman and Hodge [62] docked L-leucine hydroxamic acid into the active site of thermolysin. They used a potential that imposed no constraints between the ligand and the receptor and produced docked conformations similar to the crystallographic determined one.

A wide variety of other MD methods has been applied to protein docking and been used to incorporate a degree of flexibility into the receptor on docking. These methods are discussed in Section 5.3.1.

2.2.3. Fragment-Based Methods

Fragment-based docking methods generally involve three main steps: dividing the ligand into fragments, docking the ligand fragments, and linking and merging the ligand fragments after they have been docked.

This method relies on the choice of these fragments and on the importance associated with certain functional groups for docking: this can, however, result in the omission of certain ligand-docking solutions and can sometimes affect the quality of results.

FlexX [22, 63] is a fragment-based docking method that has been validated on 19 test cases, giving good results (i.e., similar to crystal structure with low RMSD) for 10 of the cases, but these results were not necessarily the structures with the best energy [64]. The method was improved to include explicit waters in the binding pocket [63]. The method first uses a torsion angle database to identify fragments that will make the primary interaction with the receptor, and hence from which other fragments will be joined and conformations formed. The interactions between the fragments and the receptors are aligned and optimized on the basis of hydrogen bond and hydrophobic interactions. Three sites on the fragment are mapped onto three sites on the receptor, and then the fragments are clustered, superimposed, and filtered. The ligand is then built by adding new fragments to the original fragment, and these new fragments are then ranked and assessed using an empirical scoring function.

The DOCK [19] algorithm uses spheres to describe the ligand and the receptor cavity. These spheres are then matched, orientated, and scored. The method has been improved by the incorporation of bipartite graphs, whereby the ligand atoms and receptor site points are separate sets of nodes. Clique detection has also been incorporated and was found to be the most efficient method for finding maximal pairs of interactions between matching sites [65-67]. Each of these processes generates many orientations that are then scored using the AMBER [46] force field. This algorithm is further discussed in a later section in which the method has been developed to incorporate flexibility using an ensemble of structures.

The docking algorithm LUDI [26] is a module of the INSIGHT suite of programs. LUDI calculates the points of interaction in the active site of the receptor. The points of the interaction sites are derived from nonbonded contact distributions based on the Cambridge Structural Database, from the output of the program GRID [68], or from a set of geometric rules. Fragments are then predicted on the basis of hydrogen donors/acceptors and lipophilic interactions. Using this information, a new ligand can be developed by joining fragments using a database of bridging fragments. The algorithm has been successful at predicting modifications to an HIV protease inhibitor and identifying an inhibitor to DHFR [26].

Another fragment-based docking algorithm is called ADAM [69]. The method was found to be extremely efficient at predicting ligand interactions in two enzyme systems—methotrexate to DHFR and 2′-GMP to ribonuclease T-1. The first step of the algorithm is to characterize the ligand-binding modes in terms of hydrogen bonding in the receptor pocket and the ligand. The hydrogen-bonding sites of interaction between the ligand and the receptor are determined by systematic rotations of torsion angles and superimposed. The AMBER [46] force-field is used to minimize the resulting structures and assess the interaction energy.

HAMMERHEAD [70] also uses a fragment-based approach in which the fragment of the ligand was assumed to make contact with the receptor is referred to as the "head." Each head fragment is then aligned to the receptor, and the best scoring pairs are retained. The remaining fragments that are not head regions (referred to as "tails") are then aligned onto the head fragment and optimized for hydrogen bonding and hydrophobic contacts. The best-scoring head and tail merged complexes are then merged with other fragments to improve the binding. The algorithm was found to be very accurate compared to the crystal structure for the following complexes: benzamidine/trypsin, biotin/streptavidin, methotrexate/DHFR, and NAPAP to thrombin. The algorithm was also successfully used as a virtual screening tool to screen for binders to streptavidin.

The SLIDE [29] algorithm is a fragment-based docking algorithm that is discussed later in this chapter, in Section 5.2, as it has been successful at incorporating flexibility into the receptor as well as the ligand. The SLIDE algorithm was the precursor to the docking program, called SPECITOPE [29].

2.2.4. Genetic Algorithms

Genetic algorithms (GA) are programmed to mimic the evolution process and have been widely used in protein simulations [71, 72]. The GAs manipulate a set of variable values that define a possible solution, such as a certain binding mode in protein docking, rotatable torsion angles of the ligand, and relative orientation of the ligand and receptor.

The variables are refered to as "chromosomes" or binaries, and the algorithm begins with a random set of chromosomes, which then they evolve by "mutations" and "crossovers," maintaining advantageous combinations of various values according to a predefined function for the system. The next stage, which is sometimes referred to as "the survival of the fittest," involves scoring and ranking the new population. The "fittest" chromosomes according to the score and rank survive and continue into the next stage of the algorithm. The "fittest" chromosomes become the parents of the next generation, these spawn new chromosomes, and the process is repeated a defined number of times.

The GOLD [24] algorithm uses a GA to represent both rotatable dihedral and ligand–receptor hydrogen bonds. Intermolecular hydrogen bonds between the ligand and the receptor are optimized by using a least-square fitting method. Each complex is scored on the basis of the number of hydrogen bonds, a 4–8 intermolecular dispersion potential, and a 6–12 intramolecular potential for the energy of the ligand. Each complex is ranked according to this scoring function. The GOLD docking algorithm is discussed in more detail in relation to flexibility within the receptor later in this chapter (see Section 5.2).

The DARWIN [73] docking algorithm is a GA-based method that incorporates a molecular mechanics process involving the CHARMM [54] force-field. The two approaches (GA and MD) are run in parallel, which enables the ligand orientation and energy of the complex to be optimized together. DARWIN was successfully tested on three carbohydrate-binding proteins: Concanavalin A, Fab Se155-4, and Fab ME36.1. Some false-positive effects were observed on docking, which highlighted the problems of solvent-effects in ligand docking.

The AUTODOCK [74] algorithm uses two-point crossover and mutation operators, and the fitness function is made up of the following five terms: a coulombic electrostatic potential, a Lennard–Jones hydrogen bond term, a desolvation term, and an entropy term based on the number of sp^3 bonds in the ligand. These terms are then weighted using linear regression analysis from a test set of 30 protein–ligand complexes with known binding affinities. This docking algorithm was tested on seven complexes, which represented a range of complexity and chemical properties. The algorithm was able to redock the structures to within 1.14 Å of the original crystal structure.

Other successful GA docking algorithms include the algorithm named DIVALI [75, 76] and methods developed by Judson et al. [77] and Gehlhaar et al. [78].

2.2.5. Points of Complementarity

Points of complementarity can be used between the protein receptor and the ligand to aid docking. For example, FTDOCK [79] uses shape and electrostatic models to perform rigid docking. The translational space of two rigidly rotating molecules is scanned using the Fast Fourier Transform and the Fourier correlation theory [13]. The algorithm discretizes two molecules A and B, with A being the larger of the two molecules. The molecules are discretized into 3D grids, with every node assigned a value. A grid node, which is within 1.8 Å of an atom, is considered to be inside a protein. Grid spacing is determined by the size of the respective molecules plus 1 Å. The size of the molecule is defined as the approximate radius from the geometric center of the protein to the most distal atom.

The FTDOCK software also incorporates electrostatic interaction of the binding site as well as shape complementarity. The charge–charge interactions at the binding face are calculated by Fourier calculations, using a Coulombic model. FTDOCK measures the point charge of one protein interacting with the electric field of the other, acting as grid points. The electrostatic calculations are then performed in a similar way to the shape complementarity calculations described earlier. Charges are assigned to the larger protein A, and the molecule is placed in a grid. The electric field at each grid node (excluding that of the protein core) is calculated.

In 1998, Burkhard et al. developed a program called SANDOCK [80], which uses a matching algorithm to fit ligand atoms into the binding pocket. The ligand-binding pocket is represented by a dotted surface, whereby each dot represents the chemical property and accessibility. Each matching dot between the ligand and the receptor has to meet certain criteria regarding shape and chemical complementarity. A scoring function was developed on the basis of van der Waals, hydrogen bonding, and hydrophobic interactions. The method was successfully tested on a thrombin complex to within 0.7 Å of the crystal structure.

Other methods include LIGIN [81, 82], which uses the surface to dock a rigid ligand to a rigid receptor, and the algorithm FLOG [23], which selects ligands from a database based on shape. The use of FLOG in virtual screening is discussed in Section 6.1.

2.2.6. Other Docking Algorithms

Other docking algorithms of interest are PRO_LEADS [27] (see Section 4.2), which uses a tabu search with a scoring function, and EUDOC [21], which systematically searches using rigid body rotations and translation with a rigid active site.

2.3. Docking Scoring Methods

Docking algorithms generally produce a large number of conformations for a ligand within a binding pocket, and most solutions can be rejected because of clashes with the receptor; however, the remainder of the solutions have to be assessed with a scoring method of some kind. The primary goal of any scoring method in ligand docking is to be able to differentiate between potential binders and nonbinders. This is achieved by ranking conformations according to some defined scoring function. When applying a docking algorithm to the virtual screening of many thousands of compounds, it is imperative to achieve a balance between speed and accuracy. This is why the need for rapid scoring inevitably results in some approximations to the calculations. Many scoring functions are trained on certain classes of proteins but can cause problems and errors such as false positives when using the function on other types of protein–ligand systems.

Scoring functions approximate the binding free energy between a ligand and the receptor. To achieve the balance between speed and accuracy, scoring functions consider the free energy of binding as an additive function of different energy terms that reflect the different contributions to binding. Various conditions within the receptor-ligand complex can affect binding, and these need to be incorporated into a scoring function. The free energy of binding can be represented in an additive way by [83, 84].

$$\Delta G_{\text{bind}} = \Delta G_{\text{solvent}} + \Delta G_{\text{conf}} + \Delta G_{\text{int}} + \Delta G_{\text{rot}} + \Delta G_{\text{t/r}} + \Delta G_{\text{vib}} \tag{3}$$

where $\Delta G_{\text{solvent}}$ is the contribution because of solvent effects, representing the change in the interactions with the solvent of the ligand, receptor, and complex. The term ΔG_{conf} is the contribution because of conformational changes in the protein and ligand. Most of the docking methods consider the receptor as a rigid entity, ignoring flexibility and thus introducing large errors in docking and virtual screening (see Section 4.2). The term ΔG_{int} is the free energy resulting from specific interactions between the ligand and the receptor. The variable ΔG_{rot} is the contribution mainly resulting from entropy loss associated with the freezing of rotatable degrees of freedom in the protein and the ligand, usually estimated as $RT \ln 3$ (~0.7 kcal/mol) per rotatable degree of freedom. The term $\Delta G_{\text{t/r}}$ is the free energy change resulting from rotational and translational entropy loss arising from the association of two bodies (the ligand and the protein) in one body (the ligand complexed with the receptor). Finally, ΔG_{vib} is the free energy resulting from changes in the vibrational modes on ligand binding. This contribution is very difficult to estimate and is usually ignored in the calculation. The weighting of each of these contributions is based on a regression model of a test set of protein–ligand complexes with known binding affinities [85–88].

Binding affinity is usually quantified by the following equation:

$$\Delta G^{0} = -RT \ln K_{i} \tag{4}$$

where the standard Gibbs free energy ΔG^0 is related to the binding constant K_i in thermodynamic equilibrium conditions.

It is clear, therefore, that the test set on which these methods are based needs to be variable and large if the algorithm is going to perform well on protein–ligand complexes outside of the original test set. This is also a factor to consider when comparing the performance of docking algorithms; that is, whether the docking is accurate because the complex is one similar to the test set on which the scoring algorithm is based.

The docking algorithms LUDI [26] and FlexX [22] use a similar scoring method to that described above and are based on the SCORE algorithm by Bohm [85].

Many scoring functions are determined by interaction terms based on a force-field such as AMBER [46], CHARMM [54], and the Merck force-field [89]. This approach is very time consuming, and the method relies on a representation of the energetic terms that are difficult to approximate.

A knowledge-based scoring method has proved popular with some groups developing protein-docking algorithms. This method is based on statistical analysis of pairs of atoms within known ligand–protein complexes. Verkhivker et al. [90] used a knowledge-based pair potential approach incorporating hydrophobicity and conformational entropy scales to determine binding affinity within HIV-1 proteases. They used a data set of 30 protease (HIV-1, HIV-2, and SIV) structures, and their estimates for binding affinity were reasonably accurate. Another method, called SmoG [91], used interatomic interaction free energies for a variety of atom types from 126 complexes taken from the PDB.

In 1999, Muegge and coworkers [92, 93] also developed a knowledge-based scoring function called potential of mean force score (PMF-score) and implemented it in the program DOCK. They derived distance-dependent Helmholtz free-interaction energies of protein–ligand atom pairs of 77 varied complexes from the PDB. The calculated score produced a standard deviation of 1.8 log K_i units, and the method was further tested by docking ligands to the FK506-binding protein.

Gohlke et al. [94] developed a knowledge-based scoring algorithm called DrugScore based on information extracted from the PDB by ReLiBase. The method incorporates implicit solvation and entropy considerations and docked ligands to within <2.0 Å, compared with their original crystal structure.

A sensitive discrimination potential for virtual screening has been described [95] and validated on 63 ligands and 23 diverse receptors, using ICM. Scores derived from a weighted sum of five physical terms are calculated. These terms include hydrophobicity, solvation, electrostatics, hydrogen bonding, ligand internal energy, and the van der Waals ligand–receptor interaction energy.

The ICM scoring function consists of the following terms:

$$E_{\text{score}} = \Delta E_{\text{IntFF}} + T\Delta S_{\text{Tor}} + \alpha_1 \Delta E_{\text{HBond}} + \alpha_2 \Delta E_{\text{HBDesol}} + \alpha_3 \Delta E_{\text{SolEl}} + \alpha_4 \Delta E_{\text{HPhob}} + \alpha_5 Q_{\text{Size}} \quad (5)$$

where ΔE_{IntFF} is the van der Waals interaction of the ligand with the receptor in addition to the internal force-field energy of the ligand; $T\Delta S_{\text{Tor}}$ represents the ligand conformational entropy loss contribution when the ligand binds to the receptor, which is assumed to be proportional to the number of free torsions; ΔE_{HBond} is the term relating to hydrogen bonding; $\Delta E_{\text{HBDesol}}$ represents the disruption of hydrogen bonds because of solvent effects on ligand binding (desolvation of hydrogen bond donors and acceptors); ΔE_{SolEl} is the electrostatic energy change resulting from solvation when the ligand binds to the receptor, which is calculated by solving the Poisson equation using the boundary element method as implemented in the Rapid Exact Boundary Electrostatics (REBEL) [96] module of ICM; ΔE_{HPhob} is the hydrophobic free energy gain that is assumed to be proportional to the accesible surface area of the hydrophobic atoms buried when the ligand–receptor complex forms; and Q_{Size} avoids bias toward larger ligands by implementing a size correction term. The values related to α_1–α_5 are weights that give each interaction the appropriate strength. These weights were benchmarked and optimized on a diverse set of ligand–receptor complexes.

3. IMPORTANCE OF FLEXIBILITY IN LIGAND BINDING

3.1. Different Views of the Mechanisms of Ligand Binding to Protein Receptors

Protein flexibility has been more or less ignored so far in ligand docking and screening. This would be correct in cases in which ligand binding could be described by the lock-and-key theory, by which the protein can be represented by a single minimum energy state, to which the ligand fits as the key in a lock. In this case, little if any conformational change can be expected in the receptor on ligand binding. However, this is not the case in a vast majority of ligand–protein complexes, in which the protein exhibits structural changes on ligand binding. These changes explain why sometimes very different ligands can bind to the same receptor at the same binding site.

These ligand-induced changes were described as the "induced fit" effect [7], whereby the ligand forces a deformation on the receptor binding site, which is assumed to be in a single-state conformation previous to ligand binding. A more recent way to describe how the process of binding takes place is by extending the concept of folding funnels to ligand binding [97]. According to this theory, a protein in solution is assumed to exist in a statistical ensemble of conformationally different low-energy states separated by low-energy barriers. Figure 4 shows a correlation between the different types of energy landscape and the type of binding process. It is clear that the width of the funnel near the minima measures the conformational entropy, and thus the size of the equilibrium ensemble. Figure 4A depicts a protein described by a single well-defined conformational state, which would correspond to the lock-and-key case, with little conformational change on ligand binding. Figure 4B represents the rugged bottom of the energy landscape, with the equilibrium state of the protein being represented by multiple states separated by low-energy barriers. A ligand may bind favorably, although not necessarily, to the state with lowest energy, thus shifting the equilibrium and the distribution of populations toward that state. In this way, the average structure of the protein will change, reflecting the conformer selection by the ligand. Experimental crystal structures in fact correspond to an average of the conformational ensemble and might reflect a state with higher free energy than the minimum, depending on the energy landscape. The degree of ruggedness correlates with the structural diversity of the conformational ensemble that characterizes the protein in equilibrium, and thus its flexibility. The novelty of this approach is that a flexible protein or binding site is flexible not because it can adapt to the different types of ligands on binding but because it is capable of preexisting in a number of conformational states with similar energy, and different ligands can choose within this ensemble the most favorable conformation (conformer selection).

The binding of the ligand to a conformation of higher energy and thus sampled occasionally should correspond to a reduced rate of association. This is indeed the case in the

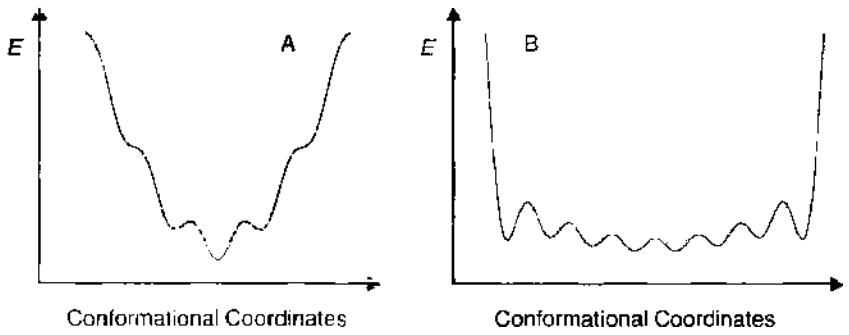


Figure 4. Representation of the native structure of protein in the vicinity of the bottom of the funnel and its relation to ligand binding. (A) The native state is defined by a single, well-defined conformational state. Little conformational change is expected on ligand binding. This would correspond to the lock-and-key case. (B) The native state corresponds to a multiminima landscape, with low-energy barriers separating the equilibrium states. Different ligands can bind to structurally different states, thus shifting the conformational equilibrium toward those states, which are ultimately selected by the ligand (conformer selection)

picomolar inhibitor BIRB 796 binding to p38 MAP protein kinase, in which the loop including the Asp-Phe-Gly (DFG) motif, usually in the DFG-in conformation, with the Phe located in the binding pocket, moves to the DFG-out conformation on ligand binding, thus displacing the Phe by ~10 Å to a solvent-exposed conformation [98]. This conformation had been never observed in any Ser/Thr kinase, although it seems to be more stable in protein tyrosine kinases. The association rate (k_{on}) is 8.5×10^4 $M^{-1}s^{-1}$, which is 500-fold slower than that of the nanomolar inhibitor SK&F 86002. In a binding experiment, BIRB 796 required fourfold more incubation time to achieve maximal potency than other inhibitors, which are supposed to bind to lower energy conformations. The dissociation rate for BIRB 796 is $\sim 10^6$ slower than that of SK&F 86002, showing that the ligand is slow to come out of the pocket, mainly because of the large number of hydrophobic contacts within the binding pocket. BIRB 796 has been claimed to be one of the most potent and slowest-binding inhibitors of human p38 MAP protein kinases known to date.

Other inhibitors also slow inhibition kinetics. Huperzine binds to acetylcholinesterase with marked slower association/dissociaton rates than other inhibitors of similar potency, mainly because of an increase of hydrophobic interactions around the ligand and cation-π interaction [99]. The low-nanomolar inhibitor oseltaminir carboxylate of neuraminidase also displays slow-binding kinetics [100, 101].

Slow-binding kinetics may have a potential effect on *in vivo* efficacy, because a slower dissociation rate may translate into extended dosing protocols.

The evidence about the mechanisms of ligand binding has strong implications in computer-aided drug design, as it shows that we have to incorporate flexibility in such a way as to be able to consider potential ligands that may differ in shape and size from the known binders.

3.2. Protein Flexibility and Function

There is a wealth of experimental evidence that shows that flexibility is needed for proteins to perform their biological function, and very often, this function requires sequential binding of ligands and substrates. Nuclear receptor transcriptional activity is controlled by the binding of agonists, antagonists, and cofactors. ATP binds to protein kinases, phosphorylates substrates, and triggers a downstream signaling cascade. Membrane proteins bind ligands as a way to transmit signals from outside the cell to inside. Crystal structures of ligand–protein complexes show that usually, most of the ligand is buried within the receptor, which can only be the result of protein flexibility.

However, it is interesting to examine the question of how flexibility is distributed within a flexible ligand binding site. By performing a structure-based thermodynamic stability analysis of 16 nonstructurally homologous proteins, Luque and Freire found that binding sites have a dual character, with the presence of regions with high structural stability and regions with low structural stability [102]. In many, but not all, of the cases, the flexible parts are loops. In enzymes, the catalytic sites are usually characterized by high structural stability, although it should be noted that extreme stability can be adverse for biological activity [103, 104].

On the basis of observations that broader-range binding proteins may bind ligands with different shapes and sizes than natural ligands with equal or even higher affinities [105, 106], and that this could be also the case for specific enzymes or receptors [107]. Ma et al. argued that, in fact, specific and broad-range ligand binding are the outcome of the preexisting population of conformers within the statistical ensemble described above, reflecting a dynamic situation of large-scale hinge-bending-type movements between relatively rigid domains or subdomains [108].

In fact, according to Van Regenmortel, the structural conformation of the binding site is dictated by the ligand, as in the process of binding, the conformer that makes the most favorable interaction with the ligand is selected [109]. In an illustrative example, the crystal structure of recombinant plasmepsin II complexed with pepstatin A contained two independent proteins in the asymmetric cell unit [110], exhibiting different domain displacements [111], and thus showing that the protein preexists at least in the two observed states.

4. WHY IS RECEPTOR FLEXIBILITY AN IMPORTANT CONSIDERATION IN LIGAND DOCKING?

To most of the modern pharmacologists the receptor is like a beautiful but remote lady. He has written her many a letter and quite often she has answered the letters. From these answers the pharmacologist has built himself an image of this fair lady. He cannot, however, truly claim ever to have seen her, although one day he may do so

D. K. deJongh, in *Introductory Remarks to Molecular Pharmacology*, Vol. 1, Academic Press, edited by E. J. Ariens, 1964

4.1. Examples of Protein Mobility on Ligand Binding

There are many examples of receptors that have one or more critical flexible residues, examples of which will be discussed here. The most common example of residue flexibility within the binding pocket are residues involved in catalysis. Single flexible residues are observed by x-ray crystallographic analysis to move and open out, allowing ligands of varying sizes to bind with the receptor.

In 2000, Najmanovich et al. [112] undertook an interesting study to try and analyze the extent amino acid side chains lining the binding pocket undergo conformational changes on ligand binding. From a study of two nonredundant databases (980 holo-proteins and 353 apo-proteins from the PDB), they found that in ~85% of cases, three or fewer residues belonging to the binding pocket showed conformational changes. Although this seems a small number, any degree of flexibility is critical to ligand binding and needs to be addressed using ligand-docking algorithms.

An example of a single flexible residue being important for ligand binding was published in 1994 by Smith et al. [113]. They solved the x-ray crystallographic structure of human carbonic anhydrase II bound to three inhibitors that only differed in the length of the substituent on the four-amino group. There is an observed difference in binding affinity between the three inhibitors, with the K_i ranging from 1.88 to 0.37 nM. It was identified from the crystal structures that a histidine residue (His64) takes up alternate conformations on binding, accounting for the wide variation of affinity between the relatively similar inhibitors. In the weakly bound inhibitors, His64 is interacting with a water molecule, but in the complex with the tighter-bound inhibitor, there is no water present, and the His64 has changed its pose slightly. It is hypothesized that the increase of entropy when the bound water molecule is released and the subsequent movement of the histidine results in tighter binding.

In 1999 Kryger et al. [114] published the crystal structure of the *Torpedo californica* acetylcholinesterase (AChE) protein complexed with the Alzheimer drug E2020 (Aricept) at 2.5 Å resolution. This is an interesting structure because it highlights the effect of a single flexible amino acid. The flexible residue Phe330 is in stark contrast to the rest of the structure, which is rigid. The flexibility was observed by studying the structures of AChE bound to six other inhibitors in the ligand gorge. The authors described Phe330 as a "swinging gate" guiding acetylcholine toward the active site. This can be seen in Figure 5, where it is shown the superposition of four different crystal structures of acetylcholinesterase (PDB codes 2ACE, 1EVE, 1VOT, and 1ACL) bound to four different ligands: acetylcholine, Aricept, huperzine, and decamethonium.

As well as examples of flexible single residues influencing ligand binding, there are also examples in which a group of flexible side chains play a part in interacting with inhibitors. In 2000 [115], once again, the crystal structure of AChE was solved, but this time from the species *Drosophila melanogaster*. The structure was solved at 2.7 Å with two potent reversible inhibitors bound, an iodo analogue and a tacrine analogue, with binding affinity K_i of 1.09 and 4 nM, respectively. The structure highlighted that the active site gorge is much narrower and shifted several angstroms compared with the *Torpedo californica* structure. They showed that nine aromatic side chains within the active site gorge change their conformation on inhibitor binding. In Figure 6, the comparison of the binding pocket between the unbound native structure of *Drosophila melanogaster* AchE (PDB code 1QON) and the structure bound to a tacrine analog (1DX4) shows that rearrangement of key aromatic residues is necessary for the ligand to bind the receptor.

the ceiling of the binding pocket. These flaps close once the inhibitor is bound. Bouzida et al. docked two flexible ligands into 10 rigid protein conformations of HIV protease complexed with specific inhibitors. There results showed that it would have been almost impossible to identify these inhibitors by ligand docking or virtual screening because out of the 10 crystal structures, one of the ligands scored well in only one crystal structure (the one in which it was crystallized), and the other ligand scored well in only four structure within 2-Å RMSD of its position in the crystal structure.

The study of DHFR highlighted another example of why flexibility needs to be incorporated into protein-docking protocols. DHFR is a widely studied enzyme because of its role in DNA synthesis; it is also an interesting protein for understanding protein flexibility because it undergoes ligand-induced conformational changes and is in two conformational states when unbound (a loop movement of more than 7 Å has been reported [118–120]). Obviously, with such fundamental ligand-induced and structural mobility structure-based drug design attempts, using rigid receptor techniques will be futile.

In a recent report, Cavasotto and Abagyan [121] performed extensive flexible ligand/grid receptor docking experiments against 33 crystal structures of four different protein kinase subfamilies: three serine/threonine protein kinases—the cAMP-dependent protein kinase (cAPK), the cyclin dependent kinase 2 (CDK-2), the mitogen activated protein P38 (P38)—and one tyrosine protein kinase—the lymphocyte-specific kinase (LCK). Protein kinases have been implicated in proliferation, invasion, and metastasis of many types of cancer (see Ref. [122] for a recent review on the therapeutic importance of protein kinases and the progress in developing new protein kinase inhibitors). Protein kinases are atractive test cases for the effect of receptor flexibility on ligand binding, as binding-pocket side chains may adopt different conformational states, and even in some cases loop rearrangements are observed. The ligand is partially solvent exposed and interacts with the receptor through many hydrophobic contacts. The fact that the common ATP binding sites are good drug targets has been rationalized in terms of the induced-fit effects in protein kinases together with hydrophobicity of the binding pocket [123]. Also, some ligand–protein interactions are mediated by water molecules. The authors performed a small-scale ligand docking against the protein kinase receptors. A 1000-compound library of random molecules seeded with the corresponding protein kinase cocrystalllized ligands was screened against each of the receptors, using the ICM flexible ligand/grid receptor docking algorithm [34, 95]. During the energy optimization of the ligand in the field of the receptor, a conformational set of low-energy states is generated, and the best energy conformation is scored. In this way, the reported RMSD values for ligands in flexible ligand/rigid receptor docking always refer to the best energy-ranked solution. The screening was repeated four times and the best score of each ligand was kept.

Native ligands were docked to the protein kinase structures with remarkable accuracy. The average RMSD is 0.74 Å, and 100% of the native ligands in the 29 holo complexes are docked within 1.5 Å RMSD (individual RMSD values for each receptor are detailed in Figs. 7 and 8). RMSD values were calculated between heavy atoms of the docked ligand with the ligand complexed in the crystal structure, after superposition of the backbone atoms within the ligand binding pocket.

The influence of the ligand-induced conformational changes on docking results becomes evident when a ligand is docked to a receptor complexed with another compound. Because sometimes small variations in the structure of the binding pocket might have a large effect in docking geometries, cross-docking experiments are useful to assess the magnitude of this influence. On average, only 70% of the ligands were cross-docked correctly in the protein kinase benchmark (detailed cross-docking matrices for CDK-2 and c-APK are detailed in Figs. 7 and 8).

In a very recent work, Erickson et al. [124] used four different docking methods, DOCK [19], FlexX [22], GOLD [24], and CDOCKER [125, 126], to examine the effect of several factors on ligand docking. On a first test set, composed of 41 ligand-protein complexes representing a diversity of size, flexibility, and polarity with respect to the ligand, the authors assessed the effect of force field parameters and ligand flexibility on the ability of the docking algorithm to reproduce the experimental pose (within 2 Å RMSD). A second test composed

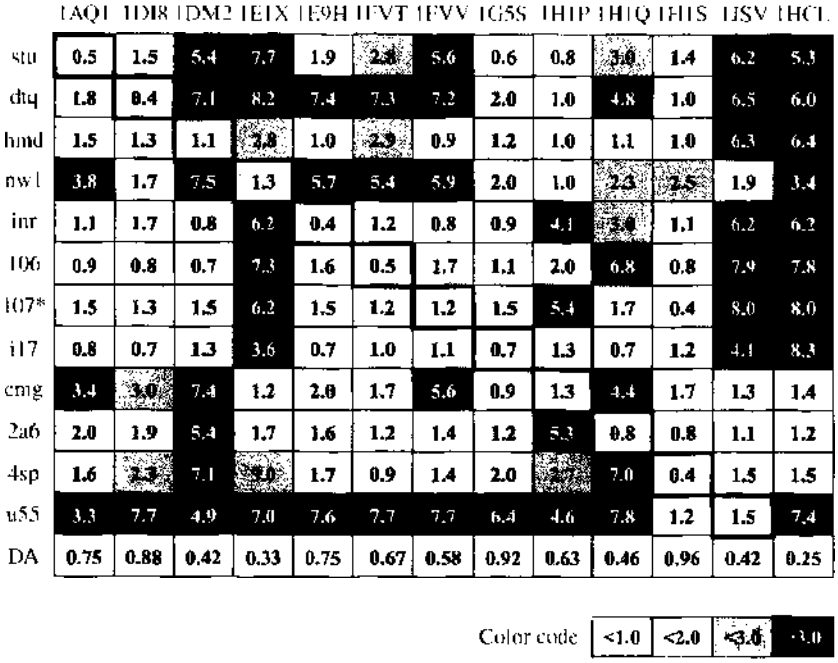


Figure 7. Cross-docking on CDK-2. RMSD (Å) for heavy atoms is calculated after superposition of the backbone atoms within the ligand binding pocket. RMSD values refer to the best energy solution. Thick border cells indicate native complex. DA: docking accuracy. The disordered and solvent-exposed benzyl moiety of ligand 107 was excluded in the RMSDs calculations. Reprinted with permission from [121], C. N. Cavasotto and R. A. Abagyan, *J. Mol. Biol.* 337, 209 (2004). © 2004, Elsevier.

of structures of trypsin, thrombin, and HIV-1 protease was used with CDOCKER to assess the effect of protein flexibility on the geometric prediction accuracy. For each protein, they considered a set of experimental structures bound to different ligands as well as the unbound (apo) structure. They docked ligands to their original structures, to the apo one, and to an "average" structure. The "average" structure was selected as the crystal structure whose binding site coordinates were closest to the average position with respect to the RMSD of all the structures. Adding more evidence to what has been already said, the authors found that docking accuracy falls dramatically, using the "average" or apo structure. In fact, the drop in accuracy correlates with the extent of protein flexibility on ligand binding.

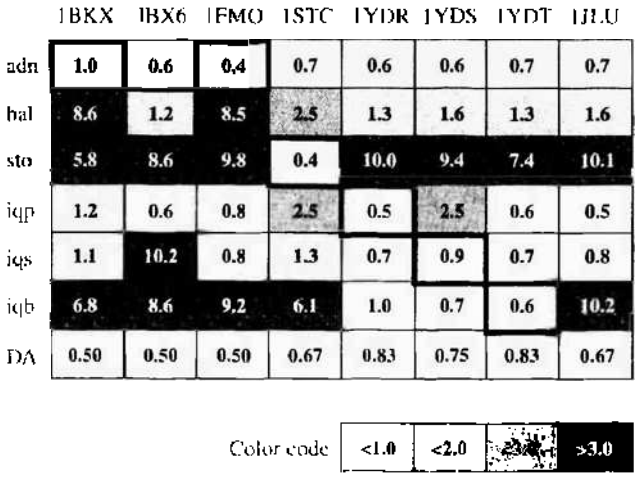


Figure 8. Cross-docking on cAPK. RMSD (Å) for heavy atoms is calculated after superposition of the backbone atoms within the ligand-binding pocket. RMSD values refer to the best energy solution. Thick border cells indicate native complex. DA: docking accuracy. Reprinted with permission from [121], C. N. Cavasotto and R. A. Abagyan, *J. Mol. Biol.* 337, 209 (2004). © 2004, Elsevier.

These examples emphasize the need for docking methods to incorporate flexibility of side chains and backbone within the binding pocket, and allowing protein mobility within a docking protocol will improve virtual screening and docking accuracy.

5. LIGAND DOCKING METHODS INCORPORATING PROTEIN FLEXIBILITY

Here we discuss a number of different docking methods that attempt to incorporate flexibility to a certain degree. The methods can be classed into the following categories: soft docking, allowing overlapping atoms; rotomeric libraries that allow rotatable bonds to sample possible conformations; and multiple receptor ensembles generated by nuclear magnetic resonance, x-ray crystallography, or computational means.

5.1. Soft Docking

Since 1990, various attempts have been made to incorporate protein flexibility into the ligand-docking procedure. The early methods used a soft-docking approach, which is one of the simplest ways to consider flexibility. Soft receptors are generated by overlapping atoms in the ligand-receptor complex and relaxing the high-energy penalty that would occur via this high van der Waals interaction. A soft docking method, in which the surface structures of the ligand and receptor are matched and implicit, minor conformational changes were allowed on the basis of complementarity in size and shape, close packing, and the absence of steric hindrance, was developed. This approach was first undertaken by Jiang et al. in 1991 [127] by incorporating three steps. First, develope a cube representation of the molecular surface and volume of the two interacting proteins. Second, dock by geometrically matching the surface representations, which allows conformational changes such as complementarity of size and shape, close packing, and the effects of steric hindrance. Third, screen complexes with energetically favorable interactions between the buried surface area.

In 1992, Walls and Sternberg used a soft potential for the docking of an antigen to an antibody [128], and a simple electrostatic model was used to remove clashes after docking. The docking method was undertaken on three x-ray crystallography structures: antibody-lysozyme complexes, HyHEL-10, D1.3, and HyHEL-5. For each complex, between 15 and 40 possible orientations of the ligand were identified. The RMSD in the binding patch between the docked and crystal structure complexes was between 1.9 and 4.0 Å.

Gschwend et al. concentrated on improving scoring and ranking of soft-docked complexes [129]. More recently, soft docking and improved scoring methods were included in an approach called SPECITOPE (the precursor to this algorithm was SLIDE [29]) to screen a peptidyl database [29]. Side-chain rotations were used to resolve clashes between the ligand and the receptor. The method of side-chain rotations is discussed further in the next section.

Another approach, using a combination of gradient minimization and short Monte Carlo runs on the most promising solutions, has been undertaken [130]. Solvation was implicitly taken into account when evaluating the structures with a continuum model. A study of ligand-induced receptor rearrangement of three systems was undertaken: progesterone and 5 β-androstane-3,17-dione to the antigen-binding fragment of a steroid-binding antibody, and Nα-2-naphthyl-sulfonyl-glycyl-D-*para*-amidino-phenyl-alanyl-piperidine to human α-thrombin. All three are challenging tests because of protein rearrangements on binding, but each was redocked to within <1.5 Å in the ligand atoms, taking the best minima in terms of free energy.

All these soft-docking methods produced a limited set of reasonable results, with the best ranking complexes within 1–3 Å of the native structures. The softening of geometric criteria used in these soft-docking approaches allows some overlap of the interacting surfaces. In general, each molecule is digitized into grids according to its volume and surface. A fast translation algorithm is then applied to two sets of volume and surface grids (ligand and receptor), and an exhaustive sampling of rotation and translation space is made. The softness of the method is implemented by varying the grid size and cone angle cutoff in calculating the local surface complementarity.

The main advantages of the soft-docking method is that it is less computationally expensive, and therefore faster, than more vigorous docking algorithms, and generally the initial docking algorithm does not need a lot of alteration to incorporate a degree of flexibility.

5.2. Sampling Side-Chain Conformers

Sampling side-chain conformations of the residues within the ligand-binding pocket is one popular approach for incorporating flexibility into the docking protocol. To reduce the complexity of the system, only a few degrees of freedom are chosen, such as rotations around single bonds. Selection of which torsional degrees of freedom to model is the critical step and is usually reliant on extensive structural data such as a crystal structure in two forms.

Partial side-chain flexibility was first implemented by Leach in 1994 [131]. Leach used a rotameric library in which amino acids were given a number of rotomeric states depending on their rigidity and size. The number of rotomeric states ranged from 0 for small amino acids, such as alanine, to 55 for acids such as arginine. These rotomeric states are then searched for the best energy against the ligand in the binding pocket. The method is based on two theorems, the first is called Dead End Elimination (DEE) [132], and the second's called A Star, or A* [133]. These two algorithms combined control the number of rotamers analyzed, and therefore allow a large set of residue side chains to be analyzed.

The DEE method eliminates rotamers that cannot contribute to the local energy minima according to some predetermined cutoff. Eliminated rotamers are classified as dead ending, and if enough rotamers can be removed, it is possible that the global minimum can be found. If the number of remaining rotamers is not too large, then convergence can be achieved. Using the A* method in combination with DEE allows a small subpopulation of structures to be determined. Leach and Lemon used this combined DEE and A* method, which was further developed to include energy cutoffs and enable the identification of the global minimum energy of all the side chains [134].

In addition, a rotomeric library of side chains has been used in combination with DEE to study the partial flexibility of the side chains in the binding pocket of HIV-protease mutants [135]. After a full search of the library, the ligand and receptor are minimized using a Monte Carlo optimization procedure. This method was successful, with the predicted complex for the mutant structures being ~1 Å different from the experimental crystal structures.

Jones et al. [136] used a genetic algorithm approach called GOLD to perform fully flexible ligand and partially flexible receptor docking. The method allows full rotation for hydrogens and lone pairs of hydrogen bond donors and acceptors. This allowed favorable hydrogen-bonding networks to be developed between the ligand and its receptor, with the fundamental requirement that the ligand must displace loosely bound water on binding. The main limitation of this work is that flexibility is confined to the terminal bonds, and a few problems were encountered when trying to dock hydrophobic ligands as a result of their reduced contribution to the hydrogen-bonding network. However, GOLD performed very well with hydrophilic ligands. In 1997, additions to this algorithm were reported and the program was tested on 100 ligand–receptor complexes from the PDB. The GOLD program was successful within certain defined limits at accurately redocking the ligand in 71 of the complexes [24].

A method has also been applied to simulating the binding of several viral peptides to murine major histocompatibility complex class I H-2Kb and to calmodulin and an 18-residue-long helical peptide from calmodulin-dependent protein kinase II alpha [137]. The method allows partial flexibility of side-chains by applying a combinatorial method that permits all single bonds of the ligand and those in the receptor-binding pocket to be fully rotated.

A docking program called SPECITOPE [29] (the precursor to this algorithm was SLIDE [29]) uses side-chain rotations to resolve docking clashes between the ligand and receptor. As discussed in the previous section, this algorithm uses a soft-docking initial approach that resolves clashes between the ligand and the receptor. If a clash is identified, the side chains of the ligand are rotated through the minimal angle that resolves the clash. The single bond closest to the clash is rotated first, and if this fails to resolve the clash, then the next

rotatable bond closer to the ligand backbone is rotated. If rotation of ligand side chains is unsuccessful, then the same procedure is undertaken on the receptor. Although this method resolves side-chain clashes, it does not attempt to find the minimum energy conformation. The method was tested on four distinct targets: a serine protease, a DNA repair enzyme, an aspartic proteinase, and a glycosyltransferase.

A docking method called the Mining Minima optimizer allows any number of rotatatable bonds to be treated as continuous degrees of freedom [138, 139]. The residue side chains to be treated as flexible can be user defined, and the energies of the flexible side chains are optimized, along with the ligand. The algorithm was tested on the hypothetical protein YecO (HI0319) from *Haemophilus influenzae*, which showed that by allowing side-chain relaxation, candidate substrates of various sizes can be accommodated into the ligand-binding pocket.

Two algorithms called SOFT SPOTS and PLASTIC were developed by Anderson et al. in 2001 [140]. The authors studied the enzyme thymidylate synthase, which is known to undergo significant ligand-induced conformational changes when an inhibitor binds. Conventional rigid docking algorithms failed to accurately dock two known inhibitors, BW1843U89 and CB3717. The program SOFT SPOTS identified regions in which rotation of side chains may be necessary to accurately dock the inhibitor. This prediction was based on the role of hydrophobic residues in conformational changes, as seen in enzymes such as thymidylate synthase. The algorithm PLASTIC is then used to vary the rotomers of the side-chains identified by SOFT SPOTS. This algorithm incorporated two structures of thymidylate synthase in different forms. This technique of considering multiple structures to incorporate flexibility is described in the next section.

The internal coordinate mechanics (ICM) global optimization method has also been used to incorporate side-chain and backbone flexibility for the analysis of the structural binding determinants of a retinoid X receptor (RXR) antagonist to the receptor [141–143]. Compound SR11179 (see Fig. 9) functions as an antagonist. Flexible-ligand-flexible-receptor docking to the antagonist-bound conformation of the human nuclear receptor Retinoid X Receptor α (hRXRα), PDB code 1DKF, explains the binding determinants of this ligand, although originally it was assumed that its antagonist behavior could be explained in terms of clashes with the side chains of the receptor in the agonist-bound conformation. It was interesting, therefore, that SR11179 could be docked into the hRXRα-binding pocket in the agonist-bound conformation (PDB codes 1FM6 and 1FBY) with no clashes, but with a diminished electrostatic interaction between the carboxylate moiety and arginine 316 and less hydrophobic contacts. Thus, this subtle energetic difference in binding SR11179 to one or the other conformation might explain its antagonist preference.

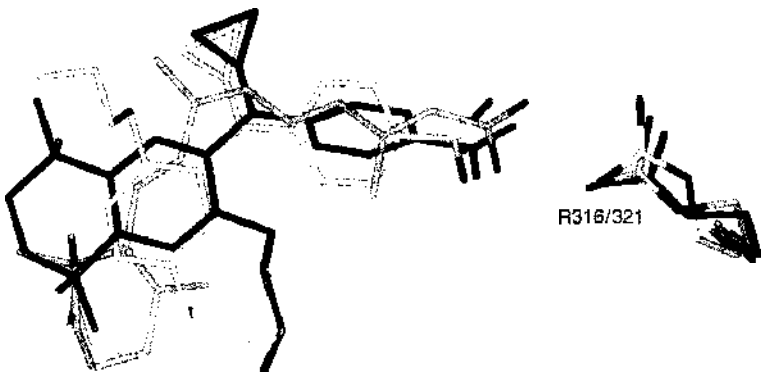


Figure 9. Superposition of ligand SR11179 within the hRXRα ligand binding pocket in agonist- (light gray) and antagonist-bound (dark gray) conformation. As found in crystal structure 1FBY (agonist bound), 9-cis retinoic acid is displayed in yellow. Arginine 316/321 (agonist/antagonist numering) side-chain-forming salt bridges with the carboxylate moiety of the ligands is also displayed. SR11179 in the antagonist bound conformation makes a stronger salt bridge with arginine than in the agonist-bound conformation, which may explain why SR11179 functions as an antagonist, even though SR11179 makes no clashes within the binding pocket in the agonist-bound conformation and its scaffold superimposes well with high-affinity agonist 9-cis retinoic acid. Oxygens are displayed in red, and nitrogens in blue. Hydrogens are omitted for the sake of clarity.

Using the same methodology, a similar explanation was found to explain the lack of agonist activity of SR11179 toward the human retinoid acid receptor γ (hRARγ). The salt bridge with arginine 276 is less strong than that of trans-retinoic acid, the native ligand.

More recently, Frimurer et al. [144] used a rotamer library of four key residues to improve predictions of binding geometry and affinities in protein tyrosine phosphatase 1B (PTP1B). The library was built based on observations on three crystal structures of PTP1B bound to different inhibitors and resulted in 96 models. Docking of the three inhibitors to these models using FlexX (see Section 2.2.3) improved their geometry and binding-energy predictions. The method was able to successfully discriminate between known correct and incorrect ligand poses, and the energy ranking of the conformations correlated well with observed binding affinities. However, because of the combinatorial explosion of binding pocket models thus generated, this method does not seem the best to be used in a virtual screening task, where hundreds of thousand compounds should be docked and scored in an affordable computing time.

The concept of induced fit has also been modeled by incorporating a hinge-bending concept that allows all angular rotations while the receptor automatically closes onto the binding ligand by matching surfaces. Atoms on either the ligand or the receptor define hinge points and, depending on the ligand–protein complex, can be located on loops, domains, or subdomains, or between any flexible part of a protein [145, 146].

A side-chain sampling method has recently been used to simulate the problems associated with modeling G-protein-coupled receptors. As discussed earlier, modeling ligand–receptor interactions of GPCRs is critical because GPCRs are a major drug target, influencing a wide range of diseases. However, there is a lack of structural information, with only one structure of a GPCR being solved to date and very low sequence homology between other members of the family. Therefore, attempting any molecular modeling study of a GPCR is a very challenging exercise, and so a technique that uses fully flexible receptor–ligand docking would enable a clearer picture of a GPCR to be developed.

A global energy optimization procedure was undertaken on the GPCR bovine rhodopsin (Rho), where a fully flexible ligand was docked into a fully flexible receptor. The side-chain torsional angles of 21 residues surrounding the retinal-binding pocket of Rho were randomized and two different receptor conformations A and B were generated (see Fig. 10). This process was undertaken to mimic a GPCR model in which very low sequence homology with respect to the template of the pocket, which is a very realistic scenario when attempting to model ligand–receptor interactions in GPCRs. The three rotational angle variables of the Rho retinal ligand were also randomized.

The optimization procedure was implemented by ICM [25], consisting of four steps, as described earlier. The conformational sampling is based on the BPMC procedure [35, 36], which randomly selects a variable in the internal coordinate space [25] and then makes a step to a new random position independent of the previous one, but according to a predefined continuous probability distribution. It has also been shown that after each random step, full local minimization greatly improves the efficiency of the procedure [37, 38]. However, because some energy terms might have no derivatives or might be very expensive to compute, a double-energy Monte Carlo minimization scheme [37, 38] circumvents these obstacles by minimizing the energy with respect to the differentiable terms but calculating the full energy, also using the nondifferentiable terms. This double-energy scheme allows for the incorporation of complex energy terms, such as surface-based solvation energy, into the global optimization process.

This is a global energy optimization procedure that has as many as 30–50 essential variables and that therefore is very computationally challenging. This sort of procedure is key for predicting the binding mode and side-chain rearrangement for new ligands to the same GPCR or any ligands to a homology model, provided the backbone changes are small. This procedure was able to predict the correct conformation of retinal in the Rho binding pocket very accurately. The RMSD for ligand heavy atoms in both A and B pockets was 0.19 Å, and the RMSD for the binding-pocket side-chains for A and B was 0.14 and 0.19 Å, respectively.

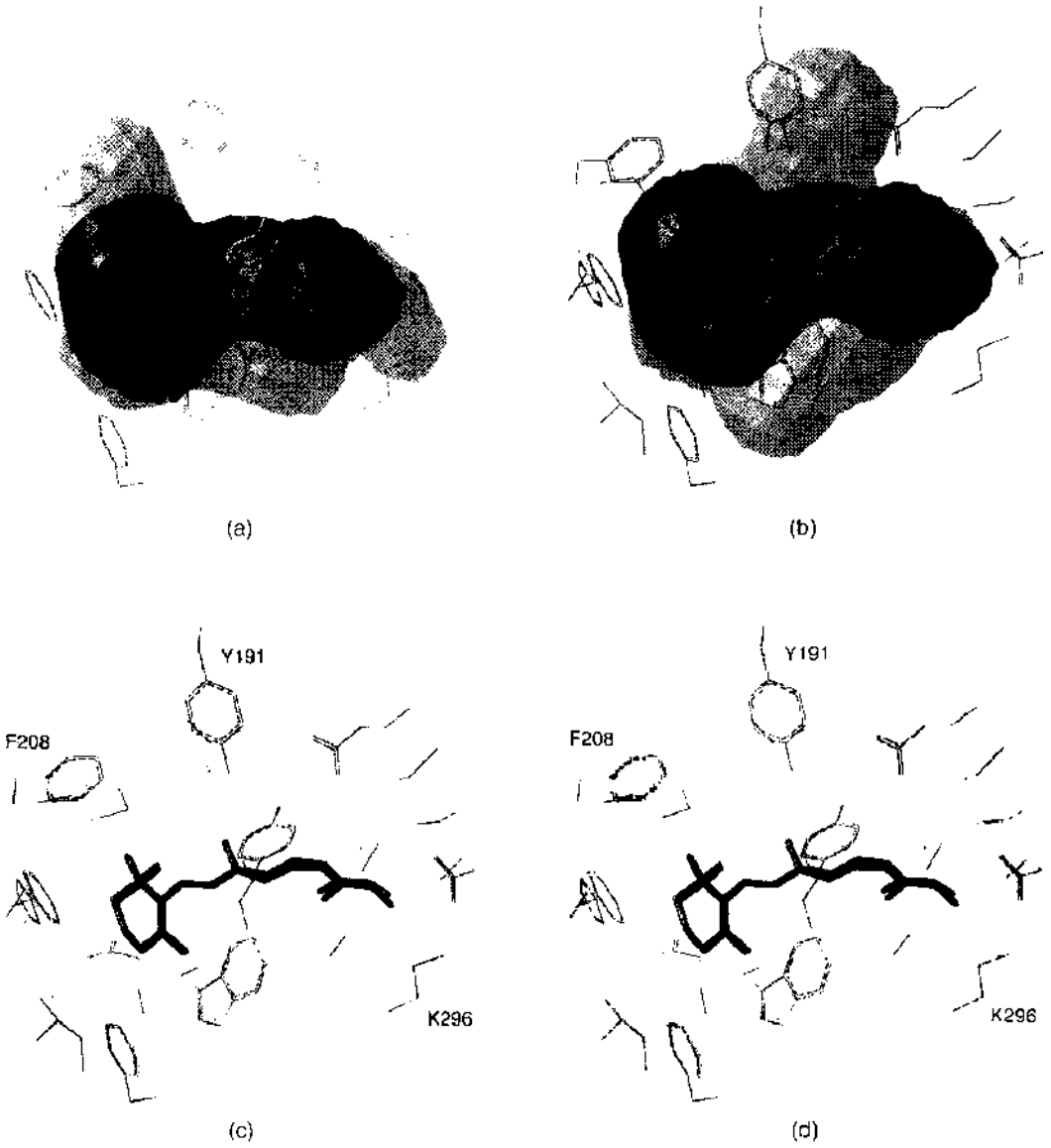


Figure 10. The shape of the binding pockets (yellow) and the random positions of the side chains (yellow) for the two different starting conformations A and B before unbiased global energy optimization are shown in parts (a) and (b). The position of the side chains and ligand for the two conformations A and B after global energy optimization are shown in parts (c) and (d). The crystal structure binding pocket, ligand, and side chains are colored blue in all parts of this figure (pictures generated using ICM). Reprinted with permission from [45], C. N. Cavasotto et al., *Proteins-Struct. Funct. Genet.* 51, 423 (2003). © 2003, Wiley.

5.3. Use of Multiple-Receptor Conformations

More recently, theories about protein conformational states have moved on beyond Koshland's [6, 7] 1958 theory of induced-fit, and the use of soft-docking and rotameric side chains. The binding process is now thought of as an ensemble of a selection of energetically favorable substates, as has been described in Section 3.1. All the methods previously described concentrate on small, simple side-chain rearrangements; however, recently, new methods are being developed to incorporate not only side-chain flexibility but backbone as well.

In two excellent reviews [147, 148], the concept of combining multiple protein structures to incorporate protein flexibility was analyzed. Carlson [148] pointed out that these methods have prompted many important questions, such as, Which should be the source of these structures, MD, MC, x-ray crystallography, or NMR? How many ensemble structures are needed? How should the structures be used or results be combined? How should this flexibility be incorporated in the virtual screening procedure?

Here we describe the methods that have been applied so far for generating multiple conformations to incorporate protein flexibility. The methods fall into two main categories: computationally and experimentally generated ensemble of structures. This field of research is very new, and each method has its own merits and disadvantages and has been used to varying degrees of success.

5.3.1. Generating an Ensemble of Receptor Structures by Computational Means

A popular and potentially accurate method for generating a subensemble of states is by using computational simulations. This approach can be very computationally expensive, but recent advances in computational speed and knowledge-based approximations are making these methods a more realistic approach for generating an ensemble of conformations.

There are two main ways to generate an ensemble of structures: Monte Carlo and molecular dynamics (MD). Both Monte Carlo and MD computational methods use a force field to generate a simulation of protein flexibility. A Monte Carlo method randomly generates conformations of a receptor by sampling conformational space, whereas a Monte Carlo method uses the laws of classical mechanics to simulate the motion of the atoms in the protein.

Pang and Kozikowski [149] were the first to generate a set of protein conformations by MD for use in docking studies. They generated 69 conformations of an acetylcholinesterase from a 40-ps MD simulation and docked the Huperizine A ligand to each of the conformations. From this study, there were able to identify a preferential catalytic binding mode.

A simplified MD simulation has been described using an implicit solvation model [31]. The ligand and active site region is fully flexible, whereas the remaining parts of the receptor are fixed. The problems that may occur as a result of the active site being flexible and the rest of the protein being rigid is in part overcome by using a buffer region that incorporates a small harmonic potential. They tested their method with the benzamidine ligand and trypsin, 100 independent simulations were calculated, and many of the solutions matched the low-energy conformation observed in the crystal structure of the complex. This method incorporated a rapid simulation but was limited in its conformational sampling.

Conformational sampling was enhanced by Mangoni et al. [56] by studying MD simulations at a higher temperature, and therefore a higher velocity. The docking was performed between phosphocholine and an immunoglobulin in explicit water, and the center of mass of the ligand was separated from its internal and rotational motions by coupling it to different temperatures.

A method called multiple-copy simultaneous search (MCSS) [150], in collaboration with locally enhanced sampling (LES), has been applied to multiple protein conformers generated by MD. MCSS uses quenched MD and finds functional groups that favorably bind to regions of a binding pocket. Energetically favorable positions are extended to include flexibility within the ligand-binding pocket, and these favorable groups can be extended by connecting building blocks. The method was tested on HIV-1 protease and was found to be more accurate than the previous uses of MCSS as a result of the incorporation of LES.

A novel approach to ligand docking using MD was described by Given and Gilson in 1998 [151]. The method does not aim directly for the global energy minimum, but instead explores the position of the ligand in a binding site, thoroughly producing a hierarchical ensemble of local-minima conformations. The method treats all ligand degrees of freedom as flexible and can be adapted to incorporate receptor flexibility.

In 2000, a new, very successful "dynamic" approach was developed for incorporating an ensemble of MD structures to form pharmacophore models of a receptor [152]. This approach identified ligands that would fit into the binding pocket of HIV-1 integrase, causing the least disruption to the shape and flexibility of the binding site, and therefore reducing any entropic penalties of binding brought about by the protein. The method used a series of snapshots generated by an MD simulation in an explicit solvent molecule environment to represent the flexible nature of the protein without a ligand bound. A process called multiunit search for interacting conformers (MUSIC) is used to search and probe the surface of each MD snapshot (with the solvent and counterions removed) for binding regions of

different functional groups. The MUSIC procedure performs multiple MC simulations, followed by a clustering of the probes that define the ideal position of a particular functional group within the binding pocket. The results of this analysis for each MD snapshot are then overlaid to form a consensus for particular functional groups in each region of the binding pocket. Using this approach, several new inhibitors were identified for HIV-1 integrase from the available chemical database (ACD).

MD simulations have also been used to generate an ensemble of different receptor conformations as input for the generation of a composite interaction weight-averaged grid [153]. This method was applied to virtual screening against dihydrofolate reductase and found improvements in the top-ranked 10% of a database of druglike molecules. This will be discussed in more detail in the section describing virtual screening and flexibility.

To speed up an MD simulation for the sampling of peptides, Nakajima et al. [154] used the multicanonical method. This method was originally developed for MC [155] simulations and keeps the temperature constant during an MD simulation on a deformed potential energy surface. Using a five-residue peptide, metenkephalin, it was found that this method sampled more conformational space than conventional MD.

Another method for enhancing the sampling rate during a docking simulation has been described by Pak and Wang, using the Tsallis transformation [156, 157]. They initially used the Tsallis transformation to fold a 16-residue peptide and noticed a significant improvement in conformational searching efficiency. This transformation was then applied to the docking problems of streptavidin/biotin and protein kinase C/phorbol-13-acetate. The method allows partial flexibility in the receptor side-chain binding pocket. The Tsallis transformation is applied to the nonbonded interaction potential of the chemistry at Harvard macromolecular mechanics (CHARMM) [54] force field and dynamic simulations are run with infrequent q-jumping and q-relaxation between the normal and smooth energy.

A multiple-copy ligand stochastic method for improving the efficiency of sampling during MD docking simulations has also been described [158]. In this method, called F-DycoBlock, the protein moves in the mean field of the multiple copies of the ligand, incorporating four different types of receptor flexibility.

The relaxed complex method [159, 160] uses an ensemble of structures generated through MD simulations of the receptor (without ligand) to dock a minilibrary of binders using a fast, rigid receptor docking method. The goal is to search for the correct ligand–receptor conformation. The rapid-docking procedure is used as a preliminary filter, and the selected complex conformations are then reevaluated with a more accurate energy function, using the Molecular Mechanics/Poisson Boltzmann Surface Area approach. The authors observed the experimental conformation within the lowest free-energy states.

A procedure to generate an ensemble of receptor conformations through flexible-ligand-flexibile-receptor docking has been recently presented [121]. This algorithm to generate multiple-receptor conformation and its use in docking and virtual screening is described in detail in Section 6.3.

So far in this section, we have only discussed MD methods for incorporating flexibility. As mentioned briefly in the previous section, Caflisch et al. [49] have described an MC method for finding the binding mode of the blocked tetrapeptide N-acetyl-Leu-Pro-Phe-methylamide to the FK506 binding protein (FKBP). The authors used the standard method of MC, involving random displacements of the substrate with minimization of the complex but incorporated a rapid method for calculating polar and nonpolar solvation effects, using a finite-difference Poisson–Boltzmann and weighted solvent accessible area terms. The CHARMM [54] force field was used for energy minimization, and the Lennard–Jones and coulombic potentials are softened and increased gradually during each iteration of 1000 seed structures. The Monte Carlo method is then applied to the 20 structures with the lowest energy.

As described in Section 2.2.1, the Monte Carlo docking algorithm called PRODOCK [20] has been developed with a grid-based energy evaluation technique, using Bezier spines to increase efficiency.

MD methods have been successful, but a study by Philippopoulos and Lim [161] indicates that an ensemble of structures derived by NMR or x-ray crystallography is preferable to

an ensemble of MD or Monte Carlo structures. There are a number of limitations using NMR and x-ray structures, and with advances in computer speed, both approaches have the potential to be accurate in the prediction of flexibility on ligand binding.

With the speed of computers improving every year and the development of parallelization techniques getting ever better, the use of computationally derived ensemble models should not be overlooked. They are an accurate, cheap, and fast method for identifying an ensemble of structures to use in protein docking to incorporate flexibility within the pocket.

5.3.2. An Ensemble of Receptor Structures Generated by Nuclear Magnetic Resonance and X-ray Crystallography

X-ray crystallography and NMR are the two ways to obtain atomic data about the structure of a protein. However, each method provides different insights into the structure of a protein as a result of NMR being undertaken with the protein in solution, and x-ray in a solid form determined by the nature of the conditions in which the protein was crystallized.

In general, the result of an NMR spectroscopy structure determination of a receptor is not one single structure but a set of multiple structures all in agreement with J-coupling data and the nuclear Overhauser effect. This set of structures may provide a more accurate representation of the receptor than, for example, a crystal structure that represents a snapshot of the receptor in the state in which it was crystallized. However, a set of multiple x-ray structures in a variety of different forms could be equally as useful as an ensemble of structures generated by NMR.

It is hypothesized that at least part of the conformational variability in the ensemble of structures solved by NMR is a result of the flexible nature of a protein. Comparison studies between NMR and x-ray crystallographic structures seem to give evidence to this [162, 163]. Structures solved by these two techniques usually differ in the mobile and flexible loop regions and surface residues.

As mentioned in the previous section, Philippopoulos and Lim [161] report that an ensemble of structures derived by NMR or X-ray crystallography is preferable to an ensemble of MD or Monte Carlo structures. The authors report that NMR structures sample more conformational space than computer simulations and that NMR structures highlight more flexibility than MD in both the side chains and the backbone. A comparison was made with the thermal factors in the x-ray structures that correlate well with the NMR flexibility. It was therefore concluded that NMR structures would be useful for ensemble structure-based docking.

In 1997, Knegtel et al. [164] developed two averaging methods for docking a ligand into an ensemble of NMR structures. The method develops an average of all the receptor conformations in the NMR ensemble and docks to that average rather than exploring new potential flexible conformations. The first method, referred to as an "energy-weighted average," combines the conformational variability from the ensemble into an average based on the interaction energy between a ligand and each of the structures in the ensemble. The second method produces an average of the force-field score. DOCK 3.5 (see Section 2.2.3) was used for the docking process, and weights were developed that were stored on a scoring grid. The weights allowed the weighting to meet two criteria: first, ligand should score well when docked to its native receptor structure and should also score well against the ensemble; second, the weighting should limit the amount of time it takes to dock one ligand and therefore enable a large database to be screened. The method was tested on five HIV protease inhibitor crystal structures, five ligand-bound crystal structures of ras p21, five cocrystals of bovine retinal binding protein, and 25 NMR structures of polychlorobiphenyl metablolite protein. Using these methods, the known binding energies and ligand pose can be reproduced accurately. Knegtel et al. note that there are obvious limitations in using a small set of conformational states but add that just a small set of states is going to contain more information on potential flexibility than just one structure. Other grid-averaging methods have been developed based on this method and are discussed in Section 6.1 because they enable virtual screening to be more efficient by docking large databases to one average structure rather than multiple structures.

Sudbeck et al. [165] described a method for qualitatively evaluating the binding site of HIV reverse transcriptase by using multiple structures. They superimposed the Cα backbone of the least-flexible region of nine inhibitor complexes. This allowed a picture of the general flexibility within the binding pocket to be viewed and provided a reference point for qualitatively evaluating predicted conformations.

In 2000, Kastenholz et al. [166] developed an addition to the GRID algorithm that enables multiple conformations of a receptor to be included in the algorithm. The algorithm can be primarily used for identifying selective regions within a binding pocket, and thereby focusing drug design projects. The algorithm incorporates a consensus principle component analysis (CPCA) to analyze grids of multiple protein structures. From this analysis, contour plots are produced that can be studied to enable a compound that may produce specificity to the receptor to be designed. The plots highlight not only the region in the binding pocket where interactions take place, but also the type of interaction. The algorithm was tested on the x-ray crystal structures of three serine proteases of the chymotrypsin family: thrombin, factor Xa, and trypsin. Overall, the binding pockets of each of these enzymes are very similar; however, this algorithm was able to visualize the subtle differences that relate to the specificity of their ligands.

The docking algorithm FlexE [63] is a useful tool for analyzing an ensemble of structures because, unlike a lot of algorithms using ensemble approaches, it can take into account receptor and ligand flexibility at the same time and does not rely on any postdocking optimization of the ligand. First, the program creates a superposition of all the structures in the ensemble, and an average of the nonvariable parts of the structures such as the backbone regions and some rigid side chains is made. Each ensemble contains the definition of the protein atoms, the resolution of any potential alternative positions of atoms based on the PDB file, and the definition of the active site. All atoms are selected within 6.5 Å of any part of the ligand within the binding pocket. The water molecules are removed, the side chains of lysine and arginine are protonated, and the carboxylate groups of aspartic and glutamic acid are ionized. The remaining flexible regions are then sampled by bond rotation, as described in the sampling side-chain conformers section of this chapter (Section 5.2). FlexE is able to perform very well when there is side-chain flexibility within a receptor and in some cases when there are variations within the loop. However, any larger movements such as backbone or domain movements are not predicted as accurately.

FlexE has been tested on 10 ensembles of proteins containing 105 crystal structures from the PDB and one homology model of aldose reductase. Sixty of the structures have ligands that are used to judge the correct binding modes. The test set has very similar backbone structures but different conformations as a result of point mutations and differences within side chains and loops regions. The analysis showed that for 50 of the ligands, 83% were docked within 2 Å RMSD of the correct conformation and recombined to a new protein conformation that fits the ligand. The ligands also demonstrated the need for flexibility within docking, using the aldose reductase as an example. If flexibility were not considered, it would be impossible to identify all potential inhibitors to this receptor.

Another program, named AUTODOCK 3.0 [18], has been used to treat the problem of flexibility in the receptor by using multiple structures. Four different methods for combining an ensemble of proteins were tested on complexes of 21 peptidomimetic inhibitors of HIV-1 protease. In this test set, many of the structures show motion in an arginine residue that is critical for ligand binding. Two weighted average methods were developed and performed well. They were effective for the incorporation of side-chain motion into docking simulations and were also successful in a case in which structural waters showed heterogeneity in different complexes. One of the two weighted-average methods applied was an adaptation of the method described by Knegtel et al. [164], in which a weighted average of all the energies are computed at each point. The least successful method was a method that took a mean of the grid at each point and then took the overall minimum.

The main limitations of the experimental generation of an ensemble of structures is that NMR has a constraint on the size of structures that can be solved, and x-ray crystallography has drawbacks in determining the structures of proteins such as membrane proteins, and sometimes crystallizing in many conformational states can be problematic.

The evidence is clear (see Section 5.3.1): An ensemble of structures generated by either computational means or NMR is more advantageous than a snapshot image produced by a single crystal structure of a receptor.

5.4. Normal Mode Analysis

Collective degrees of freedom determined by normal mode analysis can be used to represent protein flexibility and study the dynamics of proteins. The normal modes are the outcome of the diagonalization of the second-derivative of the energy (Hessian) using an atom mass metric matrix. The eigenvectors provide the preferred direction of collective motion, whereas the associated eigenvalues are related to the extent of the energy change on deforming the structure following the eigenvector. If the energy function were harmonic, any energy change could be exactly expressed as a combination of normal modes. Even though this is not the actual case, small perturbations near equilibrium can be accurately described by normal modes.

Protein flexibility can be represented by low-frequency normal modes (excluding the six modes with zero eigenvalue, which characterize rigid motion), which characterize protein large-scale vibrations (they are also called soft degrees of freedom, as they induce a small energy change on perturbation along those modes). High-frequency modes are characteristic of small movements around the equilibrium position (hard degrees of freedom; a small displacement may induce a large energy change).

The excellent correlation between normal mode analysis and functional movements has been described [167], and recently a comparison between protein motions and those analyzed with normal mode analysis has been published [168].

The advantage of using normal mode analysis for ligand docking is that the flexibility of the receptor can be represented by a few low-frequency modes. This approach has been applied to model the binding of a Hoechst 33258 analog to DNA [169]. This is an interesting example, as comparison of the crystal structures of unbound DNA with DNA bound to a Hoechst 33258 analog shows small but important conformational changes in the minor groove-binding site. The authors used a DNA model limited to the central six base pairs [d(GAATTC)$_2$], and solvation effects were taken into account by scaling the atomic partial charges. The ligand was kept rigid (in the conformation found in the experimental complexed DNA) during all energy minimization procedures, and the authors used only a van der Waals interaction term (Lennard–Jones 6–12) to score ligand-binding position, consistent with the rigid ligand approach. They also showed that the conformational difference between the free and bound DNA structures can be described in terms of 5–40 low-frequency normal modes (depending on the model for partial charge, improper torsion, and angle force constants scaling). Then, using the structure of free DNA and placing the ligand at the position found in the ligand-bound DNA crystal structure, they found that relaxation of the complex in a subset of low-frequency normal modes significantly improved the fit of the ligand into the pocket. Because of small changes in bond lengths and angles caused by normal mode relaxation, full-atom energy minimization on the receptor should be performed to restore covalent geometry. Naturally, this problem could be at least reduced by performing normal mode analysis in internal coordinates [170, 171]. The authors also verified that, although normal mode relaxation leads to improved geometries for the misplacing of the ligand, the observed experimental binding pose could still be identified as the one with correct fit. This is an important contribution because it shows that normal modes could be used to represent receptor flexibility in ligand docking.

A method called low-mode conformational search procedure (LMOD) [172] complemented with conformational searching (c-LMOD) was used to study ligand docking to purine nucleoside phosphorylase (PNP) [173]. An improved version of LMOD, called L-LMOD, was used in the loop optimization problem. The authors optimized the position of the glycine-rich loop of the homology model of the *c-jun* N-terminal kinase JNK3, using other three-protein kinases as templates the P38 MAP kinase, ERK2, and HCL [174]. The authors also argued that the new structure compared well to the x-ray one. They also pointed out and illustrated a very important point regarding the dangers of overrefinement of the modeled structures.

The LMOD method was also used in the case of HIV integrase as a way to incorporate receptor flexibility in docking [175]. In an attempt to validate their procedure, the authors

tried to reproduce the experimental geometry of 5-CITEP bound to HIV integrase (PDB code 1QS4), where the important loop Ile141-Asn144 is missing in the structure. They found the pose of the ligand with least RMSD (0.39 Å) in the fifth-lowest energy conformation, 2 kcal/mol above the global minimum. However, some calculated distances of key contacts between receptor and ligand show differences of ~2 Å when compared to the experimental ones. Then, they applied the same methodology to predict the binding mode of low-micromolar inhibitor L-731,988, indicating that 5-CITEP and L-731,988 bind in a different way, and that the *in vivo* activity of L-731,988 might be caused by a better protein kinase profile. Although the methodology and the achievements are excellent, at least in the test case, because of computing time, this approach clearly can be used only in binding-site modeling or refinement and ligand geometry prediction, but not in a large-scale ligand docking and screening of virtual chemical libraries.

Very recently, a novel methodology to use normal mode analysis as a way to incorporate receptor flexibility in ligand docking and virtual screening was presented [176]. Assuming that the use of multiple receptor conformations is the best approach to incorporate protein fluctuations in ligand docking and scoring, an ensemble of representative receptor backbone conformations of the binding pocket was generated based on normal mode analysis. Each ensemble conformation is then complexed with known binders spanning different chemotypes, and the binding-pocket side chains + ligand are subjected to a stochastic global energy optimization using ICM. Complexes are ranked by energy, including terms for entropy and solvation, and the lowest energy complex per ligand is used as a starting point to perform ligand docking and virtual screening. It is observed that in the cyclin AMP kinase (c-APK), there is a ligand-induced displacement of the glycine-rich loop. For example, comparison of c-APK crystal structures 1FMO (bound to adenosine) and 1BX6 (bound to balanol) shows that balanol can not dock into 1FMO because of clashes with the glycine-rich loop. Using the outlined methodology, the authors were able to dock balanol into normal-mode-generated conformations of 1FMO, discriminating the correct binding geometry from misdocked ones. A small-scale virtual screening was also performed on 1FMO and the generated balanol-bound structure, using a random virtual library of 1000 compounds seeded with c-APK known binders. The improvements in the enrichment factors were comparable to what has been already reported [121].

6. PROTEIN MOBILITY AND VIRTUAL SCREENING

6.1. Grid Averaging

The biggest challenge of incorporating flexibility into docking predictions is to make the method computationally viable. For many ensemble methods, the time constraints involved for docking a large, compound database is very large. Therefore, a method that can reduce an ensemble of structures into one structure would be very beneficial for virtual screening.

As described in Section 5.3.2, Knegtel et al. [164] used grid averaging to combine multiple receptors into one. This approach has been adapted and analyzed recently to be incorporated into virtual screening of large compound databases.

In 2000, Broughton [153] described a method for incorporating protein flexibility within a docking protocol that would enable. Large databases of compounds to be screened by virtual screening. The method was tested on the highly refined crystal structure of DHFR with water molecules, using a database of 7000 compounds. The method would also enable accurate prediction of rank-order activity in a homologous series. This was tested on cyclooxygenase-2 (COX-2), which has critical flexibility within certain regions of the active site. Also, it is known that inhibitors binding to this protein are sensitive to this flexibility.

The FLOG [23] docking method was adapted to incorporate flexibility into the receptor and allow for water mobility. Rather than dock ligands to multiple receptors to incorporate flexibility, which as discussed earlier, is very impractical for the large databases used in virtual screening, Broughton used statistically weighted descriptors of the target active site to average the multiple structures. Although this ensemble approach is similar to that of Knegtel et al. [164] it differs in two ways: first, ensemble generation, and second, ensemble averaging.

MD was used to generate the ensemble, which is important because many pharmacologically important receptors only have one or very few crystal structures, and in theory, MD should explore a wider conformational space. An average model was constructed on the basis of statistically weighting the mean potential protein–ligand interaction energies computed from the MD run.

Each grid point on each of the receptor conformations generated by MD was specified by four indices, and the following values for each point in each of the structures were calculated: minimum value, maximum value, mean value, standard deviation, and a weighted mean.

Using the average and weighted average methods, the number of known DHFR ligands found in the top-ranked 10% of the screened database was improved from the standard nonflexible receptor approach. Also, the method was able to rank a homologous series of inhibitors in a manner that correlates to their binding activity.

In 2002, a study was undertaken to assess the most accurate and computationally efficient means of combining several 3D structures into a single docking simulation [18]. The single combined grid should represent the range and motion of the side chains and loops, hence incorporating ligand binding site flexibility. Four different approaches were analyzed to determine whether a single representative grid of interaction energies can represent the inherent flexibility of the HIV-1 protease binding pocket.

The test set consisted of 21 diverse peptidomimetic inhibitors of HIV-1 protease. Of these 21 complexes, 20 rely on a structural water for stability, and many of the structures show motion in an arginine residue that is critical for ligand binding.

Here is a summary of the four methods for combining the interaction energy grids and the way they performed with the HIV-1 protease test set. A "mean" grid method is one that takes a point-by point average across the grid. Each point within the grid is averaged. The main drawback of this method is that the repulsive areas are enhanced. In the test, this method performed poorly compared to the other three methods. The method predicted poor binding energy, and the RMSD between the crystal and docked structures was unacceptably high. A "minimum" grid method takes the minimum value at each point in the grid. The main drawback of this method is that the favorable regions are overestimated, and therefore, ligand specificity becomes an issue. However, this method overcomes the problems associated with repulsive areas dominating and performed better overall than the "mean" grid method. A method called "clamped grid" is based on the Knegtel et al. [164] method described earlier. This method is a weighted average of all the energies across all the structures at each point of the grid. A threshold energy value is decided on, and each point of the grid is weighted on this value. For example, if the interaction energy of a point is less than a given threshold, then the structure is given a value of 1. If the energy is greater than the threshold value, then a very small weight is applied. An average is then taken, and the weights are normalized to 1. The main drawback is that the values are artificially weighted and the determination of the threshold is critical. This method performed well with low RMSD between experimental and docked and good binding energies.

A method called an energy-weighted grid is similar to the clamped grid method but uses a Boltzmann assumption to calculate the weight. Hence, overcoming the arbitary nature of the "clamped grid" method. This method also performed well and performed the best along with the "clamped grid" method for generating good RMSD and binding-energy values.

The main advantages of being able to successfully incorporate many grids into one map is that it lowers the computational time required to dock and therefore enables the method to be used for high-throughput virtual screening. Both the "energy-weighted" and "clamped-grid" methods successfully predicted the multiple conformations of Arg8, which is critical for ligand binding, and they were able to incorporate multiple solvation methods. However, the authors warned that the method has the potential to produce artifacts that may result in incorrect docking with some systems.

6.2. Pharmacophore Models

In 2000, Carlson et al. [152] were the first to develop a "dynamic" pharmacophore model that accounts for the inherent flexibility of a ligand-binding site. They developed their method on the pharmacologically important HIV integrase receptor, which is involved in the early

stages of the viral lifecycle. They used the method to identify several new inhibitors, by screening the ACD, that complement two important residues within the active site (D64 and D116).

The "dynamic" pharmacophore method uses a MD simulation to create an ensemble of snapshots in a similar way as that described in Section 5.3.1. The development of the MD simulation snapshots was described by the same group a year before the "dynamic" pharmacophore model was published [177]. HIV integrase is ideal for this method because the ligand-binding site is shallow and solvent exposed and is therefore less restricted in conformational sampling.

The snapshot structures represent the dynamic nature of the active site of the protein. The solvent molecules and counterions are removed from each snapshot, leaving the active sites exposed. A procedure called MUSIC, which employs a MC procedure, was used to determine the binding region surface of various functional groups. After the MC minimization, which incorporates a Boltzmann-weighted Metropolis sampling algorithm, the probe molecules (functional groups) are clustered within the local minima and then analyzed depending on their specific functionality. For example, Carlson et al. used methanol probes to define binding positions and orientations for hydroxyl groups.

Once the MUSIC simulation is complete, the results for each snapshot are superimposed, and any trends for a functional group at a particular part of the active site of the ensemble of structures is noted. Any conserved functional group-binding region within the ensemble of snapshots can indicate a high propensity for a certain group in a dynamic environment, such as the active site of HIV integrase. These conserved regions define the dynamic pharmacophore model.

Although the identification of several new inhibitors was more than enough evidence that the method is very powerful, a separate validation using a static model was undertaken. It was found that inhibitors that bind to the "dynamic" model were unable to fit into the static model. Hence, this highlights the need for developing a method such as this to represent an inherent flexibility within the active site.

6.3. Internal Coordinates Mechanics–Flexible Receptor Docking Algorithm Method

6.3.1. Effect of Receptor Flexibility in Virtual Screening Against Protein Kinases

In Section 4.2, we cited the results of the report of Cavasotto and Abagyan [121] on protein kinases. In that section, the effect of protein flexibility in ligand docking to protein kinases was shown, mainly in its effect on geometry prediction of known binders. However, geometry–prediction accuracy, although being a necessary component in the screening procedure, is not enough to achieve accurate scoring and high enrichment factors, which are the ultimate goal of virtual screening.

The authors performed a small-scale virtual screening of the 1000-compound library (seeded with the cocrystalized binders) on the 33 protein kinase structures. About 80% of the native binders were in the top 1.5% ranking of the screened database when docked to their cocrystallized structure, showing the accuracy of the docking and scoring protocol (the scoring function has not been optimized or tuned for protein kinases). The relation between RMSD values and ranking can be used as a tool to assess the effect of structural diversity on ligand docking and scoring. The correlation between scoring and RMSD deviation is plotted in Figure 11. Only 4% of the points are in the upper-left corner (bad pose–good score), and half of these are close to the borderline of 2.5 Å. Although the common threshold to assess the correctness of docking is 2 Å, because of variations in the position of the backbone when overlaying the structures for RMSD calculation, the authors used a threshold value of 2.5 Å. The lower-right corner corresponds to the good pose—bad score compounds (17%).

The percentage of ligands that are both docked correctly and ranked within the top 10% is ~49% (lower-left corner). This should be compared to the ~70% of compounds that are cross-docked correctly. Thus, it is seen that the protein kinase scoring and ranking is more

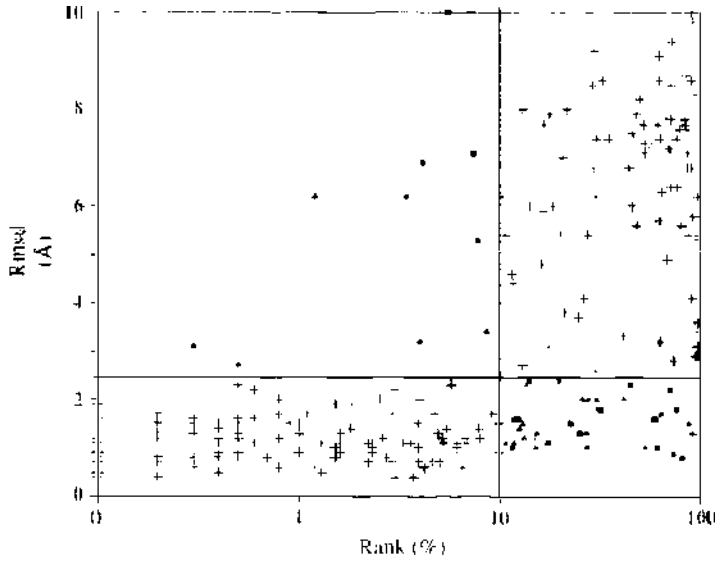


Figure 11. RMS deviations versus ranking of protein kinase ligands. Colored marks represent outliers: yellow circle, cAPK; blue triangles, CDK-2; red squares, P38. No outliers were found for LCK. The black lines delimit the areas of 10% ranking and 2.5-ÅRMSD deviations. RMSD values refer to the best energy docking solution. Reprinted with permission from [121], C. N. Cavasotto and R. A. Abagyan, *J. Mol. Biol.* 337, 209 (2004). © 2004, Elsevier.

sensitive to the induced-fit effects than ligand docking. However, 49% of correctly ranked compounds, and ~70% of correct poses, is a remarkable result.

6.3.2. Using Multiple Receptor Conformations to Improve the Enrichment Factor: The Merging and Shrinking Procedure

Cross-docking RMS deviations and enrichment factors can be used as tools to select which crystal structure should be used in the screening or whether there is a need to use more than one receptor structure to better represent conformational flexibility. From two different cAPK structures with adenosine binding (PDB codes 1BKX and 1FMO), the cross-docking results provide no clue as to which structure to use, but inspection of the enrichment factors shows that 1FMO might be a better target for virtual screening (16.7 versus 0.0 for the top 2% screened database). CDK-2 structure 1H1S has both good enrichment factor (58.3 for the top 1% selection) and high docking accuracy.

The small-scale virtual screening including nonnative known binders could be very useful when more than one experimental structure is available. The combination of screening results against two and three different protein kinase structures can yield better RMSD and enrichment factors values than each one alone, by using the merging and shrinking procedure, by which the screening against each conformation are merged and the best rank for each compound is kept, thus, shrinking the scoring list to a size corresponding to a virtual screening against a single receptor. It should be noted that the better performance is not self evident because the merging–shrinking procedure can also dilute the correct results. The improvement in enrichment factor within the merging–shrinking procedure was characterized by the enrichment factor fold-increase, calculated as:

$$EF\ fold - \text{increase} = \frac{EF_{\text{merged}}}{EF_{\text{individual}}} \tag{6}$$

The merging–shrinking procedure was evaluated on 21 groups of two or three protein kinase complexes by measuring the enrichment factor fold-increase for the top 1%, 2%, and 10% of the screened database. The average enrichment factor fold-increase was 1.85 ± 0.65, using 141 enrichment factor fold-increase values. The enrichment factor increase considering the top 1%, 2%, and 10% of the screened database was 1.89 ± 0.60, 1.83 ± 0.60, and 1.83 ± 0.74, respectively, showing that the enrichment factor fold-increase was roughly independent of the size of the focused selection. The enrichment factor fold-increase was less

than one in only four out of the 141 cases (<3%). It was also shown that combining even two structures is enough to have most of the ligands docked correctly (more than 75% of the ligands can be considered as correctly docked on merging and shrinking). The merging–shrinking procedure represents a more significant improvement on the RMSD values than on enrichment factors, showing again that scoring is more sensitive to protein flexibility than ligand docking. Still, an average enrichment factor fold-increase of ~2 is a significant improvement.

The computing time for the flexible ligand/rigid receptor docking and scoring is ~1–2 minutes, using a 700 MHz processor (1 Mb RAM for the dual-processor node). Therefore, screening against a few structures is affordable and has the advantage of using real actual receptor conformations, which may differ not only in side-chain conformation but in backbone rearrangement.

6.3.3. The ICM-Flexible Receptor Docking Algorithm

A small-scale virtual screening followed by analysis of cross-docking results and enrichment factors can help to determine the strategy to include potential structural diversity of the receptor binding pocket in the virtual screening. The merging and shrinking procedure, using the results of two or three screenings, could be a solution to improve the enrichment factor. The ICM-flexible receptor docking algorithm (IFREDA) approach described below allows both side-chain rearrangements and essential backbone movements, thus sampling adequately the conformational space of the receptor, even in cases of large loop movements. The IFREDA is especially useful in cases in which only one holo (or apo) crystal structure and known binders are available. If this is the case, a *de novo* receptor structure ensemble is generated by performing flexible ligand docking to a flexible receptor of the known binders, where side-chain and essential backbone movements are taken into consideration. Then a virtual screening against the generated mutliple conformations using flexible ligand–grid receptor docking is performed, and finally, a merging and shrinking of the screening results.

This methodology for *de novo* generation of an ensemble of structures consists of seeding, soft van der Waals structure relaxation, and stochastic global energy optimization.

First, in seeding, is the placement of the ligand within the ligand binding pocket. Second is the generation of four conformations by flipping the ligand 180 degrees with respect to its principal axis of inertia. Third is the generation of 40 conformations performing 10 random displacements and rigid rotations of the ligand in each of the original four complexes.

In the soft van der Waals structure relaxation, each complex conformation is relaxed through *in vacuo* minimization, using a variable-weight soft van der Waals algorithm. The ligand and selected backbone loops and pocket side chains were regarded as fully flexible in this stage.

In stochastic global energy optimization, first, stochastic global energy optimization performed, using the double-energy minimization scheme [35] of each of the 40 complexes with respect to the side chain and ligand variables. Second comes clustering of a set of geometrically diverse low-energy states generated during the energy optimization (by comparing the RMSD of the heavy atom coordinates of the ligand). Third, Full-energy minimization of the top-ranking conformations by keeping flexible the selected backbone loops, pocket side chains, and ligand occurs. Finally, reevaluation of the energy with a more accurate solvation energy term, by solving the Poisson equation using the boundary element algorithm, takes place [178].

6.3.4. De Novo Receptor Generation of Accurate Protein Kinase Structures Using Internal Coordinates Mechanics–Flexible Receptor Docking Algorithm

The IFREDA procedure was validated with eight protein kinase complexes and was able to generate the correct receptor and ligand-bound conformation and to energetically discriminate them from misdocked ones within the accuracy of the energy function, even in cases in which loop rearrangement (~2 Å) was necessary for the ligand binding. The complexes selected for this validation were, A: 1FMO + balanol (1BX6). B: 1JLU + staurosporine

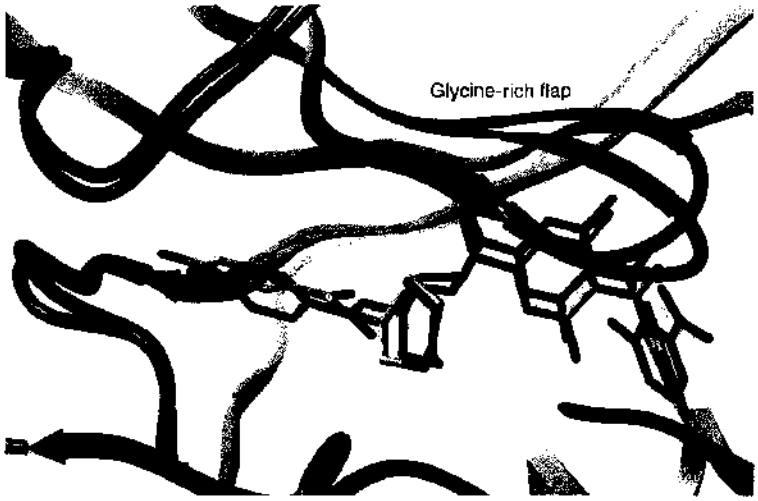


Figure 12. Flexible docking of balanol into the 1FMO binding pocket (model A). The modeled structure is in grey, whereas 1FMO is displayed in yellow. Balanol carbon atoms are displayed in yellow (corresponding to native structure 1BX6) and white (modeled complex). Notice the displacement in the modeled structure of the glycine-rich flap ~2 Å upward to allow balanol to bind (picture constructed using ICM version 3.0). Reprinted with permission from [121], C. N. Cavasotto and R. A. Abagyan, *J. Mol. Biol.* 337, 209 (2004). © 2004, Elsevier.

(1STC), C: 1JLU + H7 (1YDR), D: 1DM2 + NU6027 (1E1X), E: 1E1X + oxindole 16 (1FVT), F: 1JSV + oxindole 16 (1FVT), G: 1BMK + 14e (1M7Q), and H: 3LCK + PP2 (1QPE). Ligand-associated PDB codes are displayed in parentheses after the ligand name.

Binding sites have regions of very high and very low stability, as shown in this publication [102]. The side chains within the binding pocket were flexible, and portions of the backbone were considered to be rigid. Flexible backbone regions were chosen by undertaking an inspection of the mobile parts in the available crystal structures of the corresponding protein kinase family. In some cases, these regions correspond with high B-factor regions. In Figures 12 and 13, we compare the structures generated through the *de novo* generating protocol of IFREDA with the ligand-coerystallized structure for models A and B.

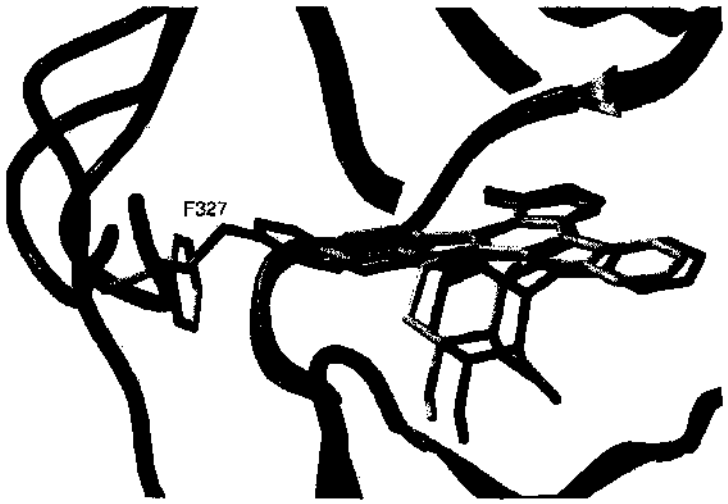


Figure 13. Flexible docking staurosporine into the 1JLU binding pocket (model B). The modeled structrure is in grey, whereas 1JLU is displayed in yellow. The glycine-rich flap has been cut for the sake of clarity. Staurosporine carbon atoms are displayed in yellow (corresponding to native structure 1STC) and white (modeled complex). Notice the displacement in the modeled structure of the C-terminal loop that contains Phe 327 (picture constructed using ICM version 3.0). Reprinted with permission from [121], C. N. Cavasotto and R. A. Abagyan, *J. Mol. Biol.* 337, 209 (2004). © 2004, Elsevier.

Merging and shrinking procedures were used to condense the results from the small-scale virtual screening against the generated set of structures. An enrichment factor fold-increase of 1.89 ± 0.60 was observed by the authors, which is slightly better than that achieved using multiple experimental structures. This clearly demonstrates that multiple receptor conformations generated through IFREDA represent the structural space of the binding pocket. Therefore, these conformations can be used when multiple experimental structures are not available, or when the experimental structures that are available do not represent the conformational space of the receptor. Finally, an interesting observation is that the *de novo* structure generation does not create a custom pocket that is suitable only for the ligand used in pocket generation.

7. SUMMARY

The correct functioning of a cell is in part a result of a multitude of critical interactions between molecules. It is typical of scientific endeavor to try and improve and alter these interactions. The development of modern day drugs illustrates that it is sometimes possible to identify molecules that can bind with greater affinity to those evolved in nature. It is clear from this, however, that each receptor must be able to interact with many different kinds of molecules and change their shape accordingly. In this chapter, we have described many different computational methods that try and predict these intereactions.

The main aim of protein–ligand docking is to identify a small subset of compounds from a large compound database, which will then be tested experimentally for binding affinity to a target receptor. This not only saves money and time but may also identify compounds that may never have been identified by more conventional experimental means. The desired compounds may be agonists or antagonists depending on the nature of the receptor and disease under investigation. In the past, docking has been successful for identifying compounds to receptors solved by x-Ray, NMR, and modeling. Once a ligand has shown some binding activity in experimental assays, the scaffold of that compound could then used in cycles of docking studies to identify better or more specific binders.

There are a number of challenges associated with attempting to predict ligand receptor interactions, such as accurate scoring, extensive sampling, and explicit consideration of ionization states of ligands. These factors can lead to inaccuracies in a docking and virtual screening methods and result in wasted time and money and cause embarrassment to the computational chemist or bioinformatician relying on the accuracy of these methods. One of the main challenges in protein–ligand docking, as described in this chapter, is the ability to incorporate flexibility into the receptor on complex formation.

When using conventional nonflexible or pseudo-flexible docking algorithms, problems associated with lack of flexibility within a receptor come to the surface. These problems present themselves when docking a known inhibitor or agonist to a receptor. Clashes with the receptor can be observed, and poor ligand–protein contacts that are not strong enough to hold the ligand in place are common. This, of course, has the consequence of incorrect scoring and poor ranking of compounds. When undertaking a docking procedure, it is expected that a good fit will be observed between the crystal structure and the ligand with which it was crystallized. However, the docking scenario becomes more complicated when larger or additional groups to the ligand scaffold interact with the same binding pocket.

Within this chapter, we have highlighted some examples of the substantial evidence that has been published that there is inherent flexibility within most ligand-binding sites. For example Fig. 5 highlights the experimental evidence shown with acetylcholinesterase ligand-complex, in which the movement of a single amino acid is critical for ligand binding. In addition, the same receptor, though crystallized from a different species, shows that ligand binding is dependent on the movement of many amino acid side chains within the ligand-binding pocket (Fig. 6).

Bouzida [50] and Murray [117] demonstrate that docking algorithms can fail when flexibility is not taken into consideration. Recently, Cavasotto and Abagyan [121] have presented an analysis of cross-docking of ligands to protein kinases. Ligands can be redocked to their

native crystal structure with remarkable accuracy, but the influence of the ligand-induced conformational changes on docking results becomes evident when a ligand is docked to a receptor complexed with another compound. It was shown that on average, only 70% of the ligands were cross-docked correctly in the protein kinase benchmark. This percentage was dramatically improved with the use of flexible receptor docking. It is clear these small variations in the structure of the binding pocket have a great effect on docking geometries.

We have described how proteins exist as an ensemble of conformers with similar energies. The challenge for scientists undertaking structure-based drug design is to be able to identify compounds that have high affinity to less populated receptor states or to identify compounds with affinity to highly populated states. The receptor-binding pocket can be visualized as an energy landscape whereby when a ligand binds it causes a shift in the dynamic population of energies in favor of that particular complex.

It is evident that docking algorithms need to incorporate flexibility into the receptor as well as the ligand. Ideally, the algorithm will be computationally fast enough to perform virtual screening on large databases of compound structures. All the methods described in Section 5, many an extension of the early docking algorithms discussed in Section 2.2, have accomplished one or another of these requirements, with differing degrees of success. There are a number of problems associated with early flexible docking methods such as soft docking and sampling side-chains conformers. Although soft docking may be a solution to avoid or diminish the influence of van der Waals clashes, it is unable to reproduce the network of interactions that are determinant for ligand binding. As it has already been stated, small movements in the receptor structure can lead to errors up to ~3.3 kcal/mol in the binding-energy prediction. This sensitivity shows the clear limitations of the soft-docking approach. Sampling side chains is an attractive solution for systems if no significant backbone rearrangement were observed on ligand binding. However, efficient sampling of the degrees of freedom renders this method unaffordable because of the necessary computing time.

Docking methods incorporating flexibility are developing rapidly, in part because of the improvement of computer processor speed, available RAM, and disk capacity. However, it is inconceivable to imagine direct incorporation of receptor flexibility sometime in the near future because of the extraordinarily massive computational demand that would be involved. Therefore, methods incorporating ensembles of protein structures are currently the best approximation that can be made at this time. It should be noted, however, that the generation of multiple receptor conformations by combining different rotamer states of binding-pocket side chains could be of little use in docking hundreds of thousands of compounds to the receptor ensemble, because of the combinatorial explosion of conformations.

Although it is clear that there is currently no computationally efficient docking method that is able to screen a large database of flexible ligands against a flexible receptor, we believe that further developments of methods like FlexE [63], IFREDA [121], and the pharmacophore model [152] could eventually lead to significant improvements in this field.

REFERENCES

1. A. H. Machle, C. R. Pruft, and R. F. Halliwell, *Nat. Rev. Drug Discovery* 1, 637 (2002).
2. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, *Nucleic Acids Res.* 28, 235 (2000).
3. C. R. Beddell, P. J. Goodford, F. E. Norrington, S. Wilkinson, and R. Wootton, *Br. J. Pharmacol.* 57, 201 (1976).
4. I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin, *J. Mol. Biol.* 161, 269 (1982).
5. E. Fischer, *Ber. Dtsch. Chem. Ges.* 2985 (1894).
6. D. E. Koshland, Jr., Angewandte Chemie-International Edition 33, 2375 (1995).
7. D. E. Koshland, Jr., *Proc. Natl. Acad. Sci. USA* 44, 98 (1958).
8. B. K. Shoichet, S. L. McGovern, B. Q. Wei, and J. J. Irwin, *Curr. Opin. Chem. Biol.* 6, 439 (2002).
9. R. D. Taylor, P. J. Jewsbury, and J. W. Essex, *J. Comput.-Aided Mol. Design* 16, 151 (2002).
10. R. Abagyan and M. Totrov, *Curr. Opin. Chem. Biol.* 5, 375 (2001).
11. R. L. Desjarlais, R. P. Sheridan, J. S. Dixon, I. D. Kuntz, and R. Venkataraghavan, *J. Med. Chem.* 29, 2149 (1986).
12. R. L. Desjarlais, R. P. Sheridan, G. L. Seibel, J. S. Dixon, I. D. Kuntz, and R. Venkataraghavan, *J. Med. Chem.* 31, 722 (1988).

13. E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A. A. Friesem, C. Aflalo, and I. A. Vakser, *Proc. Natl. Acad. Sci. USA* 89, 2195 (1992).
14. D. J. Bacon and J. Moult, *J. Mol. Biol.* 225, 849 (1992).
15. A. R. Leach and I. D. Kuntz, *J. Comput. Chem.* 13, 730 (1992).
16. M. Helmer-Citterich and A. Tramontano, *J. Mol. Bio.* 235, 1021 (1994).
17. D. Fischer, S. L. Lin, H. L. Wolfson, and R. Nussinov, *J. Mol. Biol.* 248, 459 (1995).
18. F. Osterberg, G. M. Morris, M. F. Sanner, A. J. Olson, and D. S. Goodsell, *Proteins-Struct. Funct. Genet.* 46, 34 (2002).
19. T. J. A. Ewing and I. D. Kuntz, *J. Comput. Chem.* 18, 1175 (1997).
20. J. Y. Trosset and H. A. Scheraga, *J. Comput. Chem.* 20, 412 (1999).
21. E. Perola, K. Xu, T. M. Kollmeyer, S. H. Kaufmann, F. G. Prendergast, and Y. P. Pang, *J. Med. Chem.* 43, 401 (2000).
22. M. Rarey, B. Kramer, T. Lengauer, and G. Klebe, *J. Mol. Biol.* 261, 470 (1996).
23. M. D. Miller, S. K. Kearsley, D. J. Underwood, and R. P. Sheridan, *J. Comput.-Aided Mol. Design* 8, 153 (1994).
24. G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor, *J. Mol. Biol.* 267, 727 (1997).
25. R. Abagyan, M. Totrov, and D. Kuznetsov, *J. Comput. Chem.* 15, 488 (1994).
26. H. J. Bohm, *J. Comput.-Aided Mol. Design* 6, 61 (1992).
27. C. A. Baxter, C. W. Murray, D. E. Clark, D. R. Westhead, and M. D. Eldridge, Abstracts of Papers of the American Chemical Society 216, U693-U693, 1998.
28. C. McMartin and R. S. Bohacek, *J. Comput.-Aided Mol. Design* 11, 333 (1997).
29. V. Schnecke, C. A. Swanson, E. D. Getzoff, J. A. Tainer, and L. A. Kuhn, *Proteins-Struct. Funct. Genet.* 33, 74 (1998).
30. N. A. Metropolis, A. W. Rosenbluth, N. M. Rosenbluth, A. H. Teller, and E. Teller, *J. Chem. Phy.* 21, 1087 (1953).
31. B. A. Luty, Z. R. Wasserman, P. F. W. Stouten, C. N. Hodge, M. Zacharias, and J. A. McCammon, *J. Comput. Chem.* 16, 454 (1995).
32. M. Liu and S. M. Wang, *J. Comput.-Aided Mol. Design* 13, 435 (1999).
33. G. Nemethy, K. D. Gibson, K. A. Palmer, C. N. Yoon, M. G. Paterlini, A. Zagari, S. Rumsey, and H. A. Scheraga, *J. Phys. Chem.* 96, 6472 (1992).
34. Molsoft LLC, 3.0 ed. Molsoft LLC, La Jolla, CA, 2003.
35. R. Abagyan and M. Totrov, *J. Mol. Biol.* 235, 983 (1994).
36. R. A. Abagyan and M. Totrov, *J. Comput. Phys.* 151, 402 (1999).
37. R. Abagyan and P. Argos, *J. Mol. Biol.* 225, 519 (1992).
38. Z. Li and H. A. Scheraga, *Proc. Natl. Acad. Sci. USA* 84, 6611 (1987).
39. M. Schapira, B. M. Raaka, H. H. Samuels, and R. Abagyan, *Proc. Natl. Acad. Sci. USA* 97, 1008 (2000).
40. M. Schapira, B. M. Raaka, H. H. Samuels, and R. Abagyan, *BMC Struct. Biol.* 1, 1 (2001).
41. M. Schapira, B. M. Raaka, S. Das, L. Fan, M. Totrov, Z. U. Zhou, S. R. Wilson, R. Abagyan, and H. H. Samuels, *Proc. Natl. Acad. Sci. USA* 100, 7354 (2003).
42. K. Palczewski, T. Kumasaka, T. Hori, C. A. Behnke, H. Motoshima, B. A. Fox, I. Le Trong, D. C. Teller, T. Okada, R. E. Stenkamp, M. Yamamoto, and M. Miyano, *Science* 289, 739 (2000).
43. S. Goto, T. Nishioka, and M. Kanehisa, *Nucleic Acids Res.* 27, 377 (1999).
44. S. Goto, Y. Okuno, M. Hattori, T. Nishioka, and M. Kanehisa, *Nucleic Acids Res.* 30, 402 (2002).
45. C. N. Cavasotto, A. J. W. Orry, and R. Abagyan, *Proteins-Struct. Funct. Genet.* 51, 423 (2003).
46. W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, *J. Ame. Chem. Soc.* 117, 5179 (1995).
47. J. Y. Trosset and H. A. Scheraga, *Proc. Natl. Acad. Sci. USA* 95, 8011 (1998).
48. J. Y. Trosset and H. A. Scheraga, *J. Comput. Chem.* 20, 244 (1999).
49. A. Caflisch, S. Fischer, and M. Karplus, *J. Comput. Chem.* 18, 723 (1997).
50. D. Bouzida, P. A. Rejto, S. Arthurs, A. B. Colson, S. T. Freer, D. K. Gehlhaar, V. Larson, B. A. Luty, P. W. Rose, and G. M. Verkhivker, *Int. J. Quant. Chem.* 72, 73 (1999).
51. T. Halgren, Abstracts of Papers of the American Chemical Society 220, U168-U168 (2000).
52. R. Friesner, T. A. Halgren, P. Shenkin, R. B. Murphy, J. Banks, H. Beard, J. Klicic, D. Mainz, J. Perry, and T. F. Hendrickson, Abstracts of Papers of the American Chemical Society 223, U465-U465 (2002).
53. T. A. Halgren, R. B. Murphy, J. Banks, D. Mainz, J. Klicic, J. K. Perty, and R. A. Friesner, Abstracts of Papers of the American Chemical Society 224, U345-U345 (2002).
54. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, *J. Comput. Chem.* 4, 187 (1983).
55. A. Di Nola, D. Roccatano, and H. J. C. Berendsen, *Proteins-Struct. Funct. Genet.* 19, 174 (1994).
56. R. Mangoni, D. Roccatano, and A. Di Nola, *Proteins-Struct. Funct. Genet.* 35, 153 (1999).
57. A. Miranker and M. Karplus, *Proteins-Struct. Funct. Genet.* 11, 29 (1991).
58. R. Rosenfeld, Q. Zheng, S. Vajda, and C. DeLisi, *Genet. Anal.* 12, 1 (1995).
59. R. Rosenfeld, S. Vajda, and C. DeLisi, *Annu. Rev. Biophys. Biomol. Struct.* 24, 677 (1995).
60. P. J. Rossky, J. D. Doll, and H. L. Friedman, *J. Chem. Phy.* 69, 4628 (1978).
61. R. E. Kozack and S. Subramaniam, *Protein Sci.* 2, 915 (1993).
62. Z. R. Wasserman and C. N. Hodge, *Proteins-Struct. Funct. Genet.* 24, 227 (1996).
63. H. Claussen, C. Buning, M. Rarey, and T. Lengauer, *J. Mol. Biol.* 308, 377 (2001).
64. M. Rarey, B. Kramer, and T. Lengauer, *Proteins-Struct. Funct. Genet.* 34, 17 (1999).

65. C. Bron and J. Kerbosch. *Commun. ACM* 16, 575 (1973).
66. A. T. Brint and P. Willett. *J. Chem. Info. Compu. Sci.* 27, 152 (1987).
67. H. M. Grindley, P. J. Artymiuk, D. W. Rice, and P. Willett, *J. Mol. Biol.* 229, 707 (1993).
68. P. J. Goodford. *J. Med. Chem.* 28, 849 (1985).
69. M. Y. Mizutani, N. Tomioka, and A. Itai. *J. Mol. Biol.* 243, 310 (1994).
70. W. Welch, J. Ruppert, and A. N. Jain. *Chem. Biol.* 3, 449 (1996).
71. L. B. Booker, D. E. Goldberg, and J. H. Holland, Artificial Intelligence 40, 235 (1989).
72. J. H. Holland, *Sci. Amer.* 267, 66 (1992).
73. J. S. Taylor and R. M. Burnett, *Proteins-Struct. Func. Genet.* 41, 173 (2000).
74. G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson, *J. Comput. Chem.* 19, 1639 (1998).
75. K. P. Clark and Ajay, *J. Comput. Chem.* 16, 1210 (1995).
76. Ajay and K. P. Clark, Abstracts of Papers of the American Chemical Society 209, 102-COMP (1995).
77. R. S. Judson, E. P. Jaeger, and A. M. Treasurywala, *Theochem-J. Mol. Struct.* 114, 191 (1994).
78. D. K. Gehlhaar, G. M. Verkhivker, P. A. Rejto, C. J. Sherman, D. B. Fogel, L. J. Fogel, and S. T. Freer, *Chem. Biol.* 2, 317 (1995).
79. H. A. Gabb, R. M. Jackson, and M. J. E. Sternberg, *J. Mol. Biol.* 272, 106 (1997).
80. P. Burkhard, P. Taylor, and M. D. Walkinshaw, *J. Mol. Biol.* 277, 449 (1998).
81. V. Sobolev, R. C. Wade, G. Vriend, and M. Edelman, *Proteins-Struct. Funct. Genet.* 25, 120 (1996).
82. V. Sobolev, T. M. Moallem, R. C. Wade, G. Vriend, and M. Edelman, *Proteins-Struct. Funct. Genet.* 210 (1997).
83. Ajay and M. A. Murcko, *J. Med. Chem.* 38, 4953 (1995).
84. H. J. Bohm and G. Klebe, *Angewandte Chem. int.* 35, 2589 (1996).
85. H. J. Bohm, *J. Comput.-Aided Mol. Design* 8, 243 (1994).
86. H. J. Bohm, *J. Comput.-Aided Mol. Design* 12, 309 (1998).
87. A. N. Jain, *J. Comput.-Aided Mol. Design* 10, 427 (1996).
88. C. W. Murray, T. R. Auton, and M. D. Eldridge, *J. Comput.-Aided Mol. Design* 12, 503 (1998).
89. T. Halgren, *J. Comp. Chem.* 17, 490 (1995).
90. G. Verkhivker, K. Appelt, S. T. Freer, and J. E. Villafranca, *Protein Eng.* 8, 677 (1995).
91. R. S. DeWitte and E. Shakhnovich, Abstracts of Papers of the American Chemical Society 214, 6-COMP (1997).
92. I. Muegge and Y. C. Martin, *J. Med. Chem.* 42, 791 (1999).
93. I. Muegge, Y. C. Martin, P. J. Hajduk, and S. W. Fesik, *J. Med. Chem.* 42, 2498 (1999).
94. H. Gohlke, M. Hendlich, and G. Klebe, *J. Mol. Bio.* 295, 337 (2000).
95. M. Totrov and R. Abagyan, in "Drug-Receptor Thermodynamics: Introduction and Experimental Applications" (R. B. Raffa, Ed.). Wiley, New York, 2001.
96. M. Totrov and R. Abagyan, *Biopolymers* 60, 124 (2001).
97. B. Ma, S. Kumar, C. J. Tsai, and R. Nussinov, *Protein Eng.* 12, 713 (1999).
98. C. Pargellis, L. Tong, L. Churchill, P. F. Cirillo, T. Gilmore, A. G. Graham, P. M. Grob, E. R. Hickey, N. Moss, S. Pav, and J. Regan, *Nat. Struct. Biol.* 9, 268 (2002).
99. Y. Ashani, J. O. Peggins III, and B. P. Doctor, *Biochem. Biophys. Res. Commun.* 184, 719 (1992).
100. M. Z. Wang, C. Y. Tai, and D. B. Mendel, *Antimicrob. Agents Chemother.* 46, 3809 (2002).
101. A. Graul, P. A. Leeson, and J. Castaner, *Drugs Future* 1189 (1999).
102. I. Luque and E. Freire, *Proteins Suppl* 4, 63 (2000).
103. P. Zavodszky, J. Kardos, Svingor, and G. A. Petsko, *Proc. Natl. Acad. Sci. USA* 95, 7406 (1998).
104. B. K. Shoichet, W. A. Baase, R. Kuroki, and B. W. Matthews, *Proc. Natl. Acad. Sci. USA* 92, 452 (1995).
105. W. L. DeLano, M. H. Ultsch, A. M. de Vos, and J. A. Wells, *Science* 287, 1279 (2000).
106. N. Vazquez-Laslop, E. E. Zheleznova, P. N. Markham, R. G. Brennan, and A. A. Neyfakh, *Biochem. Soc. Trans.* 28, 517 (2000).
107. D. Tondi, U. Slomczynska, M. P. Costi, D. M. Watterson, S. Ghelli, and B. K. Shoichet, *Chem. Biol.* 6, 319 (1999).
108. B. Ma, M. Shatsky, H. J. Wolfson, and R. Nussinov, *Protein Sci.* 11, 184 (2002).
109. M. H. Van Regenmortel, *J. Mol. Recognit.* 12, 1 (1999).
110. A. M. Silva, A. Y. Lee, S. V. Gulnik, P. Maier, J. Collins, T. N. Bhat, P. J. Collins, R. E. Cachau, K. E. Luker, I. Y. Gluzman, S. E. Francis, A. Oksman, D. E. Goldberg, and J. W. Erickson, *Proc. Natl. Acad. Sci. USA* 93, 10034 (1996).
111. A. Y. Lee, S. V. Gulnik, and J. W. Erickson, *Nat. Struct. Biol.* 5, 866 (1998).
112. R. Najmanovich, J. Kuttner, V. Sobolev, and M. Edelman, *Proteins-Struct. Funct. Genet.* 39, 261 (2000).
113. G. M. Smith, R. S. Alexander, D. W. Christianson, B. M. McKeever, G. S. Ponticello, J. P. Springer, W. C. Randall, J. J. Baldwin, and C. N. Habecker, *Protein Sci.* 3, 118 (1994).
114. G. Kryger, I. Silman, and J. L. Sussman, *Structure* 7, 297 (1999).
115. M. Harel, G. Kryger, T. L. Rosenberry, W. D. Mallender, T. Lewis, R. J. Fletcher, J. M. Guss, I. Silman, and J. L. Sussman, *Protein Sci.* 9, 1063 (2000).
116. C. Oefner, A. Binggeli, V. Breu, D. Bur, J. P. Clozel, A. D'Arcy, A. Dorn, W. Fischli, F. Gruninger, R. Guller, G. Hirth, H. P. Marki, S. Mathews, M. Muller, R. G. Ridley, H. Stadler, E. Vieira, M. Wilhelm, F. K. Winkler, and W. Wostl, *Chem. Biol.* 6, 127 (1999).
117. C. W. Murray, C. A. Baxter, and A. D. Frenkel, *J. Comput.-Aided Mol. Design* 13, 547 (1999).

118. B. I. Schweitzer, A. P. Dicker, and J. R. Bertino, *FASEB J.* 4, 2441 (1990).
119. V. Cody, N. Galitsky, D. Rak, J. R. Luft, W. Pangborn, and S. F. Queener, *Biochemistry* 38, 4303 (1999).
120. J. Feeney, Angewandte *Chem. Int.* 39, 290 (2000).
121. C. N. Cavasotto and R. Abagyan, *J. Mol. Biol.* 337, 209 (2004).
122. J. Dancey and E. A. Sausville, *Nat. Rev. Drug. Discov.* 2, 296 (2003).
123. S. J. Teague, *Nat. Rev. Drug. Discov.* 2, 527 (2003).
124. J. A. Erickson, M. Jalaie, D. H. Robertson, R. A. Lewis, and M. Vieth, *J. Med. Chem.* 47, 45 (2004).
125. M. Vieth and D. J. Cummins, *J. Med. Chem.* 43, 3020 (2000).
126. G. Wu, D. H. Robertson, C. L. Brooks, 3rd, and M. Vieth, *J. Comput. Chem.* 24, 1549 (2003).
127. F. Jiang and S. H. Kim, *J. Mol. Biol.* 219, 79 (1991).
128. P. H. Walls and M. J. E. Sternberg, *J. Mol. Biol.* 228, 277 (1992).
129. D. A. Gschwend, A. C. Good, and I. D. Kuntz, *J. Mol. Recog.* 9, 175 (1996).
130. J. Apostolakis, A. Pluckthun, and A. Caflisch, *J. Comput. Chem.* 19, 21 (1998).
131. A. R. Leach, *J. Mol. Biol.* 235, 345 (1994).
132. J. Desmet, M. Demaeyer, B. Hazes, and I. Lasters, *Nature* 356, 539 (1992).
133. P. E. Hart, N. J. Nilsson, and B. Raphael, *IEEE Trans. Syst. Sci. Cybernet.* SSC4, 100-& (1968).
134. A. R. Leach and A. P. Lemon, *Proteins-Struct. Funct. Genet.* 33, 227 (1998).
135. L. Schaffer and G. M. Verkhivker, *Proteins* 33, 295 (1998).
136. G. Jones, P. Willett, and R. C. Glen, *J. Mol. Biol.* 245, 43 (1995).
137. J. Desmet, I. A. Wilson, M. Joniau, M. De Maeyer, and I. Lasters, *FASEB J.* 11, 164 (1997).
138. L. David, R. Luo, and M. K. Gilson, *J. Comput.-Aided Mol. Design* 15, 157 (2001).
139. V. Kairys and M. K. Gilson, *J. Comput. Chem.* 23, 1656 (2002).
140. A. C. Anderson, R. H. O'Neil, T. S. Surti, and R. M. Stroud, *Chem. Biol.* 8, 445 (2001).
141. C. N. Cavasotto, G. Liu, S. Y. James, P. J. Hobbs, V. J. Peterson, A. A. Bhattacharya, S. K. Kolluri, R. Abagyan, X.-K. Zhang, M. Leid, R. C. Liddington, and M. I. Dawson, *J. Med. Chem.* 47, 4360 (2004).
142. C. N. Cavasotto and R. Abagyan, in "Flexible Receptor Model Explains Structural Determinants for Agonist and Antagonist Activity in Retinoid Receptors." Proceedings of the "Era of Hope" Meeting, p. P16–02. Department of Defense, Orlando, FL. 2002.
143. M. I. Dawson, C. N. Cavasotto, L. Jong, K.-C. Feng, M. Leid, H. Li, R. Abagyan, and X.-K. Zhang, in "Synthesis and Biological Activity of Retinoid Antagonist Selective for the Retinoid X Receptor (RXR)." Poster 1019, San Francisco, CA, 2002.
144. T. M. Frimurer, G. H. Peters, L. F. Iversen, H. S. Andersen, N. P. Moller, and O. H. Olsen, *Biophys. J.* 84, 2273 (2003).
145. B. Sandak, H. J. Wolfson, and R. Nussinov, *Proteins* 32, 159 (1998).
146. B. Sandak, R. Nussinov, and H. J. Wolfson, *J. Comput. Biol.* 5, 631 (1998).
147. H. A. Carlson and J. A. McCammon, *Mol. Pharmacol.* 57, 213 (2000).
148. H. A. Carlson, *Curr. Pharm. Design* 8, 1571 (2002).
149. Y. P. Pang and A. P. Kozikowski, *J. Comput.-Aided Mol. Design* 8, 669 (1994).
150. C. M. Stultz and M. Karplus, *Proteins-Struct. Funct. Genet.* 37, 512 (1999).
151. J. A. Given and M. K. Gilson, *Proteins-Struct. Funct. Genet.* 33, 475 (1998).
152. H. A. Carlson, K. M. Masukawa, K. Rubins, F. D. Bushman, W. L. Jorgensen, R. D. Lins, J. M. Briggs, and J. A. McCammon, *J. Med. Chem.* 43, 2100 (2000).
153. H. B. Broughton, *J. Mol. Graph. Model* 18, 247, 302 (2000).
154. N. Nakajima, H. Nakamura, and A. Kidera, *J. Phys. Chem. B* 101, 817 (1997).
155. B. A. Berg and T. Neuhaus, *Phys. Rev. Lett.* 68, 9 (1992).
156. Y. Pak and S. M. Wang, *J. Chem. Phys.* 111, 4359 (1999).
157. Y. S. Pak and S. M. Wang, *J. Phys. Chem. B* 104, 354 (2000).
158. J. Zhu, H. Fan, H. Y. Liu, and Y. Y. Shi, *J. Comput.-Aided Mol. Design* 15, 979 (2001).
159. J. H. Lin, A. L. Perryman, J. R. Schames, and J. A. McCammon, *J. Am. Chem. Soc.* 124, 5632 (2002).
160. J. H. Lin, A. L. Perryman, J. R. Schames, and J. A. McCammon, *Biopolymers* 68, 47 (2003).
161. M. Philippopoulos and C. Lim, *Proteins-Struct. Funct. Genet.* 36, 87 (1999).
162. M. Billeter, *Quart. Rev. Biophys.* 25, 325 (1992).
163. G. Wagner, S. G. Hyberts, and T. F. Havel, *Annu. Rev. Biophys. Biomol. Struct.* 21, 167 (1992).
164. R. M. Knegtel, I. D. Kuntz, and C. M. Oshiro, *J. Mol. Biol.* 266, 424 (1997).
165. E. A. Sudbeck, C. Mao, R. Vig, T. K. Venkatachalam, L. Tuel-Ahlgren, and F. M. Uckun, *Antimicrobial Agents Chemother.* 42, 3225 (1998).
166. M. A. Kastenholz, M. Pastor, G. Cruciani, E. E. J. Haaksma, and T. Fox, *J. Med. Chem.* 43, 3033 (2000).
167. F. Tama, W. Wriggers, and C. L. Brooks, 3rd, *J. Mol. Biol.* 321, 297 (2002).
168. W. G. Krebs, V. Alexandrov, C. A. Wilson, N. Echols, H. Yu, and M. Gerstein, *Proteins* 48, 682 (2002).
169. M. Zacharias and H. Sklenar, *J. Comput. Chem.* 20, 287 (1999).
170. M. Levitt, C. Sander, and P. S. Stern, *J. Mol. Biol.* 181, 423 (1985).
171. N. Go, T. Noguti, and T. Nishikawa, *Proc. Natl. Acad. Sci. USA* 80, 3696 (1983).
172. I. Kolossváry and W. C. Guida, *J. Am. Chem. Soc.* 118, 5011 (1996).
173. I. Kolossváry and W. C. Guida, *J. Comput. Chem.* 1671 (1999).
174. I. Kolossváry and G. M. Keserü, *J. Comput. Chem.* 22, 21 (2001).

175. G. M. Keseru and I. Kolossváry, *J. Am. Chem. Soc.* 123, 12708 (2001).
176. C. N. Cavasotto, J. Kovacs, M. Totrov, and R. Abagyan, in "Incorporating Receptor Flexibility in Ligand Docking and Virtual Screening through a Normal Mode Analysis-Based Procedure." Proceedings of the 48th Meeting of the Biophysical Society. Biophysical Society, Baltimore, MD, 2004.
177. H. A. Carlson, K. M. Masukawa, and J. A. McCammon, *J. Phy. Chem. A* 103, 10213 (1999).
178. M. Totrov and R. Abagyan, *J. Struct. Biol.* 116, 138 (1996).

CHAPTER 6

Enzyme Reactivity Studied by Computer Simulations

Giorgio Colombo, Massimiliano Meli, Giacomo Carrea
Istituto di Chimica del Riconoscimento Molecolare, Milano, Italy

CONTENTS

ISBN: 1-5888.-048-9

Handbook of Theoretical and Computational Nanotechnology
Edited by Michael Rieth and Wolfram Schommers
Volume 6: Pages (259–303)

1. INTRODUCTION

The study of chemical processes in biological systems has long been a subject of intense experimental and theoretical studies. Understanding the mechanisms of enzyme reactivity is a fundamental challenge for the complete unraveling of biochemical processes, with important consequences in several aspects of (bio)molecular sciences. In recent years, enormous progress has been made thanks to the accumulation of enzyme three-dimensional structures, the development of molecular biology and genetic engineering, and the improvements in analytical techniques. However, despite these advances, the correlation of enzyme structures to reactivity or the examination of the feasibility of different catalytic mechanisms cannot yet be clearly understood in all cases. That is why the support of theoretical and computer modeling approaches has been acquiring an increasingly growing importance. In this chapter, we will discuss the study of enzymatic mechanisms through the application of computer simulation methods. The topic will be treated by first illustrating the basic methodologies (Section 2), followed by an overview of selected studies and results (Sections 3 and 4), and finally by possible future developments (Section 5).

2. BASIC METHODOLOGIES

In this chapter, attention will be focused on the application of molecular dynamics (MD) and mixed quantum mechanics/molecular mechanics (QM/MM) techniques to the investigation of the machine-like motions that provide the energy necessary to mount catalytic barriers. We will briefly discuss and outline the main features of the two methodologies, for which more complete treatments can be found in several dedicated books and reviews [1–3].

2.1. Molecular Dynamics

In MD, successive configurations of the system under study are generated by integrating Newton's equations of motion. The result is a trajectory that specifies how the positions and velocities of the atoms in the system vary with time, as a function of the position of other atoms. The trajectory is obtained by solving the differential equations of Newton's second law

$$\frac{d^2\boldsymbol{r}_i(t)}{dt^2} = \frac{\boldsymbol{F}_i}{m_i}$$

and the force is generally calculated as the derivative of the potential function V describing different atomic interactions within the system.

$$\boldsymbol{F}_i = \frac{-dV(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots \boldsymbol{r}_N)}{d\boldsymbol{r}_i} \tag{2}$$

$$\begin{aligned} V(r_1, r_2, \ldots r_N) = {} & \sum_{\text{bonds}} \frac{1}{2} K_b [b - b_0]^2 + \sum_{\text{angles}} \frac{1}{2} K_\theta [\theta - \theta_0]^2 \\ & + \sum_{\substack{\text{improp.} \\ \text{dihedrals}}} \frac{1}{2} K_\zeta [\zeta - \zeta_0]^2 + \sum_{\text{dihedrals}} K_\varphi [1 + \cos(n\varphi - \delta)] \\ & + \sum_{\text{pairs}(i,j)} [C_{12}(i,j)/r_{ij}^{12} - C_6(i,j)/r_{ij}^6 + q_i q_j/(4\pi\varepsilon_0 \varepsilon_r r_{ij})] \end{aligned} \tag{3}$$

Under the influence of the continuous potential, the motions of the particles in the system are coupled together, giving rise to a many-body problem that cannot be solved analytically. Under these conditions, the equations of motion are integrated using finite difference methods. One of the most used algorithms for integrating the equations of motion is the leapfrog algorithm [3].

The initial velocities for the atoms are, in general, determined from a Maxwellian velocity distribution corresponding to the simulation temperature. The typical integration time step for all atom enzyme simulations is generally in the femtosecond time range. Typically,

nowadays simulations are run for hundreds of nanoseconds, requiring millions of integration steps.

Several algorithms have been developed to allow the control of temperature and pressure conditions, but their discussion is beyond the goals of this chapter. For complete references, see, for instance, (Ref. [3]).

Once the trajectory has been generated at the suitable temperature and pressure conditions, structural, statistical, and thermodynamics methods can be applied to analyze the generated trajectory and extract the main features regarding the system under study. MD simulations have been applied to a variety of problems ranging from spontaneous aggregation of lipids to protein folding processes to complex formation.

2.2. Quantum Mechanics/Molecular Mechanics Methods

The basic strategy for this approach was laid out in a seminal paper by Levitt and Warshel [4], in which a classical potential was combined with an early semiempirical quantum mechanical method (MINDO/2). Formally, the method can be described as follows (independent of the QM or MM potential function used): the Hamiltonian used within the QM/MM formulation is taken to be an effective Hamiltonian (H_{eff}), which operates on the wavefunction (ψ) of the system. The wavefunction is dependent on the position of the quantum mechanical nuclei, R_{QM}, and the molecular mechanical nuclei, R_M, as well as the positions of the electrons, r.

$$H_{\text{eff}}\psi(r, R_{\text{QM}}, R_{\text{M}}) = E(r, R_{\text{QM}}, R_{\text{M}})\psi(r, R_{\text{QM}}, R_{\text{M}}) \tag{4}$$

The effective Hamiltonian can be divided into three terms [see Eq. (5)], which are generated from the interactions that occur within and between the components of the system (see Scheme 1). The contributions considered here include the completely quantum mechanical interactions, H_{QM} (QM in Scheme 1), the purely molecular mechanical interactions, H_{MM} (MM in Scheme 1), and the interactions between the OM and MM portions of the system, $H_{\text{QM/MM}}$ (indicated by the arrow in Scheme 1).

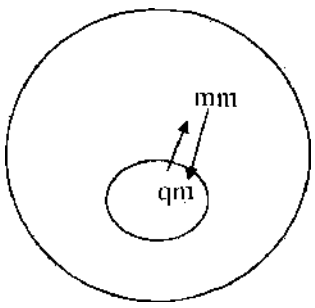
$$H_{\text{eff}} = H_{\text{QM}} + H_{\text{MM}} + H_{\text{QM-MM}} \tag{5}$$

The total energy of the system can likewise be divided into three component parts.

$$E_{\text{eff}} = E_{\text{QM}} + E_{\text{MM}} + E_{\text{QM/MM}} \tag{6}$$

These component energies can be obtained by solving either the Roothaan–Hall equations (associated with Hartree–Fock (HF)-based methodologies) or the Kohn–Sham equations (associated with density-functional-theory-based methodologies) for H_{eff}. Another way to express the total energy of the system is as the expectation value of H_{eff}. The purely MM term can be removed from the integral because it is independent of the electronic positions. Thus, the total energy of the system can be given as,

$$E_{\text{eff}} = \langle\psi|H_{\text{QM}} + H_{\text{QM/MM}}|\psi\rangle + E_{\text{MM}} \tag{7}$$



Scheme 1. Partition between QM and MM regions in an enzyme.

In Eq. (7), H_{QM} is the Hamiltonian given by semiempirical [5, 6], Hartree–Fock [7], or density-functional theory, whereas E_{MM} is an energy obtained using a classical force field of some type (see following for more discussion regarding the choices) [8–10]. The key remaining term is $H_{QM/MM}$. This term represents the interaction of the MM atom "cores" with the electron cloud of the QM atoms when interacting with MM atoms, as well as the repulsion between the MM and QM atomic cores. Finally, it was found to be necessary to add a Lennard–Jones term to the QM atoms to obtain good interaction energies, as well as good geometries for intermolecular interactions. The form of this term is

$$H_{QM/MM} = -\sum_{iM} \frac{qM}{r_{iM}} + \sum_{A} \frac{q_M Z_A}{|R_{AM}|} + \sum_{AM} \left(\frac{A_{AM}}{R_{AM}^{12}} + \frac{B_{AM}}{R_{AM}^{6}} \right) \tag{8}$$

where, q_M is the atomic point charge on the MM atom, r_{iM} is the QM-electron-to-MM-atom distance, Z_A is the core charge of QM atom A, R_{AM} is the QM-atom-A-to-MM-atom-M distance, and A_{AM} and B_{AM} are the Lennard–Jones parameters for QM atom A interacting with MM atom M. The critical term that allows the QM region to "see" the MM environment is the first term in Eq. (8), where the summation is over all interactions between MM atoms and QM electrons. This represent the core–electron interaction between MM and QM atoms and is incorporated into the QM Hamiltonian explicitly. Thus, the QM electronic structure can respond to its environment through the interaction of its electrons with the surrounding solvent/protein. The last two terms of Eq. (8) are added on to the total energy once the electronic energy has been determined by a self-consistent field (SCF) procedure and does not affect the electronic distribution of the system directly, but it does affect the geometry of the system through the computed gradients and, hence, the resulting electronic energy on subsequent SCF cycles.

The original 1976 paper describing this method by Warshel and Levitt [4] was clearly ahead of its time, as this approach was not widely used again until the late 1980s and early 1990s, when it was reexamined by several research groups using a number of different quantum mechanical methods, including semiempirical (S) [11–17], density-functional theory (DFT) [18–21], and *ab initio* methods [22–24]. Warshel also pioneered an alternative strategy to study reactive processes in solution and in enzymes, which he has termed the empirical valence bond (EVB) method [22]. This approach has been applied to a wide range of problems [22] and has some advantages and disadvantages relative to QM/MM methods that use molecular-orbital or density-functional QM methodologies. We will comment on this approach further in the next subparagraph, but the interested reader is directed toward Warshel's recent book on this subject [22]. In addition to the various potentials, a number of different sampling techniques has been used in conjunction with QM/MM studies including energy minimizations [11, 23], Monte Carlo (MC) simulations [12], and MD simulations [18, 24].

2.3. Practical Aspects

There are a number of decisions required when designing and carrying out a QM/MM study. The first is the way in which conformational sampling is to be done. In many ways the choice of sampling technique and choice of model Hamiltonian are interdependent. For example, using a purely *ab initio* approach on a large biomolecular system is too computationally expensive to effectively use a sampling technique like MD, but it is more compatible with an energy minimization approach. However, it must be kept in mind that the energy minimization techniques, although less computationally demanding because only a few thousand energy and gradient evaluations are involved, are prone to being trapped in local minima that are near the initial starting configuration. Thus, care must be exercised when using energy minimization techniques when one is interested in sampling the very rich conformational space of a protein.

To get an effective sampling of a protein system, an MC or MD approach is preferred to minimization, and, because of their generality, MD techniques are more typically used in QM/MM studies. MC methods can be very powerful tools to study solution phase systems [25], but they are not as powerful as MD techniques when it comes to the study of enzymatic

systems [22]. MD and MC methods are much more computationally demanding than energy minimization because, by their very nature, they require the evaluation of the energy and forces (MD only) of a system many thousands of times. For example, in a protein system, to sample 100 ps using an MD simulation requires at least the evaluation of the energy and forces 100,000 times, with a typical time step of 1 fs. Indeed, Merz and coworkers [27–29] have found that in some cases a timestep of 0.5 fs or less may be more appropriate in some QM/MM/MD studies of protein systems, which further increases the number of times the energy and gradient must be evaluated. Early studies using the basic QM/MM approach as applied to enzymes used energy minimization techniques [4, 26] whereas more recent studies have begun to use MD simulation techniques [27–29].

However, in all cases, the QM part of the system used a semiempirical method because of the expense associated with a fully *ab initio* or DFT method. Indeed, at this time, it is unlikely that a "long" MD simulation using an *ab initio* or DFT Hamiltonian will be carried out on a protein system, given the inherent computational expense of this type of calculation. Thus, we are at present limited to MD simulations of proteins using semiempirical models and using energy minimization (or even so-called single-point calculations, in which the energy is evaluated for a single geometric configuration) with *ab initio* or DFT Hamiltonians. Energy minimization, MD, and single-point calculations are all possible within, for instance, the ROAR program, but only for a semiempirical Hamiltonian. Future releases will include DFT and *ab initio* capabilities.

The next necessary choice is the selection of the quantum mechanical level to be employed. The choice of MM model will be discussed briefly below, but it should be pointed out that, because the QM portion of a QM/MM calculation so dominates the MM portion in computational expense and energy, the critical computational choice that will affect the speed of the algorithm has to do with selecting the QM potential and not the MM model. However, it must be noted that recent efforts to incorporate a polarized MM model into a QM/MM calculation increase the expense of the MM portion of the calculation significantly [15]. Semiempirical QM models (e.g., MNDO [30], [31], AMI [6], and PM3 [32–34]) formally scale as O(N3), where, because of semiempirical approximations, N is typically no greater than 4 for a "heavy" atom like carbon), whereas local density approximation (LDA) DFT methods also scale as O(N3), [6], but N is typically much greater than that found in semiempirical theory (e.g., for a typical good-quality basis set like 6-31G** [7], contracted basis functions are used to represent a carbon atom [7]. HF *ab initio* methods formally scale O(N4), and correlated HF approaches can scale as O(N5) or higher [7]. Thus, in terms of computational expense, semiempirical methods are the cheapest, followed by LDA-DFT, HF, and finally correlated HF. However, the accuracy of these methods roughly follows the reverse order. It is important to note that with the advent of new linear-scaling methodologies [35–37] that introduce improved algorithms for Fock matrix generation and solution, the scaling properties for QM calculations are changing, but for our purposes, the above analysis will suffice. Although in principle one would prefer to use the most accurate level of theory in all cases, this is not practical at this point in time. Thus, using contemporary computers and current numerical approaches, it is only possible to carry out "long" (e.g., 100 ps or more) MD simulations using semiempirical Hamiltonians on a typical protein active site (e.g., about 20–30 heavy atoms), whereas it is possible to carry out energy minimization studies using all QM levels except for highly correlated HF approaches.

Typically, a simple MM potential is used in QM/MM studies [8–10, 38, 39]. These potentials contain harmonic bonds and angles along with a truncated Fourier series expansion to represent the torsion potential. Nonbonded interactions (i.e., those interactions beyond the 1 and 4 positions along a polymer chain) are typically represented with a Lennard–Jones "6–12" potential function, and electrostatic interactions are handled using atom-centered charges. Examples of force fields of this type include OPLS [10], AMBER [9, 38, 39], and CHARMM [8]. This class of potential function has been quite successful in modeling protein structure and dynamics. The major deficiency in these models is the lack of explicit polarization. This typically leads to an unbalanced model in QM/MM studies because the QM region is polarized, and the MM region is not. However, recent work has gone into incorporating polarization effects into the MM region of a QM/MM study (the so-called

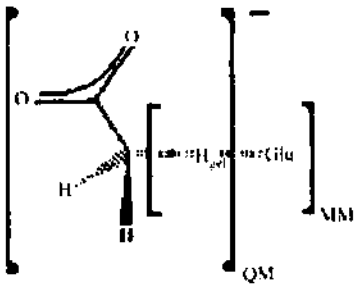
QM/MMpol approach) [15, 16]. This is a promising approach, but for now it will be confined to explicit solvent studies as opposed to studies of enzymatic systems [15, 16].

2.4. Van der Waals Parameters for Quantum Mechanics Atoms

One of the more difficult decisions that needs to be made when setting up a QM/MM simulation is the proper value for the van der Waals (VDW) parameters to be used on the QM atoms for their interaction with MM atoms [term 3 of Eq. (8)]. Although one of the strengths of QM/MM methods is their ability to accurately model a system without extensive reparameterization of either the QM or MM methodology, it has been noted empirically that a 5–10% scaling of the VDW parameters normally used within a classical force field can greatly improve the accuracy of the free energies of solvation calculated using QM/MM methods. Some researchers will also scale the MM core charges seen by the QM partition of the system [terms 1 and 2 of Eq. (19)] to improve calculated radial distribution functions and solvation-free energies calculated for a solute. Although researchers have worked to definitively establish the VDW parameters to be used in QM/MM calculations through in-depth parameterization on test systems, it must be noted that the necessary scaling factor is highly method dependent (i.e., S, DF, or HF) and is even basis-set dependent. This is unfortunate, as it requires individual parameterization for any particular combination of MM and QM levels of theory or acceptance of the 5–10% scaling factor approximation.

2.5. Link Atoms in Quantum Mechanics/Molecular Mechanics Studies of Enzymes

In solution phase reactions, the boundary between the solute (usually QM) and the solvent (usually MM) is very clear, and typically it is possible to avoid introducing the QM/MM interface between atoms that are covalently linked. However, in the case of enzymes, this is not the case, and it becomes necessary to introduce the concept of link atoms that covalently connect the QM and MM regions used in the representation of the protein. This interfacial region can be quite arbitrary and should be chosen with care. For example, consider modeling a glutamic acid residue within the QM/MM framework (see Scheme 2). The first consideration is where to make the QM and MM "cut" such that it does not adversely affect the electronic structure that we associate with a glutamate ion. If we treat the carboxylate as the formate anion, this will greatly alter the pKa relative to a typical carboxylate anion (4.0 vs. 4.5). Thus, although formate is computationally convenient, it is electronically the incorrect choice. A better choice is to add an extra carbon atom to generate the acetate anion, which has a better pKa match with the glutamate anion. The next step necessary when the QM/MM interface falls at a covalent bond is the introduction of an extra QM atom, for example, H_{QM} in Scheme 2. This is necessary to cap the exposed valence at the carbon atom such that a closed-shell calculation is done. There are several reasons for doing this, not the least of which being that the free radical of the carboxylate anion is not what we are attempting to model. Another import consideration is that closed-shell calculations are, from a technical aspect, far easier to carry out than are open-shell calculations. This so-called link atom does not see any atoms within the MM region because it is an extra QM atom solely included to satisfy valence considerations that normally would not be present in



Scheme 2. Partitioning of a Glu residue between QM and MM regions.

the real system. The final MM carbon atom of the Glu residue is then attached to the final QM carbon atom by a MM-type harmonic bond. This bond keeps the MM and QM regions at the appropriate distance from one another. The remaining angles and dihedrals present across the QM/MM boundary are typically represented using MM terms (i.e., a harmonic bond angle and a Fourier torsion potential), and care should be taken in selecting these such that the torsional properties around the bond at the QM/MM boundary are as accurate as possible (e.g., based on the appropriate QM or MM calculations). The concept of a link atom is an approximation inherent in the basic QM/MM approach of which one needs to be aware when treating protein systems or any system that requires the use of a covalent QM/MM boundary region. However, through careful selection of the location of this region, gross errors can be avoided. New and better ways need to be formulated to treat the link atom region, and indeed, some work has proceeded along these lines. In particular, the work of Rivail and coworkers using localized orbitals to represent the QM/MM boundary region looks promising, but it needs to be explored further to see how it performs in a variety of situations. At this time, however, ROAR only includes the capabilities to use the link atom approximation.

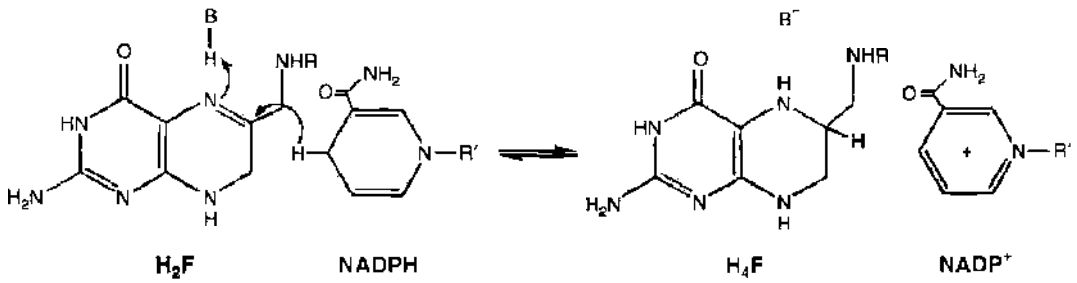
3. RELEVANT STUDIES OF ENZYME REACTIVITY

In this section, we will discuss several examples and applications of computer studies of enzyme reactivity in water and organic solvents. The examples reported represent cases that we considered particularly interesting. First, examples of studies in water are reported, followed by the description of the use of enzymes in organic solvents. The second aspect is treated more extensively as a reflection of our own expertise and experience. The fact that only a few selected studies are presented reflects our biases and knowledge of the argument.

3.1. Enzymes in Water

Enzyme systems in general imply the motions of many atoms, but the study of reaction free-energy profiles involves their projection on one single coordinate, and the transition state is identified with the configuration at the top of the free-energy barrier. Because the free-energy barrier appears in the exponential of the overall rate expression in transition state theory [40, 41], the most important contributions to enzymatic catalysis are expected to arise from the lowering of the free-energy barrier corresponding to the transition state. Computer simulations have supported this hypothesis and indicate that the reduction of the free-energy barrier mainly by electrostatic effects is a major factor in catalysis [42]. The catalytic role of enzymatic motions can be analyzed within this context; the motions influencing the activation free-energy barrier are expected to have a great effect on reaction velocity, exactly because the energy barrier is in the exponential. Numerous computational studies have identified the importance of conformational changes in enzymes reactions [43–51]. Some of these studies indicate that substrate binding influences the distribution of conformations [52, 53], and other indicate that specific modes of the protein are directly coupled to the chemical reaction [45, 54, 55]. Recent studies indicated that thermally averaged, equilibrium motions representing conformational changes along the collective reaction coordinate play an important role in enzyme reactions [49, 56–58]. These motions have the typical timescales of the enzyme reactions and reflect the conformational changes that generate transition state configurations determining reactivity and influencing the free-energy barrier.

Among the studies conduced in this theoretical framework, one outstanding example is represented by the study of dihydrofolate reductase (DHFR) by Hammes-Schiffer and coworkers [57, 58]. The reaction catalyzed by this enzyme is represented in Scheme 3 [57, 58]. Various enzyme complexes were studied by long classical MD simulations (10 ns each) to provide residue–residue-based maps of correlated motions in the backbone. These maps indicated strong correlated and anticorrelated motions involving spatially distant regions. These correlations were found in the enzyme-substrate and not in the enzyme-product complexes. Moreover, these regions included the proximal and distal regions, whose experimental perturbations led to a change in the enzyme kinetics. The authors then employed



Scheme 3. The reaction catalysed by DHFR.

QM/MM dynamics simulations to study the hydride transfer reaction catalyzed by DHFR. The hybrid simulations furnished evidence for a network of coupled promoting motions extending throughout the protein and the ligands, where promoting motions refer to equilibrium, with thermally averaged conformational changes along the reaction coordinates leading to configurations conducive to reactions. Hybrid simulations of mutants showed that mutating the network of coupled motions through structural perturbations increased the free-energy barrier of the reaction, in accordance with experimental data. The study showed that the molecular motions give rise to conformations in the enzyme complex in which the hydride transfer reaction is facilitated because of short transfer distances and suitable orientations.

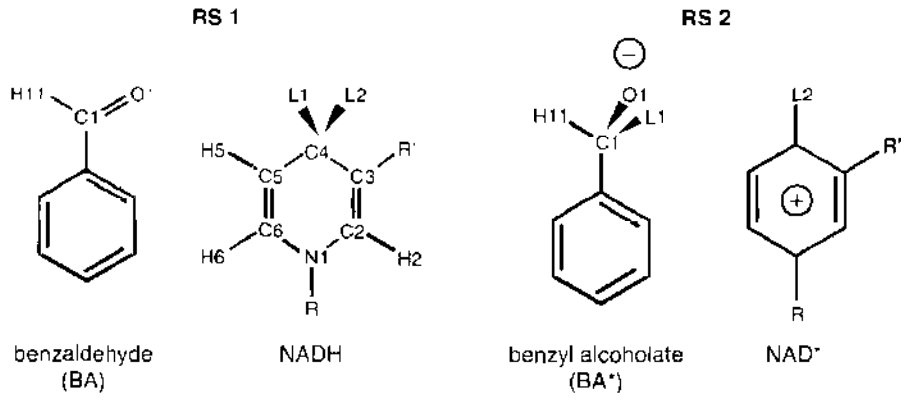
In the context of general theoretical examination of enzyme reactions, another very interesting study is represented by the investigation of the acy-lation reaction of class A beta lactamases by pennicillins [59]. Merz and coworkers, [59] in particular, concentrated on benzylpenicillin acylation of the class A TEM1 beta-lactamase via hydroxyl-only and hydroxyl/and-carboxylate-assisted processes. These mechanisms correspond to a one-step Ser130-assisted process and a to second route by which the beta-lactam carboxylate and the Ser130 hydroxyl group help the proton transfer from the hydroxyl group of Ser70 to the beta-lactam leaving N atom. The internal geometry of the reactive part of the TEMl-benzylpenicillin system is taken from a B3LYP/6 – 31 + G* computational study on the methanol-assisted methanolysis reaction of a penicillin model compound (3 alpha-carboxypenam). The 6-acylamino side chain and the 2-methyl groups of benzylpenicillin, together with the closer residues around the essential Ser70, are relaxed by carrying out geometry optimizations with a hybrid QM/MM method. The corresponding relative energies in the protein combine the B3LYP/6 – 31 + G electronic energies of the reactive subsystem with semiempirical PM3 energies of the TEMl-benzylpenicillin systems both *in vacuo* and in solution. The PM3 calculations on the TEMl-benzylpenicillin systems are performed with a divide-and-conquer linear-scaling method. The hydroxyl-and-carboxylate-assisted pathway, which is the most favored path, is in agreement with the experimentally observed kinetic isotope effects and is also compatible with the effects of mutagenesis experiments on the Ser130 residue. These results indicate that a similar mechanism for the formation of acylenzyme intermediates could be relevant to other active-site serine penicillin-recognizing enzymes.

Warshel and coworkers developed hybrid quantum mechanical molecular mechanics (QM/MM) approaches [4, 42] and, in particular, developed and advanced the EVB approach [60].

The EVB approach has been used extensively in studies of the energetics and dynamics of enzymatic reactions, including the first simulations of the dynamics of chemical reactions in enzymes [43] and solutions [61]. Here, this method is exemplified in the examination of the dynamics of the hydride transfer reaction of alcohol dehydrogenase (ADH).

This reaction is described by the two valence-bond-like (VB-like) resonance structures shown in Scheme 4. The potential surfaces of the VB structures are described by analytical potential functions that include the interactions of the charges in each VB structure with their environment. That is, the energies of the VB states are expressed as

$$\varepsilon_i = \alpha^{[i]}_{(\text{gas})} + U^i_{(\text{intra})}(R, Q) + U^i_{(\text{ss})}(R, Q, r, q) + U_{(\text{ss})}(r, q) \qquad (9)$$



Scheme 4. Resonance structures used in the EVB description of AD11.

Here, R and Q represent the atomic coordinates and charges of the VB structures, and r and q are those of the surrounding protein and solvent, is the gas-phase energy of the ith VB structure at infinite separation between the fragments; (R, Q) is the intramolecular potential of the solute system; (R, Q, r, q) represents the interaction of the solute atoms (S) with the surrounding solvent and protein atoms (s); and U (Ss) (r, q) represents the potential energy of the protein/solvent system (ss designates surrounding–surrounding). The mixing of the resonance structures gives the ground-state potential surface for the reaction in its particular environment.

The barrier for hydride transfer oscillates with the fluctuations of the environment surrounding the bound substrate. Once in a while, the environment reaches a configuration in which the barrier is small and the hydride can be transferred. Thus, the rate of hydride transfer depends on the fluctuations of the surrounding active site and the resulting fluctuations of the potential of the two VB states energies $\varepsilon 1$ and $\varepsilon 2$. A closer look at the connection between the fluctuations of the environment and the chance of a charge-transfer reaction illustrates (e.g., [42, 43, 61]) that the rate constant reflects the fluctuations of the energy gap $\Delta\varepsilon = \varepsilon 2 - \varepsilon 1$ and, in particular, the electrostatic contributions to the energy gap. However, the chance for having the proper electrostatic fluctuations probably is determined by the corresponding Boltzman probability.

EVB calculations of diverse types of enzymatic reactions (see the review in Refs. [60] and [42]) repeatedly reproduced the reduction of the activation barrier, $\Delta\Delta g^*$, when moving from water to the protein active site. The largest contribution to this $\Delta\Delta g^*$ was found to be a result of the electrostatic effects of the protein. This electrostatic effect appeared to be associated with the fact that the enzyme dipoles are preorganized toward the TS charges, and thus the reorganization energy is reduced significantly [22, 60].

These cases presented for enzymes in water are particularly representative of an ever-increasing wealth of data and analysis being applied to understanding the general principles of enzyme catalysis. These aspects are extensively discussed in several reviews, such as Refs. [1] and [2].

On the basis of our experience, the use of biomolecules as nanoscale machines mainly is in regard to the exploitation of enzyme catalytic machinery to carry out organic synthesis, which is a particularly interesting issue in the application of theoretical methods. The next part of this chapter will review developments and ideas in this field.

3.2. Enzymes in Organic Solvents

The suspension or dissolution of enzymes in nonaqueous media affords many advantages (as well as some disadvantages) over the use of proteins in aqueous solution. For example, Klibanov and coworkers have demonstrated that proteins show different thermostability [62], molecular pH memory [63], and altered substrate specificity [64] when they are placed in an anhydrous organic solvent. Furthermore, the ability of enzymes to catalyze reactions that are kinetically or thermodynamically impossible in water has also been observed [65, 66].

This led to the possibility of widespread use of enzymes as highly specific catalysts in nonaqueous solvents in the synthesis of organic compounds. For instance, a number of peptides containing D–amino acid residues were synthesized using subtilisin as a catalyst, which is impossible in water because of the enzyme's strict L-stereoselectivity [67]. The synthetic potential of this phenomenon was also exploited in the subtilisin-catalyzed acylation of carbohydrates in anhydrous dimethylformamide. Not only does subtilisin esterify a number of sugars and related compounds in this solvent but does so on a preparative scale and with a marked positional selectivity [67]. Lipases and other hydrolases in nonaqueous solvents also catalyze a variety of reactions like esterifications, transesterifications, aminolysis, and thiotransesterifications, whereas in water these processes are almost completely suppressed by hydrolysis [68]. The tyrosinase-catalyzed oxidation of phenols is another example. In water, this reaction results in negligible yields because of the rapid polymerization of the product ortho-quinones and the inactivation of the enzyme. In contrast, in chloroform, where both species are more stable, the quinones are readily obtained [69]. The possibility of applying biocatalysis to the synthesis of a wide range of compounds with high enantiomeric excesses, as well as the new thermostability characteristics that enzymes can have in organic solvents, spurred the scientific community to try and gain a deeper understanding of how protein structure and reactivity are affected by anhydrous organic solvents. The rationalization of these effects is supposed to further our ability to employ proteins as catalytic agents in nonaqueous media. This review will focus on different approaches to computationally study and rationalize the reactivity of enzymes in water and organic media. In particular, the stress will be put on reactivity issues and on the attempts to rationalize them by the application of realistic atomic-detail computational models.

3.3. Protein Conformation in Organic Media

Computer simulations have proven to be an invaluable tool in the understanding of protein structure and dynamics [70, 71]. Since the seminal work of McCammon, Gelin, and Karplus, demonstrating the dynamic nature of protein structures [70], numerous studies have appeared examining the behavior of proteins in the crystal lattice, *in vacuo*, and in aqueous solution [70]. These studies have shown that proteins possess a wealth of conformational substates of nearly equal energy [72]. The surrounding medium may, therefore, influence protein flexibility by altering the energetic barriers separating these conformational substates and thereby affect the ease with which the protein may sample these states. Merz and coworkers, in a series of papers [73], provided the first molecular-level pictures of the differential behavior of a protein in water and in a nonaqueous environment, through MD. The authors studied bovine pancreatic trypsin inhibitor (BPTI), chymotrypsin, and subtilisin in a series of different environments and showed that all of the proteins were more flexible in water than in organic solvent in a range of temperatures, and that intramolecular forces such as hydrogen bonds and salt bridges were stronger in the organic solvent simulations than in water, as shown by the increase in the number of stable hydrogen bonds. Ion pair interactions and reduced ratio of surface area to volume of protein also contributed to the stability of the proteins in hexane.

The authors concluded that the stability of the protein in organic media is a result of an increase in intramolecular stabilizing interactions such as hydrogen bonds, ion pair interactions, and reduced ratio of surface area to volume. Along the same lines, Zengh and Ornstein showed, via high-level *ab initio* calculations, that to be stable, zwitterionic ion pairs must be removed from the solvent environment and point toward the interior of the protein in a low dielectric constant medium. The relatively hydrophobic protein interior and the hydrophilic active site appear to be unaffected by the presence of the nonaqueous solvent. MD simulations thus clarify why enzymes have different structural characteristics in organic solvents, as compared to water. The higher number of intramolecular contacts, and the decreased flexibility, can also be considered an important factor in determining the lower activity of enzymes in nonaqueous media. If the enzyme is, in fact, more rigid because of all the intraprotein interactions, it will be less able to adapt to the substrates entering the active site pocket. This phenomenon is important as an explanation of the peculiar reactivity in organic solvents.

3.4. Modeling Enzyme Activity and Selectivity in Organic Media

A complete understanding of the enzyme-substrate-solvent interaction is necessary to increase enzyme utility to synthetic organic chemists [74]. Several theories have been proposed to rationalize the mechanism by which organic media can influence reactivity. In some cases, diffusional limitations have been claimed to be the main cause of different and limited reactivity of enzymes in organic solvents. However, this possibility was questioned by Klibanov and coworkers, who showed experimentally that slower catalysis was not caused by mass transfer limitations [74].

Another theory pointed out the change in conformation of the enzyme as the main cause of different reactivity compared to water. X-ray crystallography experiments on cross-linked crystal (CLC) crystal of subtilisin soaked in both acetonitrile and dioxane showed that the three-dimensional structures of these proteins were virtually identical to their counterparts in water [75, 76]. Moreover, Yennawar et al. showed that un-cross-linked chymotrypsin in hexane had the same structure as in water [77]. Through Fourier transform infrared spectroscopy, Klibanov and coworkers demonstrated that placing lyophilized subtilisin in organic solvents such as octane, acetonitrile, and dioxane had no appreciable effect on its secondary structure, as reflected by its α-helix content [74].

The attention and the efforts of many researchers were then attracted by the study of the energy of the binding of the enzyme and the substrate, and by the study of the reaction intermediate (transition state) in the presence of both the enzyme and the solvent. A model proposed by Klibanov indicated that the selectivity dependence arises because of the differences in the thermodynamics of substrate solvation, and because this model has a basis in thermodynamics, it can make some quantitative predictions. One of the first applications of this rationale involved the study of the substrate specificity of subtilisin Carlsberg in the transesterification reaction of N-Ac-L-Ser-OEt and N-Ac-L-Phe-OEt with 1-propanol in 20 different anhydrous solvents. The serine substrate was strongly preferred in some solvents (dichloromethane, chloroform, toluene, benzene, and DMF), whereas phenylalanine was preferred in alcoholic solvents (t-amyl alcohol, t-butyl alcohol, and t-butylamine).

The thermodynamic model represented specificity as a function of the solvent to water partition coefficients of the substrate [78]. However, this model was shown in a paper by our group to give only qualitative answers [79]. Other factors such as a conformational change in the enzyme molecules induced by the solvents and a differential solvent displacement from the enzyme active center by the substrate are claimed by several authors to be fundamental [79]. Thus, a more in-depth computational study was required to evaluate the influence of the solvent on enzyme reactivity, based on force-field or higher level studies of the interactions between the substrate, the enzyme and, possibly, the solvent.

Hult and coworkers [80], used information from the crystal structures of *Candida antarctica* lipase B (CALB) (incubated with the detergent Tween 80, or inhibited by a racemic mixture of n-hexylchlorophosphonate ethyl ester) to model the tetrahedral intermediates of two acyl transfer reactions into the active site to explain the high selectivity of CALB toward certain secondary alcohols. The mechanism was assumed to be identical to the general one of serine proteases. Most of these enzymes contain a Ser-His-Asp/Glu catalytic triad.

Two different model reactions were studied. The first model represented the transesterification reaction of 1-phenylethanol with an octanoate ester as acyl donor. In the second model, the hydrolysis of a glycerol butanoate derivative was studied. MD and minimization techniques were used throughout the paper. Starting models of the enzyme-substrate transition states were built as tetrahedral covalent complexes, with the Oγ atom of the active site serine residue connected to the carbonyl carbon atom of the substrate. In this reaction mechanism, the protonated active site histidine forms hydrogen bonds to the Oγ atom of the catalytic serine and to the oxygen of the alcohol group of the substrate. Furthermore, the oxyanion of the transition state was forced to interact with the oxyanion hole residues. The energies and structures of the complexes were calculated by a combined energy minimization and MD protocol [80].

In the case of 1-phenylethanol, the S-alcohol was found in such a conformation that it could not easily donate a proton to the catalytic hystidine 224 residue, which is considered

to be essential for nucleophilic attack on the acyl enzyme. As a consequence, in a hydrolysis reaction, the alcohol moiety of the substrate would be stuck in a position in which it could accept a proton from the histidine, and is therefore not efficient as a leaving group.

In contrast, the energy minima of the tetrahedral intermediate of the R-enantiomer contain a favorable hydrogen bond between the catalytic histidine and the oxygen atom of the alcohol group, which would enable rapid proton transfer for catalysis. In the glycerol derivative hydrolysis study, the fast-reacting R-enantiomer was found to have all its bonds in the carbon chain of the large group in the relaxed trans conformations. The small group was interacting with Trp104 in the small pocket. In contrast, the slow-reacting S-enantiomer had its large group twisted away from the specificity pocket, in contact with the side chains of Leu278, Ala282, and Ile285. The presence of these side chains forced the carbon chain of this group of the substrate to adopt an energetically less favorable conformation. It was hypothesized that the enantiopreference for the R-enantiomer in the hydrolysis of glycerol butanoate derivatives was caused by these strain differences in the substrate conformations.

The results of this study (together with x-ray data) provided structural explanations for the high stereselectivity of CALB toward secondary alcohols that can be extended to more general cases. All the MD and energy minimizations performed by Hult and coworkers, however, were run *in vacuo* and gave only structural-sterical information about the interaction between the enzyme and the tetrahedral intermediate. Haeffner et al. [81] took a step forward, proposing an MD-based molecular modeling method to predict the enantioselectivity of lipase-catalyzed transesterification reactions [81]. A tetrahedral intermediate with a rigid central part was used to mimic the transition state. The geometric parameters were calculated *ab initio* on the tetrahedral intermediate formed in the gas phase reaction between methyl acetate and a methoxyde ion, using the $6-31+G^*$ basis set, and the charges were calculated with an ESP fitting methodology. The same set of charges was used for both enantiomers in all the three cases.

After MD conformational steps and energy refinement of the stored structures, enantioselectivity was expressed as a function of the energy differences between the two diastereoisomeric enzyme-substrate complexes. Despite using the same set of charges on two different diastereomeric complexes, this approach was able to predict which enantiomer was the fast-reacting one. However, the issue of solvent influence, along with that of the electrostatic influence of the enzyme on reactivity, was not examined. This problem could be tackled by the use of mixed QM/MM methods, which were introduced by Warshel and Lewitt [4]. Using this approach, the reactive events can be described at a QM level for the active site, whereas the rest of the protein is treated by MM or MD calculations. The two regions can interact through an interaction Hamiltonian: in this way, the active site (i.e., the QM region) can respond to the influence of its protein and solvent environment in terms of polarization, electron correlation, and charge transfer effects.

In 1991, Daggett and coworkers [82] described semiempirical molecular orbital calculations of serine protease–catalyzed hydrolysis of amides and esters (Scheme 1) and the effect of the protein environment and dynamics on the process. They found that the PM3 Hamiltonian was much better, compared to AM1, at reproducing hydrogen bond geometries, and so the former was used to describe the reaction mechanism. The lowest-energy pathway for formation of the tetrahedral intermediate was for serine to approach the substrate, followed by coupled heavy atom motion and proton transfer to complete the reaction. The importance of the active site residues and the environment in stabilizing the substrate was addressed by evaluating particular interactions in fully solvated enzyme–substrate models. The interaction energies were calculated for both a noncovalent enzyme–substrate complex and a model for the tetrahedral intermediate in which a covalent bond was formed between the serine and substrate. The oxyanion hole and Asp greatly stabilized the active site region of the tetrahedral intermediate, and presumably the transition state. The environment itself, excluding the active site residues and the oxyanion hole, was important in stabilizing the scissile bond of the substrate in the tetrahedral intermediate versus the Michaelis complex.

This was certainly one of the first important reports of the use of information gained from QM simulations in elucidating mechanistic pathways in which the role of the protein environment was not negligible. QM/MM approaches were used successfully to describe

the structural and reactivity characteristics of metal-containing enzymes, to design a charge model to describe the active site of these enzymes, and finally, to describe the binding preferences of inhibitors to Zn(II) containing proteins [27]. Williams and coworkers were able to describe and characterize the transition state structure for the reduction of pyruvate, catalyzed by lactate dehydrogenase (LDH), by means of QM/MM calculations involving a fully flexible active region comprising 1900 atoms and 5700 degrees of freedom [83].

A striking application of this methodology to describe enzyme reactivity is from Warshel's group [84]. They developed an effective approach for *ab initio* calculations of activation energies in enzymatic reactions. Their approach used an empirical valence bond [85] potential surface as a reference potential for evaluating the free energy of a hybrid *ab initio* QM(ai)/MM potential surface. This procedure involved an authomated calibration of the EVB potential using gas-phase *ab initio* calculations. This method was used to study the nucleophilic attack step in the catalytic reaction of subtilisin. It was found that the use of the EVB potential as a reference allowed one to obtain the actual activation free energy of enzymatic reactions.

Amide hydrolysis in trypsin and water solution was studied by Stanton et al. [86], who proposed a new, more general method to combine *ab initio* QM calculations with the classical mechanical free-energy perturbation (FEP) approach to calculate the energetics of enzyme-catalyzed reactions and the same reaction in solution. This method, enabling enzyme and solution reactions to be compared without the use of empirical parameters, was applied to the formation of the tetrahedral intermediate in trypsin. Two critical aspects of this new approach were the calculation of the reaction energetics in solution and the use of the restrained electrostatic potential (RESP) [9] protocol to calculate the charge distributions of structures along the reaction pathway, which enabled the authors to circumvent the problem of partitioning the charge across a residue that is being divided into QM and MM parts. These charges were then used along with standard parameters to define the force field representations of the complex in the FEP calculations. Two sets of free energy calculations were performed to obtain the differences in free energy between points on the reaction pathway. In the first set, the QM region was static during the entire length of the simulation. In the second set of simulations, internal degrees of freedom of the QM region were allowed to relax to respond to the dynamics of the enzyme. In this way, the effect of active site flexibility on free-energy calculations can be assessed. This protocol allows us to examine the nature of enzyme catalysis by performing analogous simulations in water. Comparison of these two simulations gives insights into the role of the enzyme as a catalyst. This method is certainly general and is applicable to the investigation of the energetics of several enzyme reaction pathways. The use of *ab initio* calculated charges to describe enantioselectivity was described by Ke, Tidor, and Klibanov [87].

The transition states for the acylation and deacylation of chymotrypsin by substrates 1–4 of Figure 1 were approximated by the corresponding transition states. In this study, they determined the substrate charge distribution (determined via STO-3G electrostatic potential fits) for the *R* and *S* substrates, using an ensemble of 10 energy-minimized structures obtained from gas-phase MD simulations. Then they docked the substrates into the enzyme active site and used vacuum (modeled using a distance-dependent dielectric) MD simulations followed by energy minimization to obtain suitable structures of the enzyme/substrate

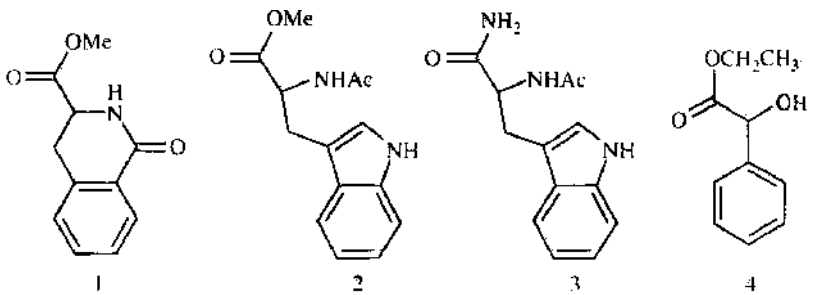


Figure 1. Substrates used by Ke et al., reported in Table 1 (Ke et al., 1998).

complexes. To mimic the presence of water, the authors used continuum electrostatic models, and the $\Delta\Delta G^*$ activation energy differences in water between the *R* and *S* substrates were calculated according to Eq. (10)

$$\Delta G_{R-S}^{\text{water}} = \Delta E_{R-S}^{\text{covalent}} + \Delta E_{R-S}^{\text{vdW}} + \Delta G_{R-S}^{\text{elec}} + \Delta G_{R-S}^{\text{nonpolar}} \tag{10}$$

where the first term is the difference in the bonded energy terms and the second is the difference in the VDW energy for MM energy-minimized structures of the two intermediates. The two ΔG values in the second term of the equation are the electrostatic contributions in continuum solvent consisting of coulombic and reaction field contributions, and the nonpolar term is computed from the solvent-accessible surface area difference between the *R* and *S* enzyme-bound tetrahedral intermediates. The results were in good agreement with experimental values, as reported in Table 1. This study was a significant advance over earlier efforts, but significant improvements could still be made, especially through the use of explicit solvents and the determination of atomic point charges that included the effect of the enzyme environment. This had been done in the past on ester cleavage by a β-cyclodextrin, but the first application to enzyme systems was by Colombo et al. [88]. Their study focused on garnering a deeper understanding of the enantioselectivity of the serine protease subtilisin in the organic solvent dimethylformamide (DMF), in the resolution of a racemic mixture of sec-phenethyl alcohol by a transesterification reaction with the acylating agent vinyl acetate.

Once again, the formation of the tetrahedral intermediate is thought to be the rate-determining step for catalysis (Fig. 2). Moreover, it is thought that the structure of the transition state for the formation of the tetrahedral intermediate closely resembles the structure of the tetrahedral intermediate itself. The behavior and the characteristics of the tetrahedral intermediates for both the *S* and *R* enantiomers were examined through the use of MD and FEP simulations, with an explicit representation of the organic solvent DMF. A critical aspect of this study was the determination of the charge distribution of the two (*S* and *R*) tetrahedral intermediates through the use of a mixed QM/MM electrostatic potential fitting methodology. In designing the active site model, the *S* and *R* tetrahedral intermediates were found to have significantly different charge distributions as a result of the presence of the stereodifferentiating environment presented by the enzyme.

In contrast, the charge distribution obtained for models of the tetrahedral intermediate in the gas–phase displays similar charge distributions. From MD simulations, both steric and, more important, electrostatic complementarity were seen to play a role in the determination of the enantios-electivity of this enzyme reaction. Table 2 displays the differential charges on the tetrahedral intermediates for the two enantiomers and on the active site residue atoms. These point charges were then used for FEP simulations to obtain a free-energy difference that was in good accordance with experiment (Table 2), which quantitatively supports the accuracy of the model and indicates that all-atom molecular simulations are capable of providing accurate qualitative and quantitative insights into enzyme catalysis in nonaqueous solvents, as well as to enzyme catalysis in general.

This approach was extended to study the differential charge distribution between the enantiomers in water and hexane as were (Colombo et al. [68]). Steric and electrostatic factors are shown to play important roles in determining asymmetric induction and selectivity

Table 1. Comparison of experimentally determined enantioselectivities in the chymotripsin catalyzed hydrolyses of Substrates 1–4 of Fig. 1 with simulation results *in vacuo* and in aqueous solution [87].

	$\Delta\Delta G^{\ddagger}$ (kcal/mol)		
		Calculated	
Substrate	Experimental	In vacuo	In water
1	−4.9	−4.4	−4.5
2	>7	6.3	5.3
3	>7	8.0	8.7
4	0.1	1.2	0.1

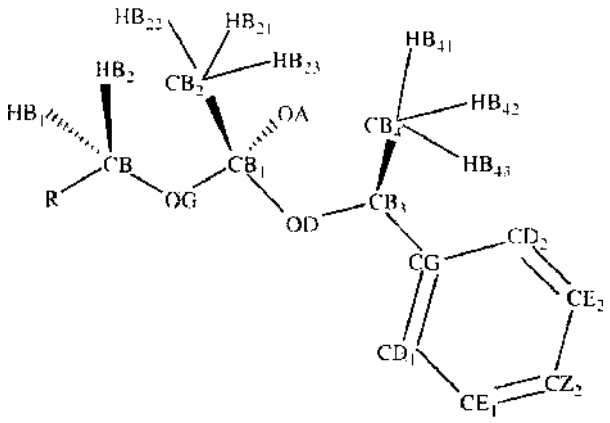


Figure 2. Tetrahedral intermediate for the Subtilisin-catalized transesterification of *sec*-phenethyl alcohol with acetate: see Table 2 (Colombo et al., 1999).

for enzymes in different environments. Steric factors determine the way different groups can fit into the pockets of the enzyme, and electrostatic factors can determine and tune the nucleophilicity and basicity (or charge concentration) of the incoming (hexane and DMF) or leaving (water) alcoholic group. The differential charge distributions on the important atoms for reactivity are reported in Table 3.

The solvent, in turn, promotes differential solvation of groups in different environments. The apolar phenyl group tends to be removed from water for both enantiomers, and in organic solvents only the phenyl group of the *R*-complex is exposed at the expense of hydrogen bonding interactions. Moreover, the solvents can determine differential charges on analogous atoms of the same enantiomer in different solvent environments, through polarization and charge transfer effects. These effects have recently been proven to be very important in the solvation of biomolecules [89].

Table 2. Average ESP charges calculated for the *R* and *S* complexes of Fig. 2 [88].

	sub *R* chg ac	std dev	sub *S* chg ac	std dev	charge diff (Δ)
CB	0.056	0.062	−0.029	0.063	−0.027
HB2	0.081	0.027	0.064	0.029	0.017
HB3	0.035	0.023	0.042	0.036	−0.007
OG	−0.305	0.061	−0.351	0.033	0.047
CB1	0.588	0.106	0.697	0.101	−0.108
CB2	0.331	0.065	−0.297	0.067	−0.034
HB21	0.075	0.038	0.045	0.033	0.030
HB22	0.092	0.027	0.046	0.026	0.046
HB23	0.060	0.027	0.063	0.034	−0.003
OA	0.876	0.036	−0.914	0.038	0.037
OD	−0.462	0.073	−0.543	0.065	0.081
CB3	0.205	0.118	0.464	0.118	−0.258
HB3	0.045	0.059	−0.044	0.038	0.089
CB4	−0.277	0.051	−0.222	0.059	−0.054
HB41	0.052	0.022	0.042	0.027	0.010
HB42	0.065	0.024	0.033	0.031	0.032
HB43	0.095	0.027	0.050	0.017	0.045
CG	0.055	0.071	−0.047	0.075	0.103
CD1	−0.154	0.048	−0.069	0.053	−0.085
HD1	0.095	0.020	0.126	0.019	−0.031
CE1	−0.109	0.054	−0.142	0.049	0.033
HE1	0.101	0.021	0.126	0.021	0.025
CZ	−0.114	0.052	−0.116	0.042	0.001
HZ	0.097	0.020	0.093	0.023	0.004
CE2	−0.107	0.037	−0.131	0.034	0.024
HE2	0.111	0.014	0.100	0.011	0.011
CD2	−0.125	0.051	−0.139	0.045	0.014
HD2	0.117	0.013	0.099	0.015	0.018

Table 3. Average charge values in three different solvents for the reactive intermediate of Fig. 2.

	H2O		Hexane		DMF	
ATOM	S	R	S	R	S	R
CB3	0.36 + 0.11	0.19 ± 0.09	0.43 ± 0.13	0.25 ± 0.08	0.46 ± 0.11	0.21 ± 0.11
OD	−0.53 ± 0.05	−0.45 ± 0.03	−0.50 ± 0.07	−0.41 ± 0.04	−0.55 ± 0.05	−0.46 ± 0.07
OA	−0.88 ± 0.03	−0.91 ± 0.01	−0.90 ± 0.03	−0.89 ± 0.03	−0.91 ± 0.04	−0.87 ± 0.03
OG	−0.33 ± 0.04	0.38 ± 0.02	−0.33 ± 0.04	0.31 ± 0.05	−0.35 ± 0.03	−0.31 ± 0.05
CB1	0.67 ± 0.06	0.61 ± 0.04	0.66 ± 0.07	0.53 ± 0.05	0.70 ± 0.01	0.58 ± 0.10
CC	0.11 ± 0.06	0.06 ± 0.06	−0.02 ± 0.10	0.04 ± 0.07	−0.04 ± 0.07	0.05 ± 0.07
CB4	−0.25 ± 0.07	−0.17 ± 0.08	−0.22 ± 0.04	−0.21 ± 0.05	−0.22 ± 0.06	−0.27 ± 0.05
HB3	−0.002 ± 0.04	0.05 ± 0.03	−0.03 ± 0.04	0.01 ± 0.03	−0.04 ± 0.04	0.04 ± 0.05
ND1	−0.16 + 0.04	−0.14 ± 0.04	−0.10 ± 0.06	−0.06 ± 0.06	−0.14 ± 0.06	−0.11 ± 0.05
NE2	0.08 ± 0.08	0.005 ± 0.05	0.05 ± 0.07	−0.02 ± 0.08	0.02 ± 0.07	−0.08 ± 0.05
HD1	0.36 ± 0.03	0.34 ± 0.03	0.32 ± 0.03	0.32 ± 0.03	0.34 ± 0.03	0.34 ± 0.03
HE2	0.22 ± 0.05	0.29 ± 0.03	0.23 ± 0.06	0.27 ± 0.03	0.26 ± 0.04	0.26 ± 0.03

Electrostatic stereodifferentiation can be important in the determination of the $\Delta\Delta G^*$ between the two transition states leading to enantiomeric resolution. Through QM/MM electrostatic potential fitting (ESP) calculations, the authors have been able to rationalize the electrostatic factors resulting from the protein chiral auxiliary on the two enantiomers. A complex interplay of factors determines selectivity and reactivity in different solvents. Whether one factor is more important than any other is still a subject of investigation and research. Nonetheless, the application of QM/MM simulation techniques can be a valuable and precious help in designing better catalyst systems to be used in organic synthesis: Identifying important steric and charge differences through simulation techniques and magnifying them by solvent engineering or molecular biology (directed evolution or site directed mutagenesis) can lead to the expression of better optimized catalysts. The critical aspect of all the QM/MM approaches reviewed here is that they allow the explicit treatment of the atomic determinants of the enzyme reaction studied. These methods have been applied with reasonable success to a wide variety of problems from the study of hydrolases to the study of metal-containing enzymes. Moreover, the use of quantum mechanics was necessary to highlight the fact that enzymes, in their catalytic activity, create a stereodifferentiating environment not only from the steric but also the electrostatic point of view. Calculating the modulation of charge distributions by solvent effects has also been a critical advancement in the study of enzyme reactivity. The use of these methods, opens the way to the application of atomic detail computational methods to systems containing a high number of atoms in realistic conditions (temperature, solvents, pressure, etc.), leading to the possibility of a very precise definition of all the factors influencing their reactivity.

4. APPLICATION OF SIMULATIONS TO REAL-WORLD CASES

In this paragraph we consider case studies in which the interplay of experimental approaches and computer modeling technologies has been applied to define possible reaction mechanisms and to drive the development of better reaction conditions for biocatalytic reactions. The first application goes more in depth in the problem of enantioselectivity modulation by organic solvents (Section 4.1), already pointed out in the previous paragraphs. The second will deal with the problem of enzyme reaction regioselectivity (Section 4.2). The third will end with the determination of stability determinants in proteins, with the final aim of designing new proteins with improved stability properties (Section 4.3).

4.1. Rationalizing Enantioselectivity in Organic Solvents Through Quantum Mechanics/Molecular Mechanics Approaches

The study of enzyme selectivity in nonaqueous solvents is a particularly intriguing field of research. In these media, the absence of a continuous aqueous layer around an enzyme makes it possible for the enzyme to interact directly with the nonaqueous solvent, which

results in modifications of the properties of the enzyme. Specifically, the biocatalyst can attain new properties in terms of stability, activity, and specificity/selectivity [62]. More important, in organic solvents, enzymes such as hydrolases and proteases can catalyze esterification and transesterification reactions readily with high product yields [90]. However, a complete understanding of the enzyme-substrate-solvent interaction is necessary to increase their utility to synthetic chemists [91, 92]. As we saw earlier, several theories have been proposed to rationalize the mechanism by which organic media influence enzymatic reactivity.

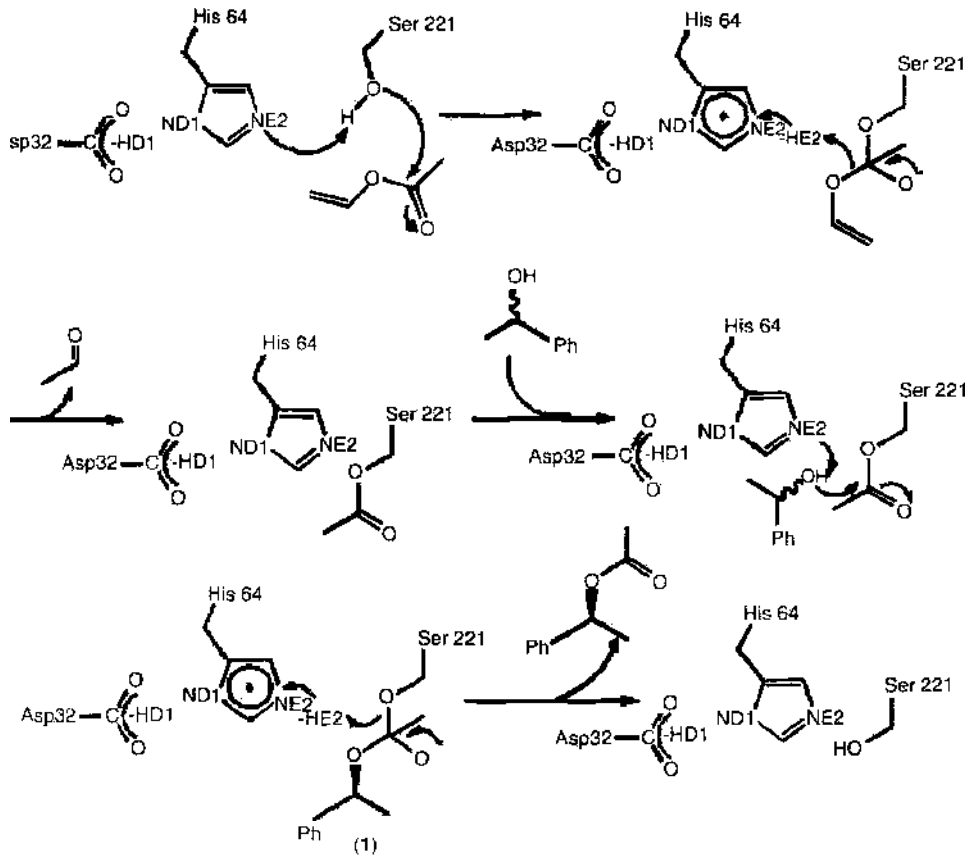
Previous work from our laboratory has focused on the effect of nonaqueous solvents on protein structure and dynamics [93, 94]. To extend this work to how enzyme function is affected, we became interested in enantioselectivity in organic solvents (as well as in water). Enantioselectivity is an appealing characteristic of enzyme catalysts, and we were interested in investigating experimentally and theoretically the mechanism by which the solvent environment might influence enzyme enantioselectivity. The ultimate goal of this work is to aid in the rational design of biocatalytic systems suitable for organic synthesis.

In particular, we were interested in garnering a deeper understanding of the enantioselectivity of the serine protease subtilisin in the organic solvent DMF. The reaction of interest is shown in Scheme 5. This is the resolution of a racemic mixture of *sec*-phenethyl alcohol by a transesterification reaction with the acylating agent vinyl acetate, catalyzed by subtilisin in anhydrous DMF. The enantioselectivity of an enzyme following Michaelis–Menten behavior can be expressed in terms of the parameter E (which depends on the $(k_{cat}/K_M)_S/(k_{cat}/K_M)_R$ ratio) [95] which, in turn, can be related to the free energy $\Delta\Delta G^*$ by the expression [96]

$$\Delta\Delta G^* = -RT \ln E$$

Because two enantiomeric substrates in an achiral environment (DMF) have the same ground-state free energy, $\Delta\Delta G^*$ can be expressed as the free-energy difference between the rate-determining transition states leading to the S and R products

$$\Delta\Delta G^* = \Delta G_{S-R}$$



Scheme 5. Enantioselection of *sec*-phenethyl-alchol by Subtilisin in organic solvent.

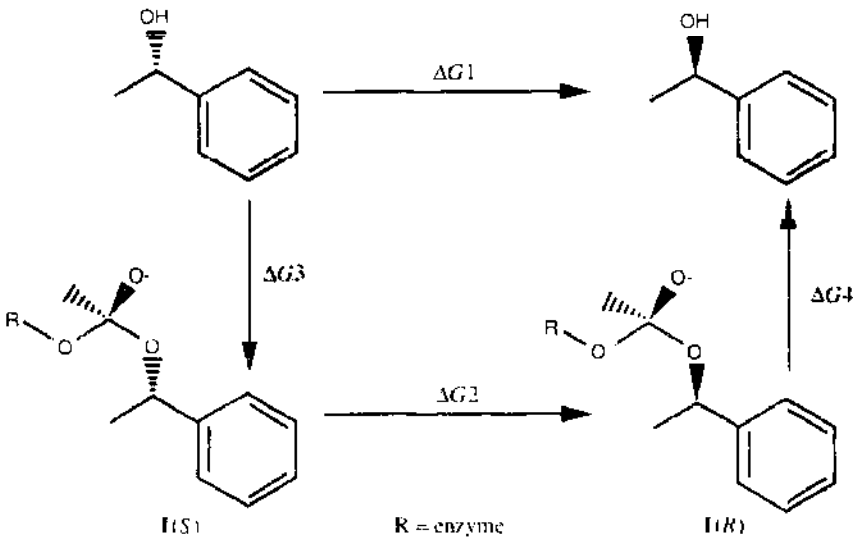
The experimental value for enantioselectivity (E) of the reaction shown in Scheme 5 was found by chiral GC to be 2.2 ($\Delta\Delta G^{\ddagger} = \Delta G_{(S\ R)} = 0.4$ kcal/mol), with the S enantiomer reacting faster than the R enantiomer [97].

The formation of the tetrahedral intermediate is thought to be the rate-determining step in the catalysis of serine proteases [99]. Moreover, it is thought that the structure of the transition state for formation of the tetrahedral intermediate closely resembles the structure of the tetrahedral intermediate itself [99]. Thus, in the past, workers interested in studying serine proteases used the structure of the tetrahedral intermediate to approximate the transition state structure [87, 100]. In this chapter, we determine the free-energy difference between the two tetrahedral intermediates I(R) and I(S) given in Scheme 6, which have been used to represent the R and S transition states for the transesterification of *sec*-phenethyl alcohol catalyzed by subtilisin (see Scheme 5). We also report studies that address the role that structural, energetic, and electrostatic factors play in influencing the enantioselectivity of this reaction when carried out by subtilisin in DMF.

In particular, the importance of the electrostatic interactions in the differentiation between two enantiomers, to our knowledge, has not been analyzed in detail. Clearly, two enantiomeric molecules in an achiral environment must have the same charge distribution on the corresponding atoms. The same enantiomeric substrates, complexed or bound to an enzyme, however, experience a chiral environment, which gives rise to two diastereoisomeric complexes. In this case, analogous atoms of the two substrates will be perturbed by different electrostatic fields, generated by different electrostatic environments. This "electrostatic stereodifferentiation" is certainly a factor in the determination of the free-energy difference defining enzyme enantioselectivity. The availability of QM/MM ESP fitting methods [11–101] has allowed us, for the first time, to carefully take into consideration this aspect of enzymatic reactivity by explicitly considering polarization and charge-transfer effects. In particular, we show differences in the charge distributions on the atoms of the substrate and on the atoms of the catalytically important residues and differential fluctuations of these charges arising during MD simulations. Most critically, the accurate treatment of the electrostatic problem has allowed us to evaluate a $\Delta\Delta G^{\ddagger}$ value that is in good accord with the experimentally determined enantioselectivity value.

4.1.1. Computational Methods

The starting enzyme coordinates were obtained from the crystal structure (PDB entry 1 sen) [102]. The intermediates, I(S) and I(R), were linked to the Oγ (OG in Fig. 2) of Ser-221. The O^- atom of the intermediate was placed in the oxyanion hole defined by the side chain $CONH_2$ group of Asn-155 and the backbone NH group of Ser-221. Three chloride atoms



Scheme 6. Thermodynamic cycle used for the free-energy perturbation calculations in subtilisin reactivity calculations, Section 3.1.

were added to neutralize the excess charge on the enzyme/substrate complex. The 50 most strongly bound water molecules (as determined from experimental *B*-factors) were retained. This number of water molecules has been found to be approximately the amount of water required for the protein to retain its catalytic activity in nonaqueous environments [67, 103]. The enzyme/substrate complex was then solvated in a cubic box of 1515 DMF molecules. The solvent and the enzyme were modeled using the AMBER force field [9]. All-atom AMBER parameters were used for the enzyme/substrate complex [9], TIP3P for the water molecules [104], and OPLS [105] united-atom parameters for the DMF. This system was then AMBER [32] minimized using ROAR 1.0 [106]. The AMBER-minimized system was then QM/MM minimized using ROAR 1.0 [106]. The QM region included the substrate and the residues involved in catalysis (Ser-221, Asp-32, His-64, Asn-155), and the PM3 Hamiltonian [36] was used. The total charge of the QM region was set to −1. The solvent and the rest of the enzyme were treated at the MM level, using the potential functions described above.

The structures of the complexes obtained after the first QM/MM minimization were used to evaluate the atomic charges of the intermediate and of the active site residues in the enzyme-bound state, with QM/MM (MNDO Hamiltonian) [101, 107] methods coupled with ESP fitting (the total charge of the QM system fixed to −1). We used the MNDO Hamiltonian for ESP fitting because MNDO has been shown to give ESP fitted charges that are well correlated to HF/6 − 31G^* ESP-derived charges, whereas PM3 does not [101, 107]. It is important to notice that the set of charges obtained from these calculations includes the influence of the enzyme solvent environment on the substrate. Using the charges obtained in this way, we ran a MD equilibration of the enzyme/substrate complex in a DMF solvent box.

The MD simulations were performed using the SANDER module of AMBER [108]. The temperature of the system was slowly raised from 0 to 313 K (the experimental reaction temperature) over 9 ps, followed by equilibration for 120 ps at 313 K and at a constant pressure of 1 atm. The temperature and pressure were controlled using the methods of Berendsen and coworkers [109]. Periodic boundary conditions and a time step of 1.5 fs were used in all the simulations. In all cases, the bond lengths were constrained using the SHAKE algorithm with a tolerance of 0.0005 Å [110]. The structure obtained at 120 ps was used to recalculate the charges on the groups in the QM region, as described previously. The final results are summarized in Table 2 for the substrate, Table 4 for Asp 32, Table 5 for His 64, and Table 6 for Asn 155. The final charge models were then used for all subsequent MD (one run of 300 ps total) and FEP simulations.

The free-energy calculations were carried out using the GIBBS module of AMBER 5 [108]. We used the slow-growth method, coupled with the dual topology representation for the groups undergoing changes during the FEP simulation [111]. Thus, the topologies for both the *R* and *S* enantiomers were simultaneously defined, and as the slow-growth FEP simulation proceeds, the phenyl group of the *S* enantiomer of *sec*-phenethyl alcohol slowly disappears, while the corresponding group of the *R* enantiomer slowly appears. A similar run on intermediate **II** (Scheme 7) was carried out in the gas phase (using gas-phase MNDO-calculated charges) and in DMF to better quantitate the effect the enzyme environment has on the atomic point charges relative to those obtained in the gas phase.

Table 4. Calculated charges for Asp 32 using QM/MM calculations.

Atom	AMBER	First QM/MM	120-ps QM/MM
CB	−0.0303	(*S*) −0.3453	(*S*) −0.3534
		(*R*) −0.3523	(*R*) −0.3274
HB2	−0.0122	(*S*) 0.1034	(*S*) 0.1268
		(*R*) 0.1034	(*R*) 0.0737
HB3	−0.0122	(*S*) 0.0915	(*S*) 0.0979
		(*R*) 0.0915	(*R*) 0.0922
CG	0.7994	(*S*) 0.7103	(*S*) 0.7035
		(*R*) 0.7150	(*R*) 0.6865
OD1	−0.8014	(*S*) −0.7494	(*S*) −0.7810
		(*R*) −0.7587	(*R*) −0.7467
OD2	−0.8014	(*S*) −0.7056	(*S*) −0.7552
		(*R*) −0.7116	(*R*) −0.7117

Table 5. Calculated charges for His 64 using QM/MM calculations.

Atom	AMBER	First QM/MM	120-ps QM/MM
CG	−0.0012	(*S*) 0.0974	(*S*) 0.1274
		(*R*) 0.0955	(*R*) 0.1569
ND1	−0.1513	(*S*) −0.0430	(*S*) −0.1812
		(*R*) −0.0679	(*R*) −0.2278
HD1	0.3866	(*S*) 0.2939	(*S*) 0.3700
		(*R*) 0.3025	(*R*) 0.4014
CE1	−0.0170	(*S*) 0.0369	(*S*) −0.0071
		(*R*) −0.0161	(*R*) 0.0150
HE1	0.2681	(*S*) 0.2827	(*S*) 0.2963
		(*R*) 0.2961	(*R*) 0.2538
NE2	−0.1718	(*S*) −0.1093	(*S*) 0.1853
		(*R*) 0.0366	(*R*) 0.1038
HE2	0.3911	(*S*) 0.3042	(*S*) 0.1605
		(*R*) 0.2155	(*R*) 0.1910
CD2	−0.1141	(*S*) −0.2033	(*S*) 0.2759
		(*R*) −0.2219	(*R*) −0.2660
HD2	0.2317	(*S*) 0.2461	(*S*) 0.2528
		(*R*) 0.2572	(*R*) 0.2667

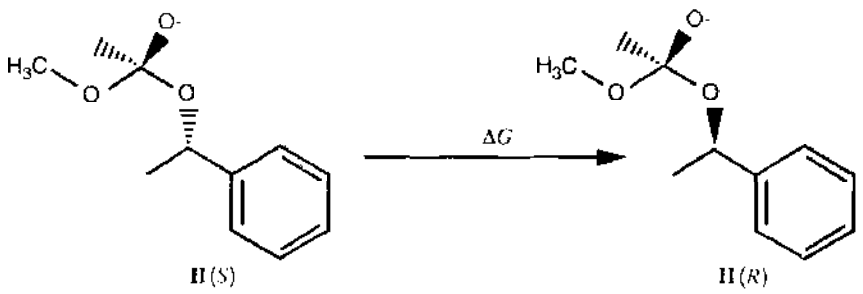
The FEP simulations were carried out over three different timescales (120, 450, and 750 ps), starting from the structure obtained after another 90 ps of equilibration, using the charges obtained after 120 ps of MD simulation (see earlier for details). The reason for using three different FEP simulation lengths was to determine whether the simulations were converged. We also tested convergence by starting the FEP simulations from a different starting structure obtained after another 30 ps of equilibration (120 ps total) on the structure and with charges obtained from the 120 ps MD simulation described earlier. In this case, the calculations were only carried out over two different timescales (450 and 750 ps). Periodic boundary conditions, a 1.5-fs time step, a constant T of 313 K, and a constant P of 1 atm, were used in all FEP simulations [109]. The SHAKE algorithm was used to constrain all bonds with a tolerance of 0.0005 Å [110].

4.1.2. Results and Comparison with Experiments

A primary aim of this study was to examine the structural factors that influence the stereoselectivity of subtilisin in DMF. One we can focus our attention on is the hydrogen-bonding pattern between the substrate and the active site of the enzyme. We determined the average values of the parameters of interest over the last 150 ps of a 300-ps MD simulation. As can be seen from Fig. 3, in the case of the *S* complex, the HE2 of His-64 forms hydrogen bonds with both the OD and OG oxygens of the enzyme/substrate complex. In the case of the *R* complex, the hydrogen bond between OD of the substrate and HE2 of His-64 is weakened, as indicated in Fig. 4. Hence, in the case of the *R* complex, this interaction is completely missing during the MD simulation, while it is present in the *S* complex. The distance analysis

Table 6. Calculated charges for Asp 32 using QM/MM calculations.

Atom	AMBER	First QM/MM	120-ps QM/MM
CG	0.7130	(*S*) 0.6157	(*S*) 0.5491
		(*R*) 0.6223	(*R*) 0.5861
OD1	−0.5931	(*S*) −0.5902	(*S*) −0.5719
		(*R*) −0.6140	(*R*) −0.6207
ND2	−0.9191	(*S*) −0.7413	(*S*) −0.6941
		(*R*) −0.7488	(*R*) −0.6787
HD21	0.4196	(*S*) 0.4197	(*S*) 0.4036
		(*R*) 0.4317	(*R*) 0.4041
HD22	0.4196	(*S*) 0.2637	(*S*) 0.3206
		(*R*) 0.2761	(*R*) 0.2645



Scheme 7. Free-energy perturbation of the tetrahedral complex carried out *in vacuo*.

also shows that the interaction between HD1 of His-64 and the two oxygens of Asp-32 are equally favorable for the *S* and *R* complexes.

From the structural analysis, we also noticed that once the complex was formed and equilibrated, the phenyl group of the *S* substrate fit very nicely into a hydrophobic pocket defined by residues Leu-126, Gly-127, Gly-128, and Asn-155. In the case of the *R* enantiomer, this region is occupied by a DMF molecule, and the phenyl ring is oriented toward the surrounding solvent. The methyl group of the alcoholic moiety of the *R* complex is now pointing into this hydrophobic pocket, and in total, these arrangements contribute to disrupt the catalytically essential hydrogen bond (OD to HE2) to His-64. These observations suggest that, in the transesterification reaction, the *R* alcohol cannot readily donate its proton to the catalytic histidine residue (His-64), which is essential for the catalysis [111]. This, in itself, favors the reactivity of the (*S*)-*sec*-phenylethyl alcohol over that of the corresponding *R* enantiomer.

This hypothesis is further confirmed by an analysis of the RMS fluctuation per residue applied to the active site amino acids. Table 7 shows that Asp-32, His-64, and Ser-221 in the *S* complex were more flexible than the corresponding amino acids in the *R* complex. This higher flexibility may favor reactivity by allowing the catalytic residues to adopt a more favorable orientation for reaction in the case of the *S* complex. An examination of the solvent-accessible surface areas (SASA) for the catalytically important amino acid residues (in particular, Ser-221 bound to the intermediate) indicates that, in the case of the *R* complex, the substrate is more exposed to solvent (Table 8). This SASA calculation supports the previous observation of a differential penetration of the solvent into the active site region. In particular, in the case of the *R* complex, the solvent occupies the hydrophobic pocket occupied by the substrate phenyl ring in the *S* complex case. In Table 2 the various ESP charges used in this study for the *R* and *S* substrates are given. We observed significant differences between the QM/MM ESP charges for enzyme-bound *R* and *S* substrates. On further reflection, we determined that the differences observed in the *R* and *S* substrate charges arise because of the presence of a chiral environment that allows each diastereomeric enzyme/substrate complex to experience a differing electrostatic environment. Thus, the solvent and enzyme atoms can have an "electrostatic influence" on the enzyme enantioselectivity by presenting an electrostatic field that alters the charge distribution of the

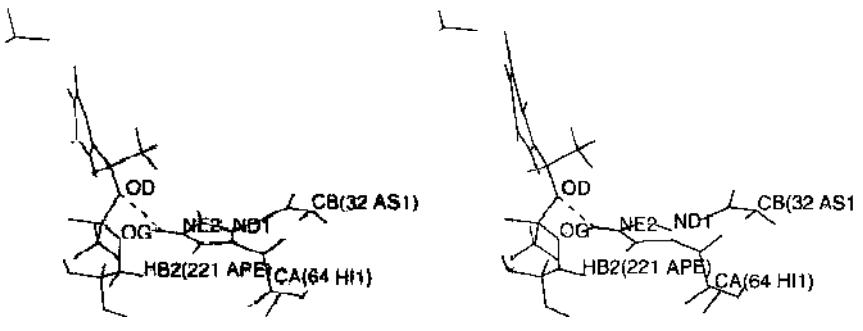


Figure 3. Active site of Subtilisin with the *S* enantiomer bound. The phenyl ring is directed into a hydrophobic pocket.

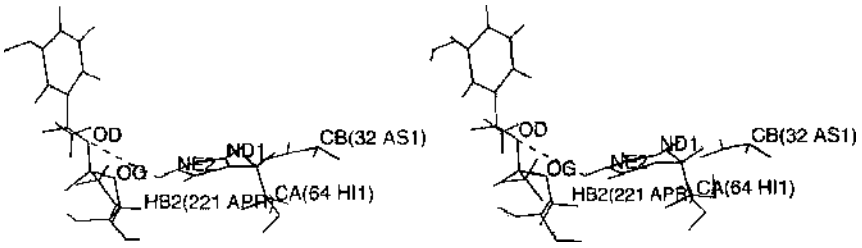


Figure 4. Active site of the enzyme with the *R* enantiomer bound. The phenyl ring is now directed outside, and the hydrophobic pocket is occupied by a DMF molecule.

incoming *R* and *S* substrate molecules through polarization and charge-transfer effects [89]. To further examine this effect, we examined the charge distribution (through QM/MM ESP fitting) in the *R* and *S* substrate/enzyme complexes at a number of points (12 total, every 15 ps) during the last 180 ps of the corresponding MD simulations. From these calculations, we determined the average and standard deviations of the charges on each of the QM atoms.

The ESP-derived atomic charges show rather substantial variations during the MD trajectories. The instantaneous fluctuations of the partial charges were examined for all the atoms in the QM region. Examples of these variations are given in Figs. 5 and 6, where we report the variation versus time of the charges on the oxygen atoms of the tetrahedral intermediates and on the *N* atoms of the catalytically essential His-64. From the observations of these figures and of the values reported in Tables 9 and 10, we find that the largest differences and deviations are is the atoms that are "involved" in the reactive process and in the H-bond network. For example, the NE2 atom on His-64 has a higher positive charge for the *S* complex case. As we have pointed out in the structural analysis, this atom is bound to HE2, which is involved in H-bond formation with the OG and OD oxygen atoms on the substrate. Interestingly, HE2 itself does not vary too much between the *R* and *S* complexes. The charges on the oxygen atoms (OA, OD, and OG) also appear to be important. From Fig. 5 and Table 10 we find that these atoms have a higher concentration of negative charge, in the case of the S complex. This factor helps in the stabilization of the H-bonding pattern in the active site, thereby lowering the energy of the transition state leading to preferential reactivity of the *S* enantiomer. Significant charge variation was also observed for the stereogenic carbon (CB3) along with the bound hydrogen (HB3), methyl carbon (CB4), and aromatic carbon (CG). This reflects the fact that this part of the molecule is placed in differing environments in the *R* (the methyl group is in the hydrophobic pocket, and the phenyl ring is solvent exposed) and *S* (the methyl group is solvent exposed, and the phenyl is in the hydrophobic pocket) cases.

These results also indicate that environmental influences observed in our MD simulations give rise to changes and fluctuations in the calculated partial charges. One very important factor we noticed was the asymmetric charge distribution on the substrate, reflecting the fact that the two enantiomers of *sec*-phenethyl alcohol in the transition state are in a stereodifferentiating environment not only from the steric but also from the electrostatic point of view. These results confirm the suggestion that environmental changes during the course of the simulation play a significant role in causing charge changes and reorganization in the model reacting system [29].

To investigate this point further, we calculated the atomic point charges of the intermediate model compound II in the gas phase in which the carbon CB1, which has three oxygens (OG, OA, and OD) bound to it, is forced to have an *S* configuration, as is required by the location of the oxyanion hole. As can be seen from Table 11, the differences between the

Table 7. RMS fluctuation per residue of the active site (Å).

Residue	*S*-Complex	*R*-Complex	Residue	*S*-Complex	*R*-Complex
Asp32	1.4485	0.9351	Asn155	1.5107	0.8533
His64	1.5076	1.0451	Ser221	1.5016	1.0316

Table 8. SASA for active site residues.

Residue	*S*-Complex	*R*-Complex	Residue	*S*-Complex	*R*-Complex
Asp32	2.746	1.427	Asn155	48.136	50.785
His64	27.719	33.608	Ser221	96.137	135.707

charges calculated in this way are smaller than in the enzyme-bound solvent case. This is particularly true for the atoms highlighted above. Thus, the charge variation for the oxygen atoms OG, OA, and OD is less for the gas-phase model complex relative to the enzyme-bound charge set. The most dramatic difference is observed around the stereogenic carbon (CB3), where in the gas phase there is only a modest variation of the charges around this center, while in the enzyme-bound case this region is strongly affected by the surrounding environment. Overall, this clearly indicates that the presence of the enzyme/solvent environment strongly determines the asymmetric distribution of the atomic charges on the substrate.

4.1.3. Results and Comparison with Experiments

To place our qualitative observations from the MD simulations on a more quantitative footing, we performed FEP simulations for the conversion of the substrate II(*S*) to II(*R*) *in vacuo* (with gas-phase calculated charges, which are reported in Table 11), and the free-energy difference was, as expected, 0.0 kcal/mol. The same perturbation was performed for compound II in DMF (again using the gas-phase charges in Table 11), which resulted in a free energy difference of −0.08 kcal/mol. A third FEP calculation in DMF, but using the 120 ps calculated charges for each of the two complexes (see charges in Table 2; these were slightly modified to ensure that each substrate had a net −1 charge), resulted in a free-energy difference of −5.74 kcal/mol, almost entirely as a result of the differences in the electrostatic contribution to the free energy. Finally, we carried out FEP simulations using the gas-phase charges given in Table 11 in the presence of the enzyme (90 ps equilibration and 150, 450, and 750 ps FEP simulations). From these simulations we obtained $\Delta\Delta G^{\ddagger}$ values of 11 ± 3 kcal/mol (150 ps), 9.2 ± 2.2 kcal/mol (450 ps), and 9.6 ± 2.5 kcal/mol (750 ps), which are all in poor agreement with the experimental value of 0.4 kcal/mol. This is in contrast to the QM/MM-derived charges, which give $\Delta\Delta G^{\ddagger}$ values that are in much better agreement

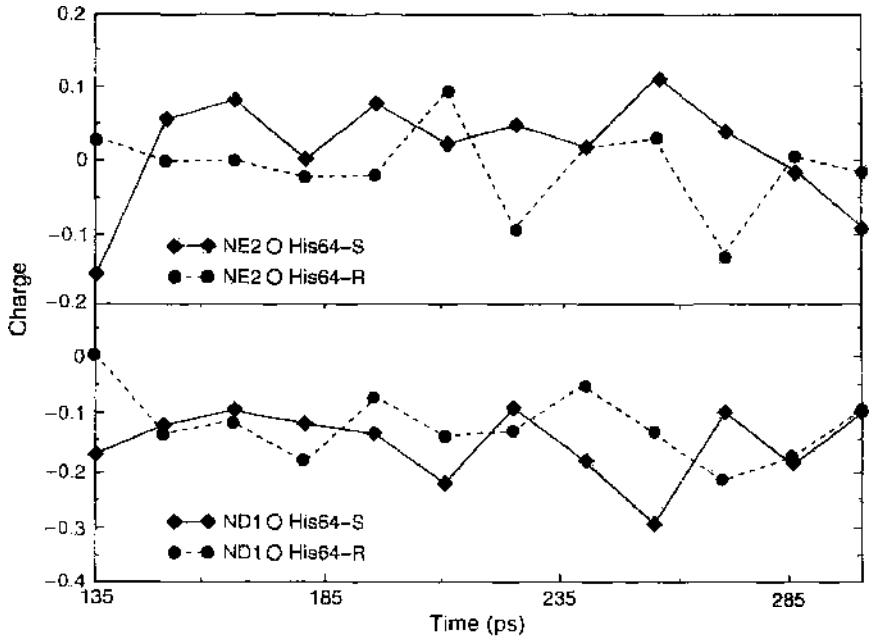


Figure 5. Charge fluctuations versus time for ND1 (bottom graph) and NE2 (top graph).

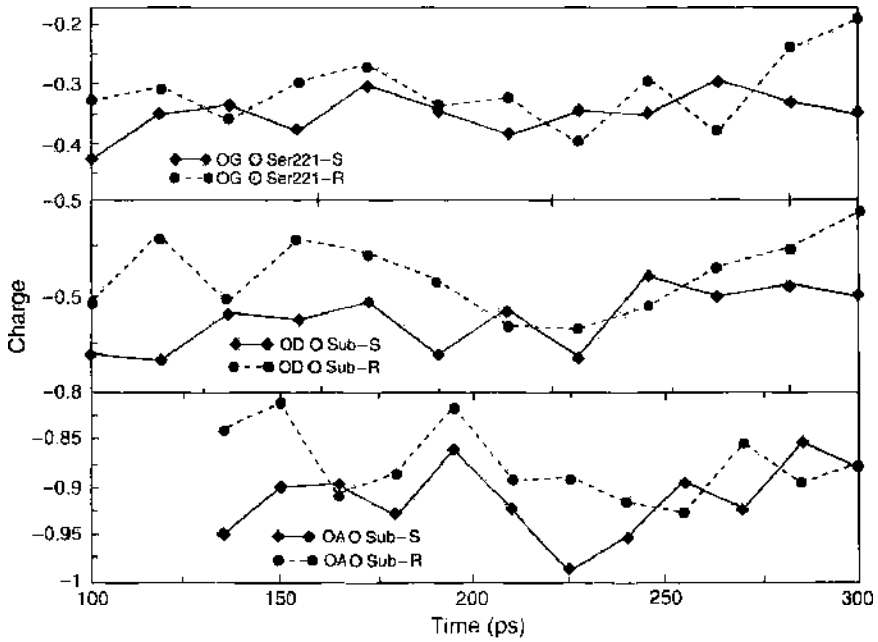


Figure 6. Charge fluctuations versus time for the oxygen atoms on the substrate. From the bottom to the top we report OA, OD, and OG, respectively. The solid line refers to the *S* complex, and the dashed line refers to the *R* complex.

with experiment (see following). Thus, the importance of taking electrostatic differentiation into account becomes evident in these examples: Charge values have a fundamental influence on the determination of $\Delta\Delta G^{\ddagger}$, and using the same charge set for chiral molecules in a stereodifferentiating environment may not give an accurate electrostatic representation [87].

The thermodynamic cycle used for our FEP simulations was presented in Scheme 6. The two enantiomers in the achiral solvent DMF have the same ground-state free energy; thus,

Table 9. Averaged QM/MM ESP Charges Calculated for Asp-32, Asn-155, and His-64.

	R Asp Av	*SD*	*S* Asp Av	*SD*	Diff
CB	−0.331	0.023	−0.343	0.020	0.012
HB1	0.090	0.014	0.100	0.013	−0.010
HB2	0.075	0.019	0.088	0.014	−0.013
CG	0.691	0.018	0.682	0.032	0.008
OD1	−0.735	0.018	−0.745	0.034	0.010
OD2	−0.727	0.015	−0.723	0.029	−0.00
	R Asn Av	*SD*	*S* Asn Av	*SD*	Diff
CG	0.586	0.031	0.568	0.021	0.018
OD	−0.605	0.016	−0.576	0.016	−0.025
ND	−0.698	0.043	−0.682	0.055	−0.017
HD21	0.382	0.032	0.396	0.030	−0.013
HD22	0.299	0.021	0.276	0.031	0.024
	R His Av	*SD*	*S* His Av	*SD*	Diff
CG	0.082	0.037	0.107	0.042	−0.025
ND1	−0.123	0.065	0.149	0.061	0.025
HD1	0.346	0.042	0.343	0.035	0.002
CE1	−0.001	0.051	0.029	0.060	−0.031
HE1	0.269	0.016	0.262	0.021	0.007
NE2	0.001	0.064	0.031	0.087	−0.030
HE2	0.259	0.038	0.254	0.054	0.005
CD2	−0.166	0.058	−0.204	0.057	0.038
HD2	0.253	0.016	0.247	0.013	0.005

Table 10. Averaged ESP Charges Calculated for the *R* and *S* Substrates.

	R Sub Av	*SD*	*S* Sub Av	*SD*	Diff
CB	−0.056	0.062	−0.029	0.063	−0.027
HB2	0.081	0.027	0.064	0.029	0.017
HB3	0.035	0.023	0.042	0.036	−0.007
OG	−0.305	0.061	−0.351	0.033	0.047
CB1	0.588	0.106	0.697	0.101	0.108
CB2	−0.331	0.065	−0.297	0.067	−0.034
HB21	0.075	0.038	0.045	0.033	0.030
HB22	0.092	0.027	0.046	0.026	0.046
HB23	0.060	0.027	0.063	0.034	−0.003
OA	−0.876	0.036	−0.914	0.038	0.037
OD	−0.462	0.073	−0.543	0.065	0.081
CB3	0.205	0.118	0.464	0.118	−0.258
HB3	0.045	0.059	−0.044	0.038	0.089
CB4	−0.277	0.051	−0.222	0.059	−0.054
HB41	0.052	0.022	0.042	0.027	0.010
HB42	0.065	0.024	0.033	0.031	0.032
HB43	0.095	0.027	0.050	0.017	0.045
CG	0.055	0.071	0.047	0.075	0.103
CD1	−0.154	0.048	−0.069	0.053	−0.085
HD1	0.095	0.020	0.126	0.019	0.031
CE1	−0.109	0.054	−0.142	0.049	0.033
HE1	0.101	0.021	0.126	0.021	0.025
CZ	−0.114	0.052	−0.116	0.042	0.001
HZ	0.097	0.020	0.093	0.023	0.004
CE2	−0.107	0.037	−0.131	0.034	0.024
HE2	0.111	0.014	0.100	0.011	0.011
CD2	−0.125	0.051	−0.139	0.045	0.014
HD2	0.117	0.013	0.099	0.015	0.018

ΔG_1 is equal to zero. Hence, the problem of defining enantioselectivity is reduced to the evaluation of the free-energy difference between the two diastereoisomeric complexes **I**(*S*) and **I**(*R*), in which the alcohols are part of the acyl intermediate.

The results obtained from a series of FEP simulations for this interconversion are given in Table 12. To assess convergence, we ran FEP simulations ranging from 150 to 750 ps (forward and backward), and we find that in all cases the calculated free energy is on the order of 1 kcal/mol. The best estimate comes from the 750 ps runs (from two temporally separated starting structures), which is 1.25 ± 0.5 kcal/mol. The experimental value for this change has been determined to be 0.4 kcal/mol. Thus, our value is in the right direction, but too strongly favors the *S* enantiomer. Nonetheless, the results are in good accord with experiment [98] and lend credence to the qualitative MD results discussed earlier.

Table 11. Charges for tetrahedral intermediate *in vacuo*.

Atom	S_{chg}	R_{chg}	Diff	Atom	S_{chg}	R_{chg}	Diff
CB	0.1575	0.0145	0.143	HB41	−0.0032	0.0494	−0.0526
HB1	−0.0354	0.0138	−0.0216	HB42	0.0304	0.0034	0.027
HB2	−0.0386	0.0141	−0.0527	HB43	0.0452	0.0659	0.0207
HB3	−0.0189	0.0372	−0.0561	CG	0.059	−0.0459	0.1049
OG	−0.4392	−0.4056	−0.0336	CD1	−0.1038	−0.0519	−0.0519
CB1	0.8315	0.7976	0.0339	HD1	0.1434	0.2046	−0.0612
CB2	−0.3664	−0.2269	−0.1395	CE1	−0.0912	−0.1546	0.0634
HB21	0.0579	0.0326	0.0253	HE1	0.0804	0.0849	−0.0045
HB22	0.0554	0.0179	0.0375	CZ	−0.1718	−0.116	−0.0558
HB23	0.0594	0.0235	0.0359	HZ	0.0861	0.0738	0.0123
OA	0.8059	−0.8129	0.007	CE2	−0.0715	−0.1526	0.0811
OD	0.5847	−0.5965	0.0118	HE2	0.0697	0.0781	−0.0084
CB3	0.4117	0.3895	0.0222	CD2	−0.2351	−0.1457	−0.0894
HB3	−0.0443	−0.0335	−0.0108	HD2	0.1164	0.084	0.0324
CB4	−0.1941	−0.2151	0.021				

Table 12. Calculated free-energy difference (kcal/mol) for the thermodynamic cycle given in scheme 6.[a]

Start	FEP time (ps)	DDG[a]	Start	FEP time (ps)	DDG[a]
90 ps	150	1.5 ± 0.5	120 ps	450	0.8 ± 0.4
90 ps	450	0.6 ± 0.3	120 ps	750	$1.3 + 0.5$
90 ps	750	1.2 ± 0.5			

[a] Experimental value: $\Delta\Delta G^{\ddagger} = 0.4$ kcal/mol.

4.1.4. Concluding Remarks on Enantioselectivity

Through the use of FEP calculations and MD simulations, we have been able to rationalize the observed enantioselectivity of subtilisin in DMF. A number of qualitative factors come into play in determining enantioselectivity, and through the use of an all-atom modeling approach, we have garnered insights into some of these factors. Steric fit factors play a role in governing enatioselectivity, and in this case the more reactive *S* enantiomer is able to place its phenyl ring into an active site pocket, and the less-reactive *R* enantiomer is unable to do so. Because of the steric complimentarity of the *S* enantiomer with the active site of subtilisin, we also observe better hydrogen-bonding complimentarity, especially for the atoms involved in catalysis (OA, OG, and OD) and their hydrogen-bonding interactions with His-64.

We also found that electrostatic complimentarity is critical in determining which enantiomer is favored. In general, it is assumed that the charge distribution of enantiomers is identical in the gas phase or a homogeneous aqueous phase. However, in a stereodifferentiating environment like an enzyme active site, this is not necessarily the case. Indeed, in this example, the stereodifferentiating nature of the active site alters the charges on the *R* and *S* enantiomers to such an extent that their solvation free-energy difference in a homogeneous (i.e., DMF) environment is quite large (almost 6 kcal/mol). Clearly, this effect needs to be accounted for when enzyme enantiselectivity is modeled.

We have described one of the first quantitative attempts to evaluate enantioselectivity through molecular modeling methods. Our model and approach have shown promise in predicting the reactivity of biocatalytic systems in nonaqueous environments, and indeed, we can evaluate the major factors (structural and electrostatic) that contribute to determination of enzyme enantioselectivity.

4.2. The Problem of Regioselectivity: A Modeling Approach

4.2.1. Problem Definition

Selectivity is surely one of the key factors, if not **the** key factor, of the growing popularity of biocatalysis among organic chemists, as testified to by the hundreds of reports describing striking enzymatic performances. Among these reports, probably the most impressive ones are the few dealing with the ability of enzymes to recognize remotely located stereocenters, as it is presently impossible to match these transformations with classical chemical catalysts. Examples of discrimination of a stereogenic center three or more bonds apart from the reaction center have been reported, describing the use of hydrolases (proteases or lipases) or oxidoreductases. For instance, Hedenstrom [112] and coworkers have recently disclosed the highly enantioselective *Candida rugosa*–catalyzed esterification of a series of methyldecanoic acid with 1-hexadecanol in cyclohexane. Using 8-methyldecanoic acid (in which the stereogenic center is located eight bonds apart from the reactive carboxylate) as a substrate, they still measured a significant and unexpectedly high *E* value (25) [112]. Remote control of enantioselectivity is the common feature of all these papers, a property that we have also exploited for the kinetic resolution of racemic *N*-Boc-piperidine-2-ethanol, a building block for the stereoselective syntheses of piperidine alkaloids [113]. Extending the remote control of enzymatic site- and regioselectivity is a point of major interest in using enzymes as biomolecular machines. In this section, we focus our attention on the remote control of enzyme selectivity in the case of the acylation of stevioside and steviolbioside.

4.2.2. Enzymatic Modifications: The Interplay Between Synthesis, Analytical Chemistry, and Computer Modeling

Enzymatic modification of natural glycosides has been a major research focus of our lab since the late 1980s [114], as these complex and often labile molecules are ideal targets for mild and selective biotransformations. Specifically, in the past we have reported on the use of hydrolases and glycosyltransferases for the modification of flavonoid, terpene, and alkaloid glycosides, and we are presently adding glycosidases to our repertoire [115]. As proposed years ago, the efficient preparation of random libraries of new derivatives (so-called combinatorial biocatalysis) [116, 117] is a useful tool to improve the properties of these biologically active compounds.

Stevioside (Compound 1, Fig. 7) is a sweet diterpenoid glycoside isolated from the South American plant *Stevia rebaudiana* Bertoni. The presence of three glucopyranosyl moieties makes this compound an interesting model for enzymatic transformations. Accordingly, a few years ago, we submitted compound 1 as well as its congener steviolbioside (compound 2, Fig. 8)—easily prepared from compound 1 by alkaline hydrolysis—to the action of the β-1,4-galactosyltransferase from bovine colostrum (GalT) [118]. Despite the presence in compound 1 of three possible galactosylation sites (HO-4'; HO-4''; HO-4'''), this was not only biotransformation stereo- and regio-selective but it was also completely site-selective, producing the tetrasaccharide derivative compound 3 (see Fig. 7) as the only product. Similarly, GalT-catalysis on compound 2 gave an almost quantitative conversion to the HO-4''-galactosylated compound 4.

To expand the number of stevioside and steviolbioside derivatives in our hands, we investigated the regioselective acylation of these diterpenic glycosides. A preliminary screening indicated that lipase PS was the best catalyst for this biotransformation. Accordingly, stevioside was dissolved in THF containing vinyl acetate, the enzyme was added and the suspension shaken at 45°C for 3.5 h. An almost 70% conversion to a major product and to a minor less polar by-product was observed, and both compounds were isolated by flash chromatography (51% and 4% isolated yields, respectively). At the beginning only the most abundant compound was examined and characterized as 6''-*O*-acetyl stevioside (compound 6, Fig. 7) by usual spectroscopic techniques.

The ^{1}H NMR (nuclear magnetic resonance) spectra in DMSO-d_6(^{1}H- and ^{13}C NMR spectra of compounds 1 and 2 were previously examined in pyridine [119, 120] and methanol–pyridine [188]) revealed three anomeric protons at 5.24, 4.42, and 4.37 ppm, respectively. The low field signal was attributed to H-1'''; that is, to the anomeric proton of the glucose moiety connected to the carboxyl function. Analysis of the extended spin connections (TOCSY) indicated that the acetate group was not located on this glucose because all proton

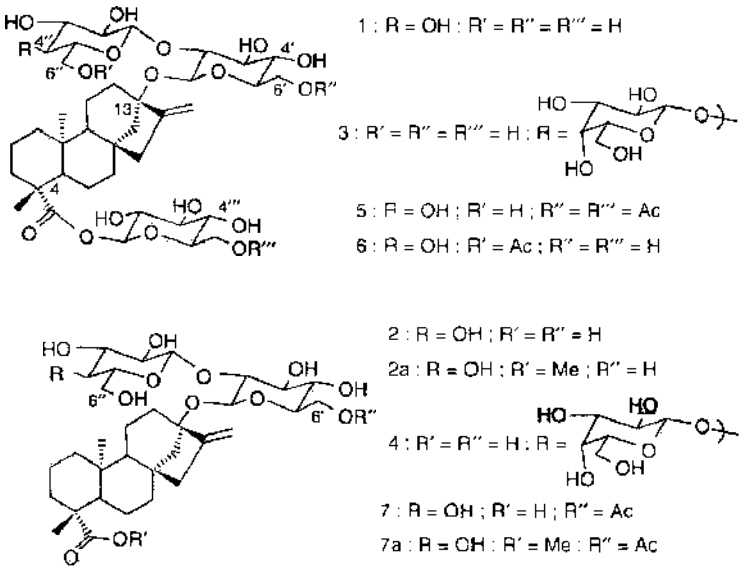


Figure 7. Stevioside, steviolbioside, and their different derivatives.

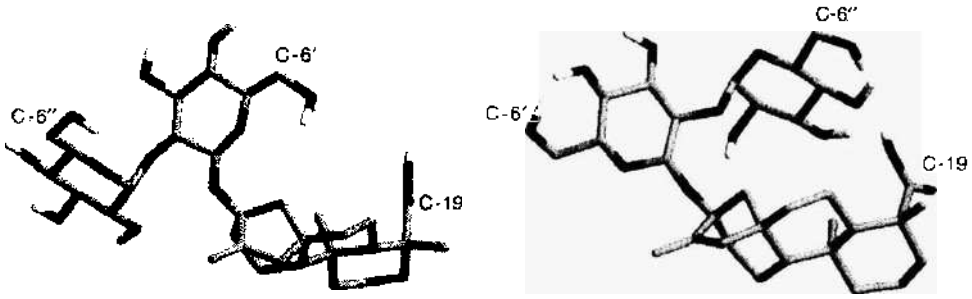


Figure 8. Conformational energy minima of steviolbioside in water (left) and chloroform (right).

and carbon resonances were in agreement with those of an unsubstituted saccharide. The anomeric proton at 4.42 ppm was attributed to the H-1′ of the sophorose disaccharide because of its vicinal relationship to H-2′, in turn directly linked to the low-field substituted C-2′ at 82.91 ppm, and because of the long-range interaction, with the quaternary C-13. That the acetate group was located on the external glucose of the sophorose unit was easily deduced by the *J*-network (vicinal, long-range, and spatial coupling) connecting the last anomeric proton at 4.37 ppm to the low-field C-6″ methylene protons at 4.19 and 3.95 ppm via H-2″ at 3.03, H-3″ at 3.10, H-4″ at 3.19 and H-5″ at 3.30 ppm. This result was consistent with the data obtained in the GalT-catalyzed reaction, as acylation took place on the external glucose of the sophorose moiety.

Enzymatic acylation of steviolbioside (compound 2) was also very efficient: Conversion was almost quantitative to a single product that was isolated and characterized. However, to our surprise, lipase PS directed its action to the inner glucose of the sophorose moiety of compound 2, producing the monoacetate 7. A similar result was obtained with the methyl ester 2a, which was quantitatively converted into the monoacetate 7a. In this case, too, the attribution was unequivocally supported by NMR data. The H-1′ at 4.53 ppm was connected to the low-field H_a-6′ and H_b-6′ at 4.28 and 3.92, bearing the acetoxy group, whereas H-1″ at 4.36 was connected to an unsubstituted oxymethylene at 3.59 and 3.48 ppm.

Therefore, a significant effect on lipase PS site and regioselectivity was observed by changing a remote portion of the molecule: Deglucosylation at position C-19 changed the acylation site of the sophorose moiety linked at C-13. This unexpected result prompted us to examine the structure of the very minor product obtained in the acetylation of stevioside. We found that it contained two acetate groups and that one of these was on the CH_2-6‴ of the glucose attached to the carboxyl function (H-1‴ at 5.32 was connected to CH_2-6‴ at 4.23 and 4.09). In analogy with the previously described monoacetate 6, we expected the additional acetate to be located on the external glucose of the sophorose moiety. However, in spite of our reluctance, a clear connection between the second low-field methylene at 4.27 and 3.99 and the H-1′ at 4.47 convinced us that the additional acyl moiety was located on CH_2-6′, and therefore the structure of 6′,6-β-di- *O*-acetyl stevioside 5 was assigned to this diester.

As the surprising structure of this very minor by-product did not give any clue to explain the puzzling behavior of lipase PS, we decided to start a molecular modeling analysis of the conformation of the two substrates, 1 and 2, in different simulation environments. Our goal was to verify whether the deglucosylation at position C-19 of compound 1 originated such a significant conformational variation in compound 2 to justify the different enzymatic selectivity experimentally observed. The solvent in these models is treated as an analytical continuum starting near the van der Waals surface of the solute. The equilibrated continuum solvent is characterized by a dielectric constant typical for the particular solvent, (i.e., 78 D for water and 4.8 D for chloroform). The increased accuracy in modeling the solvent effects yields parallel increased accuracy in the conformational searches. Conformational searches on compounds 1 and 2 were performed using the Monte Carlo/Energy Minimization (MC/EM) approach implemented in MacroModel: the dihedrals (one to five at a time) corresponding to simple bonds are changed by random quantities, the new conformations obtained are scored on the basis of the MM energy function, and low-energy conformations are stored as probable structures for the compound in solution. In this work, the SUMM

variant of the MC/EM algorithm was used [121, 122]. For each compound in each different solvent, 30,000 steps of the MC/EM procedure were run. All the minima found within 50 kJ/mol of the global minimum were fully reminimized to allow a more accurate determination of the relative energies.

The global minimum structures of both 1 and 2 in chloroform (a model for a generic hydrophobic organic solvent) and *in vacuo* were substantially indistinguishable. Steviolbioside was found to adopt an extended conformation in water (Fig. 8, top) and a folded one in the organic solvent (Fig. 8, bottom). In water, the extended conformation allows the molecule to expose the hydroxyl groups of the sugar moieties and the COOH group to the bulk solvent, giving rise to favorable hydrogen bond–type solvating interactions. In contrast, in the organic solvent chloroform, the molecule tends to adopt a folded conformation as a result of the H-bond formation between the carboxylate group at C-19 and the OH moieties on C-6″ and C-4″. This conformation causes the hydroxyl moiety on C-6′ to point toward the solvent, and the lower steric hindrance of this group might be the reason why it suffers selective acylation when steviolbioside enters into the active site of lipase PS.

However, stevioside showed a similar extended conformation in water (Fig. 9, top) and in chloroform (Fig. 9, bottom). In water, the extended conformation is favored because of the high number of hydrogen bonding interactions with the bulk solvent. In chloroform, the presence of the glucopyranosyl ester moiety at C-19 prevents the formation of the stabilizing hydrogen bonds and forces the molecule to again adopt a completely extended conformation to minimize unfavorable steric clashes. Therefore, the hydroxyl group on C-6″ is forced to point outside the molecule, thus minimizing the steric hindrance at this center and directing the acylation mainly on this group.

In conclusion, these calculations correlate well with the experimental results and indicate that the substrate conformation might be the main factor influencing and determining the recognition by lipase PS: The less sterically hindered hydroxyl group enters the active site of the catalyst more favorably, thus directing the site- and regioselective acetylation in organic solvent alternatively to the 6′-OH in compound 2 and to 6″-OH in compound 1 (Fig. 9). Obviously, contrary to what has been proposed in the very few literature examples on simulated sugars docking inside a protease active site [123], we were focusing on the substrate that has to be acylated by the biocatalyst, and therefore the enzyme environment has not been taken into consideration. However, in the absence of a high-level *ab initio* calculation of the whole reaction pathway describing all the intermediates at an electronic level, useful and sound rationales to specific experimental outcomes can be more easily extrapolated using classical mechanics-based approaches.

Needless to say, the observed remote control of lipase PS selectivity by substrate modification makes stevioside and steviolbioside a peculiar case among the reported examples of biocatalyzed site- and regioselective elaboration of natural glycosides.

4.3. Computational Approaches to the Understanding of Protein Stability

Molecular modeling techniques have a much wider range of application than just the ones described here. In particular, going beyond the description of reactivity, theoretical methods have been applied to the design of new proteins with novel properties. However, for the

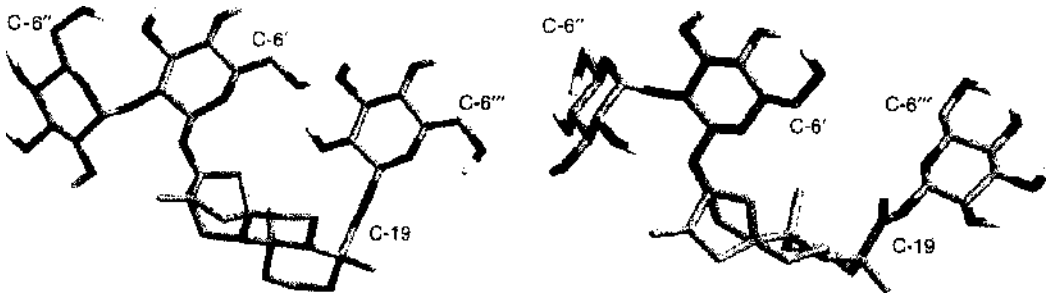


Figure 9. Conformational energy minima of stevioside in water (left) and chloroform (right).

design procedures to be effective, a deeper understanding of the determinants of stabilization of tertiary structures is needed. In this final paragraph, after an introductory overview on the problem of folding, we will present a new approach based on a method described by Tiana et al. [124] to rapidly address the folding-stability problem in small proteins.

4.3.1. The Problem of Protein Folding and Stability

The study of how the structure and stability of a protein is connected with its amino acid sequence has been, during the last years, a major focus of research. In particular, the problem of protein folding and its relation to stability has been studied through a variety of experimental and theoretical techniques. These studies have highlighted the central role the free-energy landscape plays in determining the properties displayed by proteins. The discussion has then been turned to the problem of determining this landscape for specific proteins [125].

Experimental techniques have yielded detailed information on macroscopic features of protein dynamical behavior, such as folding times and stability, and on some specific issues at the level of amino acids, such as the sensitivity to mutations [126]. However, there is still no experimental procedure capable of providing insight at the amino acid level into either the folding process in its completeness or the stabilization determinants of proteins. To gain insight into these questions, one must turn to theoretical and computational methods.

Because of the high computational costs involved, realistic models that provide an exhaustive description of the free-energy landscape have been used only in the study of short peptides. Ferrara and Caflisch [127], for instance, could reconstruct the whole free-energy landscape of a small designed β-sheet peptide with an all-atom representation of the solute and an implicit model of the solvent. Daura et al. [128] were able to demonstrate the reversible folding of a small (seven residues) helix forming β-peptide in methanol solvent using long MD simulations.

However, minimal models [129–132] have provided important results about the general features of the free-energy landscape of proteins. They give an approximate description of both the interaction energy among amino acids (usually through a contact potential encoded in a 20×20 matrix (e.g., Ref. [133]), and of the entropy. Making use of lattice models, that is, describing the protein as a chain of beads occupying the vertices of a cubic lattice, it has been possible to understand the basic energetic properties that characterize a folding sequence; that is, a large gap (measured in terms of typical contact energies) between the energy of the native and of the lowest, structurally dissimilar, state [130]. Studying the space of sequences and the role mutations have on good folders, it has been shown that the existence or less of such a gap depends on the existence of few "hot" sites in the native conformation, occupied by highly conserved, strongly interacting amino acids [134]. On the basis of these results, it has been found that the process of folding takes place in a hierarchical fashion [135–137], guided by local elementary structures (LES), formed very early in the folding process and stabilized by the interaction among "hot" amino acids [138]. The docking of LES leads to the overcoming of the highest free-energy barrier encountered by the protein in the transition between the denatured and the native conformation, giving rise to the folding nucleus [139], and thus to a conformation that sits in the energetic basin of attraction of the native state. Such hierarchical mechanism makes the folding process possible by efficiently squeezing out the entropy from the system. Although the extension of this solution to the case of real proteins is not expected to be straightforward, it depends on working out the *ab initio* interaction between single protein residues [140].

Other approaches focus on the geometric features of the protein, providing a detailed picture of the structure and using simple approximations for the potential function. From the physical point of view, this means focusing the attention on the entropic part of the energy landscape, neglecting the energetics. An example of this approach is to be found in Ref. [141], in which the folding rate of proteins is correlated with a "contact order" parameter accounting for the locality of the interactions. The Go model is another approximation that describes well the entropy of the protein chain, making a crude approximation on the interaction energy [142], being a contact function that assumes the value -1 if the contact is a native contact, and zero otherwise.

In this paragraph of the chapter, we introduce and illustrate a simple novel approach to extract information concerning the folding process and the stability of proteins from the structure of their native state. It is solely based on a detailed description of the interaction energy and neglects the protein entropy. The rationale behind this choice of approach is that the stabilization energy is distributed quite unevenly in proteins, as testified by the fact that mutations in most of the sites of a protein have little effect on its folding properties, whereas there are a few key ("hot") sites (in which the stabilization energy is concentrated) that are highly sensitive to mutations [134, 143]. These sites are those that, in the analysis performed with minimal models, build out the local elementary structures that control the folding process [144]. As a consequence, by determining these sites one can get insight not only into the mechanism that stabilizes the protein but also into the kinetics of the folding process.

The understanding of how the different amino acids determine the free energy of the protein, and in particular which are the residues that play a key role in the protein's the stability and kinetics, is of great interest, not only to unravel how the one-dimensional sequence of amino acids encodes for the three-dimensional native conformation but also to eventually design new proteins with specific tasks (see, e.g., Ref. [145]) or to modify existing proteins [88, 146].

In the following sections, we will describe the protocol used for determining the "hot" sites of the protein. We shall first test the method on a lattice designed model proteins (Section 4.3.2) and then apply it to four small proteins (Section 4.3.3), characterized by different folds and representative of diverse structural motifs. The last two sections contain the discussion of the results and the conclusions.

4.3.2. The Method and Its Validation from a Minimal Model

The basic idea of this analysis is to extract energetic information on the protein from MD simulations, and from it to gain insight into the determinants of the stability of the native protein conformation and into their influence on the folding process. The main information needed to achieve this goal is the interaction matrix M_{ij}, calculated averaging the corresponding interaction energies, comprising all the nonbonded inter-residue energy components (e.g., van der Waals and Electrostatic), over a MD trajectory starting from the native conformation. The matrix M_{ij} can be decomposed in eigenvalues, in the form

$$M_{ij} = \sum_{\alpha=1}^{N} \lambda_\alpha \mu_i^\alpha \mu_j^\alpha \tag{1}$$

where N is the number of amino acids in the protein, λ_α is an eigenvalue, and μ_i^α are the components of the associated eigenvector. We assume that the eigenvectors are normalized to unity and that as M_{ij} is symmetrical, all the eigenvalues are real.

For the sake of simplicity, we label the N eigenvalues in increasing order, so that λ_1 is the most negative. Accordingly, the different terms in the sum in Eq. (1) approximate the real interaction energy M_{ij} to an increasing extent, with the first term containing the largest contribution to the stabilization of the native conformation. The components μ_i^1 of the associated eigenvector indicate to which extent each amino acid participates to the stabilization. In other words, each term in Eq. (1) accounts for an amount of energy λ_α that is shared among the different residues according to the corresponding eigenvector μ_i^α.

If the second eigenvalue λ_2 is much higher than λ_1, one can approximate the whole interaction matrix as

$$M_{ij} \approx \lambda_1 \mu_i^1 \mu_j^1 \tag{2}$$

reducing the information needed to specify the interaction from N^2 to N numbers.

Because the interaction patterns of real protein are particularly complicated ("frustrated," in the language of Ref. [147]), one does not expect this approximation to be particularly accurate. However, it reduces a complicated two-body interaction into an effective interaction determined solely by a simple number (a generalized "charge," as in the case of the electrostatic interaction), which can be ascribed to each amino acid. Thus, the approximation given in Eq. (2) states that there are some amino acids that are strongly interacting, and

others that are weakly interacting *tout court*, depending on the "charges" μ_i^1 and μ_j^1 of the two amino acids. As a consequence, if amino acid A strongly attracts amino acid B (e.g., both with large, positive "charge" [unlike the case of electrostatic interaction, here "charges" with the same sign attract each others, whereas "charges" with opposite signs repel each other because AI in Eq. (2) is negative, being defined as the most negative of the eigenvalues.]), and amino acid B strongly attracts amino acid C (e.g., both with positive, large charge), in the present approximation A strongly attracts C. The opposite is also true: If the interactions A–B and B–C are weak, then the interaction A–C also will be weak. In contrast, the real interaction M_{ij} could be more complicated than this, being able to account for any combination of attractive and repelling pairs (e.g., A attracts B and B attracts C, but A repels C). As a consequence, the approximation introduced in Eq. (2) is, in principle, rather crude. However, from the analysis of minimal models, it emerged that the folding and stabilization of small, single-domain proteins is essentially controlled by a folding nucleus displaying a rich network of contacts among the most strongly interacting residues [138, 139, 144]), and thus containing a large fraction (~30–50%) of the stabilization energy of the system. This network arises from the fact that the folding nucleus (FN) results from the docking of the LES, in which each residue interacts favorably with many other residues belonging to the same or to one of the other LESs. As a consequence, essentially each residue in the nucleus attracts each other and can be easily described in terms of "charges."

Residues belonging to the nucleus display a large "charge," whereas the other residues carry a much lower "charge." In other words, the presence of the nucleus is the physical reason why $\lambda_1 \ll \lambda_2$.

Summing up, the analysis of the eigenvalues spectrum provides information about the folding features of the protein. If λ_1 is much smaller (i.e., more negative) than the other eigenvalues, it means that the approximation introduced in Eq. (2) is good for our purposes and the protein displays a well-defined nucleus. Again, we should state the caveat that interactions in real proteins are very complicated, and even if the folding is essentially controlled by a folding nucleus, the approximation given in Eq. (2) can never be quantitative.

In keeping with this caveat, in what follows, we will use the approximation introduced in Eq. (2) not to describe the whole network of interactions within a given protein but only to investigate the possible presence of a stabilizing nucleus and to describe the energetics of this nucleus. As a consequence, it is not necessary for the approximation to be useful to provide small error bars on all the sites of the protein, but only to do so on those participating in the folding nucleus. In fact, the validity test for the method will be its ability to identify which sites of the protein are important for the stability of the native conformation of the protein and for the dynamics of the associated folding process (as measured by φ-value analysis, see below).

In the text that follows, we will also show that, for all of the proteins studied, the energy of the nucleus is well approximated by Eq. (2). In contrast we find that the overall correlation coefficient between eigenvector components and φ-values is, as a rule, only qualitative, as a consequence of the fact that we are concentrating on the energetics of the folded state of the protein, neglecting the entropic contribution because of the distribution of folding–unfolding events displayed by proteins in solution.

Furthermore, the method used in this chapter that is centered around the approximated relation introduced in Eq. (2), although being an application of the well-known principal component analysis (PCA), is not related in any way to the various kinds of conformational PCA, or "essential dynamics," present in the literature [148], the goal of which is to effectively decrease the dimensionality of the conformational space.

To test the above procedure, we have studied a 36mer lattice model sequence designed (making use of the contact energies given in Table VI of Ref. [133], which in the rest of the paper we call MJ) to fold to a unique conformation, whose nucleus is known to be built out of the three local elementary structures 3–6, 11–14, and 27–30 [138], which contains 47% of the stabilization energy. The spectrum of eigenvalues of the interaction matrix calculated in the native conformation is displayed as a solid curve in the upper part of Fig. 10. For comparison, the spectrum associated with a random sequence is displayed in the lower part of the same figure. In the case of the designed sequence, the lowest eigenvalue is well

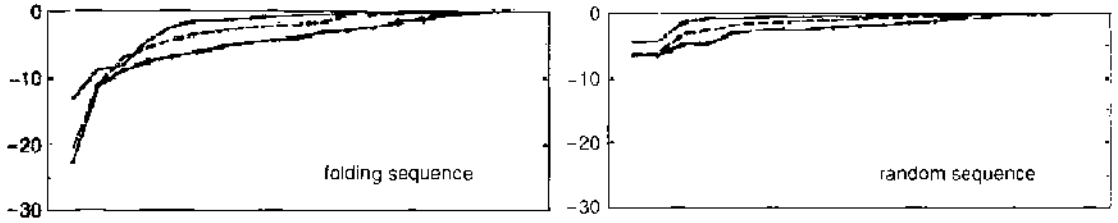


Figure 10. Negative eigenvalues of the interaction matrix in the lattice model. The eigenvalues are plotted in ascending order for a folding sequence and for a random, nonfolding sequence. The solid curve is associated with the interaction in the native conformation. The dashed curve displays the eigenvalues of the average interaction matrix, calculated starting from the native conformation, and is allowed to fluctuate for 10^4 steps at the folding temperature, whereas the dotted curve is associated with the average interaction starting from a random conformation and letting the system evolve for 10^4 steps.

separated from the others (i.e., $\Delta\lambda = \lambda_2 - \lambda_1 = 113.6$ in units of RT_{room}, where the average spacing between the others is $\bar{\lambda} = 13.3$, leading to the ratio $\Delta\lambda/\bar{\lambda} = 8.5$), whereas in the case of the random sequence, this is not true (i.e., $\lambda_2 - \lambda_1 = 0.4$, with the average spacing between the others being 4.0). As discussed above, this difference reflects the fact that the designed sequence displays a nucleus of mutually interacting residues, where the stabilization energy is concentrated (i.e., 50% of the native energy is carried by 30% of the residues, or those belonging to the LESs that form the FN). In contrast, in the random sequence there is a complicated pattern of attractive and repulsive contacts.

Because for real proteins the equilibrium state, at the length scale of single amino acids, is an ensemble of different conformations [149] sharing the same overall topology, we have repeated the calculation, letting the system fluctuate among these conformations. This is done by performing a MC search (10^4 MC steps) at folding temperature and calculating the interaction energy between pairs of amino acids averaged over these MC steps. The resulting eigenvalue spectrum, displayed as a dashed curve in Fig. 10, is qualitatively equal to that calculated in the unique native conformation. In particular, the folding nucleus energy is -5.90 in this case, as compared with -7.80 in the case of considering a unique conformation.

A difference between the eigenvalue spectra of the designed and of the random sequence is detectable also for interaction matrices calculated, starting from random conformations and performing the average over a time that is much smaller than the overall folding time (10^3 MC steps, compared to a characteristic folding time of 8×10^5 MC steps for the folding sequence). The results are displayed as a dotted curve in Fig. 10. For the folding sequence $\lambda_2 - \lambda_1 = 43.2$, to be compared to the average spacing of 7.6, while for the random sequence $\lambda_2 - \lambda_1 = 1.1$ and the average spacing is 2.6. The reason for this behavior is that local elementary structures, which are part of the nucleus and carry some of the stabilization energy of the protein, are formed at the very early stages of the folding process of good sequences. In contrast, random sequences undergo a disordered collapse not displaying any internal structure. The details concerning how the stabilization energy is distributed among the amino acids are contained in the normalized eigenvector μ_1^i associated with the lowest eigenvalue. In Fig. 11 are displayed the components of such an eigenvector in the case of the interaction matrix calculated in the native conformation of the designed sequence (corresponding to the solid curve in the upper part of Fig. 10). The plots corresponding to the other two curves in the upper part of Fig. 10 are similar. The plot displays six peaks in correspondence to the amino acids building the local elementary structures (residues 3, 6, 11, 14, 27, and 30) and one (residue 16) that, although not participating in a local elementary structure, interacts strongly with the folding nucleus. At any late, these seven sites are exactly those that, if mutated, lead to the denaturation of the protein and are called "hot" and "warm" sites in the paper [134].

4.3.3. Application to Proteins

Four experimentally well-characterized protein domains, displaying different three-dimensional structures, were used to test the model with realistic all-atom simulations including an atomic representation of the solvent: the α-spectrin SH3 domain (57 residues),

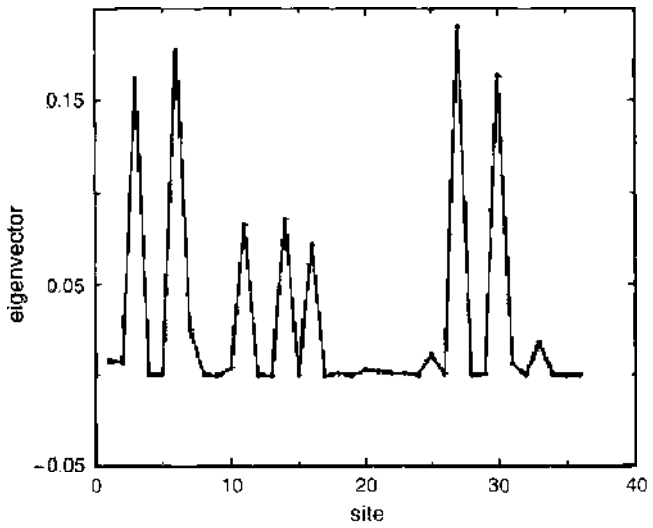


Figure 11. The eigenvector associated with the largest eigenvalue for the lattice-model folding sequence (corresponding to the first point of the solid curve in the upper part of Fig. 1.

the src SH3 domain (60), the IgG binding domain of protein G (herein called simply Protein G) (56), and Chymotrypsin Inhibitor 2 (CI2) (65). All of these systems have been studied in depth in terms of the mutations needed to (de)stabilize their folded states and to influence their folding kinetics. Most of these studies were based on an extensive experimental analysis performed analyzing the role of point mutations and on the characterization of the effects of every single mutation on the thermodynamic stability as well as on the kinetic properties of the protein. The ability to identify the mutation sites via a fast and simple computational method thus becomes an extremely appealing feature of all-atom molecular simulations. For the sake of simplicity in the discussion of the eigenvalue, and of the eigenvector properties, the numbering followed here for each protein starts from residue 1.

The four systems mentioned above were analyzed through 10-ns long, all-atom MD simulations, using an explicit representation of the solvent and the Particle Mesh Ewald method to calculate electrostatic interactions to avoid cut-off induced artifacts.

The lower part of the eigenvalue spectrum, consisting of the first 20 eigenvalues for each of the four proteins, is displayed in Fig. 12a–d. For all the proteins examined, the separation

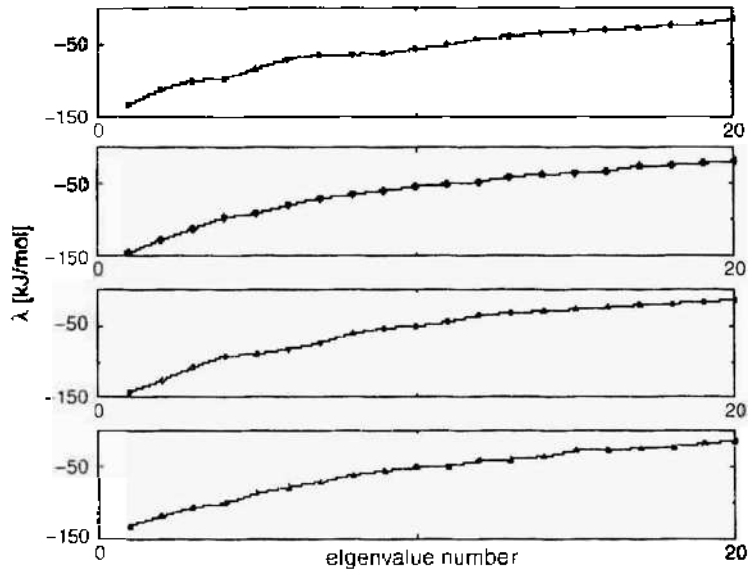


Figure 12. The negative eigenvalues of the interaction matrix for the all-atom MD simulations. The eigenvalues are plotted in ascending order: (a) filled circles refer to the α-spectrin SH3 domain, (b) diamonds refer to the src SH3 domain, (c) open circles refer to the IgG-binding domain of protein G, and (d) filled triangles refer to CI2.

$\Delta\lambda$ between eigenvalues λ_1 and λ_2 is much larger than the average spacing $\bar{\lambda}$ between the other eigenvalues (cf. Table 13), as in the case of the lattice model protein discussed above. Only in the case of CI2 does this property become marginal, and in particular, the first three eigenvalues are almost equally spaced.

The reasons why the ratio between $\Delta\lambda$ and $\bar{\lambda}$ is somewhat smaller for the four real proteins than for the idealized lattice protein are the use of a general-purpose force field, which has not been optimally tuned for the particular fold of the proteins under study; the noise present in MD simulations as a result of the numerical approximations in the integration steps; real proteins can concentrate energy in clusters of amino acids for reasons different than their folding and stabilization properties, such as for binding other molecules, for enzymatic activity, and so forth.

4.3.4. α Spectrin SH3 Domain

The spectrum of eigenvalues for α-spectrin SH3 domain is reported in Fig. 12a (filled circles). In this case, the difference between the first two lowest eigenvectors is $\Delta\lambda = 20.9$, which means five times the average spacing $\bar{\lambda} = 4.2$.

In Fig. 13 (a) are displayed the components μ_i^1 of the normalized eigenvector associated with the lowest eigenvalue, components that indicate how the stabilization energy is distributed among the amino acids of the α-spectrin SH3 domain. The plot displays particularly large values of the components in the intervals 10–20, 34–39 and 46–49 (and in particular, the maximum of the peaks are at sites 10, 14, 18, 36, 47, and 48).

This protein has been extensively studied by Martinez, Pisabarro, and Serrano (Martinez et al., 1998 [150]). By means of point mutations, they showed that residues V18, V39, and V48 are important for the stability of the protein, as their mutation leads to a strong destabilization of the native state (cf. Table 14), with $\Delta\Delta G_{V_i}$ values ranging between 7.5 and 12.9 kJ/mol. These sites are indeed among those showing, in our analysis, a large value of the eigenvector component. This is a straightforward result, considering that the components of the eigenvector express the degree to which a given residue participates in the stability of the protein.

A more striking result is that the residues that play a key role in the kinetics display large values of the eigenvector components. In fact, through the analysis of φ—values, Martinez and coworkers showed that the formation of the distal hairpin (residues 38–48) and the anchoring of the second strand of the RT loop (residues 18–19) are determinant in the kinetics of folding (cf. Martinez [150]). The highest peak in Fig. 13(a) corresponds to residue VI8 (RT loop), whereas the two peaks centered at residues W36 and V48 correspond to the sites stabilizing the distal hairpin. The summary of stability and kinetic data is listed in Table 14.

The connection between the energetic properties of a residue and its role in the folding kinetics is not unexpected and can be easily rationalized in terms of stabilization of local elementary structures and their subsequent assembly into the folding nucleus [138]. Such local elementary structures, being the elements that lead the folding process, have to be strongly stabilized at the very early stages of the kinetics and, consequently, must display large values of the eigenvector components. In the case of α-spectrin SH3, the distal hairpin plays the role of local elementary structure and, interacting with the second strand of the RT loop, forms the folding nucleus (as discussed by Martinez and coworkers [150].

Table 13. Difference $\Delta\lambda$ between the first two eigenvalues and the average spacing $\bar{\lambda}$ between the other eigenvalues and their ratio.

	$\Delta\lambda$ kJ/mol	$\bar{\lambda}$ kJ/mol	$\Delta\lambda/\bar{\lambda}$
Lattice protein	108.4	18.8	5.8
α-spectrin SH3	20.9	4.2	5.0
Src SH3 protein	18.7	4.0	4.7
CI	15.7	4.1	3.8
CI2	13.2	3.9	3.4

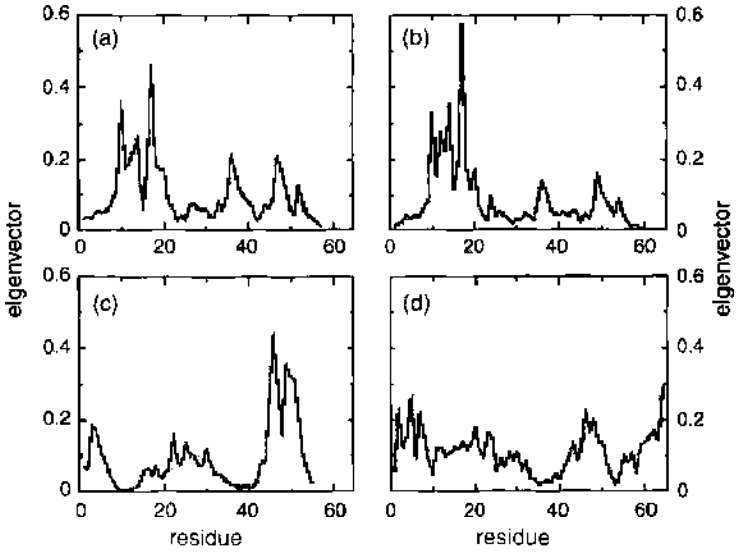


Figure 13. The eigenvectors associated with the largest eigenvalue in the all-atom MD simulations: (a) α-spectrin, (b) src SH3 domain, (c) IgG binding domain, and (d) CI2.

According to Fig. 13a, also the first strand of the RT loop (residues 10–15) also plays an important role, if not for the folding and stability of the whole protein, at least for the stabilization of its N-terminal region.

Summing up, the eigenvector components are useful in locating hot sites. The performance of the method is assessed by evaluating how many hot sites it is able to identify. To provide a consistent definition of a hot site on the basis of the components of the eigenvector (and, for comparison, of φ-values), we calculate the most probable value of the eigenvector components distribution and define as hot all the sites of the protein displaying an eigenvector component larger than the most likely value. The distribution of eigenvector components displays a peak and a long tail, and consequently the most probable value is not the average but, rather, the value associated with the top of the distribution. However, because φ-values are not calculated in all sites, the statistics are low, and it is difficult to define the most likely value from the φ-values distribution. In this case, hot sites are defined as those displaying φ-values above the average.

For α-spectrin SH3, the eigenvector method is able to identify three hot sites out of four sites totals, displaying two false positives (cf. Table 15). One of the two false positives (site 38) is at the borderline of being a hot site, displaying a φ-value of 0.37, while the threshold value we take as threshold for this protein is 0.4. The other false positive (site T19) is directly bound to a correctly predicted hot site (site 18), with its motions and its interactions strongly

Table 14. Effect of point mutations on the amino acids of α-spectrin SH3 domain, as found by Martinez, Pisabarro, and Serrano [151].

	$\Delta\Delta G_{N-U}$ [kJ/mol]	$\varphi_{\ddagger-U}$
L3	3.38	0.24
V18	7.53	0.45
T19	4.18	0.38
L28	5.18	−0.3
S31	2.8	0.27
K38	2.50	0.4
V39	12.95	0.55
V48	8.98	

The quantity $\Delta\Delta G_{N-U}$ is the unfolding free energy on mutation, and $\varphi_{\ddagger-U}$ is the φ-value associated with the transition state ‡. Note that the numeration of amino acids has been changed in such a way that the first is numbered with 1 (instead of 5).

Table 15. Effect of point mutations on the amino acids of α-spectrin SH3 domain, as found by Martinez, Pisabarro, and Serrano [151].

Protein	No. of predicted hot sites	False positives	Correct hot sites
α-SH3	3 vs. 4	2	18, 39, 48
Src-SH3	13 vs. 17	6	12, 18, 20, 24, 26, 34, 38, 39, 41, 44, 46, 49, 50
G	5 vs. 7	3	3, 6, 32, 46, 47, 49, (51)
CI2	8 vs. 12	13	12, 15, 16, 17, 18, 20, 21, 64
Protein	Energy Nucleus	Energy Nucleus (app.)	threshold
α-SH3	−1334	−1209	0.08
src SH3	−1661	−1322	0.05
G	−1001	−966	0.09
CI2	−1690	−1803	0.10

(A) for each protein (first column) is indicated the number of hot sites located by the present method over the total number of hot sites (second column), as indicted by φ values; the third column contains the number of false positives, and the fourth the hot sites. In parenthesis are hot sites that the present method does not locate. (B) The exact energy of the corresponding nucleus (in kJ/mol) as calculated by MD simulations listed in the fifth column, as can be compared with the value obtained by the largest eigenvalue approximation [Eq. (2)]. The last column indicates the threshold on the eigenvector component used to define hot sites, found as the most likely value of the component distribution.

coupled to those of residue 18. It is then conceivable that perturbing residues in the region of a hot spot should have definite consequences on the protein stability.

Note that the overall correlation coefficient is quite low, giving a value of 0.45. This is an expected result, as the approximation employed is meant to describe well only hot sites, not the rest of the protein.

It is also interesting to study the energetics properties of the folding nucleus, defined as the set of sites whose numerical value μ_j^1 of the first eigenvector is higher than the most probable value (cf. Table 15). This shares a stabilization energy of −1334 kJ/mol (see also Table 15). This is equivalent to 35% of the overall stabilization energy of the protein. The energy of the nucleus evaluated through the approximation of Eq. (2) is −1209 kJ/mol, which is equivalent to a relative error of 9%.

4.3.5. src SH3 Domain

The SH3 domain of the src protein displays the same fold as the α-spectrin domain discussed above, although their sequence homology is only 36%. Also, for src SH3 the lowest eigenvalue is well separated from the others, being $\Delta\lambda = 18.7$ kJ/mol, and the average separation being $\bar{\lambda} = 4.0$ kJ/mol (see Fig. 12c). The components of the corresponding eigenvectors are displayed in Fig. 14.

Baker and coworkers performed a mutation analysis on src SH3 domain [151, 152], concluding that mutating residues G34, W36, 150, and Y54 to alanine cause a destabilization of the native state ranging from 5.8 to 16.7 kJ/mol. Furthermore, they showed that residues D9, Y10, S12, and L18 are important for the protein stability, in that their mutations to alanine causes destabilizations ranging between 4.1 and 17.1 kJ/mol [151]. In addition, in this case

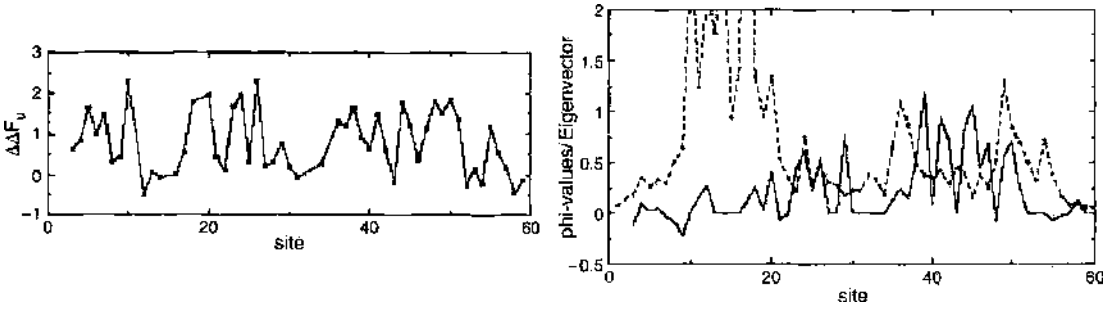


Figure 14. In the lower part, the φ-values of src SH3 (solid curve) and the corresponding components of the eigenvector corresponding to the lowest eigenvalue (dashed curve). The curve for the eigenvector components is presented in a blown up form (multiplied by 4). In the upper part, the change in free energy on mutations. Thicker axis lines indicate the areas important for stability.

there is a good agreement between the sites playing a major role in the stability and the kinetics of the protein and the peaks in the components of the eigenvector (residues 10–20, 35–37, 49–51 and 54, cf. Fig. 14). Only site 23 is found in experiments to be important for the stability of the protein, but it displays a small component in the eigenvector. This could be a result of the fact that this site is occupied by a glycine that, lacking the sidechain, may display a lower number of interactions with the flanking amino acids in MD simulations.

The comparison of the eigenvectors associated with α-spectrin and src SH3 domains shows a remarkable similarity, although the sequences are rather different (only 36% homology). This fact suggests a stronger evolutionary relationship between the two proteins than what the mere comparison of sequences would indicate. In other words, α-spectrin and src SH3 domains have diverged enough to change 74% of their amino acids, but not enough to mutate their topologies and energetic pattern and, as a consequence, their folding mechanism.

This aspect is reflected by the chemical nature of the residues at the "hot" sites of the two proteins. Both SH3 domains display peaks corresponding to acidic (D9) and hydrophobic (V18 for α-spectrin SH3 and L18 for src SH3) residues in the N-terminal region and sterically demanding hydrophobic (W36 for α-spectrin SH3, W37 for src SH3) together with aliphatic (V48 for α-spectrin SH3, I50 for src SH3) at the C terminal. Moreover, these hot sites are part of a mostly hydrophobic core that is coincident with the mechanical nucleus defined by Martinez [150], formed during the folding process of the two domains studied. These authors showed that passing through the transition state barrier in SH3 domains requires the formation of a defined structure with little conformational variability in well-defined regions, such as the ones identified here. It is worth noting that the regions rich in hot sites, involving residues 34–39 and 46–49, are the peptide ligand-binding regions, indicating that the conservation of the pattern also assumes a functional role. We can also conclude that the analysis of the energetic patterns of proteins sharing the same fold (topology) could be used to assess their evolutionary relationships in a way that is more efficient than the comparison of the sequences, because proteins have some degree of freedom in mutating their amino acids (even the key amino acids) for functional reasons, whereas the energetic pattern has to remain similar.

A summary of the parameters providing information on the validity of the method is given in Table 15. Repeating the same procedure as that employed for the analysis of α-spectrin SH3, we are able to correctly identify 13 out of 17 hot sites. In this case, there are six false positives, but three of them are consecutive to hot sites (17, 19, and 35), and this determines strong steric interactions with the hot spot regions, resulting in high component values in the first eigenvector. In the case of this protein, too, it is most likely that the introduction of a mutation in the position of one false positive can be effective in changing the stability properties of the domain. The energy of the nucleus is -1661 kJ/mol (43% of the overall energy), which has to be compared with the approximated value of -1322 kJ/mol, equivalent to a relative error of 20%.

4.3.6. Protein G

Protein G is particularly suitable for our analysis because it has been extensively studied by Baker and coworkers [153], who measured the effect of mutations in many sites of the protein.

The φ-values of protein G are shown in Fig. 15 and display a high peak in the interval 45–52 and two small peaks around residues 3 and 22. Moreover, residues D22, A26, and F30 are important for the stability of the native state, as their mutations to A cause a decrease of the stabilization free energy of 7.3, 12.3, and 5.9 kJ/mol, respectively.

The spectrum of eigenvalues for protein G, as in the case of SH3 domains, displays a gap between the lowest and the next-to-lowest level $\Delta\lambda = 15.7$ kJ/mol, larger than the average distance between the others, $\bar{\lambda} = 4.1$. The components of the associated eigenvector are also displayed in Fig. 15 (dashed curve). The highest peak in this plot matches the peak in the φ-values at sites 45–52, corresponding to the second hairpin, whose formation is critical to the folding mechanism [154]. Moreover, the other minor peaks in the eigenvector

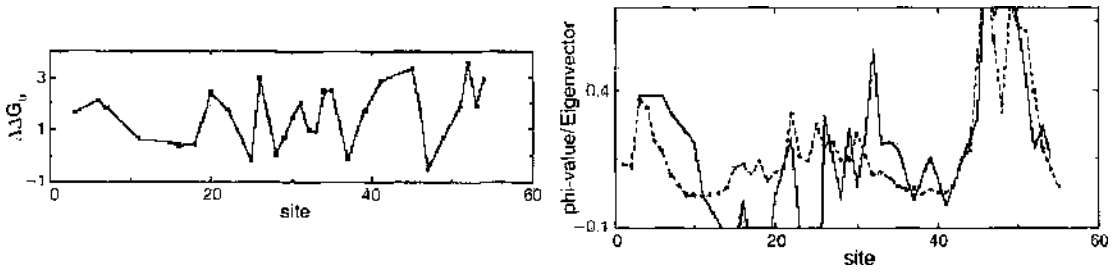


Figure 15. Same as Fig. 5, for Protein G. In the lower panel the φ-curve (solid line) is multiplied by 2 for clarity.

components at sites 3–7, D22, A26, and V29, also correspond to the other sites relevant to the stability and to the kinetics (cf. Fig. 15).

Some of the residues relevant for stabilization are also involved in the definition of the binding region of the IgG binding domain from protein G. In fact, mutations of residues in the regions 27–32 and 43–46 cause major variations in the binding constant of human immunoglobulin Fc fragment [155]. In particular, the binding properties of the region 27–32 explain the peak in the associated eigenvector components, which, from the analysis of the mere folding behavior, has to be considered a false positive.

In addition, for protein G the outcome of the model is summarized in Table 15. For protein G, the method can identify five hot sites over seven, with three false positives (sites 22, 45, and 52, the second and the third of which are consecutive to hot sites). The energy of the nucleus is −1001 kJ/mol (i.e., 28% of the overall energy), and the approximated value is −966 kJ/mol, equivalent to a relative error of 3%.

4.3.7. Chymotrypsin Inhibitor 2

Chymotrypsin Inhibitor 2 (CI2) is a 64-residue polypeptide inhibitor of serine proteases [156]. CI2 is a single module of structure: The interatomic interactions are quite uniform over the structure, and they do not segregate into regions that make more tertiary interactions within themselves than they do with neighboring atoms. As such, CI2 can be considered a basic folding unit, or foldon [157]. This is reflected by the fact that most φ-values of this protein are fractionary [158].

This peculiarity of CI2 is underlined both by the fact that the difference between the first two lowest eigenvalues is $\Delta\lambda = 13.2$ kJ/mol, with the average spacing being $\bar{\lambda} = 3.9$, and by the absence of well-defined isolated peaks in the components of the eigenvector associated to the first eigenvalue.

Nonetheless, a careful inspection of Fig. 16d shows that the stabilizing regions of the molecule are located around residues 10–20 and 40–50, corresponding to the helix of the N-terminal region, making contacts with residues located on a β-sheet at the opposite end of the polypeptide. Mutations in these regions cause destabilization of the folded state between 4.0 and 20.6 kJ/mol. The long-range interactions between these regions are those required to stabilize the helix in peptide fragments of CI2, as shown by Gay et al. [159] in a study on the stability of the secondary structure of peptides isolated from CI2. Moreover, residues

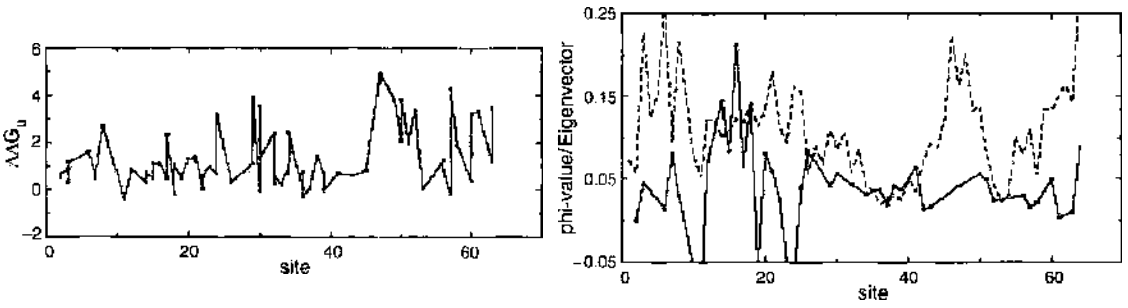


Figure 16. Same as Fig. 5, for CI2. In the lower panel the φ-curve (solid line) is multiplied by 2 for clarity.

around positions 16, 49, and 57 constitute a nucleation site whose disruption leads to loss of stability and fast unfolding of the protein at high temperature conditions [160–162].

As shown in Table 15, the method can identify eight hot sites over a total of 12 sites. The number of false positives is 13. We attribute this behavior to the conformational flexibility of CI2, which is also present in the transition state ensemble as seen by the high number of fractionary φ-values, which determines a wide network of intramolecular interactions in the folded state of the protein. The energy of the nucleus is predicted to be -1690 kJ/mol, with an error of 6%, and to be equivalent to 44% of the overall stabilization energy.

4.4. Discussion

By simulating the equilibrium fluctuations of a protein around its native conformation through MD simulations, it is possible to investigate the energetic pattern (i.e., interaction energy, not free energy) that defines the native state and contributes to make it a minimum in the free-energy landscape. A feature common to all proteins that emerge from such calculations is that in each protein, few residues build clusters of strong interactions, surrounded by the other, weakly interacting residues (see Fig. 17). This fact can be mathematically assessed, noting that the lowest eigenvalue of the residue–residue interaction matrix is well separated from the others and that the clusters can be detected when searching for peaks in the components of the associated eigenvector.

The presence of a strongly interacting nucleus is not unexpected and has been observed in minimal lattice models of proteins [132, 134, 138, 139]. This nucleus not only gives an account for the all-or-none character of the folding transition [130] and for its remarkable tolerance to point mutations [134] but also is responsible for the fast folding of the protein. In fact, the hierarchical assembly of local elementary structures, composed of strongly interacting residues lying close along the chain, has been shown to drive the protein to the native state in the case of simple models (see Section 1). This structure is necessarily associated with a cluster of strong interactions composed of those that stabilize the local elementary structure and those that stabilize the nucleus, keeping together the local elementary structures.

Minimal model systems are ideally designed such that the same nucleus residues determine both the folding mechanism and the stability of the native over the unfolded state: The kinetic and thermodynamic determinants of folding are, as a consequence, coincident in these simplified representations.

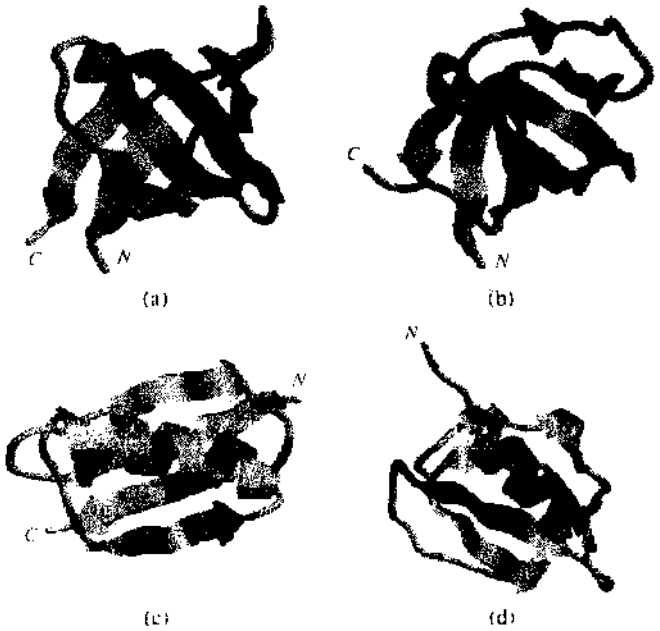


Figure 17. The hot sites mapped on the three-dimensional structure for (a) α-spectrin SH3 domain, (b) src SH3 domain, (c) protein G, and (d) CI2. The hot sites are part of the folding nucleus.

In real proteins, this might not necessarily be the case, as the pathways to the native state can be mediated by interactions with the solvent, by stereospecific interactions among the sidechains, and so forth. However, at least in the case of single-domain proteins like the ones considered here, φ-value analysis has proved that a localized cluster of residues is the determinant for both the stability and the kinetics of folding. This is not to be considered a totally unexpected phenomenon if we consider that folding transition states and intermediates show a high degree of structural similarity to the native states [163].

The analysis of the components of the eigenvector associated with the lowest eigenvalue of the native-state interaction matrix is helpful in determining such a cluster. As a consequence the sites individuated by this analysis play a major role in the kinetics and stability of the protein (Fig. 17), as shown by the good agreement between the plot of the eigenvector components and the plots displaying φ-values and that change in stability on mutations. From this analysis, it is not possible to tell which of the sites indicated by the eigenvector are connected with stability, which with kinetics, and which with both of them. The eigenvector contains a superposition of the two pieces of information, and only an analysis of the entropy of the system can separate them. From this point of view, this analysis consists of an approximation of the free energy of the native state in which the entropic term is neglected, and consequently, it is quite obvious that the results it can give are partial.

It is to be noted that the average spacing between the peaks in the eigenvector component of all the proteins we analyzed is ~19 residues. This is in accordance with the findings of Berezovsky and Trifonov [164, 165], who analyzed the size of closed loops of the backbone of a large number of proteins and observed an average ranging from 25 to 30 residues, a size that is optimal from the point of view of a fast search in conformational space. We suggest that the loops found by Berezovsky and Trifonov in Refs. [164, 165] could be related to the loops between local elementary structures discussed above [138].

Two caveats to the method presented here need to be highlighted. First, the MD simulations with explicit water that we employ produce an interaction matrix, which accounts for the van der Waals energy, for the electrostatic energy, and for the hydrogen bonds energy, but misses the hy-drophobic interaction. This is, in fact, an effective interaction that arises from the averaging of the solvent degrees of freedom and does not appear in the potential function of the system. The fact that we obtain good agreement between these model calculations and the experimental results indicates that the hydrophobic interaction does not play a major role in determining the local elementary structures but only plays one in the generic collapse of the chain into a globule (which is not considered here). Another problem is the choice of the time duration of the simulations. They have to be long enough to average the uncertainties in the structure connected with the experimental determination of the native conformation and to be equilibrated. Running simulations in which multiple folding–unfolding events could be observed would give access to the entropic term that has been neglected in the analysis, but reaching the timescales required for that purpose is currently out of reach even for small proteins. We have nonetheless performed simulations ranging from 2 to 15 ns, observing no qualitative differences between the different cases belonging to this range.

From the practical point of view, the analysis of the components of the eigenvector associated with the lowest eigenvalue of the native-state interaction matrix can be helpful in determining the positions or the regions in which mutations have to be introduced to change the stability and, connected with it, the dynamical features of the folded protein. Once the hot sites have been identified, saturation mutagenesis experiments can be focused on that region to modify the properties of the protein under study. In many cases [166], "hot sites" have been identified through several cycles of error-prone mutagenesis and analysis, which rely on the random introduction of mutations in different parts of the protein and on subsequent saturation mutagenesis on the above defined hot sites. This procedure, despite being quite efficient and successful, is not based on a rational approach and might be time consuming. Previous MD studies [167] have yielded a reasonable agreement with experimentally determined points of mutation in CI2, using a geometrical approach based on the relative number of contacts of one particular residue in the native and in the transition state

ensembles. This approach, however, relies on the generation via MD of the ensemble of transition state structures based on high-temperature unfolding simulations.

In the case of our method, the straightforward application of realistic MD simulations of the native state followed by the analysis of the eigenvectors of the energy interaction matrix can be a useful help to a rational protein design without the need to generate transition state structural ensembles. It thus is possible to get information on the sites of a protein that contribute most to the stability and to the accessibility of the native state by means of a simple energetic analysis of MD simulations, simulations that can be performed even for large proteins with just the help of a PC. This allows us not only to gain insight into the folding pathways of a given protein but also to be able to predict sites on which to focus our mutagenesis experiments.

5. FUTURE PERSPECTIVES

In the course of this review, we have tried to give an overview of the computational methods applied to the field of enzyme catalysis. It is quite clear that the highest possible level of detail is necessary to give a clear picture at atomic resolution of the phenomena involved in defining the reactivity of biomolecules. Only with the advent of mixed QM/MM methods, in fact, was it possible to define different models of charges for different problems such as the enantiodifferentiation by subtilisin or the effect of the solvent. The improvement of algorithms and the fast increase in computer performances, together with the possibility of building high-performance parallel computers based on PCs at a low price, make it possible to apply computational methods to the study of reactivity problems almost on a routine basis. The study of a particular problem at a theoretical level can help give a better understanding of the fundamentals of a certain enzyme reaction, and thus help design new experiments or new catalysists via, for example, molecular biology. The field of the computational study of enzyme reactivity is growing in two main directions. On the one hand, high-performance computing is available nowadays at reasonable costs, thanks to the previously mentioned possibility of building massively parallel PC clusters. On the other hand, QM theory and new methods for long-timescale simulations are being developed in several laboratories. The development of these two complementary aspects will allow a more realistic treatment of a series of problems by QM, using linear scaling approaches [97] or embedding DFT into MM environments and, in parallel, the simulation of slow rate phenomena (with time ranges of micro- to milliseconds) that are currently out of reach. Thus, a combi nation of hardware improvement and theory development is going to make computational approaches even more useful to the biochemist community.

REFERENCES

1. S. J. Benkovic and S. H. Schiffer, *Science* 301, 1196 (2003).
2. A. Warshel, *Acc. Chem. Res.* 35, 385 (2002).
3. A. R. Leach, "Molecular Modelling," Longman Singapore Publisher, Singapore, 1996.
4. M. Levitt, and A. Warshel, *J. Mol. Biol.* 104, 59 (1976).
5. J. A. Pople and D. L. Beveridge, "Approximate Molecular Orbital Theory" McGraw-Hill: New York, 1970.
6. M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, and J. J. P. Stewart, *J. Am. Chem. Soc.* 107, 3902 (1985).
7. J. W. Hehre, L. Radom, P. v. R. Schleyer, and J. A. Pople, Wiley, New York, 1986.
8. B. R. Brooks, R. E. Brucooleri, B. D. Olafson, D. J. S. S. Swaminathan, and M. Karplus, *J. Comput. Chem.* 4, 187 (1983).
9. W. D. Cornell, P. Cieplak, C. I. Bayly R. I. Gould, K. M. Merz, Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, *J. Am. Chem. Soc.* 117, 5179 (1995).
10. W. L. Jorgesen and J. Tirado-Rives, *J. Am. Chem. Soc.* 110, 1657 (1988).
11. M. J. Field, P. A. Bash, and M. Karplus, *J. Comp. Chem.* 11, 700 (1990).
12. J. Gao and X. Xio, *Science* 258, 631 (1992).
13. V. Luzhkov and A. Warshel, *J. Comput. Chem.* 13, 199 (1992).
14. M. A. Thompson, *J. Am. Chem. Soc.* 117, 11341 (1995).
15. M. A. Thompson, *J. Phys. Chem.* 100, 14492 (1996).
16. D. Bakowies and W. Thiel, *J. Phys. Chem.* 100, 10580 (1996).
17. G. Monard, M. Loos, V. Thery, K. Baka, and J.-L. Rivail, *Int. J. Quant. Chem.* 58, 153 (1996).
18. R. V. Stanton, D. S. Hartsough, and K. M. Merz, Jr., *J. Comput. Chem.* 16, 113 (1994).

19. T. A. Wesolowski and A. Warshel, *J. Phys. Chem.* 97, 8050 (1995).
20. D. Wei and D. R. Salahub, *J. Chem. Phys.* 101, 7633 (1994).
21. I. T. non, M. T. C. Martins-Costa, C. Milloy, M. F. Ruiz-Lopez, and J.-L. Rivail, *J. Comput. Chem.* 17, 19 (1996).
22. A. Warshel, *Acc. Chem. Res.* 14, 284 (1981).
23. U. C. Singh and P. A. Kollman, *J. Comput. Chem.* 7, 718 (1986).
24. R. V. Stanton, L. R. Little, and K. M. Merz, Jr., *J. Phys. Chem.* 99, 17344 (1995).
25. J. Gao, *Rev. Comput. Chem.* 6, 119 (1995).
26. P. A. Bash, M. J. Field, R. C. Davenport, G. A. Petsko, D. Ringe, and M. Karplus, *Biochemistry* 30, 5826 (1991).
27. K. M. Merz, Jr. and L. Banci, *J. Phys. Chem.* US 43, 17414 (1996).
28. K. M. Merz, Jr. and L. Banci, *J. Am. Chem. Soc.* 119, 863 (1997).
29. D. S. Hartsough and K. M. Merz, Jr., *J. Phys. Chem.* 99, 11266 (1995).
30. M. J. S. Dewar and W. Thiel, *J. Am. Chem. Soc.* 99, 4899 (1977).
31. M. J. S. Dewar and W. Thiel, *J. Am. Chem. Soc.* 99, 4907 (1977).
32. J. J. P. Stewart, *J. Comp. Chem.* 10, 209 (1989).
33. J. J. P. Stewart, *J. Comp. Chem.* 10, 221 (1989).
34. J. J. P. Stewart, *J. Comp. Chem.* 12, 320 (1991).
35. W. Yang and T.-S. Lee, *J. Chem. Phys.* 103, 5674 (1995).
36. S. L. Dixon and K. M. Merz, Jr., *J. Chem. Phys.* 104, 6643 (1996).
37. C. A. White, B. G. Johnson, P. M. W. Gill, and M. Head-Gordon, *Chem. Phys. Lett.* 240, 8 (1994).
38. S. J. Weiner, P. A. Kollman, D. T. Nguyen, and D. A. Case, *J. Comput. Chem.* 7, 230 (1986).
39. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner, *J. Am. Chem. Soc.* 106, 765 (1984).
40. S. J. Benkovic and S. H. Schiffer, *Science* 301, 1196 (2003).
41. J. B. Watney, P. K. Agarwal, and S. Hammes Schiffer, *J. Am. Chem. Soc.*, 125, 3745 (2003).
42. J. Villa and A. Warshel, *J. Phys. Chem. B* 105, 7887 (2001).
43. A. Warshel, *Proc. Natl. Acad. Sci. USA* 81, 444 (1984).
44. S. Piana, P. Carloni, and M. Parrinello, *J. Mol. Biol.* 319, 567 (2002).
45. Q. Cui and M. Karplus, *J. Phys. Chem. B* 106, 7927 (2002).
46. B. D. Dunietz, M. D. Beachy, Y. Cao, D. A. Whittington, S. J. Lippard, and R. A. Friesner, *J. Am. Chem. Soc.* 122, 2828 (2000).
47. D. S. Lu and G. A. Voth, *Prot. Struct. Fund. Genet.* 33, 119 (1998).
48. D. Suarez, E. Brothers, and K. M. Merz, Jr., *Biochemistry* 41, 6615 (2002).
49. S. H. Schiffer, *Biochemistry* 41, 13335 (2002).
50. M. Karplus, J. D. Evanseck, D. Joseph, P. A. Bash, and M. J. Field, *Faraday Discuss. Chem. Soc.* 93, 239 (1992).
51. J. L. Gao, *Curr. Opin. Struct. Biol.* 13, 184 (2003).
52. T. C. Bruice, *Acc. Chem. Res.* 35, 139 (2002).
53. L. Young and C. B. Post, *Biochemistry* 35, 15129 (1996).
54. S. Caratzoulas and S. D. Schwartz, *J. Chem. Phys.* 114, 2910 (2001).
55. J. K. Hwang, Z. T. Chu, A. Yadav, and A. Warshel, *J. Phys. Chem.* 95, 8445 (1991).
56. S. R. Billeter, S. P. Webb, P. K. Agarwal, T. Iordanov, and S. H. Schiffer, *J. Am. Chem. Soc.* 123, 11262 (2001).
57. P. K. Agarwal, S. R. Billeter, P. T. R. Rajagopalan, S. J. Benkovic, and S. H. Schiffer, *Proc. Natl. Acad. Sci. USA* 99, 2794 (2002).
58. P. K. Agarwal, S. R. Billeter, and S. H. Schiffer, *J. Phys. Chem. B* 106, 3283 (2002).
59. N. Diaz, D. Suarez, T. L. Sordo, and K. M. Merz, Jr., *J. Phy. Chem. B* 45, 11302 (2001).
60. A. Warshel, "Computer Modeling of Chemical Reactions in Enzymes and in Solution," Wiley, New York, 1991.
61. A. Warshel, *J. Phys. Chem.* 86, 2218 (1982).
62. A. Zacks and A. M. Klibanov, *Science* 224, 1249 (1984).
63. A. M. Klibanov, *CHEMTECH* 63, 354 (1986).
64. A. Zacks and A. M. Klibanov, *J. Am. Chem. Soc.* 108, 2767 (1986).
65. J. B. West, W. J. Hennen, J. L. Lalonde, J. A. Bibbs, Z. Zhong, E. F. Meyer, Jr., and C. H. Wong, *J. Am. Chem. Soc.* 112, 5313 (1990).
66. A. Zacks and A. M. Klibanov, *Proc. Natl. Acad. Sci. USA* 82, 3192 (1985).
67. A. M. Klibanov, *TIBS* 14, 141 (1989).
68. G. Colombo, G. Ottolina, G. Carrea, and K. M. Merz, Jr., *Chem. Commun.* 7, 559 (2000).
69. R. Z. Kazandjian and A. M. Klibanov, *J. Am. Chem. Soc.* 107, 5448 (1986).
70. J. A. McCammon, B. R. Gelin, and M. Karplus, *Nature* 585 (1977).
71. L. C. Brooks III, M. Karplus, and B. M. Pettitt, "Proteins: A Theoretical Perspective of Dynamics Structure and Thermodynamics," Wiley, New York, 1988.
72. H. Frauenfelder, S. G. Sligar, and P. G. Wolynes, *Science* 254, 1598 (1991).
73. D. S. Hartsough and K. M. Merz, Jr., *J. Am. Chem. Soc.* 114, 10113 (1992).
74. K. Griebenow and A. M. Klibanov, *Biotechnol. Bioeng.* 54, (1997).
75. P. A. Fitzpatrick, A. C. U. Steinmetz, D. Ringe, and A. M. Klibanov, *Proc. Natl. Acad. Sci. USA* 90, 8653 (1993).

76. P. A. Fitzpatrick, D. Ringe, and A. M. Klibanov, *Biochem. Biophys. Res. Comm.* 198, 675 (1994).
77. N. H. Yennawar, H. P. Yennawar, and G. K. Farber, *Biochemistry* 33, 7326 (1994).
78. C. R. Wescott and A. M. Klibanov, *J. Am. Chem. Soc.* 115, 1629 (1993).
79. G. Colombo, G. Ottolina, G. Carrea, A. Bernardi, and C. Scolastico, *Tetrahedron: Asymmetry* 3, 1205 (1998).
80. J. Uppenberg, N. Ohrner, M. Norin, K. Hult, G. J. Kleywegt, S. Patkar, V. Waagen, A. Thorleif, and T. A. Jones, *Biochemistry* 34, 16838 (1995).
81. T. Ke, F. N. Haeffner, and K. Hult, *Biophys J.* 74, 1251 (1998).
82. V. Daggett, S. Schroder, and P. Kollman, *J. Am. Chem. Soc.* 113, 8926 (1991).
83. A. J. Turner, V. Moliner, and I. H. Williams, *J. Chem. Soc.* 1271 (1997).
84. J. Bentzien, R. P. Muller, J. Florian, and A. Warshel, *J. Phys. Chem. B* 102, 2293 (1998).
85. J. Aqvist and A. Warshel, *Chem. Rev.* 93, 2523 (1993).
86. R. V. Stanton, M. Perakyla, D. Bakowies, and P. A. Kollman, *J. Am. Chem. Soc.* 120, 3448 (1998).
87. T. Ke, B. Tidor, and A. M. Klibanov, *Biotechnol bioeng.* 57, 741 (1998).
88. G. Colombo and K. M. Merz, Jr., *J. Am. Chem. Soc.* 121, 6895 (1999).
89. V. Gogonea and K. M. Merz, Jr., *J. Phys. Chem.* 103, 5171 (1999).
90. G. Carrea, G. Ottolina, and S. Riva, *Trends Biotechnol.* 13, 63 (1995).
91. A. M. Klibanov, *Trends Biotechnol.* 15, 97 (1997).
92. M. H. Vermue and J. Tramper, *Pure Appl. Chem.* 67, 345 (1995).
93. V. Luzhkov and J. Åqvist, *J. Am. Chem. Soc.* 120, 6131 (1998).
94. S. Toba and K. M. Merz, Jr., *J. Am. Chem. Soc.* 119, 9939 (1997).
95. C. S. Chen, Y. Fujimoto, G. Girdaukas, and C. J. Sih, *J. Am. Chem. Soc.* 104, 7294 (1982).
96. A. R. Fersht and W. H. Freeman, "Enzyme Structure and Mechanism," New York, 1985.
97. G. Nadig, L. C. V. Zant, S. L. Dixon, and K. M. Merz, Jr., *J. Am. Chem. Soc.* 120, 5593 (1998).
98. G. Ottolina, Personal communication.
99. A. Warshel, G. Naray-Szabo, F. Sussman, and J.-K Hwang, *Biochemistry* 28, 3629 (1989).
100. F. Haeffner, T. Norin, and K. Hult, *Biophys. J.* 74, 1251 (1998).
101. B. H. Besler, K. M. Merz, Jr., and P. A. Kollman, *J. Comput. Chem.* 11, 431 (1990).
102. A. U. C. Steinmetz, H. U. Demuth, and D. Ringe, *Biochemistry* 33, 10535 (1994).
103. A. Zacks and A. M. Klibanov, *J. Biol. Chem.* 263, 3194 (1988).
104. W. L. Jorgensen, J. Chandrasekhar, J. Madura, R. W. Impey, and M. L. Klein, *J. Phys. Chem.* 79, 926 (1983).
105. W. L. Jorgensen and J. Tirado-Rives, *J. Am. Chem. Soc.* 110, 1657 (1988).
106. A. Cheng, R. S. Stanton, J. J. Vincent, K. V. Damodaran, S. L. Dixon, D. S. Hartsough, S. A. Best, and K. M. Merz, Jr., ROAR 1.0, Pennsylvania State University, 1997.
107. S. C. Hoops, K. W. Anderson, and K. M. Merz, Jr., *J. Am. Chem. Soc.* 113, 8262 (1991).
108. D. A. Case, D. A. Pearlman, J. C. Caldwell, T. E. I. Cheatham, W. S. Ross, C. L. Simmerling, T. A. Darden, K. M. Merz, Jr., R. V. Stanton, A. L. Cheng, J. J. Vincent, M. Crowley, D. M. Ferguson, R. J. Radmer, G. L. Seibel, U. C. Singh, P. Weiner, and P. A. Kollman, AMBER 5.0, University of California, San Francisco, 1997.
109. H. J. C. Berendsen, J. P. M. Potsma, W. F. van Gunsteren, A. D. Di Nola, and J. R. Haak, *J. Chem. Phys.* 81, 3684 (1984).
110. W. F. van Gunsteren and H. J. C. Berendsen, *Mol. Phys.* 34, 1311 (1977).
111. J. L. Miller and P. A. Kollman, *J. Phys. Chem.* 100, 8587 (1996).
112. E. Hedenstrom, B.-V. Nguyen, and L. A. Silks, *Tetrahedron: Asymmetry* 13, 835 (2002).
113. M. Angoli, A. Barilli, G. Lesma, D. Passarella, S. Riva, A. Silvani, and B. Danieli, *J. Org. Chem.* 68, 9525 (2003).
114. S. Riva, *J. Mol. Catal. B: Enzymatic* 19–20, 43 (2002).
115. L. Husakova, S. Riva, M. Casali, S. Nicotra, M. Kuzma, Z. Hunkova, and V. Kren, *Carbohydr. Res.* 331, 143 (2001).
116. F. Secundo, G. Carrea, M. DeAmici, S. Joppolo di Ventimiglia, and J. S. Dordick, *Biotechnol. Bioeng.* 81, 391 (2003).
117. J. O. Rich, P. C. Michels, and Y. L. Khmelnitsky, *Curr. Opin. Chem. Biol.* 6, 161 (2002).
118. B. Danieli, M. Luisetti, M. Schubert-Zsilaveez, W. Likussar, S. Steuer, S. Riva, D. Monti, and J. Reiner, *Helv. Chim. Acta* 70, 1153 (1997).
119. M.-R Van Calsteren, Y. Bussiere, and M. C. Bissonnette, *Spectroscopy* 11, 143 (1993).
120. K.Yamasaki, H. Koda, T. Kobayashi, R. Kasai, and O. Tanaka, *Tetrahedron Lett.* 13, 1005 (1976).
121. T. Halgren, *J. Comput. Chem.* 17, 5 (1996).
122. F. Mohamadi, N. G. J. Richards, W. C. Guida, R. Liskamp, M. Lipton, C. Caulfield, G. Chang, T. Hendrikson, and W. C. Still, *J. Comput. Chem.* 11, 440 (1990).
123. J. O. Rich, B. A. Bedell, and J. S. Dordick, *Biotechnol. Bioeng.* 45, 426 (1995).
124. G. Tiana, F. Simona, Giacomo M. S. De Mori, R. A. Broglia, and G. Colombo, 13, 113 (2004).
125. J. E. Shea and C. L. Brooks III, *Annu. Rev. Phys. Chem.* 52, 499 (2001).
126. A. R. Fersht and V. H. Freeman, "Structure and Mechanism in Protein Science," New York, 1999.
127. P. Ferrara and A. Caflisch, *Proc. Natl. Acad. Sci. USA* 97, 10780 (2000).
128. X. Daura, B. Jaun, D. Seebach, W. F. van Gunsteren, and A. E. Mark, *J. Mol. Biol.* 280, 925 (1998).
129. A. Sali, E. Shakhnovich, and M. Karplus, *J. Mol. Biol.* 235, 1614 (1994).
130. L. Mirny and E. I. Shakhnovich, *Ann. Rev. Biophys. Biomol. Struct.* 30, 361 (2001).
131. H. S. Chan and K. Dill, *Annu. Rev. Biophys. Biophys. Chem.* 20, 447 (1991).
132. C. Micheletti, *Prot. Struct. Funct. Genet.* 51, 74 (2003).

133. S. Miyazawa and R. Jernigan, *Macromolecules* 18, 543 (1985).
134. G. Tiana, R. A. Broglia, H. E. Roman, E. Vigezzi, and E. Shaknovich, *J. Chem. Phys.* 108, 757 (1998).
135. S. R. Yeh, and D. L. Rousseau, *Nat. Struct. Biol.* 7, 443 (2000).
136. R. L. Baldwin and G. D. Rose, *Trends Biochem. Sci.* 24, 26 (1999).
137. Y. Levy, J. Jortner, and O. M. Becker, *J. Chem. Phys.* 115, 10533 (2001).
138. G. Tiana and R. A. Broglia, *J. Chem. Phys* 114, 7267 (2001).
139. V. I. Abkevich, A. M. Gutin, and E. Shakhnovich, *Biochemistry* 22, 10026 (1994).
140. R. A. Broglia and G. Tiana, *Proteins* 45, 421 (2001).
141. K. W. Plaxco, K. T. Simons, and D. Baker, *J. Mol. Biol.* 277, 285 (1998).
142. N. Go, *Int. J. Pept. Protein. Res.* 7, 313 (1975).
143. L. A. Mirny, V. L. Abkevich, and E. I. Shakhnovich, *Proc. Natl. Acad. Sci. USA* 95, 4976 (1998).
144. G. Tiana and R. A. Broglia, *J. Chem. Phys.* 114, 2503 (2001).
145. B. I. Dahiyat and S. L. Mayo, *Science* 278, 82 (1997).
146. G. Colombo and K. M. Merz, Jr., *J. Am. Chem. Soc.* 121, 6895 (1999).
147. H. Zhao and F. H. Arnold, *Prot. Eng.* 12, 47 (1999).
148. H. Frauenfelder and P. G. Wolynes, *Phys. Today*, 58 (1994).
149. A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, *Prot. Struct. Fund. Genet.* 17, 412 (1993).
150. H. Frauenfelder, F. Parak, and R. D. Young, *Annu. Rev. Biophys. Biophys. Chem.* 17, 451 (1988).
151. J. Martinez, M. Pisabarro, and L. Serrano, *Nat. Struct. Biol.* 5, 721 (1998).
152. V. Grantcharova, D. S. Riddle, J. V. Santiago, and D. Baker, *Nat. Struct. Biol.* 5, 714 (1998).
153. S. D. Riddle, V. P. Grantcharova, J. V. Santiago, E. Alm, I. Ruczinski, and D. Baker, *Nat. Struct. Biol.* 6, 1016 (1999).
154. E. L. McCallister, E. Alm, and D. Baker, *Nat. Struct. Biol.* 7, 669 (2000).
155. D. E. Kim, C. Fisher, and D. Baker, *J. Mol. Biol.* 298, 971 (2000).
156. D. J. Sloan and H. W. Hellinga, *Prot. Sci.* 8, 1643 (1998).
157. M. Byeroft, S. Ludvigsen, A. R. Fersht, and F. M. Poulsen, *Biochemistry* 30, 8697 (1991).
158. A. R. Panchenko, Z. Luthey-Schulten, and P. Wolynes, *Proc. Natl. Acad. Sci. USA* 93, 2008 (1996).
159. S. E. Jackson, N. elMasry, and A. R. Fersht, *Biochemistry* 32, 11270 (1993).
160. G. D. Gay, J. Ruisanz, J. L. Neira, F. J. Corrales, D. E. Otzen, A. G. Ladurner, and A. R. Fersht, *J. Mol. Biol.* 254, 968 (1995).
161. G. D. Gay, J. Ruizsanz, J. L. Neira, L. S. Itzhaki, and A. R. Fersht, *Proc. Natl. Acad. Sci. USA* 92, 3683 (1995).
162. J. L. Neira, L. S. Itzhaki, D. E. Otzen, B. Davis, and A. R. Fersht, *J. Mol. Biol.* 270, 99 (1997).
163. S. L. Kazmirski, K. B. Wong, S. E. Freund, Y. J. Tan, A. R. Fersht, and V. Daggett, *J. Mol. Biol.* 254, 968 (1995).
164. C. D. Snow, H. Nguyen, V. S. Pande, and N. Gruebele, *Nature* 420, 33 (2002).
165. I. N. Berezovsky and E. N. Trifonov, *J. Biomol. Struct. Dynam.* 20, 5 (2002).
166. I. N. Berezovsky, V. M. Kirzhner, A. Kirzhner, and E. N. Trifonov, *Proteins* 45, 346 (2001).
167. K. Liebeton, A. Zonta, K. Schimossek, M. Nardini, D. Lang, B. W. Dijkstra, and K. E. Jaeger, *Chem. Biol.* 7, 709 (2000).
168. A. Li and V. Daggett, *Proc. Natl. Acad. Sci. USA* 91, 10430 (1994).

CHAPTER 7

Simulating Enzyme-Catalyzed Reactions

Anna Bowman, Adrian Mulholland

Computational Chemistry Centre, School of Chemistry, University of Bristol, Bristol, United Kingdom

CONTENTS

ISBN: 1-58883-048-9

Handbook of Theoretical and Computational Nanotechnology
Edited by Michael Rieth and Wolfram Schommers
Volume 6: Pages (305–359)

1. INTRODUCTION

Enzymes are highly proficient and specific catalysts. Understanding the molecular basis for their outstanding catalytic efficiency, and tailoring their functions for practical applications, are major goals of modern chemistry and biochemistry. Enzymes hold great promise as potential components of nanotechnological systems and in nanoscale synthesis. The use of biological macromolecules is a most important frontier in nanotechnological research. These are areas in which computational modeling and simulation techniques are making an important growing contribution. In this chapter our aim is to describe how modeling methods can contribute to studies of enzyme mechanism. We provide an overview of the field, the current methods (including their strengths and weaknesses), practical considerations in modeling enzyme-catalyzed reactions, and a short review of some recent studies. The role of enzymes in nanotechnology is introduced, followed by a discussion of enzyme structure and function. Various modeling methods are briefly outlined and because of their importance in studying enzyme-catalyzed reactions, quantum mechanics/molecular mechanics (QM/MM) methods are covered in more detail. Some techniques for modeling reactions (e.g., optimizing transition state structures and reaction pathways) are discussed along with molecular dynamics simulations and free-energy calculations. Finally, a number of recent applications are reviewed.

For readers who may be unfamiliar with biological systems, we give a short introduction to the basic features of protein structure (e.g., the four levels of protein structure, cooperative

binding, and allosteric effects). The nature of enzyme active sites is then described. Practical methods for determining enzyme structure are briefly discussed, including X-ray crystallography and nuclear magnetic resonance. Finally (in Section 1.5) the insights afforded by experimental and computational techniques are compared.

1.1. Enzymes and Nanotechnology

Enzymes catalyze nearly all chemical reactions in biological systems [1, 2]. They have enormous catalytic power, generally increasing reaction rates by factors of a millionfold or more. They are typically highly specific both in the reaction catalyzed and in their choice of reagents. Understanding the basic principles of enzyme activity and specificity will have important potential consequences for developments in nanotechnology as outlined below. Nearly all enzymes are proteins, although RNA catalysts are also known and are referred to collectively as ribozymes [2]. A short introduction to protein structure is given in Section 1.2 below.

Nanotechnology is the creation and manipulation of useful materials and structures at the nanometer scale. Enzymes and other biological macromolecules and assemblies hold tremendous potential promise as nanotechnological components. Enzymes may also provide efficient routes to the specific synthesis and modification of nanotechnological systems. At the same time, nanotechnological techniques increasingly allow the manipulation of these biological systems at the macromolecular level. In nature, although some enzymes function alone, others associate with other proteins, cofactors, or cell components, whereas others are involved in creating highly complex nanoscale structures within living cells that make life possible. Some form specific complexes with nucleic acids or other proteins, whereas others bind to cell membranes. Enzymes are the nanomachines of cells: They efficiently catalyze chemical reactions, sometimes at rates exceeding a million reactions per second [2], without requiring elevated temperature, elevated pressure, or dangerous solvents.

Nanotechnology provides new means of analyzing and visualizing enzyme function at the molecular level. At the same time, enzymes have the potential to provide molecular scale scaffolding, cables, motors, ion pores, pumps, coatings, and chemically powered levers [3]. Although enzymes have many desirable features such as their high specificity and catalytic power, natural enzymes suffer from several drawbacks for practical applications, of which among the most important are their lack of activity for many unnatural substrates and their fragility, especially with respect to temperature and pH. Thermophilic enzymes [4], from thermophiles organisms that can function in temperatures up to and around 380 K, offer improved stability and resilience reducing this weakness. A familiar example is Taq polymerase, from the thermophillic bacterium *Thermus aquaticus*. Taq is known to retain at least partial activity at temperatures up to 97.5°C [5]. Protein design or modifications based on the ever-increasing understanding of the factors associated in protein stability will also assist in increasing the application range of many enzymes.

Protein engineering techniques make possible the production of novel proteins with improved characteristics for a particular application. For example, enhanced thermostability [6], increased enantiomeric excess [7], and change in substrate specificity [8]. Molecular biological genetic engineering tools allow easy and rapid modification of enzymes. Great progress has been made with the two main approaches of protein engineering: rational design and combinatorial methods [9]. Rational design involves the choice of a suitable scaffold, identification of the residues to change and characterization of the mutants. Combinatorial approaches consist of the creation of a library of random mutants of a particular protein. Combinatorial libraries for proteins have been created by several procedures: chemical mutagenesis, error-prone polymerase chain reaction [10], and now, most prevalently, the DNA-shuffling approach [11, 12]. Directed evolution uses random mutagenesis and multiple generations to produce new proteins with enhanced properties [13].

Nanoscale manipulation and imaging of biological macromolecules and their assemblies is possible, for example, by scanning electron microscopy (SEM), atomic force microscopy (AFM), or scanning probe microscopy (SPM). The development of laser-based technology such as the confocal microscope, laser tweezers, laser scissors, the multiphoton excitation confocal microscope, and the near field microscope offers functional imaging of living cells in

thick tissues; manipulation of single molecules, single organelles, and single cells; determination of the binding force and rate of interaction of DNA and other single molecules; surgery of chromosomes and organelles in a living cell; and the fabrication of miniature medical devices. The development of SPM (AFM, scanning tunneling, electrochemical probe, near field microscopy) has enabled the manipulation of single molecules, the preparation of novel biochips and biosensors, and the measurement of physical and spectroscopic properties of single molecules in living cells. These techniques are being applied to study single motors such as myosin [14] and kinesin [15] as well elastic proteins such as titin [16]. Such methods offer the promise of further insight into the structure and function of biological macromolecules and assemblies and also their manipulation in nanotechnological applications.

One particular current focus in nanotechnology is the use of proteins in combination with inorganic materials to create new devices and materials with exciting and useful new properties and capabilities. It can be envisaged that enzymatic nanomaterials and nanodevices will be built that will mimic or make use of biological and particularly enzyme-catalyzed reactions. For example, recombinant enzymes could be attached to nanoporous materials as long-lifetime biological reactors for a variety of applications. The possibilities for using enzymes in small-scale reactors are vast. Once a single-reaction microreactor is made, multiple reactors can be linked so that complex reactions may possibly be constructed. In addition, arrays of different enzymes could be immobilized together to catalyze a whole sequence of chemical reactions. Such a series of complex reactions may enable new possible avenues for producing energy, decontaminating waste to harmless by-products, and creating complicated synthetic reactions. Enzymatic reactions can be employed for applications such as chemical production and chemical detoxification; these are currently important in, for example, some areas of pharmaceutical synthesis.

Enzymatic micro- and nanoreactors potentially offer many advantages over conventional systems such as higher activity, increased active-site concentrations, very low pressure drops with high flow rates, and more efficient mass transfer from the solvent phase to the reacting chamber.

The potential for the nanotechnological application of enzymes is clearly vast, and this is an exciting and expanding area of research. It is obviously impossible to cover anything more than a brief description of some interesting recent developments here. Some recent examples of applications of enzymes in nanotechnology include work by Nguyen and coworkers on the design and evaluation of quasi-immortal artificial enzymes, cofactors, and arrays of enzymes via the principles underlying molecular self-assembly and supramolecular inorganic chemistry [17]. Gilardi and coworkers are working on assembling redox proteins as 'molecular Lego' and linking them to surfaces in nanodevices for biosensing and other nanotechnological applications [18]. The molecular Lego approach aims to select key domains, or building blocks, to be used, for example, to assemble artificial redox chains with the desired properties, with the aim of ultimately being able to communicate with electrode surfaces [19].

1.2. Protein Structure

In this section we will give a short introduction to the basic aspects of protein structure. This introduction should provide useful background for our subsequent descriptions of enzyme mechanism and modeling. It is aimed particularly at researchers in nanotechnological fields who may not be familiar with protein structure. What follows here is necessarily only a very brief description of some basic principles of protein structure. Many excellent detailed descriptions are available; for example, references [1, 20]. Most protein structures determined to date have been solved by the techniques of by X-ray crystallography, while nuclear magnetic resonance (NMR) methods are becoming increasingly important, particularly for smaller proteins.

Proteins are built up from a collection of 20 amino acids that are linked by peptide bonds to form a polypeptide chain (Fig. 1). There are four levels of description of protein structure. The amino acid sequence of a protein's polypeptide chain is called its primary structure. For small enzymes this chain may consist of as few as approximately 100 amino acids residues, whereas large proteins may contain over 1500 residues or more. This polypeptide chain is then folded to give the secondary structure of the protein.

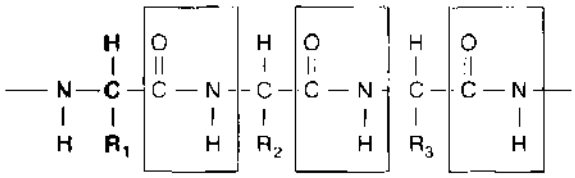


Figure 1. Most enzymes are proteins, which are built up from a basic collection of 20 amino acids. These amino acids are joined by peptide bonds to form a polypeptide chain. The peptide group is rigid and planar.

The most notable secondary structural elements are the α helix and the β sheet. Both are characterized by hydrogen bonding between groups in the main polypeptide chain. The α helix is stabilized by hydrogen bonds between the CO group of an amino acid residue and the NH group of another amino acid residue four residues ahead, where as β sheets are stabilized by hydrogen bonds between the NH and CO groups between residues that are far more distant, or even in different polypeptide chains.

The torsion angle about the imido-nitrogen and α carbon of the main chain is called phi (ϕ), and the torsion angle about the α carbon and carbonyl carbon is called psi (ψ). Most combinations of ϕ and ψ angles for amino acids are not allowed because of steric clashes between the main chain and side chains. Allowed angle pairs of ϕ and ψ can be plotted against each other in a diagram known as a Ramachandran plot. Regions in the Ramachandran plot are named after the secondary structure that results if the corresponding ϕ and ψ values are repeated in successive amino acids along the chain; for example, α helical.

Different amino acids show different preferences for different types of secondary structure. The formation of an α helix is favored by glutamate, methionine, and leucine, and β sheet formation is preferred by valine, isoleucine, and phenylalanine [20] (Fig. 2). The α helices and β sheets are linked by regions of various length and irregular shape. When these regions are on the surface of the protein they are termed loops. In addition to linking secondary structure elements, loop regions frequently participate in forming binding sites and enzyme active sites. Long loop regions are often flexible, and they are frequently involved in the function of the protein and can, for example, be centrally involved in the switch from an "open" conformation of an enzyme, which allows access to the active site, to a "closed" conformation, which shields reactive groups in the active site from water.

A large variety of conformational changes are known for enzymes. Short loops, usually in the core, are labeled turns. Secondary structure elements usually pack in to simple motifs. Several motifs combine to form compact globular structures called domains; this is described as the protein tertiary structure. The specific shape adopted by the chain gives an enzyme its functional properties; for example, forming specific binding sites or catalytic (active) sites. If a protein consists of only one polypeptide chain, then it is referred to as monomeric. However, there are many proteins that are composed of several subunits, which can be identical (as in citrate synthase, a homodimer [21]) or different (as in RNA polymerase [22]). These proteins are said to have a quaternary structure (i.e., the relationship of subunits in the overall structure; Fig. 3).

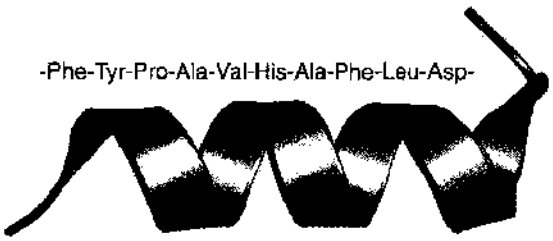


Figure 2. The primary structure of a protein is the amino acid sequence of its polypeptide chain. The polypeptide chain is folded to give the protein's secondary structure. Shown here is an α helix, one of the two main types of ordered secondary structure.

Figure 3. The secondary structure is folded to form the tertiary structure of the protein. The quaternary structure describes how subunits of a protein related to each other. The tertiary structure of the monomer (on the left) and the quaternary structure of the dimer (on the right, showing the arrangement of the two subunits; the two subunits are drawn in different colors) of the enzyme citrate synthase [21] are shown.

The visualization of proteins with computer graphics has had a dramatic effect on molecular modeling and in facilitating structural analysis. Proteins are most commonly represented using the "line" (atoms are shown as points and bonds as lines) or "ball and stick" (atoms are shown as small spheres and bonds as thin sticks) representation. However, it is not always desirable to display all the atoms individually because of the large number of atoms in a protein structure, and alternative display options such as "ribbon" and "tube" represent proteins in an uncluttered style. Proteins are also commonly represented using the "cartoon" representation: here α helices are represented by cylinders and β sheets by flat arrows. Some of the most commonly used molecular representations are shown in Fig. 4. Many molecular graphics packages are available, some commercial and some academic or more freely available. Some popular visualization packages include Sybyl [23], MOE (Molecular Operating Environment) [24], Protein explorer [25], VMD (Visual Molecular Dynamics) [26], QUANTA [27], and Insight [28].

Many of the important and interesting problems studied in enzyme systems involve noncovalent interactions between the ligand and enzyme. Visualizing molecular surfaces in terms of, for example, the van der Waals radii or solvent accessibility can greatly assist the study of such interactions. Connolly [29, 30] developed a widely used algorithm for calculating the van der Waals and solvent accessible surfaces. There are many ways to represent surfaces; some examples of typical surfaces representations are shown in Fig. 5. It is possible to calculate properties on a surface to gain a greater understanding of possible interactions.

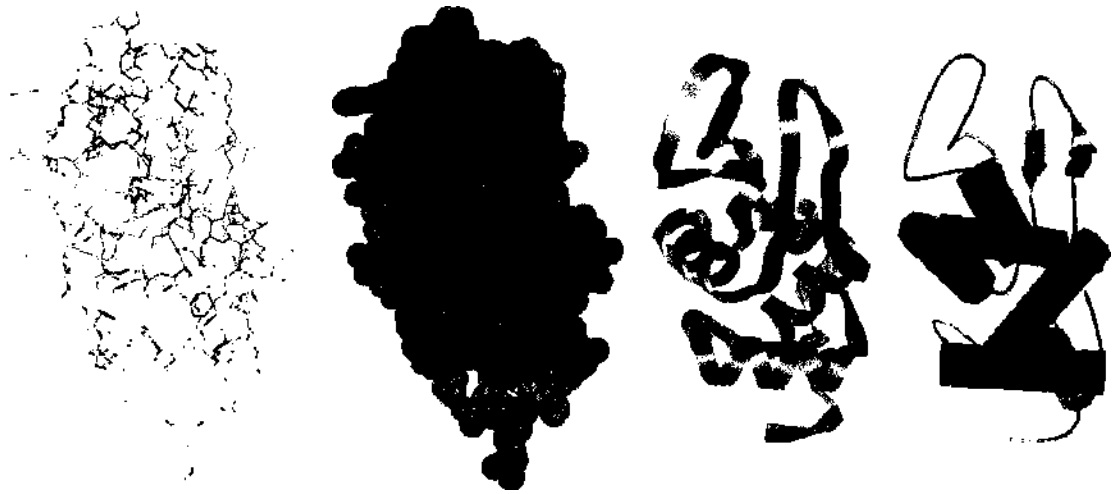


Figure 4. Some common graphical representations of proteins, illustrated using the crystallographically determined structure of hen egg white lysozyme [161]. Each representation shows different aspects of structure. Clockwise, starting from top left: "line" (in which atoms are shown as points and bonds as lines), "space filling" (in which atoms are represented by their van der Waals spheres), "ribbon" (which suggests the secondary structure of the protein with a flat tube following the protein peptide backbone), and "cartoon" (which indicates the secondary structure of the protein: α helices are represented as cylinders and β sheets as flat arrows).

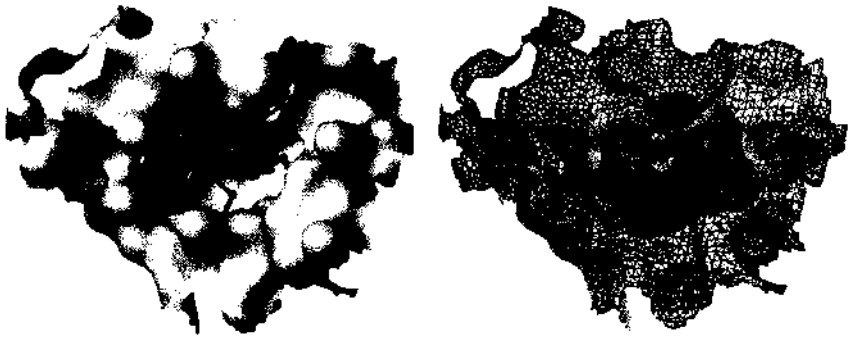


Figure 5. Graphical representations of the Connolly surfaces of proteins illustrated using the crystal structure of hen egg white lysozyme [16]. Left: part of the enzyme's surface is shown in green opaque with the ligand shown in red ball-and-stick representation. Right: part of the enzyme surface is shown in the line style and the ligand's surface is shown in opaque; both surfaces are mapped with their electrostatic potential. Mapping by electrostatic potential can indicate possible hydrogen-bonding sites in a molecule.

These properties, for example, the molecular electrostatic potential or lipophilic potential, can then be mapped on the surface using an appropriate color scheme for convenience. Many current visualization packages allow rapid calculation and visualization of molecular surfaces and properties.

1.3. Enzyme Active Sites

The active site of an enzyme is a relatively small region that binds the substrates (and cofactors in cases in which they are involved). The active site contains the residues that are directly involved in the chemical reaction that the enzyme catalyzes as well as the residues involved in binding. The substrates are typically bound at the active site by multiple weak interactions, such as hydrogen bonds, electrostatic interactions, van der Waals forces, and hydrophobic interactions. Active sites are typically found in clefts or crevices. Proteins are complex dynamic entities, exhibiting a wide range of internal motions, of which some have functional significance [31]. Many enzymes undergo large conformational changes as part of their reaction cycles [1], and their function and relationship to the chemical changes occurring in the reaction is significant.

One important model of enzyme-substrate interaction and related conformational changes is the induced-fit model proposed by Koshland [32]. This model proposes that the initial interaction between the enzyme and substrate is relatively weak, but that these weak interactions quickly induce conformational changes in the enzyme that strengthen binding and bring catalytic groups close to the substrate. After binding takes place, the mechanism of catalysis generates transition-state complexes and subsequently products of the reaction.

Many proteins have multiple binding sites, and in some cases cooperative binding of ligands to these sites is known to occur. For example, when a molecule of oxygen is bound to one subunit of hemoglobin (which is a tetramer), this enhances the binding of three additional oxygen molecules to the same protein molecule [2]. The binding of oxygen to hemoglobin is said to be cooperative. Enzymes in which the binding of a compound with a site on the protein, other than the active site, can affect the properties of active site in the same enzyme molecule are said to be allosteric. The allosteric property is involved in the regulation of enzyme activity. For example, threonine deaminase catalyzes the committed step from threonine to α-ketobutyrate. The end product of the pathway, isoleucine, serves as an allosteric inhibitor, regulating carbon flow through the pathway to maintain a proper balance of metabolites for efficient protein synthesis as a form of feedback inhibition [2].

1.4. Investigating Enzyme Structure and Function

The three-dimensional structure of an enzyme dictates its physical, chemical, and biological properties. Experimental studies have been among the most important sources of insight into protein, and particularly into enzyme structure. A wide variety of experimental methods

can afford insight into enzyme structure and so provide information about this key aspect of enzyme modeling. The most important experimental technique for studying protein structure has been X-ray crystallography. With X-ray crystallography the interaction of X-rays with electrons in the enzyme molecules arranged periodically in the lattice in a protein crystal is used via the techniques of protein crystallography to obtain an electron density map of the molecule, which is interpreted to an atomic model. A well-ordered crystal of the enzyme is required. Finding appropriate coordinates to produce suitable crystals can be challenging, particularly for membrane proteins. One indication of the precision of a protein structure determined by X-ray crystallography structures is its resolution, ranging from low resolution, where just the shape of the protein maybe revealed, to high resolution (1–2 Å), where most atomic positions can be determined. It should be remembered that the resolution is a global measure of model quality (dependent, for example, on the nature of the crystal and experimental conditions), and even in high-resolution structures there can be considerable uncertainty because of the dynamic nature of proteins, which gives rise to conformational variability.

NMR methods are increasingly important for protein structure determination, particularly for smaller proteins. In NMR the magnetic spin properties of atomic nuclei are used to build up a list of distance constraints between atoms in the enzyme from which the three-dimensional structure of the protein can be determined. This method does not require the growth of crystals, as it can be used on concentrated protein solutions. However, it is restricted to smaller enzymes. Recently, high-resolution X-ray powder diffraction has been used to solve and refine protein structure [33, 34]. Again, this method has the advantage of not needing a protein crystal.

The Research Collaboratory for Structural Bioinformatics (RCSB) makes three-dimensional biological macromolecular structural data from all experimental techniques available. The RCSB protein data bank (PDB) is the single worldwide repository for the processing and distribution of three-dimensional structure data of large molecules of proteins and nucleic acids [35].

In combination with other experimental techniques such as site-directed mutagenesis and genome sequencing, a fuller understanding of protein structure and the function of the active site residues can be achieved. However, to really understand chemical reaction mechanisms, the chemical intermediates involved in the reaction need to be investigated. Kinetic and spectroscopic techniques are widely used for this purpose. However, unstable and short-lived intermediates are generally very difficult to isolate or to study by these methods. Central to questions of reactivity and catalysis is the transition state of a reaction, but these are even more difficult to examine directly by experiments. Kinetic isotope effects can provide indirect indications of the nature of a transition state. Mutation studies of enzymes can also give insight to the roles of active site residues, but interpretation of this experimental type of study is frequently not straightforward because of the possible changes in protein structure [36] or chemical mechanism [37] that mutagenesis may cause.

1.5. Modeling Enzyme-Catalyzed Reactions

1.5.1. Role of Modeling

Computer simulation and modeling methods have the potential to make a significant contribution to studies of enzyme mechanisms, complementing experimental investigations [38, 39]. Taking experimental data, particularly structural information, as a starting point, calculations can be used to address questions that are difficult to resolve by experiment alone. For example, the structures of transition states and reaction intermediates can be modeled and analyzed, as can the interactions responsible for stabilizing these species in an enzyme [40, 41]. Computational modeling does not require any mutation of the enzyme or substrate in question, and so it is possible to study the "wild-type" reaction (i.e., the reaction as it occurs in the naturally occurring wild-type or native enzyme). The wild-type reaction can be difficult to study experimentally because of the naturally fast catalytic turnover. The interaction of modeling with experiments has dual importance: to use experiments to validate modeling methods and also to generate new suggestion for experimental work.

1.5.2. Goals of Modeling

Numerous computational molecular modeling methods are available, ranging from quantum chemistry to the interpretation of three-dimensional structure. The various techniques have different strengthens and weaknesses. It is important to choose an appropriate method for the particular type of problem under investigation. Given below is a discussion of modeling methods useful for investigating a variety of aspects of enzyme reactions. Modeling an enzyme-catalyzed reaction can enable differentiation between alternative mechanisms and examination and prediction of molecular interactions and reactions. One objective of modeling is to analyze the contributions to catalysis; for example, by identifying catalytic residues. Another purpose is the modeling of dynamic systems, enabling the study of vibrational modes and conformational change. Modeling also allows detailed study of unstable or excited species such as transition states and intermediates. As enzymes are characteristically designed to stabilize transition states, detailed information on the nature of a transition state can potentially help in the design of strong, transition state–like inhibitors. Ultimately, it should be possible to calculate the energy profile for the reaction in the enzyme and to dissect energetic contributions to lowering the activation energy. Understanding enzyme-catalyzed reactions is a detailed and complex problem, which requires many levels of analysis; therefore, computational techniques have a very important role to play.

2. MODELING METHODS

Modern modeling methods can provide detailed insight into enzyme reactions with different strengths and weaknesses have been applied to various aspects of enzyme-catalyzed reactions. What follows is a discussion of the theory and methods that are of particular importance for simulating enzyme-catalyzed reactions. The section is designed to introduce researchers in various nanotechnological fields to some basic aspects of modeling methods. Many excellent texts exist, which include descriptions of the various methods available, and interested readers may find these interesting [42–44]. The introduction will begin by looking at molecular mechanics calculations and some popular force fields. The following section will discuss electronic structure calculations, covering *ab initio* orbital calculations including Hartree–Fock theory and basis sets. It will move on to include the correlation energy. Next, a brief overview to density-functional theory (DFT), including a discussion of the relative merits of both DFT and *ab initio* methods. Then, the semiempirical molecular orbital methods will be examined, including a short review of the three most widely available methods, MNDO, AM1, and PM3. Finally, so-called linear scaling electronic structure calculations methods, and in particular the "divide-and-conquer" method, are introduced and discussed.

2.1. Molecular Mechanics Calculations

Molecular mechanics (MM) is a method of low computational cost that can handle long dynamics simulations of entire solvated proteins and can provide good structural geometries. However, this method is generally incapable of describing a chemical reaction, which is quantum mechanical in nature, methods of this type are covered in Sections 2.2–2.7.

Potential functions or force fields are mathematical functions that return energy as a function of conformation. A number of MM potential functions have been developed and successfully applied for studying the dynamics and conformational changes of enzymes [45–47]. Typically, force fields are sums of terms that correspond to bond, angle, torsion angle, van der Waals, and electrostatic interaction energies as functions of conformation (Fig. 6). Many force fields used for modeling proteins and other biological macromolecules have the basic form

$$V(r) = \sum_{\text{bonds}} \frac{k_r}{2}(r - r_0)^2 + \sum_{\text{bond angles}} \frac{k_\theta}{2}(\theta - \theta_0)^2 + \sum_{\text{torsion angles}} [k_\phi(1 + \cos(n\phi - \delta))]^2 + \sum_{\text{atom pairs}} \frac{q_i q_j}{r_{ij}} + \frac{A_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^6} \quad (1)$$

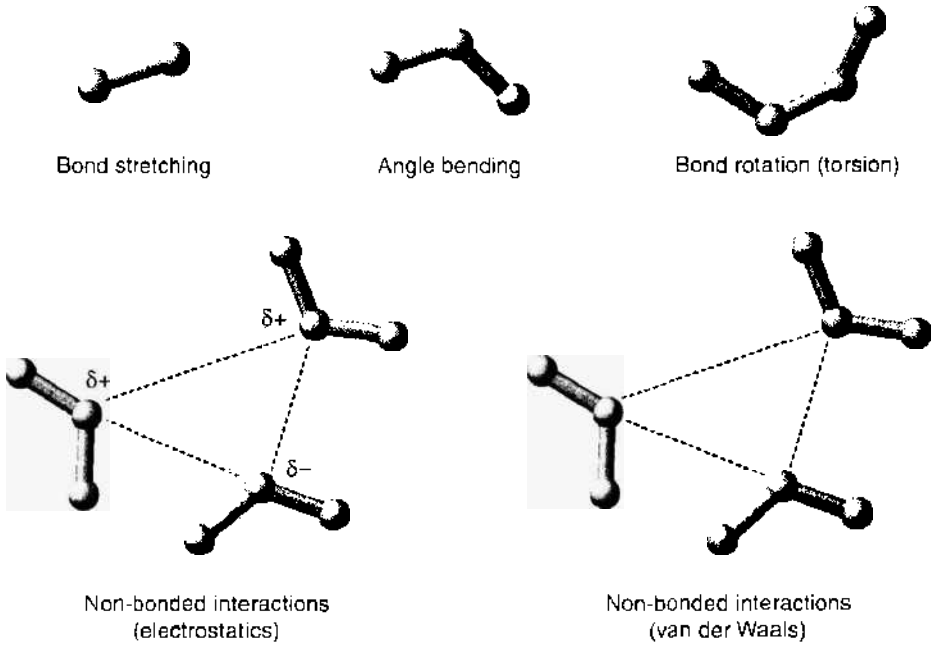


Figure 6. Empirical "molecular mechanics" potential functions have been developed that can model protein structure and dynamics effectively and efficiently. A simple functional form is generally used (see Section. 2.1). There are four key contributions to a molecular mechanics force field: bond-stretching, angle-bending and torsional terms, and nonbonded interactions (van der Waals and electrostatic interactions).

where k_r and k_θ are force constants in harmonic potentials for bonds and angles with equilibrium values r_0 and θ_0 and actual values r and θ respectively. Cosine functions are usually used for the potential associated with torsion angle variation with n between 1 and 5 and the phase angle δ. A Lennard–Jones (12-6) potential is used for attractive dispersion (with coefficient C_{ij}) and close-range repulsion (with coefficient A_{ij}) nonbonded interactions, where r_{ij} is the distance between nonbonded atoms i and j. The electrostatic interactions between nonbonded atoms depend on the atomic partial charges q_i and q_j. These charges can be determined by fitting them to electrostatic potential calculated for small molecules using *ab initio* methods; for example (as in the AMBER [Assisted Model Building with Energy Refinement] forcefield), by fitting to (scaled) *ab initio* calculated interaction energies, or by simpler schemes.

Both the functional form, for example, as shown in Eq. (1), and parameters such as k_i, and k_θ are required to define a force field. The main aim of force fields used in molecular modeling is to reproduce structural properties. All force fields are empirical, and two force fields, although having an identical functional form, may have very different parameters. A popular force field for modeling proteins and nucleic acids is AMBER [46, 48], which was developed by the Kollman research group at University of California at San Francisco. The Karplus group at Harvard devised CHARMM (Chemistry at Harvard using Molecular Mechanics) [45, 49], which is widely used and developed and can also be used to model structure dynamics. Other molecular mechanical force fields developed especially for biomolecules include GROMOS (Groenigen Molecular Simulation) [50, 51] and MMFF (Merck Molecular Force Field) [52, 53]. The MM2 force field [54] for predominantly for organic molecules is also popular. It should be remembered that parameters are different from programs and that it is possible, for example, with the TINKER program [55] to use any of several common parameter sets, such as AMBER-95, CHARMM22, MM2, MM3, OPLS-AA, OPLS-UA, and the in-house TINKER parameter set.

To reduce the computational expense of a simulation, the number of nonbonded interactions can be limited by applying a cutoff beyond which no interactions are calculated. A cutoff range is applied after which no van der Waals energies are calculated; this is normally between 8 and 12 Å. The cutoff is often used with a "switching function," which multiplies the van der Waals interactions and causes them to go smoothly to zero at the cutoff distance.

In infinite periodic systems, an alternative to the use of cutoffs is the Ewald sum technique. In the most favorable case, called "particle–mesh–Ewald" the scaling is $N \log N$ [56].

Because of their low computational expense, MM can handle long dynamics simulations of entire solvated proteins (as discussed in Sections 6.1 and 6.3). However, for modeling a chemical reaction, a quantum mechanical description is required, and this type of method is covered in the next section.

2.2. Electronic Structure Calculations

Electronic structure calculations have proven extremely useful in studies of small molecules (on the order of tens of atoms) and their reactions. In quantum mechanics, the properties of the system are exclusively dependent on the wavefunction of the system. The wavefunction can be described as a function consisting of a time-dependent part and a spatially dependent part

$$\Psi(x, y, z, t) = \Psi(x, y, z) \cdot \exp\left[-\frac{iEt}{\hbar}\right] \tag{2}$$

The Born–Oppenheimer approximation [57] states that, as the masses of the nuclei are significantly greater than that of the electron, the electrons can adjust instantaneously to any changes in the positions of the nuclei. Therefore, the electronic wavefunction depends only on the positions of the nuclei and not on their momenta. Hence, under the Born–Oppenheimer approximation the many electron wavefunction Ψ and corresponding energy may be obtained by solving the time-independent Schrödinger equation in which the nuclear positions are fixed. In its simplest form, the time-independent Schrödinger equation is

$$\hat{H}\Psi = E\Psi \tag{3}$$

where the nonrelativistic Hamiltonian $\hat{H}$ is the energy operator, and in atomic units takes the form

$$\hat{H} = -\frac{1}{2}\sum_i \nabla_i^2 - \sum_i \sum_\alpha \frac{Z_\alpha}{|\boldsymbol{r}_i - \boldsymbol{d}_\alpha|} + \sum_i \sum_{j>i} \frac{1}{|\boldsymbol{r}_i - \boldsymbol{r}_j|} + \sum_\alpha \sum_{\beta>\alpha} \frac{Z_\alpha Z_\beta}{|\boldsymbol{d}_\beta - \boldsymbol{d}_\alpha|} \tag{4}$$

where $\boldsymbol{r}_i$ denotes the electron positions, $\boldsymbol{d}_\alpha$ the nuclear positions, and Z_α the nuclear charges. Respectively, these four terms represent the kinetic energy of the electrons, the electron–nuclear attractions, the electron–electron repulsions, and the nuclear–nuclear repulsion (a constant because the nuclei are assumed to be frozen). It is only possible under the Born–Oppenheimer approximation to solve the Schrödinger equation exactly for the simplest molecular species, such as H_2^+. In polyelectronic systems, other approximations must also be made, as outlined below.

2.3. *Ab Initio* Molecular Orbital Calculations

Ab initio molecular orbital (MO) techniques [58] can be used to calculate potential surfaces and reaction pathways and to optimize the structures of stable small molecules and transition states. These methods can give a very accurate description of the electronic structure of small molecules; however, they require computationally expensive calculations.

The Schrödinger equation can be extended to polyelectronic systems by applying further approximations. Electronic structure theory comprises designing sensible approximations to the wavefunction while maintaining a balance between accuracy and computational expense. The energy of an approximate wavefunction Ψ is given by the expectation value of the Hamiltonian:

$$E = \frac{\int \Psi * \hat{H}\Psi \, d\tau}{\int \Psi * \Psi \, d\tau} \tag{5}$$

The variational principle states that the energy of an approximation to the true wavefunction of a molecule can never be less than the exact energy. Therefore, the lower the energy

calculated for an approximate solution, the better the wavefunction is. The most common methods to minimize the energy of a wavefunction are based on the Hartree–Fock approach, which is described below.

The polyelectronic wavefunction must obey the Pauli Principle by being antisymmetric with respect to the interchange of the coordinates of any two electrons [59].

$$\Psi(x_1, x_2, \ldots, x_i, \ldots, x_j, \ldots) = -\Psi(x_1, x_2, \ldots, x_j, \ldots, x_i, \ldots) \tag{6}$$

where $x_i = (r_i, \sigma_i)$ represents the space and spin coordinates of an electron. Because of the antisymmetry, no two electrons can possess exactly the same set of quantum numbers.

2.3.1. Hartree–Fock Theory

The most common basis for approximate solutions to the electronic Schrödinger equation is to make the orbital approximation. The Hartree–Fock approach takes the polyelectronic wavefunction as the antisymmetrized product of one-electron wavefunctions (spin orbitals):

$$\Psi(x_1, x_2, \ldots) = \hat{A} \prod_{i=1}^{N} \psi_i(x_i) \tag{7}$$

where $\hat{A}$ is the antisymmetry operator that ensures that the Pauli Principle is obeyed. Equation (7) can be written as a Slater Determinant [60, 61]:

$$\Psi(x_1, x_2, \ldots) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \psi_1(x_1) & \psi_1(x_2) & \cdots & \psi_1(x_3) \\ \psi_2(x_1) & \psi_2(x_1) & \cdots & \psi_2(x_1) \\ \vdots & \vdots & & \vdots \\ \psi_N(x_1) & \cdots & & \cdots \psi_N(x_N) \end{vmatrix} \tag{8}$$

The spin orbital $\psi_i(x_i)$ is the product of a spatial orbital and a spin function $\psi_i(r_i)\alpha(\sigma_i)$ or $\psi_i(r_i)\beta(\sigma_i)$, where the two spin functions $\alpha(\xi)$ and $\beta(\xi)$ correspond to spin "up" and spin "down," respectively. The linear combination of atomic orbitals (LCAO) is a further approximation that is generally made. It expresses the spatial orbitals $\psi_i(r_i)$ as a linear combination of atomic orbitals

$$\psi_i(r) = \sum_{j=1}^{n_{\text{basis}}} c_{ij}\phi_j(r) \tag{9}$$

where c_{ij} is the appropriate weighting coefficient of the atomic orbital. The atomic orbitals are normally known as basis functions, and the n_{basis} basis functions, which define a molecular orbital, are known as the basis set. Basis sets are described in more detail in Section 2.3.2.

The Hartree–Fock approach uses the Slater Determinant (Eq. [8]) and the Variational Principle to determine the basis set expansion coefficients, which minimize the energy of the wavefunction with respect to the coefficients. The coefficients c_i are determined by using a self-consistent field (SCF) procedure, which treats the electronic distributions in the molecular orbitals as independent of one another, whereas in reality the positions of electrons in different molecular orbitals are dynamically correlated. The SCF method proceeds by gradually refining the individual electronic solutions that correspond to progressively lower total energies until the point is reached at which the results for all the electrons are unchanged, when they are said to be self-consistent. The Hartree–Fock SCF method in the LCAO approximation is applied using the Roothaan-Hall equations [62, 63].

The Hartree–Fock method has a formal N^4 scaling, where N is the number of basis functions, as a result of the need to calculate the four-index integrals. However, with modern computing power, it is not quicker to compute each integral, store it, and recall it when required; instead, each integral is computed from scratch each time it is required. This method is referred to as "direct SCF," [64] and for larger systems it is a necessity. In certain instances, where data transfer is slow compared to CPU speed, the crossover between direct and conventional methods maybe so low that direct methods are always preferred.

As system size increases further, methods can be employed to reduce another of the limiting factors in the *ab initio* electronic structure calculations: calculating the Coulomb interaction between electrons. The use of fast-multipole methods [65, 66] enables the formal N^2 scaling of Colomb integral evaluation to be become linear (i.e., proportional to N). Linear methods to evaluate the exchange integrals have also been developed [67]. At present, the bottleneck in performing Hartree–Fock calculations is the diagonalization of the Fock matrix (which has a formal N^3 scaling), and research is being carried out on trying to reduce the scaling of this step, too [68].

2.3.2. Basis Sets

The identification of a suitable basis (e.g., a set of atomic orbitals or plane wave basis, as used with density-functional Car–Parrinello molecular dynamics simulations) is an important aspect of planning or assessing *ab initio* calculations. Larger basis sets give more accurate results, but they are more computationally expensive. There are two main types of basis functions that are commonly used in quantum chemical calculations on molecular systems: Slater-type orbitals (STOs) and Gaussian type orbitals (GTOs). The STO basis functions give accurate results, but the integrals involved can be difficult and time-consuming to solve. Hence, GTO basis functions are more commonly used in *ab initio* calculations. The major advantage in using Gaussian functions is that the product of two Gaussian functions can be expressed as a single Gaussian function located along the line joining the centers of the two starting Gaussian functions, which leads to more easily evaluated integrals then with STO. The mains disadvantages, in contrast to STO functions, are that GTOs lack a cusp at the nucleus, and they also have different behavior at large distances from the nucleus. To emulate the level of accuracy given by STO functions each atomic orbital is represented by one or more linear combinations of Gaussian functions known as contracted Gaussian functions (CGFs) or contracted Gaussian type orbitals (CGTO).

The basis set nomenclature reflects the number of CGFs used to represent the orbitals in a given atom. At the most basic level are the minimal basis sets; for example, STO-3G [69]. This contains one CGF for each atomic orbital (each CGF is composed of three Gaussian functions, hence 3G) whose coefficients are not varied the atomic orbital. This description is a poor one because the atomic orbitals cannot expand or contract in response to the surrounding molecular environment. Improved results are obtained with "double-zeta" or "spilt-valence" basis sets. A double-zeta basis set represents each atomic orbital as a linear combination of two contracted Gaussian functions, thus doubling the number of basis functions. This allows for a better description of the fact that the electron distribution may be different in different directions in a molecule. A more common alternative is the split valence basis set, in which only the valence orbitals are increased and a single function is used to describe inner shell or core electrons. For example, in the standard 6-31G basis set [70], a single CGF (formed from six Gaussian functions) describes a core orbital, and the valence orbitals are described by CGFs with three Gaussian functions for the contracted part plus a single Gaussian function for the diffuse part.

Further improvements can be obtained by including d-orbital type functions for first- and second-row nonhydrogen atoms; the basis set can be increased further by including the polarization (i.e., p) functions on hydrogen and helium. These polarization functions are denoted by a (d) and (d,p) respectively, for example, 6-31G(d), and allow for orbital distortion. A similar modification is the addition of highly diffuse functions to account for regions of electron density far away from the nuclei. Denoted by a +, as in 6-31+G, this is particularly important when studying anions [71] and species with lone pairs, and it can also improve results for nonbonding interactions and hydrogen bonding [72].

2.4. Correlation Energy

The Roothaan-Hall equations are not applicable to open-shell systems, which contain one or more unpaired electrons. Two important lines of attack have been developed to deal with open-shell systems, spin-restricted Hartree–Fock (RHF) and spin-unrestricted Hartree–Fock (UHF). RHF uses combinations of singly and doubly occupied molecular orbitals

and closed-shell systems are included as a subsection of RHF theory. Doubly occupied orbitals use the spatial functions for electrons of both α and β spin. The UHF theory [73] uses two distinct sets of molecular orbitals: one for spin α and one for β spin electrons (Fig. 7).

The Hartree–Fock result provides a first approximation to the wavefunction. In the SCF method, the electrons are assumed to moving in an average potential because of the other electrons; the instantaneous position of one electron is not influenced by the presence of a neighboring electron. In reality, the motions of the electrons are correlated and electrons tend to avoid each other more than the Hartree–Fock result would indicated, giving a SCF energy that is higher than the exact energy. The energy corresponding to this interaction is the energy difference between the true energy of the system and the SCF energy and is referred to as the correlation energy. The following description is only a very brief overview of correlation energy methods; many excellent texts exist on the topic, and the interested reader is referred to these [42, 44].

Changes in correlation energy are particularly large during reactions, and neglecting correlation energy can lead to misleading or incorrect results. With Hartree–Fock methods, only one electronic configuration is included. This means that the electrons in H_2 are predicted to spend equal time on each nuclei, even when they are infinitely separated. A single configuration (or single determinant) model is often accurate around the minimum energy geometry of the molecule in the ground state, where one configuration dominates the wavefunction. At higher energies, away from the minimum, other electronic configurations may become significant. Hence, when performing reaction modeling calculations there is a need to move beyond the HF level of theory and include correlation energy effects. In contrast, for systems where there is little change in correlation energy, such as in studies of conformational change Hartree–Fock can give reasonable results [74].

Electron correlation *ab initio* MO calculations can incorporate correlation effects in a number of ways. In the configuration–interaction (CI) approach the total wavefunction is written as a linear combination of configurations (SCF wavefunctions), in which excited states are included in the description of an electronic state. Minimizing the energy of the CI wavefunction gives the best values for the coefficients of the excited states in the expansion. Møller–Plesset [75] perturbation theory is an another approximate method to estimate electron correlation, which gives useful results at acceptable cost for many systems. This approach adds higher excitation as perturbations to the Hartree–Fock theory; the sum of the zero- and first-order energies correspond to the Hartree–Fock energy. Corrections to the energy at the higher orders can be then calculated. The second-, third-, and fourth-order corrections are denoted as MP2, MP3, and MP4, respectively. The Møller–Plesset energy

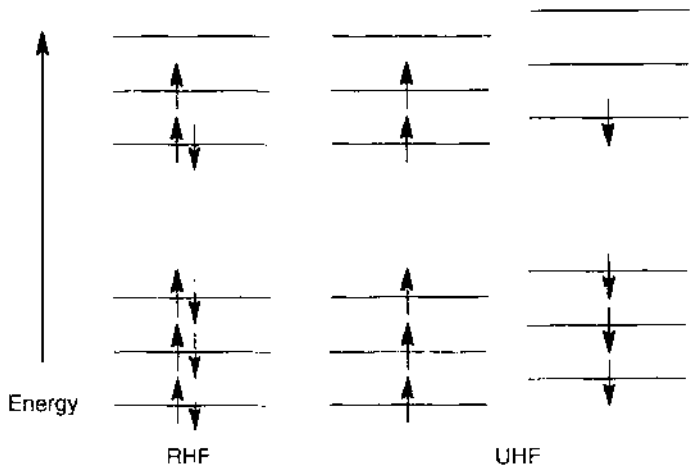


Figure 7. A simple illustration of the conceptual difference between the spin-restricted Hartree–Fock (RHF) and the spin-unrestricted Hartree–Fock (UHF) models of *ab initio* quantum chemistry. These represent two approaches that have been developed to deal with systems that contain one or more unpaired electrons.

corrections are calculated after the SCF calculation has been carried out. Møller–Plesset methods can in many cases give a reasonably accurate estimate of the correlation energy much more quickly than CI calculations; however, in contrast to the CI approach, Møller–Plesset methods do not indicate which other electronic configurations are contributing to the correlation energy. CI methods determine the energy by a variational procedure, and hence the energy is an upper bound of the exact energy. With the Møller-Plesset perturbation method, this is not always so, and the energy can be lower than the exact energy; however, this is rarely a problem, as interest is mainly in energy differences rather than in total energies.

2.5. Density Functional Theory

Although *ab initio* methods use wavefunctions to describe electronic structure, DFT [76, 77] uses electron density as the primary way to describe the system. Hohenberg and Kohn [78] proposed the basic theory underlying the DFT approach by showing that there is a relationship between the total electron energy and the overall electronic density. This electron density uniquely defines the ground-state energy and all the electronic properties of a ground state of a molecule.

Kohn and Sham [79] suggested a practical way to apply the Hohenberg–Kohn theorem. By introducing an expression for the electron density and applying the appropriate variational condition, they developed the Kohn–Sham equations. As in Hartree–Fock theory, these equations have to be solved iteratively until self-consistency is achieved. The crucial part of the DFT approach is the exchange-correlation functional. The exact form of this crucial functional is unknown. One way to estimate these exchange and correlation functionals is to use the local density approximation (LDA) [80]. However, the LDA has been shown to be inadequate for some chemical problems; for example, bond strengths are often overestimated. For this reason, extensions have been developed beyond the LDA. One common approach is to use gradient-corrected "nonlocal" functionals that depend on the gradient of the density at each point in space, and not just on its value. These corrections are added to both the exchange and the correlation terms. Becke has proposed the most popular gradient correction to the exchange functional (B or B88) [81]. The correlation functional of Lee–Yang–Parr is also widely used [82]. Hybrid methods A popular method is the B3LYP functional; this is a combination of the Becke gradient corrections (for the exchange functional) and the gradient correction of Lee, Yang, and Parr (for the correlation) with a Hartree–Fock exchange term [83]. Methods, such as B3LYP, that require a Hartree–Fock calculation in addition to a DFT calculation are known as hybrid methods. The B3LYP method is widely used for the investigation of chemical problems and often performs well.

Since 1990 there has been much debate on the relative merits of DFT compared to *ab initio* methods based on the molecular wavefunction. The main conclusion from these comparisons is that in many cases DFT is a more cost-effective method of achieving a given level of accuracy. The formal scaling behavior of DFT is N^3, where N is the number of basis functions used to represent the Kohn–Sham orbitals; this is better than Hartree–Fock, which scales formally by N^4, and much better than higher-level theories (formally MP2 scales at N^5 and CCSD at N^6 [44]). Various procedures have been used to improve the formal scaling, and with modern techniques DFT is a true linear scaling method [84]. One of the disadvantages of the DFT method is that correlation effects cannot be explicitly estimate because they are implicitly included within the functional. In addition, there is no systematic way to improve the functional by subsequently adding more sophisticated terms; in fact, the parameters used are determined experimentally.

Recently, a new approach, the self-consistent charge density functional tight-binding (SCC-DFTB) method [85, 86], has been reported. It is based on a second-order expansion of the Kohn–Sham total energy with respect to charge density variations in the LCAO (tight-binding) framework. This method has been shown to yield results with similar accuracy to MP2/6-31G* and B3LYP/6-31G* but is very fast—similar in speed to semiempirical methods such as AM1 and PM3 [87].

2.6. Semiempirical Molecular Orbital Methods

Ab initio calculations, such as those at the Hartree–Fock level and at other higher-level theories such as MP2 and CI, yield a accurate description of the electronic structure of a molecule. However, even with modern computing power, these methods can quickly become impractical as the size of the molecular system increases. They become very demanding as a result of the computational resources (cpu time, memory, and disk space) needed to evaluate a large number of two-electron integrals and carry out the SCF calculation.

Semiempirical MO methods are considerably less computationally intensive and can be practically applied to very large molecules. They are fast enough to be able to be employed for carrying out QM/MM molecular dynamics simulations, which makes them of particular relevance for applications to enzyme reaction studies of reasonably large systems. Semiempirical MO methods are characterized by their use of parameters derived wholly or partly from experimental data for some aspect of the calculation. Their speed is a result of using a minimal valence shell basis set for each atom (core atoms are not treated explicitly) and neglecting the vast majority of integrals that arise in MO theory (see below). However, the speed of semiempirical methods for large systems is controlled by the need for repeated matrix diagonalization in the SCF procedure.

One of the timesaving aspects of semiempirical MO methods is that they consider only the valence electrons explicitly. The core electrons are combined with the nuclei of the atom to form positively charged atomic "cores." The widely used semiempirical methods MNDO [88], AM1 [89], and PM3 [90] (see below) all employ the neglect of diatomic differential overlap (NDDO) approximation [91]. In the NDDO convention, the differential overlap between atomic orbitals on different atomic centers is ignored. This eliminates the need for the three- and four-center integrals that arise in the full *ab initio* SCF procedure and simplifies the solution of the Roothaan–Hall equations. However, this is a considerable approximation in modeling a variety of enzyme reactions.

Semiempirical methods are considerably less demanding of computational resources than *ab initio* calculations. They therefore can be applied to much larger systems, and in more extensive calculations. These properties have led to semiempirical methods being extensively used in studies of enzyme reaction ranging from "supermolecule" calculations of transition state structures and reaction pathways to hybrid QM/MM molecular dynamics simulations. Semiempirical methods have been applied in QM/MM and divide-and-conquer studies. However, they also have well-known weaknesses and limitations, and these must be taken into account when selecting an appropriate a method to apply to a problem and interpreting its results. Because of the importance of semiempirical MO methods in enzyme reaction modeling, the three most widely available methods, and their accuracy, are reviewed in outline below. The accuracy of results obtained using semiempirical methods is dependent on the parameters used, and the uses of system specific reaction parameters are reviewed.

2.6.1. Modified Neglect of Diatomic Overlap

Dewar and Thiel introduced the modified neglect of diatomic overlap (MNDO) method [88], which covered a wider range of elements, such as aluminum, silicon, and lead, than had been possible with earlier methods such as modified intermediate neglect of differential overlap method (MINDO) [92]. However, the original form of the method only included s and p orbitals, and so transition metals could not be treated, as they require a basis set containing d orbitals. In more recent work, d orbitals have been included for the heavier elements in the MNDO/d method of Thiel and coworkers [93].

One problem with MNDO is its inability to model accurately hydrogen bonds—a considerable limitation in the study of biological molecules [94]. This error is the result of a tendency to overestimate the repulsion between atoms when they are separated by a distance roughly equal to the sum of their van der Waals radii (see also the discussion of the AM1 and PM3 methods below). MNDO energies are too positive for sterically crowded molecules and too negative for molecules containing four-membered rings [95]. Activation barriers calculated with MNDO have found to be too high in many cases. Other limitations and difficulties encountered with the MNDO semiempirical method include that hypervalent

compounds, such as sulfoxides and sulfones, are found to be too unstable; oxygenated substitutents on aromatic rings are out of plane; the peroxide bond is systematically too short by about 0.17 Å; and the C–X–C angle in ethers and sulfides is too large by ~9° [42, 96]. Other semiempirical methods, in particular AM1 and PM3, have been developed to try and overcome some of these limitations of MNDO.

2.6.2. Austin Model 1

It was obvious after some experience with MNDO that certain systematic errors existed. This led to the Austin Model 1 (AM1) [89] being developed by Dewar's group, who were at the time based at the University of Texas, Austin. AM1 was designed to eliminate the problems in MNDO related to the repulsions at about van der Waals separations by adding two to four extra Gaussian functions for each atom. AM1 was generally an improvement on MNDO, predicting more accurate activation energies and hydrogen bond strengths and dealing better with hypervalent molecules like sulfoxides and sulfones [96]. Some of the problems with MNDO still persist, such as the underestimation of the peroxide bond length by ~0.17 Å. Furthermore, a few new defects were introduced: AM1 energies for alkyl groups were too negative and were found to be too positive for nitro groups [95]. The inability of AM1 to deal reliably with phosphorus compounds in many cases can be considered to be its main failing: AM1 phosphorus has a false and very sharp potential barrier at 3.0 Å [95]. This leads to normally symmetric geometries being distorted; for example, in P_4O_6, the supposedly equal P···P bonds differ in length by 0.4 Å [96]. AM1 also incorrectly predicts bifurcated hydrogen bonds (i.e., for molecules capable of donating two hydrogen bonds, such as water, AM1 predicts two hydrogen bonds will be formed to one acceptor atom). However, higher-level calculations show the formation of a single hydrogen bond only [97]. In general, despite these difficulties, AM1 is a distinct improvement over MNDO, particularly for modeling biological molecules, and has been used to study a large number of enzyme-catalyzed reactions.

2.6.3. PM3

The parameterization of MNDO and AM1 was done mainly by applying chemical knowledge and intuition. However, Stewart developed a method that he called PM3 [90] (being the third parameterization of MNDO, with AM1 being the second) by following a somewhat different approach to parameterization. The parameters for the PM3 method were derived using an automated parameterization procedure. In this PM3 reparameterization, the same approach as in AM1 for dealing with the repulsions at about van der Waals separation is retained, but in PM3 only two Gaussian functions per atom are used for all atom types. PM3 parameters were developed for a wider range of atom types than AM1. It has been stated that PM3 can be considered as a good local minimum in parameter space for the NDDO functional form. PM3 has been found to more accurate then AM1 for predicting hydrogen bonds geometries, for example, water dimmer (bifurcated with AM1) [98, 99], and it also improves on AM1s treatment of phosphorus [95]. However, it also has limitations: Almost all sp^3 nitrogens are predicted to be pyramidal, contrary to experimental observation; hydrogen bonds are too short by ~0.1 Å; and the charge on nitrogen atoms is often of incorrect sign and unrealistic magnitude as a result of an error in the parameterization procedure [40, 97].

MNDO, AM1, and PM3 all have problems with the treatment of some conformational properties. Bonds that have partly double bond character often give rotational barriers that are significantly too low. This is particularly prevalent for rotation around the C—N bond in amides, where values of 5–10 kcal mol^{-1} are obtained. This difficulty means that semiempirical methods may perform poorly for peptide bonds [39, 100]. A MM correction is often applied, which adds a force field rotational term to the C—N bond to raise the value to 20–25 kcal mol^{-1}, which is comparable to experimental data [95]. Another difficulty with semiempirical methods is that parameters have not been developed for all atom types, so a system that includes an atom for which parameters do not exist cannot be modeled semiempirically. However, parameters exist for most common atom types found in biomolecular molecules, so such situations are rare, although metalloenzymes can be problematic.

Both AM1 and PM3 are fast, reasonably robust over a large range of chemical functionality, and included in many molecular modeling packages [101]. They continue to be applied

for a wide range of applications. For situations in which existing methods are deemed unsuitable, the specific-reaction parameter model can be employed.

2.6.4. Specific Reaction Parameter Model

Proposed by Rossi and Truhlar [102], the specific-reaction parameter (SRP) or system-specific parameter (SSP) model is one in which standard parameters of a semiempirical model are adjusted to yield a better performance for a problem or class of problems. This method has been applied by many groups; for example, Cui and Karplus employ AM1 with specific reaction parameters (AM1-SRP) for studies of triosephosphate isomerase [103]; Tresadern and coworkers use the PM3 method with specific reaction parameters (PM3-SRP) to examine extreme tunneling in methylamine dehydrogenase [104], Ridder and coworkers use system specific AM1 parameters (AM1-SSP) for studies of glutathione S-transferase [105], in which only the sulfur was reparameterized; and AM1-SRP have also been used in studies of short-chain acyl-CoA dehydrogenase [106].

Semiempirical methods have been widely used in various types of application to study enzyme reactions, both in QM/MM and divide-and-conquer methods.

2.7. Divide-and-Conquer–Type Methods

So-called linear scaling methods aim to increase the efficiency of electronic structure computation, particularly to allow calculations on larger systems. Linear scaling implies that the method scales linearly with system size; however, scaling behavior is informative only of the relative time one system takes compared to another of a different size and cannot give information on the absolute time required for a calculation.

The divide-and-conquer approach, described by Yang and coworkers [107], is particularly suited to calculations based on atomic orbitals. In this method, partition functions are used to divide the electron density into subsystem contributions, and the subsystem densities are determined separately. There is a common Fermi energy for the entire system to ensure proper normalization of the electron density. It is then possible to carry out QM calculations using typical standard semiempirical Hamiltonians (i.e., MNDO, AM1, or PM3) and to include solvent by either explicitly incorporating quantum mechanical water molecules or by using continuum solvation models [108] such as the COSMO (conductor-like screening model) method [109]. In this way it is possible to study biological molecules in solution with up to ~15,000 atoms. Divcon [110] is a program developed in the group of Kenneth Merz Jr. (Pennsylvania State University) that uses the divide-and-conquer strategy, which has been applied to studying biological macromolecules [111, 112].

The results of divide-and-conquer calculations indicate that quantum effects such as charge transfer are important even for simple chemical processes such as solvation [113]. The divide-and-conquer method has also been used to study the benzylpenicillin acylation mechanisms of the class A TEM1 beta-lactamase [114]. Semiempirical (AM1 and PM3) divide-and-conquer methods have also been used to calculate solvation free energies for a series of small charged and neutral compounds, amino acids, several small proteins, and Dickerson's DNA dodecamer [111]. A high level of accuracy (similar to and sometimes more accurate than high level *ab initio* methods) was found, but the divide-and-conquer algorithm allows much larger systems to be studied.

This computational method may well become an important tool in investigating thermodynamic and mechanistic aspects of biomolecular compounds; for example, enzyme reactions and protein folding. In particular, it may be useful for some aspects of drug design, as it allows for quantum mechanical effects such as polarization and charge transfer, which may prove to be important in determining binding affinity. Similarly, full quantum chemical calculations on whole proteins may be useful in nanotechnological applications; for example, for calculating electronic densities of states. However, at present it is feasible only to use divide-and-conquer methods to calculate single-point energies because of the high computational demands of calculating the necessary derivates. Some improvements in both linear scaling algorithms and computing power are still needed to enable useful statistical simulations to be performed where electrons are treated quantum mechanically.

MOZYME [115, 116] is a another alternative linear scaling formulation of semiempirical molecular orbital theory designed for application to large molecules, and enzymes in particular, as the name suggests. MOZYME uses localized MOs to solve the NDDO SCF equation. MOZYME calculations on systems with up to several thousand atoms are possible, as with localized orbitals the computational time scales with the number of atoms. This method has been used to study reactions in enzymes [116, 117] and has been incorporated into MOPAC2000 [96]. However, it has limited accuracy in describing transition state and pathways because of the limitations discussed with AM1 and PM3 (see Section 2.6). The MOZYME package has recently been used (with the ONIOM [our own *N*-layered integrated molecular orbital and molecular mechanics] approach) to study the enzyme citrate synthase [118].

2.8. Empirical Valence Bond Methods

The empirical valence bond (EVB) method has been developed for simulations of enzyme-catalyzed reactions by Warshel and coworkers [119]. This method has been applied to the study of a wide range of reactions in enzymes and in solution [120]. As this method has been widely reviewed [121–123], it is only described here briefly. The EVB method represents the reacting system as a combination of (e.g., ionic and covalent) resonance forms. The corresponding Hamiltonian for each form is given by simple empirical force fields. The potential energy is given by solving the related secular equation. The electronic interaction Hamiltonian is built using terms extracted from empirical values or *ab initio* surfaces for the reaction [124]. The rest of the protein and solution are modeled by an empirical force field.

The EVB treatment offers a representation of intersecting electronic states and is useful for exploring environmental effects on chemical reactions in condensed phases [123]. However, the main advantage of the EVB approach is that it can give accurate results provided that the incorporated empirical terms are carefully chosen. This can be achieved by calibrating the free-energy surfaces by comparison with experimental data for reference reactions in solution. However, as in any valence bond representation, if the valence bond forms do not represent all important resonance forms, unusual or unexpected reaction pathways can be missed. The EVB method is an effective method for carrying out free-energy perturbation/umbrella sampling calculations for enzyme reactions [121].

An appealing feature of the EVB method is that when modeling a reaction, there is no need to follow a geometrical reaction coordinate (see Section 5). Instead, energy can be used as a reaction coordinate by following a path between valence bond optima. A mapping procedure is followed that moves gradually from the reactant to the product. During mapping, the change in both the solute structure and charge is taken into account. This EVB umbrella sampling method finds the correct transition state in the combined solute–solvent reaction coordinate. This includes the ability to evaluate nonequilibrium solvation effects [125]. Other strengths of the method have been discussed elsewhere [120].

Despite its seemingly simple appearance, the EVB method is a powerful quantitative approach that has now become a widely adopted tool for studying reactions in condensed phases [126–128]. The EVB method has given insight into the mechanisms of many different enzymes. Recently, EVB simulations have been used to study the alternative nucleotide insertion mechanisms for T7 DNA polymerase [129] and to investigate the catalytic mechanism of human aldose reductase [130].

3. QUANTUM MECHANICS/MOLECULAR MECHANICS

3.1. Methodology

A good approach to modeling enzyme-catalyzed reactions is provided by combined QM/MM methods. QM/MM methods are of current and growing interest in the field of computational enzymology. A small reactive part of the system is treated quantum mechanically (i.e., by an electronic structure method of one of the types described in Sections 2.2–2.7 above); this allows the electronic rearrangements involved in a chemical reaction, for example, during bond breaking and making, to be modeled. The large nonreactive part is described molecular

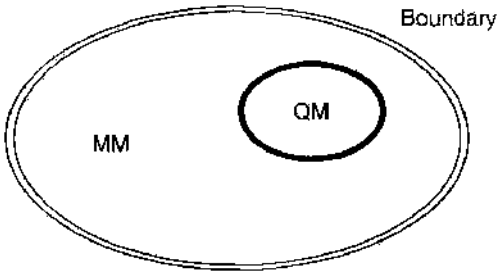


Figure 8. Combined quantum mechanics/molecular mechanics (QM/MM) methods are an attractive and practical approach for modeling enzyme-catalyzed reaction mechanisms. In the QM/MM approach, a small region at the enzymes active site is treated by a quantum chemical method while the surrounding protein and solvent is modeled by an empirical molecular mechanics potential function. It is often necessary to apply a boundary term because of the finite size of the simulation system.

mechanically, and the two regions are able to interact. The combination of the efficiency and speed of the MM force field with the versatility and range of applicability of the QM method allows reactions in large systems to be studied (Fig. 8).

The idea was first applied to an enzyme-catalyzed reaction by Warshel and Levitt in 1976 in their study of the lysozyme reaction mechanism [131]. Recently, the great potential of QM/MM calculations has been realized and implementations at the semiempirical and *ab initio* molecular orbital, and DFT QM levels have been applied to numerous enzyme systems. A variety of different approaches for partitioning the system into QM and MM regions, and QM/MM interaction schemes, have been developed and investigated. Because of its popularity and importance for studying enzyme-catalyzed reactions, the QM/MM approach is discussed in greater detail below.

The next Section (3.2) describes various QM/MM coupling schemes (for integration of the QM and MM regions). Then a brief overview of the basic theory of QM/MM methods is given, followed by a discussion of the various partitioning that can be employed. The practical aspects of modeling enzymes reactions are addressed, covering the choice and preparation of the starting structure, the various options available at the system set up, the choice of QM theory level, and the definition of the QM region. The aim is to preset an overview of current QM/MM methods and to introduce technical details of how they may be applied to model enzyme-catalyzed reactions, for example, in nanotechnology research. The interested reader can find additional description within Refs. [132, 133].

3.2. Interactions Between the Quantum Mechanical and Molecular Mechanical Regions

One of the main differences between various QM/MM models is the type of QM/MM coupling employed (i.e., in how the interactions [if any] between the QM and MM systems are treated). A useful classification system was put forward by Bakowies and Thiel [134], in which they differentiate between models of increasing levels of sophistication, from models of types A to D as described below.

A. Model A is the simplest and represents a simple "mechanical" embedding of the quantum region. The interactions between the QM and MM regions are treated in a purely classical way (i.e., the QM system is represented by [MM] point charges in its interaction with the MM environment, meaning the polarization of the QM region is not included).
B. Model B includes electrostatic interactions between the QM and MM regions, using suitably defined classical point charges in the MM region that enter the core QM Hamiltonian. The relaxation of the density matrix because of this external perturbation may be identified as the polarization of the QM region.
C. Model C is an extended version of model B; it also includes some simple polarization of the MM region in the presence of the electric field generated by the QM region.
D. Model D is the most refined and includes the self-consistent polarization of the MM and QM regions, with an iterative method.

Model A is the most straightforward implementation of a combined QM/MM model, but it is one in which the QM and MM regions do not interact in the quantum chemical calculation. In models of type A the QM/MM energy of the whole system $E_{\text{TOTAL}}^{\text{QM/MM}}$ is determined as a simple subtraction:

$$E_{\text{TOTAL}}^{\text{QM/MM}} = E_{\text{TOTAL}}^{\text{MM}} + E_{\text{QM region}}^{\text{QM}} - E_{\text{QM region}}^{\text{MM}} \tag{10}$$

where $E_{\text{TOTAL}}^{\text{MM}}$ is the MM energy of the whole system, $E_{\text{QM region}}^{\text{QM}}$ is the QM energy of the QM region, and $E_{\text{QM region}}^{\text{MM}}$ is the MM energy of the isolated QM region. This approach can be applied to all combinations of theory levels (e.g., QM/QM) and also forms the basis for the multilayer ONIOM method [135]. The ONIOM approach has recently been used (with the MOZYME package) to study the enzyme citrate synthase [118]. The polarization of the QM region by the MM environment is considered to be crucial for most studies of enzyme systems; type A models do not include polarization as a QM/MM system. Extensions can be considered, including variations of the point charges of the QM atoms, for example, to reflect chemical changes or polarization. However, as a QM/QM system in which the lower-level theory is a semiempirical method such as PM3, polarization is included at the lower level [136].

Type B models have to date been most widely used for QM/MM studies on enzyme reactions. Type B models allow polarization of the QM system by the MM system by including the charges of the MM group in the QM calculation. The electronic calculation therefore includes the effects of the MM system. Electrostatic interactions between the QM and MM regions in such models are accounted for by including the MM atomic charges in the Hamiltonian for the QM region (through the one-electron integrals). As no electrons are present on the MM atoms, the lack of steric interactions is compensated for by classical van der Waals terms (e.g., Lennard–Jones) between QM and MM atoms. This means it is necessary to assign MM van der Waals radii to the QM atoms. One limitation of current approaches of this type is that the same van der Waals parameters are used for the QM atoms throughout a simulation, as the chemical nature of the groups may change (and their interactions), use of unchanging MM parameters may be inappropriate.

When the QM region is treated by a semiempirical model it was proposed by Field et al. [137] to treat the electrostatic interactions with the MM atoms as if they where semiempirical "cores" (which represent the nucleus combined with inner electrons by an *s* orbital function [88]). In some *ab initio* QM/MM models the point charges on the MM atoms are replaced by Gaussian distributions to "blur our" the atomic charge [138].

Type C models go beyond type B models by also including polarization of the MM region (e.g., through a polarizable dipole model [139–142]). Not yet widely developed, type D models represent the most complex level of QM/MM coupling, including self-consistent polarization of the MM region through an iterative procedure. At present, extension to models of types C and D are computationally much more intensive and may not always yield better results [143]. These QM/MM methods are challenging because of the increased computational expense required for the calculation of polarization of the MM system, and also because the MM force fields that have been developed for biological macromolecules do not allow for changes in charge distribution. QM/MM methods, which include polarization of the MM system, have been developed for smaller molecules. Polarizable force fields for biological molecules are the subject of much current research and will be an important part in the direction of future QM/MM research [139–142]. QM/MM calculations can assist in the development of polarizable MM force fields; for example, in assessing polarization effects for small (QM) regions in large biomolecules [144].

3.3. Basic Theory

The basic theory of QM/MM methods has been covered by many different authors [134, 137, 145, 146], so what is presented here is simply a brief overview. The energy of the

whole system E can be written in terms of an effective Hamiltonian $\hat{H}_{\text{eff}}$ and the electronic wavefunction of the QM atoms ψ:

$$E = \frac{\langle \psi | \hat{H}_{\text{eff}} | \psi \rangle}{\langle \psi | \psi \rangle} \tag{11}$$

The effective Hamiltonian can be considered as:

$$\hat{H}_{\text{eff}} = \hat{H}_{\text{QM}} + \hat{H}_{\text{MM}} + \hat{H}_{\text{QM/MM}} + \hat{H}_{\text{Boundary}} \tag{12}$$

where $\hat{H}_{\text{QM}}$ is the pure quantum mechanical Hamiltonian, $\hat{H}_{\text{MM}}$ is the classical MM Hamiltonian, and $\hat{H}_{\text{QM/MM}}$ is the hybrid QM/MM Hamiltonian representing interactions between the QM and MM systems. Finally, $\hat{H}_{\text{Boundary}}$ is the Hamiltonian for any boundary terms applied to the simulation system to represent the effects of the bulk surroundings. The total energy of the system is then given by

$$\hat{H}_{\text{eff}} \Psi(r, R_\alpha, R_M) = E(R_\alpha, R_M) \Psi(r, R_\alpha, R_M) \tag{13}$$

where Ψ is the electronic wavefunction of the QM system. The electronic wavefunction is a function of the coordinates of the electrons r, the coordinates of the nuclei in the quantum system R_α, and the coordinates of the atoms in the MM region R_M. From the definition of the effective Hamiltonian $\hat{H}_{\text{eff}}$, the total energy of the system can be written as the sum of four contributions

$$E_{\text{eff}} = E_{\text{QM}} + E_{\text{MM}} + E_{\text{QM/MM}} + E_{\text{Boundary}} \tag{14}$$

The energy of the QM atoms E_{QM} is calculated in a standard molecular orbital or DFT calculation, applying the PM3 semiempirical molecular orbital Hamiltonian, for example. The energy of the atoms in the MM region E_{MM} is given by a molecular mechanics force field: it is defined by a potential function, such as Eq. (1) above, including terms for bond stretching, bond angle bending, dihedral and "improper" dihedral angles, electrostatic interactions (usually point partial charges from the MM force field represent the MM atoms), and van der Waals interactions (the van der Waals interactions are usually modeled by a Lennard–Jones 12-6 potential exactly as they are for most MM force fields for biomolecules). The boundary energy E_{Boundary} arises (as in MM simulations) because the simulation system can only include a finite number of atoms, so terms to reproduce the effects of the bulk must be included [147]. The system can be truncated to reduce the amount of computer time required; for example, only a part of the whole protein (e.g., a sphere around the active site) may be included in the simulation. When using this approach it is necessary to include restraints or constraints in the boundary region to force the atoms belonging to it to remain close to their positions in the crystal structure. Typically, scaled harmonic restraints are adopted for the atoms bordering on the MM region, whereas more distant atoms from the center of the system under investigation may be held fixed (e.g., in the stochastic boundary molecular dynamics approach; see Section 7.3). The atoms in the boundary region can influence the QM and MM regions; for example, through their atomic charges.

The QM/MM interaction energy $E_{\text{QM/MM}}$ is found by application of the QM/MM Hamiltonian and typically consists of terms resulting from nonbonded interactions between QM and MM atoms, the electrostatic interactions, the van der Waals interactions, and any bonded interaction terms. The van der Waals QM/MM interactions are usually calculated by a molecular mechanics procedure (e.g., through Lennard–Jones terms), exactly as the corresponding interactions would be calculated between MM atoms not interacting through bonding terms. Therefore, appropriate van der Waals parameters need to be assigned for each QM atom. The van der Waals terms describe dispersion and exchange-repulsion interactions between QM and MM regions, and these terms differentiate MM atoms, including those of the same charge, in their interactions with the QM system. The van der Waals interactions are significant at close range and can have a considerable bearing on interaction energies and geometries. The (MM) parameters for the QM atoms can be optimized to reproduce experimental or high level *ab initio* results for small complexes [148, 149]. Parameters may be optimized to reproduce interaction energies and hydrogen bond geometries.

3.4. Partitioning Schemes

In most QM/MM studies of enzymes, it is necessary to separate covalently bonded parts of the system into QM and MM regions. Some amino acid side chains may participate directly in the reaction, undergoing chemical change as part of the mechanism, and must therefore be included in the QM region. Similarly, other side chains will play binding roles, and a MM representation could be inadequate in some cases; for example, for particularly strong binding interactions. Similarly, it may be more practical to treat only the reactive parts of large cofactors or substrates by quantum chemical methods. In most enzymes, therefore, there is a need to be able to partition covalently bonded molecules into QM and MM regions. There has been considerable research into methods for QM/MM partitioning of covalently bonded systems. There are two general techniques that can be employed: firstly a "frozen" bond orbital to satisfy the valence shell of the QM atom at the QM/MM junction; for example, the local SCF (LSCF) method [150, 151] or the generalized hybrid orbital (GHO) method [152]. Alternatively, a QM atom or pseudoatom can be added to allow a proper bond at the QM/MM frontier; for example, the link atom method [137] or the connection atom method [143]. These various methods for QM/MM partitioning are described in more detail below.

3.4.1. Local Self-Consistent Field Method

In the LSCF approach [150, 151, 153], a strictly localized bond orbital, also known as a frozen orbital, describes the electrons of the frontier bond. This frozen orbital is used at the QM/MM junction (i.e., for the QM atom at the boundary). The electron density of the orbital is precalculated, often using small models. The orbitals are not allowed to adapt during the QM/MM calculation. The orbitals must be parameterized for each system, basis set, and energy calculation method. This approach has been used at the semiempirical [153] and *ab initio* [151] levels. The LSCF approach avoids the need for dummy atoms and properly describes the chemical properties of the frontier bond. However, it can be difficult to employ, especially at the *ab initio* level [154].

3.4.2. Generalized Hybrid Orbital Method

The GHO method for QM/MM calculations at the semiempirical MO level uses hybrid orbitals as basis functions on the frontier atom of MM fragment [152]. This method negates the need for reparameterization, which was necessary with the LSCF method. The classical frontier atom is described by a set of orbitals divided into two sets of auxiliary and active orbitals. The active orbitals are included in the SCF calculation, and the auxiliary orbitals generate an effective core potential for the frontier atom (Fig. 9). Parameters for the orbital

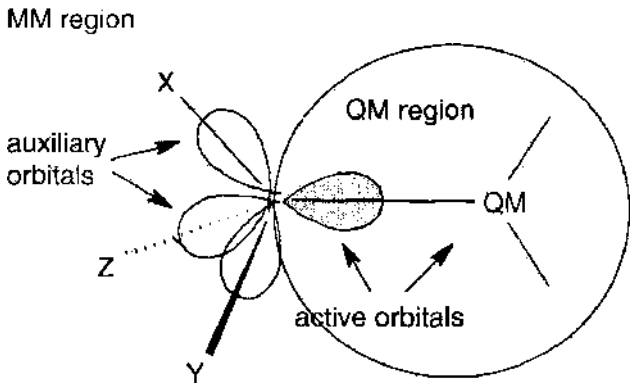


Figure 9. The generalized hybrid orbital method for partitioning a covalently bonded system for a quantum mechanics/molecular mechanics (QM/MM) calculation [152]. The covalent bond connecting the QM fragment, and the MM fragment, with a "frontier atom" F, is assumed to be a σ bond. Hybrid atomic orbitals are used as basis functions on F. These hybrid orbitals are partitioned into auxiliary and active orbitals. The active orbitals are optimized with the all the other atomic orbitals of the QM fragment in the self-consistent field calculation.

sets have been calculated at the semiempirical level, but to date the GHO method has not been extended to the DFT or *ab initio* levels.

In contrast to the LSCF method, the semiempirical parameters for the frontier atom are optimized to reproduce bonding properties of full QM systems. As a result, the GHO method enables the transferability of the localized orbitals and means that no specific parameterization of the active orbital is required for each new system. An analogous solution in DFT-based QM/MM calculations is to freeze the electron density at the QM/MM junction [155].

3.4.3. Link Atom Methods

The "dummy junction atom" or "link atom" approach introduces so-called link atoms to satisfy the valence of the frontier atom in the QM system [156] (Fig. 10). Usually this link atom is a hydrogen atom [137], but other atom types have also been used, such as a halogenlike fluorine or chlorine [157]. Despite some criticism of the link atom approach, for example, because of the additional degrees of freedom it introduces, and the fact that, for example, a C—H cannot exactly replace a C—C covalent bond, the simplicity of this method means it has been widely used in QM/MM studies of proteins. The results can strongly depend on the placement of the link atom and also on the number of MM atoms excluded from the classical electrostatic field that interacts with the QM region. Reuter and coworkers have found the LSCF and link atom approaches to perform with similar accuracy in semiempirical QM/MM calculations [158]. On the basis of these results they also recommended that the dummy atom should interact with the entire classical part apart from the few closest bonded classical atoms [158].

3.4.4. Connection Atom Method

The connection atom method aims to replace the link atom by a monovalent pseudoatom parameterized to reproduce the behavior of the covalent bond being split [143, 159]. The connection atoms have a dual character because they interact with the other QM atoms as a specially parameterized QM atom and with the other MM atoms as a standard carbon atom. This avoids the problem of a supplementary atom in the system, as the connection atom and the classical frontier atom are one. However, the need to reparameterize for each type of covalent bond at every quantum level is a considerable task [154]. The connection atom approach has been implemented at both the semiempirical (AM1 and PM3) [143] and DFT levels of theory [159]. Validation tests carried out by Antes and Thiel indicate that the connection atom approach is superior to the link atom approach [143].

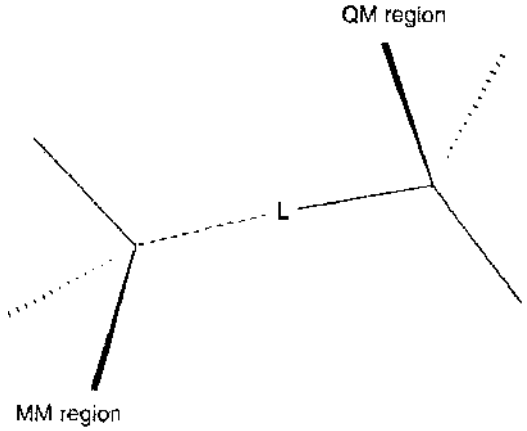


Figure 10. The link atom or dummy junction atom approach for partitioning a covalently bonded molecular system into quantum mechanical (QM) and molecular mechanical (MM) regions for a QM/MM calculation [137]. A link atom is introduced to satisfy the valence of the frontier atom in the QM system.

4. PRACTICAL ASPECTS OF MODELING ENZYME-CATALYZED REACTIONS

In this section, the practical considerations in the study and the simulation of enzyme-catalyzed reactions are outlined. It begins by looking at the variables involved in selecting a suitable starting structure and at how to set the system up using the three-dimensional structure as a starting point. Then the choices of QM theory level available to the modeler when performing QM/MM calculations are discussed. Finally, the selection of the size of the QM region and the position of any severed bonds in the QM/MM system set up is addressed.

4.1. Choice of Starting Structure

A detailed, accurate structure is necessary for a successful QM/MM simulation of an enzyme-catalyzed reaction. In general, in practice, this means a high-resolution X-ray crystallographic structure is the first requirement. A crystal structure of an enzyme with no ligands bound at the active site may, however, be a poor model of the real, reactive enzyme-substrate complex, because significant conformational changes might result from substrate binding. Reactive enzyme–substrate complexes are generally too unstable to be crystallized or trapped in crystals for native enzymes under normal conditions. In practice, structures of enzymes complexed with substrate analogs or other inhibitors usually give the vital information on which to base the binding conformation of the substrate in the model. Sometimes modification of a cofactor into its activated form is required. For example, in recent studies on *p*-hydroxybenzoate hydroxylase (PHBH) [41, 160], the structure of the flavin was modified from its crystallographic structure to the reactive C4a-hydroperoxyflavin form. This structure is a good model of the reactive complex, as no significant conformational changes are expected during the reaction cycle. Structures of the products can, of course, also be used, and the reaction modeled in reverse to give the reactants. There is also the possibility of trapping substrates and intermediates in crystallographic experiments: for example, by using less active mutant enzymes or less active substrates [161] or through rapid changes in pH or temperature (e.g., through freezing crystals).

Because of the low electron density of hydrogen atoms, protein X-ray crystallography often gives little or no information about the positions of protons. It is, therefore, important to assign correctly the protonation states of titratable amino acids residues and other groups and to decide between possible tautomers and rotamers. An error regarding the choice of charges can be damaging to the reliability of an enzyme model and can lead to disruption of the protein structure. It could well also affect the calculated reaction energetics. To check the proton allocation, either geometry minimization or MD simulations can be carried out and the structure examined to ensure that it has been maintained correctly around the residues in question [41, 162]. It can also be beneficial to carry out initial calculations of pK_as; for example, by finite difference Poison–Boltzmann methods [163]. Another potential complication when using a crystallographic protein structure is the presence of alternative conformations of side chains or ligands. This effect is the result of the fact that a crystal structure represents an average of all molecules in the crystal during the entire time of the experiment, and therefore a mixture of conformations may be observed. In this case, a single conformation must be chosen for the starting model.

4.2. Setting up the Simulation System

After selecting a crystal structure as the starting point for a simulation, thought must be directed to optimizing the model to achieve the best balance of accuracy and computer resources. The initial model complex should have any structural strain relaxed by energy minimization. Solvation should be included either explicitly, where equilibration will usually be necessary (e.g., by molecular dynamics), or implicitly by methods such as the Poisson–Boltzmann continuum solvation model. Sometimes MD simulations can help the system adjust favorably to structural alterations that may have occurred, such as moving from an inhibitor to a substrate complex.

4.3. Choice of Quantum Mechanical Theory Level for Quantum Mechanical/Molecular Mechanical Modeling of an Enzyme Reaction

Compromises will nearly always need to be made with regard to the level of QM theory when carrying out QM/MM calculations on an enzyme-catalyzed reaction. Semiempirical MO methods, despite known inadequacies (e.g., as outlined in Section 2.6), such as producing poor molecular geometries for some molecules and poorly describing hydrogen bonds in some cases, can perform well for some systems. *Ab initio* calculations at the Hartree–Fock level normally give good geometries and conformational energies but often provide poor transition state structures and activation energies. Higher-level *ab initio* methods, such MP2 and the CI approach can improve on Hartree–Fock results, as they include electron correlation, but these appreciably increase the computational expense of running the simulation. DFT-based QM/MM methods are becoming increasingly important and should in many cases offer a good combination of acceptable accuracy and speed.

Ideally, MD or Monte Carlo simulations of a reasonable length should be performed. Such simulations are possible with the use of semiempirical QM/MM methods. The performance of semiempirical QM/MM methods can be improved by being tailored for a particular reaction system by reparameterization [102] (see Section 2.6.4).

4.4. Definition of the Quantum Mechanical Region in Quantum Mechanical/Molecular Mechanical Modeling of an Enzyme Reaction

A key decision in the set-up of a QM/MM or other simulation is the choice of QM system (i.e., which groups and residues to treat quantum mechanically). When working with an enzyme reaction, it is important to decide which residues, solvent molecules, and parts of substrates and cofactors are chemically involved. If resources allow, residues hydrogen bonding to the chemically relevant atoms or having electrostatic interactions could also be included. Site-directed mutagenesis experiments and structural data can help to suggest suitable candidates. Although its is tempting to increase the QM region in the quest for a better model, the scaling of computer resources with number of QM atoms means that systems with large QM regions (>60 atoms) can be very computationally demanding. Also, enlarging the QM region does not always result in a more accurate model, especially when using low levels of QM theory.

Normally, the QM and MM regions are covalently bound, and the position of the partition can have a significant bearing on the results of QM/MM calculations. These results can strongly depend on the position of the boundary between covalently bonded QM and MM atoms; for example, the placement of the link atom. Other important considerations include the choice of which MM atoms should be excluded from the classical electrostatic field that interacts with the QM region. The QM/MM boundary should preferably be placed some distance from the site of chemical and electronic change. Also, the boundary should not disrupt conjugated systems or be placed to near a highly charged QM group. A single carbon-carbon bond is a good choice for QM/MM partitioning; for example, between the $C\alpha$ and $C\beta$ of an amino acid residue that allows the sidechain to treated QM (Fig. 11).

5. OPTIMIZATION OF TRANSITION STRUCTURES AND REACTION PATHWAYS

A useful QM/MM method should be able to calculate the energy of a system undergoing chemical change in an enzyme to a reasonable level of accuracy, but this represents only part of the requirement for modeling such a reaction. Techniques for optimizing the structures of key species in the reaction are needed, and beyond this, methods for conformational sampling, simulating molecular dynamics and calculating activation energies.

In the simplest terms, many chemical reactions can be thought of as a progression of the system across a potential energy surface, along the lowest-energy path connecting the reactants and products, passing through the transition state as the highest point. Following

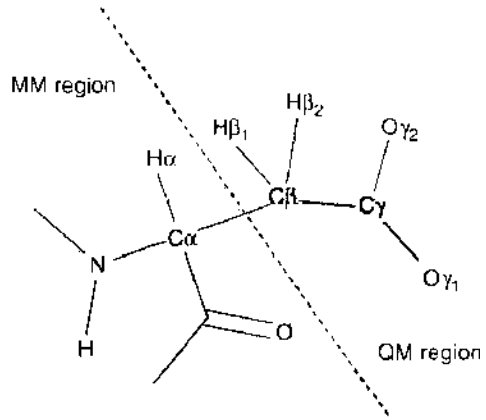


Figure 11. An example of how an aspartate amino acid sidechain can be partitioned into quantum mechanical (QM) and molecular mechanical (MM) regions in a QM/MM calculation on an enzyme. The functional group of the side chain, which may be involved in the enzyme-catalyzed chemical reaction, is treated quantum mechanically and the backbone atoms molecular mechanically. Interactions between the QM and MM regions must be included consistently.

that lowest-energy reaction path is a traditional and revealing way of calculating a reaction potential energy profile [43]. Finding the structures of the (representative) transition state, reactants and products will allow an activation energy to be calculated.

5.1. Small Model or Supermolecule Calculations

In practice, enzyme simulations (such as energy minimizations and dynamics simulations) are generally too computationally expensive to be performed with the whole enzyme treated at a quantum chemical level. However, small model systems that represent vital features of an enzyme active site can be used to examine the nature of interactions between groups and can provide valuable models of transition states and intermediates. Small models are also useful for testing the accuracy of different levels of calculations for a given application; for example, by comparing the results of semiempirical calculations with *ab initio* molecular orbital calculations, or comparing different levels of *ab initio* treatment for a particular reaction [164, 165]. Normally, the small model system would be based on an enzyme structure determined by X-ray crystallography. Residues that are considered important in the mechanism would be represented by appropriate small molecules in the same relative orientations as observed in the crystal structure. These small models should contain the same functional groups as the amino acid sidechain, cofactor, or substrate they represent; for example, typically imidazole or methylimidazole would be used to represent a histidine sidechain, and acetate for aspartate, and so on.

However, clearly such calculations do not take in to account the effects of the surrounding enzyme and solvent, which are likely to be significant in determining the reaction energetics in the real enzyme. This could lead to qualitatively incorrect descriptions of the enzyme mechanism by small model calculations. Also it may not be certain that all significant functional groups are included in the model system. From a technical point of view, it can also be difficult to find a representative optimized geometry of the model: Without the covalent and noncovalent restraints of the surrounding enzyme in place, it is possible that the molecules in the model system could drift and interact in ways that would be impossible in the enzyme. Applying constraints (e.g., to fix the backbone) to a model may prevent this unrealistic behavior, but in some cases it can lead to unsuccessful optimization, especially as there are likely to be differences between the initial starting structure and the reactive complex. For some metalloenzymes, where interesting chemistry takes place at the metal centers and coordination to metal ions holds ligand in place, small model calculations can be very useful (for examples of applications of supermolecule calculations, see Section 8.5 and the review by Himo and Siegbahn on quantum chemical studies of radical-containing enzymes [166]). Standard program packages such as Gaussian98 [167] and Jaguar [168] are available for performing small model calculations.

5.2. Local Methods

In local methods, a single conformation or configuration of a molecule is adjusted/optimized to find one stationary point at a time. These methods either use a second derivative matrix or rely on the assumption that the path near the saddle point has a predetermined shape. By design, these methods cannot find paths with multiple transition states.

5.2.1. Intrinsic Reaction Coordinate

The traditional way to follow a reaction path is to move downhill from a saddle point to the two associated (reactant and product) minima. The intrinsic reaction coordinate (IRC) is the minimum energy path when described using mass-weighted coordinates [169]. It is obtained by using a steepest descents method: The reaction path toward each minimum can be acquired directly at the transition structure from the eigenvector that corresponds to the imaginary frequency at the transition state. The steepest descent reaction path can be defined as

$$\frac{dx(s)}{ds} = -\frac{g(s)}{|g(s)|} \tag{15}$$

where x are the (mass-weighted) coordinates, s is the path length and g is the gradient. Further steps along IRC involve calculation of the gradient at each point calculated. The IRC method is a common and useful approach in gas phase reaction modeling. However, algorithms developed for small molecules are not suitable for such large systems with many degrees of freedom: direct calculation, storage, and manipulation of large Hessian matrices become extremely difficult with increasing system size. Many groups have been active in the development of geometry optimization algorithms applicable to large molecular systems such as enzymes [170–174]. One such development is incorporated in the software package GRACE, which is described in the next section.

5.2.2. GRACE

The GRACE program was designed specifically for transition state structure optimization in QM/MM calculations on enzyme reactions. In this approach, the system is divided into a core, which contains at least the QM atoms; and the environment [175, 176]. The environment is optimized to a minimum before each step with the core fixed. This fast-cycling environment subset surrounds the slow-cycling core. Partial rational function optimization is used to search for a transition structure in the core degrees of freedom. The fact that the Hessian matrix needs to be stored and maintained for the core only enables GRACE to be a much more practical method for use with large systems such as enzymes. GRACE has been used to study reactions by the optimization of transition state structures in many different enzymes including chorismate mutase [177], dihydrofolate reductase [178], and lactate dehydrogenase [175].

5.3. Global Methods

In local methods a single conformation or configuration of a molecule is adjusted to find one stationary point at a time. They rely on either using the second derivative matrix, which limits their application to low dimensional problems, or on the assumption that the path near the saddle point has a predetermined shape, which can only be valid if there is a simple potential energy surface topology. So when searching for transition states and reaction paths in large systems such as enzymes, local methods may be problematic. The main problem with local methods is that they will, by design, converge to any saddlepoint, and in macromolecules there are an extremely large number of energy minima and many transition states [31]. Ideally, a representative sample of transition states and reaction paths should be considered.

Global methods treat the entire path explicitly as a succession of points in conformational space. These points are typically simultaneously minimized under constraints that keep them from falling into nearby minima until they reach the adiabatic valley connecting the reactants to the products. Methods of this type have been applied to conformational changes in

a number of proteins. For example, a variety of techniques have been developed by Elber and coworkers to study reaction paths in proteins and other large systems [179–181]. One such approach is to minimize the average energy of a series of points along an approximate interpolated path between the initial and starting structures [179]. This path energy minimization method has been further developed by Smart for the location of transition structures [182, 183].

5.3.1. Adiabatic Mapping

Approximate reaction paths can be modeled by gradually changing a coordinate describing the reaction (the reaction coordinate) and allowing the system to relax at each stage by minimization while keeping the reaction coordinate fixed or restrained. The reaction coordinate can, for example, be a combination of internal coordinates relating to the reaction that is occurring; for example, bonds breaking and forming. The value of the reaction coordinate is gradually changed between the values representing two points on the pathway; for example, the products and reactants. The structures generated represent a sequence of points on the interconversion pathway. This procedure is described as adiabatic mapping. Adiabatic mapping will often overestimate the energy barrier if atom movements that occur during the reaction are not included in the reaction coordinate, because minimization alone will not allow the enzyme to adjust to relive the steric stain incurred. In other cases, the barrier may be underestimated, because the real transition state may not be located. One of the most challenging aspects of adiabatic mapping is the choice of reaction coordinate: It may be difficult to define a simple geometrical coordinate that accurately captures the structural changes associated with the reaction. Adiabatic mapping has been used in many studies of enzyme reactions, including QM/MM modeling of reactions catalyzed by chorismate mutase [184] and triosephosphate isomerase [103].

5.3.2. Conjugate Peak Refinement

The conjugate peak refinement (CPR) method [185] locates transition structures and reaction paths in proteins. The minimum input consists of the fully minimized reactant and product structures, between which the reaction pathway is calculated. The CPR method is optimized for finding true saddle-points and the adiabatic path that descends from them to form a continuous reaction coordinate between the reactant and product. Although CPR is a global method, it does not treat every point in the path the same way at all times. Rather, a refinement methodology is employed that recognizes and eliminates nonoptimal parts of the path. This method has the considerable advantage for large systems that only the first derivatives of the energy are required. It is available in the TRAVEL module of CHARMM [49] and has been used to determine reaction paths in a number of proteins; for example, rotamase catalysis by the FK506 protein (with a specially parameterized MM potential function for the peptide substrate) [39, 100].

6. MOLECULAR SIMULATIONS AND FREE-ENERGY CALCULATIONS

The reaction modeling techniques discussed in Section 5 are very useful when a single protein conformation suitably represents the enzyme throughout the reaction of interest. When more extensive conformational sampling is important, or when activation free energies (as opposed to potential energies) are desired, MD or Monte Carlo simulations, in combination with free-energy calculation methods are required.

Proteins are suitable candidates for dynamics simulations, as they consist of a system of interacting atoms that are compactly packed so that their motions are restricted. Internal motions form three main groups: local, which includes atom fluctuations, side chain oscillations, and loop displacements; rigid body, which includes helices, domains, and subunits; and finally large scale, which includes opening fluctuations, folding, and unfolding. Some of the most important internal motions in proteins consist of torsional oscillations of groups about the single bond that links them. These rigid groups can be atoms, small functional groups,

whole side chains, or the peptide bond of the polypeptide bond. The internal motions in proteins can be classified according to their amplitude (0.001–100 Å), energy (0.1–100 kcal), and timescale (10^{-15}–10^{3} sec) [31]. The timescale of the motion being studied is a major factor in the feasibility of the dynamic simulations and should determine the selected dynamic simulation method. For example, motions of whole domains and of the opening and closing of clefts occur on a nanosecond timescale; this means MD must be performed, as other dynamics methods are not able to sample for such a long period of time because of the associated computational expense. However, vibrations and oscillations of individual atoms and groups occur on the femto- to picosecond timescale and are open to study with methods such as Car–Parrinello dynamics.

6.1. Molecular Dynamics Simulations

Classical MD simulations are a well-established method to study the internal motions and conformational behavior of proteins [186]. MD methods solve Newton's equation of motion for atoms on an energy surface:

$$-\frac{dV}{dr} = m\frac{d^2r}{dt^2} \qquad F = ma \tag{16}$$

The available energy for the system is distributed between potential and kinetic energy, and an energy barrier can be overcome if the barrier height is less than the total energy minus the potential energy. However, processes with large energy barriers will occur only rarely, necessitating special techniques for their study. MD simulations using MM potential functions may be useful for generating a range of representative structures for subsequent QM/MM calculations on the reaction. MM dynamics simulations are useful because they can access significantly longer timescales; for example, in the nanosecond range [45–47, 187], than is feasible with QM/MM simulations allowing much better sampling.

Employing nonbonded cutoffs of normally around 10 Å (as mentioned in Section 2.1) can reduce the time required to carry out MD calculations by reducing the number of nonbonded interaction terms that need to be computed. However, applying cutoffs can introduce errors into a simulation, and there are improved ways to treat long-range electrostatics such as the Ewald sum technique [56] and fast-multipole methods [65, 66].

Given a high enough energy, normally implying a high temperature, MD will sample the whole energy surface; however, because of the length of simulation and appropriate temperatures, normally around 300 K, only the local area around the starting point is sampled and only small barriers can be crossed. Different local minima can be generated by selecting configurations along the simulation and minimizing these structures.

6.2. Monte Carlo Simulations

With the Monte Carlo method, new configurations are generated by adding a random "kick" (i.e., a "move" to change the structure) to a starting geometry. This new geometry is accepted as the starting point for the next step if it is lower in energy than the current one. Otherwise, the Boltzmann factor $e^{-\Delta E/k_BT}$ is calculated and compared to a random number between 0 and 1. If the Boltzmann factor is less than this number, the new geometry is accepted; otherwise, a new step is taken from the previous geometry.

Monte Carlo calculations are well suited to studying small molecules in solution. The random nature of the method allows the system to move ideally from one potential well to another without having to overcome the barrier. However, for simulations of proteins, Monte Carlo methods present some practical difficulties. In particular, this is a result of the large number of torsion angle moves that would be rejected on energetic grounds. Clearly, unlike MD time-dependent properties such as diffusion coefficients, and rotational and translational correlation functions, cannot be monitored. Monte Carlo simulations have been carried out to characterize the major interactions in the active site of chorismate mutase [188].

6.3. Dynamics with Stochastic Boundary Conditions

In QM/MM dynamics simulations of enzyme reactions, the system is often simplified, for example, by using the stochastic boundary method [189]. In this stochastic boundary molecular dynamics (SBMD) approach the solvated system is divided into a reaction region, a buffer region, and a reservoir region. Atoms in the reaction region are treated by standard MD algorithms, and their motions are governed by Newton's equations of motion.

$$m_i \frac{d^2 \mathbf{r}_i}{dt^2} = F_i + F_i^B \tag{17}$$

where F_i is the systematic force on atom i derived from explicit interactions with other atoms and F_i^B is the boundary force. The atoms in the buffer region are also influenced by mean boundary forces tending to hold them in position, and Langevin forces correct for the exchange of energy with those in the reservoir region.

$$m_i \frac{d^2 r_i}{dt^2} = F_i + F_i^B - \beta_i m_i \frac{dr_i}{dt} + f_i(t) \tag{18}$$

where β_i is the frictional coefficient (typical value of 250 ps^{-1} for nonhydrogen protein atoms) and f_i is the random force for atom i.

Those atoms in the reservoir region are not treated explicitly, as their presence (as fixed atoms) has been found to cause excessive rigidity of the protein [190]. A solvent deformable boundary potential prevents evaporation of water.

The saving in computation expense that stochastic boundary molecular dynamics affords enables QM/MM dynamic simulations to be carried out. This method can give more accurate information about the structure and interactions in the active site.

6.4. Umbrella Sampling

Theoretically, reactions can occur in the QM region of a QM/MM simulation system, but because of the limited simulation time and the height of many reaction barriers, reactions are unlikely to occur during the pico- to nanosecond timescale of a typical simulation. Multiple transition state configuration or barrier crossing events are required to obtain reliable results for an activation energy. In umbrella sampling [105, 191, 192], a potential is applied to direct the system into regions of a reaction coordinate that would otherwise not be adequately sampled. The probability distributions are corrected for the effect of the umbrella potential.

Free energies provide fundamental measures of the stability of the system. The statistics obtained from the various umbrella sampling simulations are combined into a single free-energy profile; for example, by using the weighted histogram analysis method (WHAM) [193].

6.5. Free-Energy Calculations Based on *Ab Initio* Calculations

Ideally, a simulation of an enzyme reaction should provide the free-energy profile (not only potential energies) for comparison with experiments (e.g., with activation free energies for kinetic studies). Free energies can be calculated by MD or Monte Carlo simulations as described above, and they ideally should be carried out on the entire QM/MM system to ensure full statistical mechanical sampling. However, as mentioned above, this is currently feasible for most enzyme reaction only for cheaper computational methods such as semiempirical QM/MM techniques and EVB. Full dynamics simulations with *ab initio* QM/MM methods are currently not feasible for typical enzyme reactions because of the computational demands of full *ab initio* calculations at every step. Several approaches have been proposed for calculating the free-energy profile for enzyme-catalyzed reactions involving *ab initio* QM or *ab initio* QM/MM calculations (as discussed below).

Warshel and coworkers have developed an EVB approach based on QM parameterization and energy corrections described as (QM(ai)/MM) [124]. In this method, EVB parameters are optimized to reproduce the results of gas phase *ab initio* calculations of model reactions along a given reaction path. MD simulations of the enzyme are then performed with

the resulting EVB potential energy function to sample conformations along the reaction path. The QM(ai)/MM potential energies are calculated for each sampled conformation. Finally, free-energy changes associated with the reaction path are calculated based on the differences between the EVB and the QM(ai)/MM potential energies.

The main advantage of the QM(ai)/MM approach is that it does allow for fluctuations of the QM part, in contrast to other approaches such as QM/FE, which do not allow for motion of the QM part. However, these fluctuations are only included through the EVB model, and hence the approach is critically dependent on the accuracy of the EVB potential energy function.

Another approach is combining gas phase quantum mechanical calculations with free-energy simulations (the QM/FE approach). When first conceived by Jorgensen and coworkers, the approach was used mainly for calculating free-energy changes associated with reactions in solution [194–196]. Recently, Kollman and coworkers have extended the method to study of enzyme-catalyzed reactions [197]. With the QM/FE method, the reaction paths in the gas phase and the charge distributions of different structures along the reaction paths are calculated using *ab initio* calculations. Then the structures along the gas phase reaction paths are solvated or mapped into enzyme environment, and the differences in free-energy for the interactions between the reacting parts and the solution or enzyme environment are determined using molecular mechanics and the free-energy perturbation method. As the QM/FE approach determines the structures and energies of the QM part in the gas phase, this approach is only valid when the solution or enzyme environment have little influence on the reaction. Recently, the QM/FE method has been used to study the proposed intramolecular proton transfer mechanism in triosephosphate isomerase [198], aliphatic hydrogen abstraction in citrate synthase [199], and the methyl transfer reaction catalyzed by the enzyme catechol O-methyltransferase [200].

Yang and coworkers more recently proposed another approach combining *ab initio* QM/MM calculations with the free-energy perturbation method [201]. An iterative procedure is used to determine optimized structures and minimum energy paths for the *ab initio* QM/MM system. With the determined minimum-energy paths, free-energy perturbation calculations are carried out to determine the change in free-energy along the reaction coordinate. This method is relies on the connection atom QM/MM model (Section 4.4) and has been used to study the reaction catalyzed by triosephosphate isomerase [201].

7. CAR–PARRINELLO MOLECULAR DYNAMICS

In 1985, Car and Parrinello proposed a new approach to *ab initio* molecular dynamics [202, 203]. The scheme combines MD simulation and DFT. Car–Parrinello MD is the integration of fictitious wavefunction coefficient dynamics with classical MD by a single extended Lagrangian. The electronic wavefunctions are treated as dynamical variables. After a converged wavefunction is determined at the first point, the orbitals evolve simultaneously with the changes in nuclear position. The orbital parameters are included as variables with fictive masses in the dynamics, analogous to the nuclear positions and masses. Except for the initial point the nuclear forces are not correct, as the electronic wavefunction is not converged in the orbital parameter space. However, this error can be controlled by the choice of dynamic parameters [43]. Simulated annealing was proposed as a mechanism for preventing the system becoming trapped in a local minimum. The SHAKE algorithm is used to impose constraints on the system to ensure that the orbitals remain orthonormal.

$$\ell = \frac{1}{2}\mathrm{Tr}\left[\mathbf{V}^{\mathrm{T}} M \mathbf{V}\right] + \mu \sum \int \left|\delta\phi_i/\delta t\right|^2 dr - E(R, \phi_i) - \sum \Lambda_{ij}\left(\int \phi_i^* \phi_j \, dr - \delta_{ij}\right) \qquad (19)$$

Because of the relatively large computational expense associated with conducting these simulations that of all potential *ab initio* electronic structure approaches, only DFT methods are currently in use with Car–Parrinello dynamics. The great advantage over force-field-type functions is that electronic structure methods are able to describe bond breaking and forming reactions, and so Car–Parrinello methods allow the direct simulation of chemical reactions.

If run in parallel, Car–Parrinello simulations can be carried out for up to a few thousand atoms. This, however, is too limited for studying most enzyme reactions, because dynamics simulations of reasonable length are currently too computationally demanding (e.g., in a recent study of HIV −1 reverse transcriptase, dynamics were performed for 1.5 ps [204]). Parrinello and coworkers [138, 205] have developed a scheme for Car–Parrinello MD simulations with a QM/MM method, with the CPMD and EGO programs (EGO is based on the CHARMM force field for proteins). Using these interfaced programs, efficient and consistent QM/MM Car–Parrinello simulations of very large systems can be performed in which the steric and electrostatic effects of the environment are accounted for explicitly. Most applications of the Car–Parrinello dynamics have been based on density functional methods; for example, to study carbonic anhydrase [138] and galactose oxidase [206, 207].

8. RECENT APPLICATIONS

In this section, some interesting and recent applications of enzyme-catalyzed reactions are discussed, highlighting the important conclusions and providing some brief description of how the calculations were performed, and details of the methods to illustrate the approaches used are given. These applications represent only a selected small number of the many enzyme-catalyzed reaction studies that have been performed, and the interested reader is urged to refer to the publications cited here and above for further details.

8.1. Phenol Hydroxylase and *p*-Hydroxybenzoate Hydroxylase

Phenol hydroxylase (PH) catalyses the hydroxylation of phenol, specifically at the *ortho* position of the aromatic phenol ring of the substrate. *para*-hydroxybenzoate hydroxylase (PHBH) catalyzes the 3-hydroxylation of *para*-hydroxybenzoate. Both enzymes are flavin-dependent monooxygenases that are involved in the microbial degradation of a wide range of aromatic compounds; for example, lignin, a major component of wood [208]. These reactions are potentially useful in the biodegradation of aromatic pollutants. The enzymes PH and PHBH are believed to share a similar reaction mechanism, involving initial substrate binding, two-electron reduction of the FAD cofactor by NADPH, and a subsequent reaction with molecular oxygen, giving a C4a-peroxyflavin intermediate form of the cofactor [209].

Protonation gives C4a-hydroperoxyflavin, which is believed to be the reactive species that hydroxylates the substrate. A number of mechanisms have been proposed for this hydroxylation step [209–212]. It has not been possible to determine the nature of this step experimentally, despite extensive experimental studies [213]. The final product is thought to be formed by keto-enol tautomerization in solution.

High-resolution structures of the substrates bound to the enzyme (with the cofactor in the reduced form) are available [214, 215] and provide good models for calculations. Recent QM/MM studies on PH [216] and PHBH [41, 160, 208, 217] aimed to determine the nature of the mechanism of the hydroxylation step. These investigations were initially at the semiempirical QM/MM (AM1/CHARMM) level [41, 160, 217], but in the case of PHBH, they have recently been extended to the *ab initio* QM/MM level [208]. The systems were treated initially with the semiempirical QM/MM method in the CHARMM program [49], with the QM region (shown in Fig. 12) modeled at the AM1 level of theory. Through comparisons with higher-level calculations on gas-phase models, AM1 was shown to be a reasonable method for use with these systems [41]. The reaction was modeled in each case as an electrophilic aromatic substitution, a mechanistic proposal that had received some experimental support.

As noted above in Section 2.6, semiempirical methods such as AM1 suffer from some well-known limitations. These QM/MM studies therefore further tested and validated the AM1/CHARMM method by comparison of calculated energy barriers with experimental activation energies derived from overall rates of enzymatic conversion of a series of substrate analogs. In the case of PHBH, fluorinated *para*-hydroxybenzoates were known experimentally to be converted by the enzyme at lower rates than the native substrate. For this series of substrates, full QM/MM (AM1/CHARMM) reaction pathway calculations were calculated by an adiabatic mapping approach. These pathways involved energy minimization of the whole

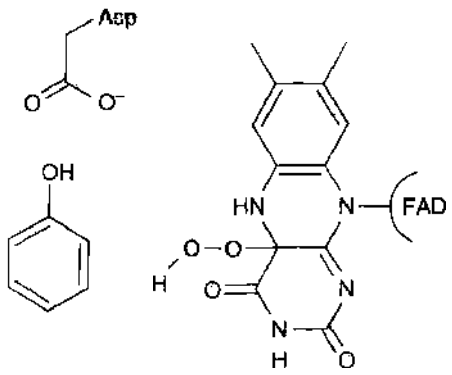


Figure 12. The quantum mechanical (QM) region used to study the hydroxylation step in phenol hydroxylase. The system was studied at the semiempirical quantum mechanical/molecular mechanical AM1/CHARMM22 [216]. Substrate deprotonation by Asp54 was found to be necessary for subsequent hydroxylation of the substrate.

QM/MM system (with boundary restraints) along an approximate reaction coordinate, which was defined as the difference of two bond lengths: $O_{C4a-hydroperoxyflavin} \cdots OH_{C4a-hydroperoxyflavin}$ (the breaking bond) and $OH_{C4a-hydroperoxyflavin} \cdots C3_{substrate}$ (the forming bond). The simulation system included 50 QM and 4840 MM atoms, including 330 crystallographic waters. The resulting energy barriers, which were found to be higher than the barrier for the nonsubstituted *para*-hydroxybenzoate, were shown to correlate linearly (correlation coefficient $R = 0.96$) with the logarithm of the experimental rate constants. This correlation strongly supports the hypothesis that the hydroxylation step proceeds via electrophilic aromatic substitution (the mechanism modeled) and also indicates that this step is rate limiting. A further mechanistic insight was that the calculations also indicated that deprotonation of the *para*-hydroxybenzoate substrates is necessary for efficient reaction in PHBH [41].

QM/MM studies of PH investigated both the hydroxylation step and the possibility of substrate deprotonation by an active site aspartate base. Because it was known experimentally that phenol binds in the neutral form, the precise mechanism was more uncertain than in the case of PHBH. The QM/MM potential energy surface for both reaction steps in the enzyme was calculated. The deprotonation step was found to precede hydroxylation and appeared to be necessary for subsequent hydroxylation. As with the QM/MM modeling of PHBH described above, the barriers to hydroxylation were calculated for a series of halogenated substrate analogues [216]. PH accepts a wide variety of substituted phenols as substrates. As a result of this large number of substrates, only the deprotonated reactant and hydroxylation transition states were optimized for the analog. As was found for PHBH, the energy barriers obtained correlated linearly ($R = 0.98$ for symmetric phenol derivatives and $R = 0.91$ for all substrates) with the logarithm of the experimental rate constants [218], as shown in Fig. 13. These correlations again lend support to the modeled electrophilic aromatic substitution mechanism and to the hydroxylation step being the rate-limiting step of the reaction. They also further validate the QM/MM method for this application and show that QM/MM results can be predictive. In absolute terms, the QM/MM energy barriers for PH and PHBH were too high because of the use of the AM1 method.

The contributions of individual amino acids to the calculated barrier were analyzed for both enzymes. This analysis divides the influence of the MM region on the QM region into residues by recalculating the energies of the reactant and transition states with fragments of the MM region present [162, 219, 220]. This is a simple and cheap method that can give useful information about the potential catalytic contributions of individual residues. However, the information is only qualitative because the effects of dielectric screening and entropic contributions are neglected. This analysis showed that many interactions contribute to the reaction energetics in determining the relative energies of the TS and product. In particular, one interaction was found in both the PH and PHBH systems, which specifically stabilizes the transition state for hydroxylation. In the transition state, the transferring hydroxyl group forms a hydrogen bond with a backbone carbonyl moiety of a nearby proline residue.

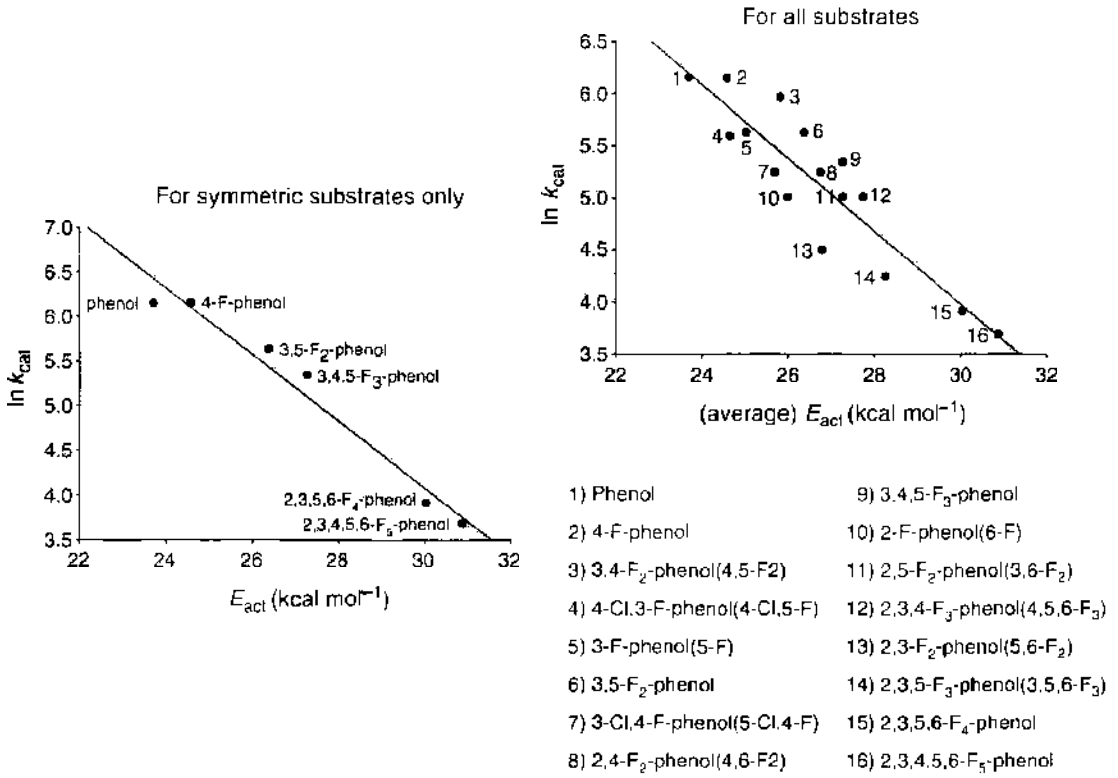


Figure 13. Phenol hydroxylase: Linear correlations between the natural logarithm of the experimental rate constants (min^{-1}) for overall conversion of symmetric phenol derivatives [218] and the (average) calculated energy barriers (E_{act}) obtained for the symmetric substrates only; and for all substrates. In the case of the asymmetric substrates, alternative substituent positions are shown in brackets.

The interaction is not present in either the products or reactants, but only in the transition state, indicating its catalytic effect. It had not been identified as catalytically important from the crystal structure. This provides a clear example of the identification of an important, novel catalytic motif by modeling alone.

Recently, *ab initio* modeling of PHBH reaction has been carried out to test conclusions from lower-level modeling [208]. QM/MM adiabatic mapping of the reaction was carried out using CHARMM27 [49] interfaced to GAMESS-US [221]. The same reaction coordinate as defined above was employed, the difference of two bond lengths: $O_{C4a\text{-hydroperoxyflavin}}$—$OH_{C4a\text{-hydroperoxyflavin}}$ (the breaking bond) and $OH_{C4a\text{-hydroperoxyflavin}}$—$C3_{substrate}$ (the forming bond). The simulation system included 49 QM and 4841 MM atoms, including 330 crystallographic waters. The QM region was treated at the *ab initio* HF level with the 3-21G(d) or 6-31G(d) basis sets. These calculations were performed using a Cray T3E-1200 supercomputer, and to complete the pathway with the larger basis set took 870 hours on 32 processors of this parallel machine (39,680 hours total unit processor time). This demonstrates how time-intensive *ab initio* QM/MM calculations can be, despite the fact that they may in some cases give little or no improvement over semiempirical methods for some reaction barriers (e.g., at the Hartree–Fock level of *ab initio* theory). The effects of electron correlation were included in this study on PHBH with single-point energy corrections that were applied at the B3LYP/6-311+G(d,p) and LMP2/6-31+G(d) levels and that were found to be required to give rate constants in agreement with experimental data. This work has supported the earlier findings described above and confirmed the significance of the hydrogen bond with the proline backbone [208].

PHBH has also been studied by QM/MM MD [222]. Three different substrates were modeled: *p*-oxybenzoate dianion, *p*-hydroxybenzoate anion, and *p*-oxybenzoic acid anion (shown in Fig. 14). The system included the enzyme, flavin-C4a-hydroperoxide, the substrate, and crystal waters for a total of ~7000 atoms. The QM region comprised ~100 atoms of the active site and was modeled with AM1. The MM region was represented by

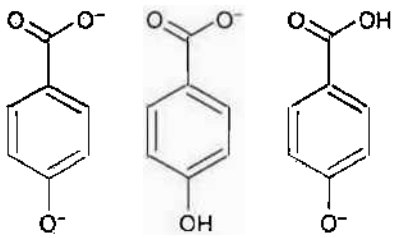


Figure 14. In a recent study [222], three possible forms of the substrate (reactants) were modeled in the enzyme *p*-hydroxybenzoate hydroxylase. Shown left to right: *p*-oxybenzoate dianion, *p*-hydroxybenzoate anion and *p*-oxybenzoic acid anion. Quantum mechanics/molecular mechanics molecular dynamics simulations of the reaction confirmed that the dianion mechanism is favored.

the GROMOS87 force field [223]. Initially, classical MD was performed to equilibrate the system, and then QM/MM MD was carried out at 300 K. Thermodynamic integration techniques [224] were used to determine the free-energy profiles for the reactions studied. In each of the three simulations, the reaction coordinate was defined as the distance between $OH_{C4a-hydroperoxyflavin}$—$C3_{substrate}$. The QM/MM MD confirmed that the dianion mechanism is favored over both monoanion mechanisms, with a free-energy barrier of 12 kcal mol^{-1}.

8.2. Glutathione S-Transferase

Glutathione S-transferases (GSTs) are responsible for metabolizing a large range of xenobiotic compounds, including steroids, pharmaceutical drugs, and environmental pollutants [225]. GSTs catalyze the addition of the cysteine thiolate of the tripeptide glutathione to electrophilic substrates. In a recent study, the conjugation of glutathione to phenanthrene 9,10-oxide, catalyzed by the GST M1-1 isoenzyme (a chemically distinct form of an enzyme that perform the same biochemical function) from rat, was modeled with a QM/MM (AM1/CHARMM) umbrella-sampling MD approach [105]. Phenanthrene 9,10-oxide is a standard model substrate for studying epoxide ring-opening reactions catalyzed by GST; for example, in studies of the stereospecificity of epoxide ring opening.

As with many enzymes, the active site of GST is highly solvent accessible. Furthermore, the solvent has been shown to have a significant effect on the reaction, as shown by large solvent kinetic isotope effects found experimentally. It was therefore particularly important to include solvent in the case of GST and to allow for its mobility. In the QM/MM model, the solvent was included explicitly using the CHARMM version of the TIP3P water model, using a stochastic boundary MD approach for simulation of a (roughly) spherical region around the active site [189, 190]. The simulation system included all whole residues with at least one atom within an 18 Å radius around the active site. The QM region contained the Cys sidechain of glutathione, including one link atom (equivalent to methyl thiolate) and the entire phenanthrene oxide molecule. The sulfur of the thiolate nucleophile was expected to pose a particular challenge for the semiempirical AM1 method. It is known that MNDO type methods such as AM1 perform poorly for sulfur and other second-row elements. Sulfur is a versatile, and therefore difficult, element for parameterization within semiempirical MO theory [226]. Therefore, for these simulations, a genetic algorithm was used to develop a specifically semiempirical model based on AM1 calibrated for treatment of the QM region. This approach, which has been shown to be useful in QM/MM studies of enzyme reactions, is the use of "specific reaction parameters" for NDDO-type methods such as AM1 designated, for example, as AM1-SRP or AM1-SSP. The parameters of the semiempirical method are optimized to fit experimental or *ab initio* data (as described above in Section 2.6.4) for a given reaction [102]. This approach can provide a fast treatment capable of calculating barriers and reaction energies accurately [227]. Crucially, semiempirical methods are fast enough for dynamics simulations to be computationally feasible. Therefore, reliable free-energy profiles can be calculated by such techniques.

In the specific case of GST, it was shown that reparameterization only of the sulfur parameters could significantly improve the accuracy of the calculations [105]. Other elements

present (C, O, and H) were modeled with standard AM1 parameters. This partial reparameterization approach is completely valid because the standard AM1 parameters for C, O, and H [90] were developed separately and prior to those for sulfur [226]. This type of methodology seems to work well for large QM systems that contain elements in different chemical configurations, for which it may be difficult to improve on the general AM1 parameters.

Free-energy profiles for the reactions shown in Fig. 15 were obtained by using the umbrella sampling molecular dynamics method [31] along a reaction coordinate, which was defined as the difference of two bond lengths: $C_{epoxide}$—$O_{epoxide}$ (the breaking bond) and $S_{thiolate}$—$C_{epoxide}$ (the forming bond). The M1-1 isoenzyme of GST was studied. Each simulation along the reaction coordinate consisted of 10 ps of equilibration followed by 20 ps of sampling dynamics. The results from each simulation were combined using the weighted histogram analysis method [193, 228]. The calculated free-energy barriers (approximately 20 kcal mol^{-1}) for formation of either diastereomeric product agree well with the activation energy derived from the experimental rate constant for the overall reaction (18 kcal mol^{-1}; Fig. 16). This agreement supports the QM/MM model of this reaction as being reasonable and confirms that the epoxide ring opening is the rate-limiting step in the enzyme-catalyzed reaction. The barriers to reaction for each possible diastereomeric product are very similar, which agrees with the experimental finding that for the M1-1 isoenzyme the ratio of products is 50:50.

Analysis of the model provided detailed insight into the catalytic mechanism of GST. Of particular interest is the interaction in the transition state structures with Tyr115, which stabilizes the epoxide ring-opened intermediate. In addition, analysis of hydrogen bonding with solvent molecules along the reaction coordinate showed a significant change in solvation of both the thiolate moiety of glutathione and the epoxide oxygen of phenanthrene 9,10-oxide, indicating that the solvent significantly affects the energetics of the reaction in agreement with conclusions drawn from experimental solvent kinetic isotope effects [229, 230].

An important goal in relating simulations of enzyme-catalyzed reactions to experiment is to reproduce and interpret the results of site-directed mutagenesis experiments. In particular, modeling methods can help to explain and predict changes in catalytic efficiency or specificity, caused by site-directed mutagenesis. Similarly, differences between closely related isoenzymes can be analyzed. The work on GST provides an example of the use of modeling to analyze effects of genetic differences on reaction specificity. Four mutations were modeled with the aim of understanding the determinants of stereospecificity of epoxide ring opening. One mutation, Asn8Asp, represents a difference in amino acid sequence between the GST isoenzymes M1-1 and M2-2, which are generally very similar in primary structure [231]. Modeling this mutation results in an increased reaction barrier toward the (9R, 10R) product, which agrees with the fact that the M2-2 isoenzyme preferentially forms the (9S, 10S) diastereomer (no crystal structure was available for the M2-2 isoenzyme, and it was therefore not possible to study it directly). The QM/MM calculations identified a single residue

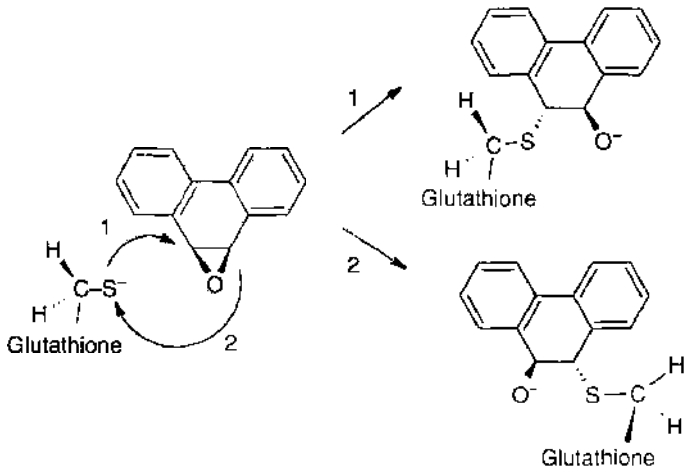


Figure 15. The conjugation of glutathione to phenanthrene 9,10-oxide catalyzed by glutathione S-transferase can lead to two diastereomeric products [105].

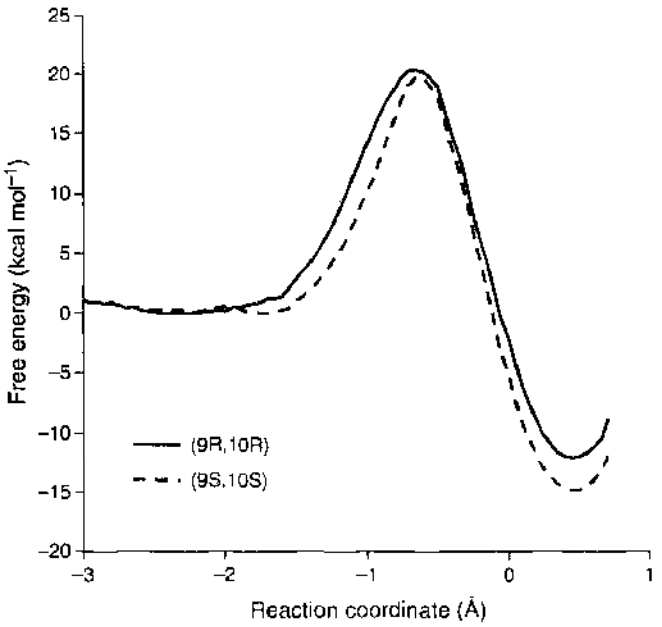


Figure 16. Quantum mechanical/molecular mechanical free-energy profiles for the reaction of glutathione with phenanthrene 9,10-oxide, catalyzed by M1-1 glutathione S-transferase, leading to the two potential diastereomeric products the 9S, 10S and 9R, 10R diasteromers. The free-energy barrier (calculated by AM1-specific reaction parameter/CHARM22 umbrella sampling molecular dynamics simulations [105]) is around 20 kcal mol^{-1}, which is in close agreement with experiment.

that is predicted to play a key role in determining the diastereoselectivity of the ring opening reaction. This illustrates the potential of QM/MM modeling in enabling understanding of the metabolic consequences of genetic variations and in the identification of stereospecificity determinants.

8.3. Chorismate Mutase

Chorismate mutase plays an important role in the biosynthesis of aromatic amino acids. It is located at a key branch point of the shikimate biosynthetic pathway [2]. Chorismate mutase catalyzes the rearrangement of chorismate to prephenate. This pericyclic reaction is formally a Claisen rearrangement (Fig. 17). The reaction within the enzyme is not believed to involve chemical catalysis and is known also to occur readily in solution via the same chemical mechanism. Chorismate mutase is a rare example of an enzyme that does not covalently bind to the substrate at any stage in the reaction. In QM/MM modeling the substrate alone can be chosen as the QM region with no need for linking atoms or other forms of QM/MM partitioning of covalently linked atoms. This technical advantage is also accompanied by a straightforward comparison between the enzyme reaction and its solution counterpart, allowing direct analysis of the causes of catalysis.

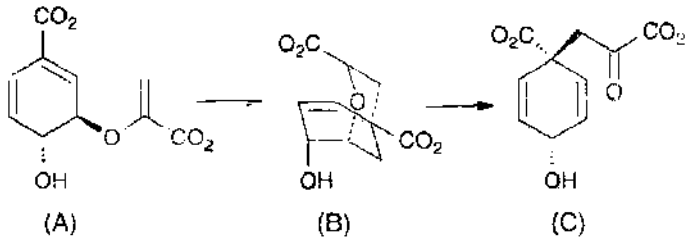


Figure 17. The Claisen rearrangement of chorismate to form prephenate, showing (a) diaxial chorismate, (b) the chairlike transition state, and (c) prephenate. This reaction, part of the shikimate biosynthetic pathway, is catalyzed by the enzyme chorismate mutase. Quantum mechanics molecular mechanics studies of the enzyme reaction show transition state stabilization to be central to catalysis [184, 243, 244].

The enzyme-catalyzed reaction has been studied widely by a variety of modeling techniques, particularly following the first QM/MM study by Lyne and coworkers in 1995 [232]. In one recent study [177], the CHARMM [49] and GRACE [175] programs were used to characterize stationary points for the reaction in the enzyme at the AM1/CHARMM level. In another QM/MM study, Gaussian94 and AMBER [46] were used to model the reaction at the B3LYP/6-31G(d) level, with the QM and MM regions optimized separately and asynchronously [233]. A MD/QM approach has also been used to model the reaction [234]. In this MD/QM study, "snapshots" of the active site were taken from MD simulations (with the AMBER MM forcefield) and optimized at the HF/4-31G/SBK (Stevens, Basch, Krauss [235]) QM level using effective fragment potentials for some important active site residues. Conformational studies in the gas phase, solution, and enzyme environment have been carried out at *ab initio* and semiempirical QM and QM/MM levels [236–238]. The reaction has also been studied in solution using Monte Carlo methods [188].

The enzyme is the focus of much current debate relating to the basic principles of enzyme catalysis. Early QM/MM studies indicated transition state stabilization by the enzyme and distortion of the substrate when bound to the enzyme [232]. Both were proposed to contribute to catalysis. In contrast, it has been argued by Bruice and coworkers [239] (based on MM MD simulations of the substrate) that transition state stabilization is not important and that catalysis instead arises from binding of a so-called 'near attack conformation' (NAC) of the substrate. NACs are substrate conformations in which bond-forming atoms are at van der Waals distance from each other [239].

These proposals have been tested in a number of recent investigations. In one recent QM/MM study [184], energy profiles were calculated using an adiabatic mapping approach similar to that described above for PHBH (Section 8.1). The system in this case was a large, fully solvated model (including all whole residues with at least one atom within a 25-Å radius around the active site). The structure used for modeling was based on a high-resolution crystallographic structure of *Bacillus subtilis* chorismate mutase with a transition state analog bound. The system was solvated and equilibrated by a MD procedure (see Section 4.2), with the protein and substrate fixed during MD. The reaction coordinate was defined as the difference of two bond lengths central to the Claisen rearrangement: C···O, the breaking bond and C···C, the forming bond. The reaction was mapped sequentially from the reactants to products, and the modeling was also run in the reverse direction to test the reversibility of the approximate reaction pathway. The effects of differences in the solvation procedure, and the details of the reaction path calculations, were tested. The potential energy barrier for the enzyme reaction was estimated at 24.5–31.6 kcal mol^{-1} (at the AM1/CHARMM level), and 2.7–11.9 kcal mol^{-1} with corrections [e.g., B3LYP/6-31+G(d)]. The results clearly show that the transition state is stabilized by the enzyme, which is in accordance with Pauling's hypothesis [1, 240] that stabilization of the reacting system by the enzyme is greatest at the transition state. Arg90, Arg7, and Glu78 were found to be the main contributors to transition state stabilization, along with one or two water molecules [184]. These findings are in agreement with experimental results, which show the importance of Arg90 in catalysis.

An additional possible catalytic contribution arises from strain (distortion) of chorismate by the enzyme. The possible role of substrate strain in lowering the reaction barrier in enzyme-catalyzed reactions was proposed many years ago by Haldane [241], and this line of thought has been drawn on in hypotheses for mechanisms such as that proposed for lysozyme [242]. However, this theory has had a varying degree of acceptance [1]. The structure of chorismate bound to the enzyme is found to be significantly altered from its minimum energy geometry in the gas phase [184]. A geometry farther along the reaction coordinate, closer to the transition state structure, is found for the substrate when bound to the enzyme. The substrate is therefore destabilized on binding, and this apparent strain/distortion could aid catalysis. This conclusion was suggested by work by Lyne and coworkers [232], whose QM/MM results at the AM1/CHARMM level for chorismate mutase provided the first indication of a role for strain in catalysis by this enzyme.

This reaction has also more recently been studied by *ab initio* QM/MM adiabatic mapping methods [243]. The system (based on a 25-Å radius selection around the active site) comprised 4211 protein atoms and 947 water molecules, which were treated MM, and

the substrate chorismate (23 atoms), which was treated QM (as the dianion). The reaction coordinate was defined as the difference of two bond lengths: C—O3, (the breaking bond) and C3—C9, (the forming bond). A series of minimizations were performed along this reaction coordinate at the RHF/6-31G(d)/CHARMM QM/MM level using CHARMM [49] interfaced with the GAMESS-US electronic structure package [146, 221]. To test the effects of electron correlation, the energies of the QM atoms taken from the RHF/6-31G(d)/CHARMM QM/MM pathway were calculated in the gas phase with higher-level methods B3LYP/6-311+G(2d,p) and MP2/6-31+G(d). These energies were used to correct the RHF/6-31G(d) QM/MM energy profile. The corrected estimates of the potential energy barrier were 7.4–11.0 kcal mol^{-1} [MP2/6-31+G(d)//6-31G(d)/CHARM22], and 12.7–16.1 kcal mol^{-1} [B3LYP/6-311+G(2d,p)//6-31G(d)/CHARMM22], which compare well with the experimental value of $\Delta H^{\ddagger} = 12.7 \pm 0.4$ kcal mol^{-1}. The results provided further clear evidence of transition state stabilization by the enzyme, in particular by Arg90, Arg7, Arg63, and Glu78.

Examination of the same pathway in solution at the same *ab initio* levels gives useful comparison between the catalyzed and uncatalyzed reactions. Again, estimates for the energy barrier for the reaction in solution were comparable to experimental value. The substrate was found to be distorted when bound to the enzyme; however, this substrate strain may only have a small contribution to catalysis. Chorismate mutase has also recently been studied using the DFT QM/MM approach [244]. The findings with this method are similar to those found with the *ab initio* QM/MM work discussed above. Although conformational effects may also contribute these QM/MM, studies in chorismate mutase show that transition-state stabilization is central to catalysis. These findings show how modeling methods can analyze fundamental principles of enzyme activity. The major driving force for catalysis in chorismate mutase is transition-state stabilization.

8.4. Methylamine Dehydrogenase

The majority of enzymes catalyze the transfer of hydrogen in some way; for example, proton movement, hydride transfer, and radical transfer. Experimental evidence shows that quantum tunneling is significant in many such enzyme-catalyzed reactions [245, 246]. It has been found that in some cases, kinetic isotope effects (KIEs) are very large, and in some cases show unusual temperature dependence [247]. Subsequent experimental work by a number of groups has shown the wide importance of hydrogen tunneling in enzymes [248, 249]. There has also been considerable debate on the contribution of tunneling to catalysis (i.e., whether tunneling may be enhanced compared to an equivalent reference reaction in solution). The possible role of protein dynamics in potentially "driving" tunneling has been widely discussed. Simulations allow these detailed and important questions to be investigated, through atomic-level analysis of enzyme reactions.

Methylamine dehydrogenase (MADH) catalyses the oxidative conversion of primary amines to aldehyde and ammonia (as shown in Fig. 18). This enzyme is found in several methylotrophic bacteria that use amines as their principal source of carbon and energy. Experimental studies have shown that there exist unusually large primary KIEs for the rate-limiting proton transfer step in MADH [250, 251]. This implies a large tunneling contribution in the reaction. It has also been shown experimentally that there is almost no dependence of the KIE on temperature for the methylamine substrate [247, 248].

Methods have been developed in the last couple of years for calculating the contributions to rates of reactions of quantum effects, such as quantization of vibrations and tunneling. One such method, which has been used in QM/MM modeling of enzyme reactions, is the variational transition state theory with small curvature approximation tunneling corrections; these methods have been developed and applied by Truhlar and coworkers [252], and they allow KIEs to be calculated (e.g., the change in rate for a proton transfer when hydrogen is replaced by deuterium or tritium). Comparison with experimentally measured KIEs allows a link to be made to experiments. However, it should be remembered that the complexity of multistep enzyme reactions may make direct comparison difficult.

Recently, calculations have been performed to determine KIEs for the proton transfer step in the methylamine-to-formaldehyde transformation in methylamine dehydrogenase [253].

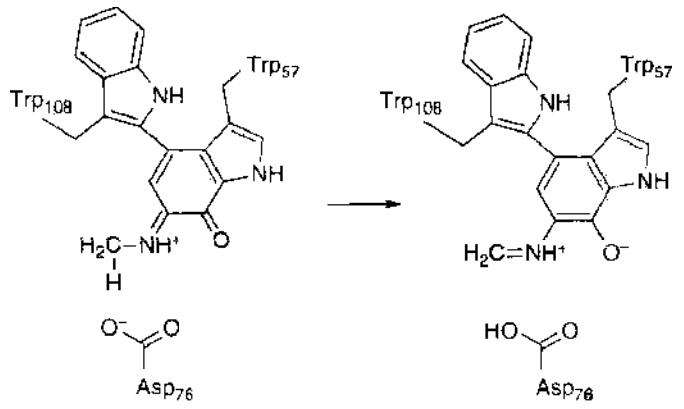


Figure 18. The proton transfer reaction from the methyliminoquinone intermediate to Asp76 in the enzyme methylamine dehydrogenase. A large tunneling contribution in the reaction is implied from unusually large primary kinetic isotope effects from experimental studies of this rate-limiting proton transfer step [250, 251].

The reaction was modeled using QM/MM methods and variational transition-state theory with multi-dimensional tunneling contributions for the dynamics [254]. The MM region was represented with the CHARMM22 force field [45] and consisted of 7248 protein atoms and 1243 water molecules. The QM region was treated using the semiempirical PM3 model with SSPs (specifically optimized for this reaction). The two regions were linked by means of the GHO QM/MM partitioning method [152]. Umbrella sampling was used to calculate a classical mechanical potential of mean force along an approximate reaction coordinate. The reaction coordinate was defined as the difference of two bond lengths: C···H, the breaking bond, and O—H, the forming bond. The difference between the quantal and classical vibrational free energies was then calculated for the reactants and the variationally optimized transition state [255]. This generates a quantal-corrected potential of mean force, and thus a new estimate of the rate constant [256]. Then a reaction path calculation was carried out to evaluate the ensemble-averaged transmission coefficient. The estimated classical mechanical free-energy of activation was 20.3 kcal mol^{-1}, but these workers found that if the quantum mechanical vibrational energies were included, the barrier height was reduced to 17.1 kcal mol^{-1}. When the quantum mechanical tunneling contributions were also included, the barrier was found to be even lower, at 14.6 kcal mol^{-1}, a finding that agrees well with the experimental value of 14.2 kcal mol^{-1}.

KIEs were also evaluated in this work by replacing all three protons on the methylimino carbon with deuterium. A hydrogen/deuterium KIE of 18.3 was calculated, which is good agreement with the experimental value of 17.2, but if tunneling is omitted then the calculated KIE is reduced to 5.9. The approach followed in this work allows different contributions (e.g., of tunneling) to be calculated separately. The results show the central role of proton tunneling in MADH.

In another study [104], the hydrogen transfer step in MADH for two substrates (methylamine and ethanolamine) was studied. Large KIE values have been measured experimentally for both of these substrates [247, 248]. However, the temperature independence of the KIE value for the methylamine substrate was not observed in experiments for the reaction with ethanolamine [247]. These reactions were studied with the QM/MM method, and the rates of proton and deutron transfer (and hence kinetic isotope effect values) were calculated using variational transition-state theory combined with multidimensional tunneling corrections [252]. The QM region was treated PM3 with specific reaction parameters with *Gaussian94*, whereas the MM region was described using the AMBER force field. The two regions were connected using the link atom approach in this work. These calculations yielded kinetic isotope effect values close to the experimental values. It was found that different hydrogen bonding patterns involving the substrate hydroxyl were possible for the ethanolamine substrate, an effect that is obviously absent for methylamine. These two different binding modes of the ethanolamine substrate generate quite different kinetic behaviors. One configuration exhibited large degrees of tunneling, and the other considerably less. The possibility

of two orientations of the ethanolamine substrate meant that the PESs for methylamine and ethanolamine are very different with the KIE values covering a wide range. This is believed to be the cause of the differing temperature behavior observed for the KIEs of the two substrates.

8.5. Cytochrome P450

Cytochrome P450s are an important class of heme iron enzymes that act as mono-oxygenases in steroid biosynthesis and a wide range of biological reactions, including the metabolism of potentially toxic hydrocarbons [257]. A large number of cytochrome P450 proteins are known, with different specificities and activities. P450s perform crucial bioregulatory functions such as detoxification and biosynthesis of sex hormones, as well as of compounds having anti-inflammatory and anti-hypertensive roles. The heme unit in the active site consists of an iron protoporphyrin IX complex with a cysteine ligand linking it to the protein. The catalytically active form of the enzyme is thought to be an oxyferryl (Fe=O) species called compound I. Reactions catalyzed by P450 enzymes include the hydroxylation of alkanes and aromatic compounds and the epoxidation of alkenes. Many of these catalytic reactions have the potential to be useful in synthetic and other practical applications. Important aspects of the mechanisms involved, and the determinants of P450 specificity, are not resolved. A detailed understanding of the mechanism of P450 enzymes is vital for predicting biotransformations of pharmaceuticals and other xenobiotics, and hence for drug development.

8.5.1. Alkane Hydroxylation

It is widely believed that the mechanism for alkane hydroxylation is the so-called rebound mechanism proposed by Groves [258–260] (Fig. 19). Shaik and coworkers have performed various studies using DFT based quantum chemical techniques to examine various aspects of the alkane hydroxylation mechanism [261, 262]. The PES for the rebound mechanism with methane as a substrate was calculated for two spin states, the high-spin (HS) quartet state and low-spin (LS) doublet state. Calculations were carried out on compound I using SH to represent the SCys ligand. Although methane is not known to be a P450 substrate, it was used to model the alkane substrate in these investigations. Calculations in the gas phase were performed at the B3LYP hybrid density functional theory level using the *Gaussian* (for HS and LS) and *Jaguar* (for the HS state only) programs. Geometry optimization was carried out with the Los Alamos effective core potential (ECP) coupled with the double-ξ LACVP (Los Alamos core valence pseudopotential) basis set for iron and a 6-31G basis for all other atoms (denoted as ECP+LACVP-6-31G). Geometries for all the species in the HS and LS rebound mechanisms were fully optimized without geometrical constraints on the system. For the first step of the reaction (see Fig. 19), the two spin states were found to show nearly identical reactivity (both gave a barrier to reaction of approximately 27 kcal mol^{-1}). However, for the second rebound step the HS state was found to have a barrier of approximately 5 kcal mol^{-1}, whereas there was no barrier to reaction for the LS state. When ethane was explored as a substrate, using very similar methods, the two spin states were found to be nearly degenerate, with a barrier to the second step of approximately 5 kcal mol^{-1} [263]. A recent study using camphor as a substrate [264] showed the barrier

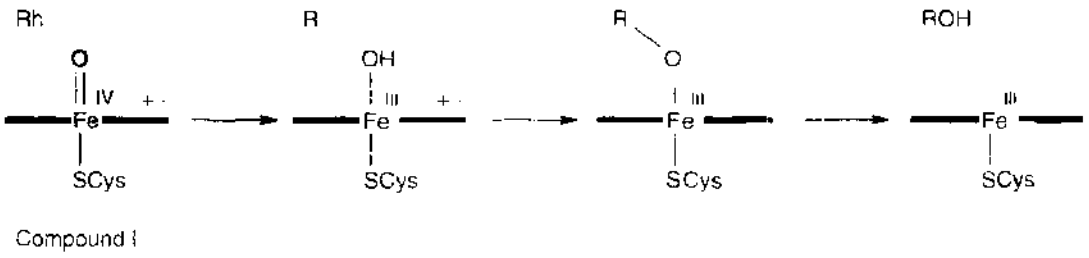


Figure 19. The rebound mechanism for alkane hydroxylation by cytochrome P450s enzymes. Initially, compound I abstracts a hydrogen from the alkane. This step is followed by recombination of the hydroxo-radical on the iron with the alkyl radical, which generates the ferric-alcohol complex.

to reaction with the HS state to be 3.3 kcal mol^{-1}, whereas that for a reaction of the LS state was much lower, only 0.7 kcal mol^{-1}. The behavior of the two spin surfaces has been proposed as an explanation of recent ultrafast radical clock experimental results [265].

8.5.2. Alkene Epoxidation

Shaik and coworkers have also studied the mechanism of ethane epoxidation by compound I [266, 267] (Fig. 20). Calculations were performed using the B3LYP hybrid density functional method with the LACVP/6-31G basis set with the *Jaguar* program. It was found that whether starting from the HS or LS state of compound I, the barrier for the C—O bond formation leading to the radical intermediate (see Fig. 20) was 14–15 kcal mol^{-1}. The second step, to form the epoxide, proceeds from the LS state with no barrier, and the HS state has a barrier of approximately 3 kcal mol^{-1}. Therefore, the reaction of the LS state is effectively concerted, whereas that of the HS state reaction is stepwise [266]. It has also been shown that the concerted reaction in the HS state has a barrier only around 4 kcal mol^{-1} higher than with the stepwise reaction [267]. The presence of intermediates with differing electronic configurations and significantly different lifetimes may go some way to explaining several experimental observations such as occasional *cis/trans* isomerization [268] and the production of aldehydes [268, 269].

8.5.3. Aromatic Hydroxylation

Although many studies have been conducted on the hydroxylation of alkanes and the epoxidation of alkenes in P450s, the mechanism of aromatic hydroxylation is much less well understood. The first step is generally agreed to be an addition of compound I to a substrate carbon, leading to a tetrahedral intermediate [270]. Subsequent rearrangement to the phenol product can proceed directly or via an epoxide. A recent study [271] applied hybrid DFT calculations at the B3LYP level with the LACV3P/6-311+G** basis set to determine energy profiles for benzene (and substituted benzene) reacting with compound I. Constrained geometry optimizations were performed at a set of fixed O—C distances between benzene and the oxygen atom of compound I. Further reactivity was considered by locating transition states for rearrangement of the tetrahedral adduct to a complex of the ferric P450 system with benzene epoxide or the ketone tautomer of phenol. The low barriers to product formation indicate that the C—O bond formation between compound I and the substrate is the rate-limiting and selectivity-determining step. Substituent effects were also investigated, and the computed activation barriers were correlated with different Hammet σ-scales. The best correlation was with a combination of radical [272] and cationic [273] σ-scales. It was shown that aromatic hydroxylation proceeds via an electrophilic pathway with a transition-state structure of mixed radical and cationic nature [271].

8.6. Dihydrofolate Reductase

Dihydrofolate reductase (DHFR) was the first enzyme to be targeted for cancer chemotherapy. It is a small enzyme that plays a supporting, but essential, role in the biosynthesis of DNA and in other cellular processes. One drug commonly used in cancer chemotherapy is methotrexate, which binds to dihydrofolate reductase far more tightly than its substrate folate, thus blocking the action of the enzyme. This drug blocks a key step in the cellular production of DNA, and it therefore tends to kill cells that are actively growing rather than

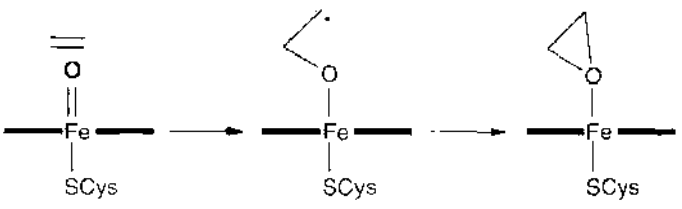


Figure 20. Mechanism of alkene epoxidation by compound I in cytochrome P450s. The concerted reaction in the high spin state has a barrier only around 4 kcal mol^{-1} higher than for the stepwise reaction [267].

cells that are not growing. Cancer cells are often the most rapidly reproducing cells in a patient, and so the drug should have the strongest effect on the cancer cells, although clearly other cells may be affected and therefore methotrexate.

DHFR catalyzes the reduction of folate to dihydrofolate and dihydrofolate to tetrahydrofolate (as shown in Fig. 21). These reductions take place via the transfer of a hydride ion from the nicotinamide adenine dinucleotide phosphate (NADPH) cofactor bound to the enzyme. The reaction also requires that a proton be transferred to the substrate. Although DHFR has been intensively studied, many details of its catalytic mechanism remain unclear. For example, it is not known how the substrate becomes protonated, and the ionization state of an important conserved acidic residue in the active form of the complex and its specific role in the catalytic mechanism have also been unclear.

Early computational work on this enzyme involved relatively low-level quantum chemical modeling of analogous reactions in the gas phase and did not include the effects of the enzyme [274, 275]. A primary aim of these early investigations was the determination of the transition-state structure. Subsequent investigations have included the enzyme using a semiempirical QM/MM approach. In one study [178], the transition-state structure was characterized using the GRACE software [175] at the AM1/CHARM22 level. In another [276], MD simulations and free-energy perturbation methods were used to investigate the energetics of the hydride–ion transfer step.

In recent work, several modeling methods have been used to study DHFR from *Escherichia coli*, with the aim of investigating the effect of a particular conserved acidic residue (Asp27); specifically, the effects of its ionization state on the hydride–ion transfer step [277]. As the ionization state of Asp27 was unclear from experimental data, it was modeled both protonated (neutral) and unprotonated (ionized). Cluster calculations were carried out to compare *ab initio*, DFT, and semiempirical (PM3) optimized structures and energetics for the hydride–ion transfer reaction. The DFT and *ab initio* reaction energies were endothermic for models representing the unprotonated Asp27 complexes but exothermic for the protonated form. Hence, the small model system indicated that reduction is only possible if Asp27 is protonated. PM3 wrongly predicted endothermic reactions for both protonation states.

Recent work [278] employed the GRACE software to study the importance of the Asp27 residue in DHFR. The system was studied with the QM/MM (AM1/CHARMM24) method. This work supported the hypothesis of an indirect proton transfer from Asp27 to the substrate via a series of water molecules. The simulations also explain the protonated state of the Asp27 and its role in stabilizing the cationic pyrazine ring.

The effects of the remaining substrate, cofactor, protein, and solvent interactions were studied using a QM/MM (PM3/AMBER) approach [277]. Dynamics simulations were

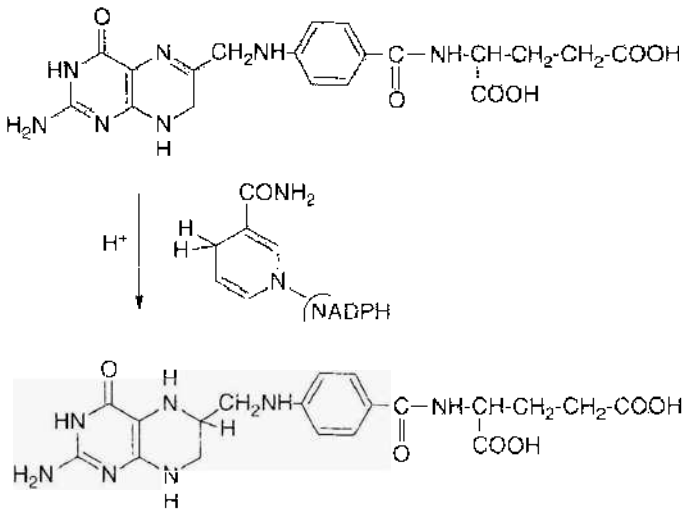


Figure 21. Dihydrofolate reductase catalyzes the reduction of 7,8-dihydrofolate to 5,6,7,8-tetrahydrofolate. The reduction mechanism requires addition of both a proton and a hydride ion from NADPH to dihydrofolate.

used to characterize the transition state complexes using a coordinate driving technique [279]. Using the results of additional cluster calculations at the DFT basis and MP2 basis level, the errors arising from the use of PM3 were estimated and corrected for. These calculations showed that the MM environment stabilized the products more than the transition state. The QM/MM calculations did not show the same dependence on the protonation state of Asp27, as indicated with the small model calculations. In fact, comparative estimates of reaction free energies indicate that the catalytic importance of the ionization state of Asp27 is small. These results reinforce the importance of modeling large, realistic systems to obtain correct estimates of reaction energetics for enzymes. Small gas phase models can in some cases give qualitatively incorrect results, as key effects of the enzyme environment may not be included.

DHFR has also been studied with a two-state EVB method [280]. This work studied a mutation of Gly 121 to valine; although this residue is found more than 12 Å from the transferring hydride, experimental results have shown that the Gly121Val mutation results in reduction in the rate of hydride transfer by a factor of 163. The calculated free-energy barrier was 3.4 kcal mol^{-1} higher for the mutant than for the wild-type DHFR; this corresponds well with the experimentally observed rate reduction. The simulations indicate that the increase in free-energy barrier arises from the interruption of a network of coupled promoting motions in DHFR [281]. The influence of a residue so removed from the active site has implications for protein engineering and drug design. This work is furthered by recent simulations using the QM/MM approach to model mutations to DHFR [282].

8.7. Citrate Synthase

Citrate synthase catalyzes the reaction between acetyl-coenzyme A (acetyl-CoA) and oxaloacetate that forms citrate and Coenzyme A. This reaction is the first step of the citric acid (or krebs) cycle a central metabolic cycle. This enzyme has been widely studied by experiment, but it is only through modeling that details of the mechanism have become clear. The reaction is believed to proceed via deprotonation of acetyl-CoA to form an intermediate, which subsequently attacks oxaloacetate. Citryl-CoA is then thought to be formed as another intermediate in the reaction and is then hydrolyzed to give the products (citrate and Coenzyme A).

One of the central mechanistic questions pertinent to citrate synthase has been the nature of the nucleophilic intermediate formed in the first step (i.e., by deprotonation of acetyl-CoA). The enol and enolate forms of acetyl CoA have been proposed, as well as an "enolic" structure [283], which is a strongly hydrogen bonded structure intermediate between the enolate and enol (Fig. 22). The enolic form was postulated as an example of a "low-barrier hydrogen bond" [283]. The relative energies of these three possible intermediates have been studied with various techniques including a semiempirical (AM1) treatment of a small active site model [40], as well as both semiempirical (AM1/CHARMM) [284] and *ab initio* QM/MM [MP2/6-31G(d)/CHARMM//HF/6-31(d)/CHARMM] [285] and QM-FE (MP2/6-31+G*//HF/6-31+G*: AMBER free-energy calculations) [199]. Although these studies applied different theoretical methods, they consistently predict the enolate of the acetyl CoA to be the true intermediate.

In a recent study [118], the reaction mechanism of citrate synthase was studied using the ONIOM method. A 10-Å-radius model of the active site and surrounding protein was modeled using MOZYME with the PM3 Hamiltonian [115, 116], and higher level *ab initio* methods were used to model the active site.

The ONIOM-MOZYME results support the earlier QM/MM findings that predict an enolate intermediate. The calculated activation energy for the enolate formation was 14.8 kcal mol^{-1}, which compares very well with the experimental value of 14.7 kcal mol^{-1}. The calculated energies for the enol and enolic intermediate are 10–13 kcal mol^{-1} higher than the enolate. The enolate of acetyl-CoA has been found to be significantly more reactive than the enol in *ab initio* calculations on small models of the nucleophilic attack in citrate synthase [286]. These calculations provide further evidence that an enolate is formed as the first intermediate in citrate synthase.

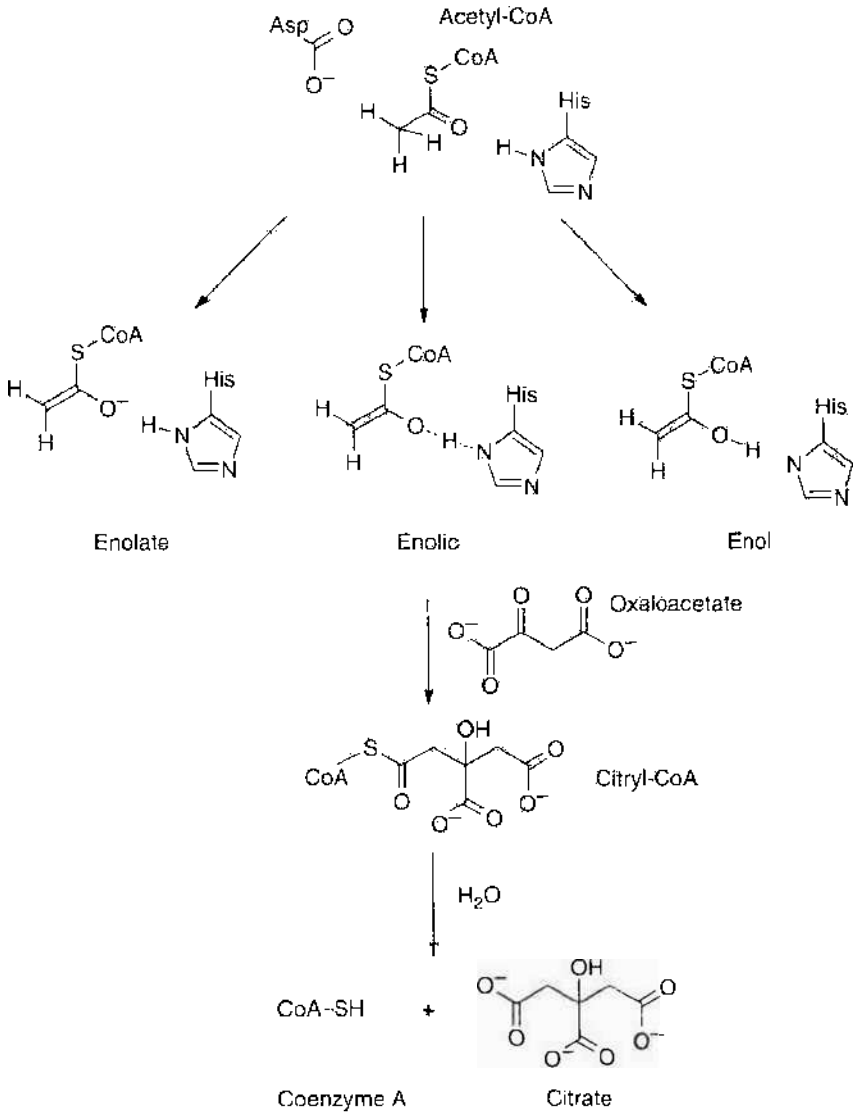


Figure 22. The reaction mechanism of citrate synthase proceeds via a nucleophilic intermediate, which has been variously proposed to be either the enolate, enol or "enolic" form of acetyl-Coenzyme A.

8.8. Human Immunodeficiency Virus Protease

HIV protease catalyses the hydrolysis of peptide bonds and is an important enzyme in the lifecycle of HIV. It is unique in that it can cleave the peptide link between a phenylalanine and tyrosine or proline, which is significant because no human enzyme can cleave between these amino acids. HIV proteases are members of the aspartic protease family, meaning there are two catalytic aspartic acid residues in the active site. Asp25 and Asp25' are the two catalytic Asp residues in HIV protease (Fig. 23). Their close interaction indicates that only one is protonated, but it is unclear from the X-ray crystallographic structures which Asp residue is protonated. Another important question is to determine the binding mode of the lytic water molecule in the enzyme–substrate complex.

Liu and coworkers studied the mechanism of HIV protease using QM/MM MD simulations [287]. These simulations were carried out with the biomolecular simulation program package GROMOS [47] interfaced with the MOPAC semiempirical molecular orbital package [95]. The PM3 semiempirical method was combined with the GROMOS87 MM force field [223] and the SPC water model. The QM region consisted of the two catalytic Asp residues, the scissile peptide and the lytic water molecule. QM/MM dynamics simulations were performed for a total of a few hundred picoseconds. Several possible reaction pathways

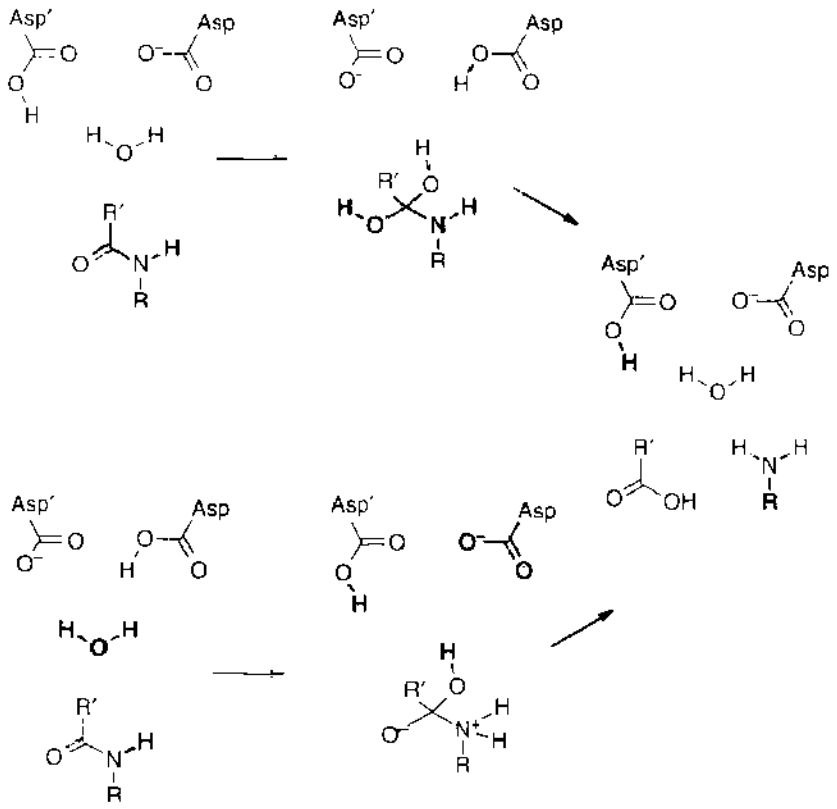


Figure 23. Two possible catalytic mechanisms for human immunodeficiency virus protease. The lytic water molecule is directly involved in the reaction. The upper scheme shows Asp52' to be protonated. Asp52 acts as a general base that activates the lytic water molecule attacking the carboyl carbon atom of the scissile peptide. Asp52' provides a proton to stabilize the negative charge built up on the carbonyl oxygen atom. In the lower scheme, Asp52 is protonated. The lytic water molecule is activated by Asp52' and the attack by the lytic water molecule results in an intermediate anion. Asp52 acts as an acid that provides a proton to the nitrogen atom of the scissile peptide bond to complete catalysis.

were investigated by umbrella sampling. The results show that the HIV protease works by the mechanism of general acid/general base catalysis. In the enzyme–substrate complex, Asp52' is protonated and the lytic water is bound at a different position than that indicated from bound inhibitor structures.

8.9. Triosephosphate Isomerase

Triosephosphate isomerase (TIM) catalyses the conversion of dihydroxyacetone phosphate (DHAP) to *R*-glyceraldehyde 3-phosphate (GAP), which is an important step in the glycolytic metabolic pathway. Many computational studies of this enzyme have been carried out with a variety of different methods [36, 220, 288, 289]. Modeling investigations have been central in advancing understanding, and generating debate, about the mechanism of this important model enzyme. However, some details of the mechanism are still not completely understood. One interesting feature is the central role of a neutral histidine (imidazole) sidechain, potentially as a general acid. A central mechanistic question is whether an enediol is formed as an intermediate and, if it is formed, whether the source of the proton is a neutral histidine (His95) or a glutamic acid (Glu165). There are three main proposals for the mechanism catalyzed by TIM, shown in Fig. 24. Pathway A is the more generally accepted mechanism, although there is no direct experimental evidence for enediol formation. It is supported by QM/MM calculations carried out at the AM1/CHARMM19 level [220]. Pathway A has also been studied using the empirical valence bond method with free-energy perturbation techniques [290]. Pathway B was proposed by Alagona and coworkers and is supported by small model calculations with *ab initio* methods (SCF/6-31+G* and MP2/6-31+4G*) [289]. In pathway C Glu165 is the only catalytic residue. Initially proposed

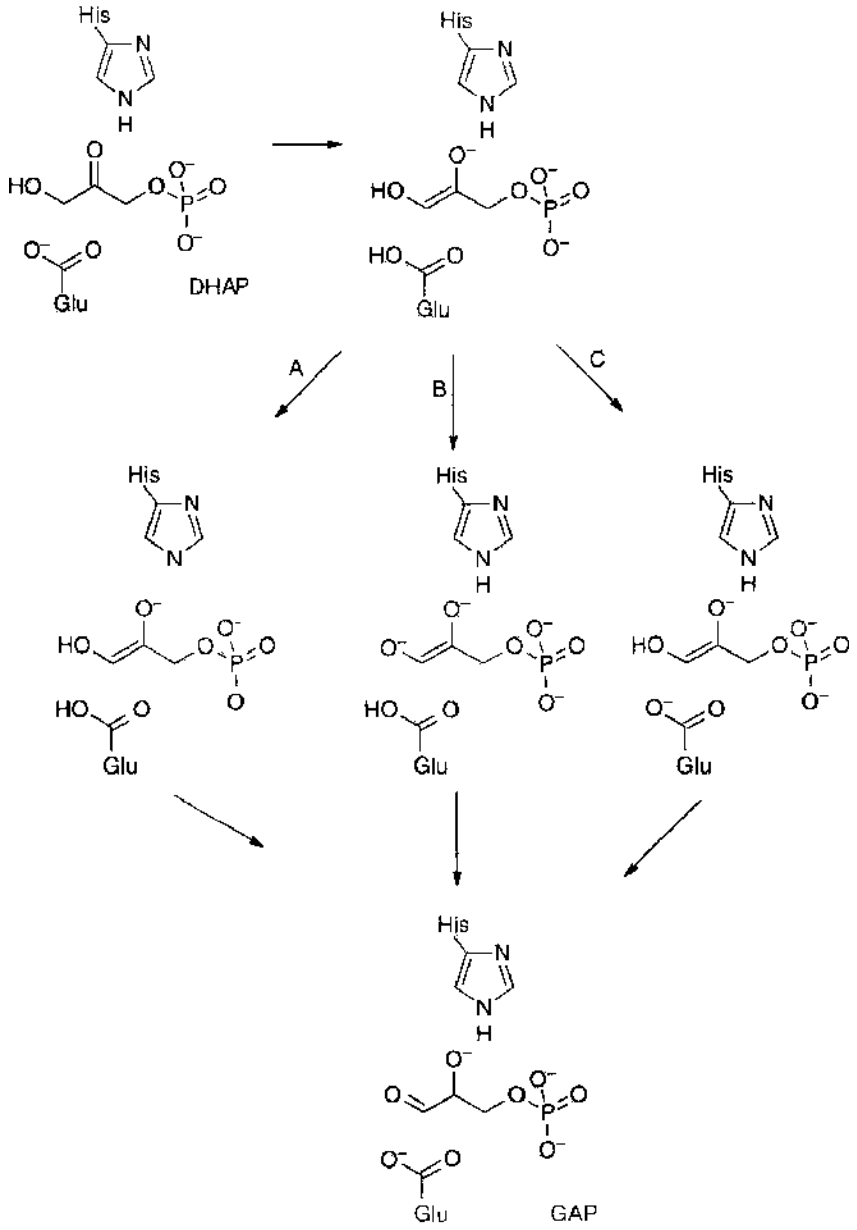


Figure 24. Three possible intermediates in the mechanism of isomerization of glyceraldehyde 3-phosphate and dihydroxyacetone phosphate catalyzed by triosephosphate isomerase. The source of the proton for enediol formation, if it occurs, is suggested to be His95 or Glu165.

only for a mutant enzyme [37], this mechanism was extended to cover the wild-type enzyme following NMR experiments on yeast TIM [291, 292].

In a recent study [103], the three proposed mechanisms were studied using the QM/MM approach at various levels of QM including AM1, AM1 with reaction specific parameters, and B3LYP/6-31+G(d,p). These simulations showed pathway B to be highly unlikely because of unfavorable interactions between His 95 and the transition state. However, from the calculations performed, it was impossible to discriminate between pathways A and C. The reaction has also been modeled recently using the SCC-DFTB/CHARMM QM/MM method [293].

9. CONCLUSIONS

One of the most important emerging areas of nanotechnology is bionanotechnology. The natural self-assembly, specificity, and potential replication properties of biological systems (combined with other desirable properties such as low toxicity and biodegradability) make

them highly attractive as components or inspirations for nanotechnological devices. Bionanotechnology is widely recognized as one of the key scientific growth areas of the twenty-first century. The interface with medicine is clearly another tantalizing frontier for developments in bionanotechnology. Techniques for creating arrays of proteins on solid substrates promise enzyme arrays that could be used for diagnostic or synthetic purposes. Highly specific and efficient microreactors for complex chemical and biochemical transformations can be envisaged. The potential applications of bionanotechnology in nanoscale electronic and electrochemical devices are clearly also vast, with the use of redox enzymes, for example, being a particular focus [17–19].

The catalysts of biological systems are enzymes. Enzymes possess many desirable catalytic properties for practical applications, and large enzymes also undergo conformational changes that may be coupled to mechanical motions (as in muscle proteins, for example). The techniques of molecular biology allow facile engineering of enzyme proteins, and experimental methods such as X-ray crystallography and NMR spectroscopy allow enzyme structures to be determined (see Section 1.4). Nanoscale manipulation and measurement techniques such as AFM allow the behavior of individual enzyme macromolecules to be analyzed in detail. Enzymes are currently used in many developments in nanotechnology: Both natural enzymes and engineered variants have been used, whereas in the future, wholly designed protein catalysts or biomimetic systems may be employed. For such developments in molecular engineering, detailed understanding of enzyme catalytic mechanisms and principles is needed. Further, techniques for predicting specificity and catalytic properties are required.

Molecular modeling and simulation methods have a vital role to play in developments in using enzymes in nanotechnology. Modeling can potentially give a level of detailed analysis of enzyme catalytic processes that cannot be approached by any current experimental technique for these complex biological macromolecules. Simulations should allow an atomic-level description of the reaction, including the dynamics and motions of the system.

Central to understanding and predicting catalytic behavior are the transition state of a reaction and other unstable species along the reaction pathway. Modeling allows transition states and reaction intermediates to be studied directly, something that is generally impossible experimentally for enzymes. Molecular modeling allows analysis of the interactions that stabilize these unstable species. The resulting detailed understanding should help to establish catalytic principles, which will help in the design of protein or biomimetic catalysts. Molecular simulations can also be used to study the conformational behavior and dynamics of enzymes. These are important considerations because proteins show complex dynamic behavior, and many proteins undergo large conformational changes during their catalytic cycles. Computer simulations can complement experimental investigations in many ways; for example, in analyzing and predicting the effects of amino acid mutations on activity and stability.

Enzymes are large macromolecules, and they thus present several major challenges for simulation techniques. A variety of modeling and simulation methods is available for studying different aspects of enzymes and their behavior. These methods have different areas of applicability and different strengths and weaknesses. For studies of protein dynamics and conformational changes, a number of simple empirical MM potential functions have been developed and proven to be very valuable (as discussed in Section 2.1). The simple form of these empirical force fields means that energy and force evaluations can be performed rapidly, and the calculations are computationally efficient, thus allowing simulations of small proteins in solution on the nanosecond timescale with current computational resources (see Section 6.1). Quantum chemical molecular electronic structure calculations, employing techniques such as *ab initio* molecular orbital methods [58] and DFT-based methods [76, 77], are extremely useful for studying small molecules and their reactions. However, as the size of the system increases, electronic structure calculations become extremely demanding computationally and the computer requirements increase significantly as a result of their scaling properties.

An important application for these methods in enzyme reaction modeling is that of supermolecule calculations, where small molecules represent important functional groups (see Section 5.1). Such calculations can be used to examine the nature of interactions between

groups at the active site and can provide useful models of enzyme transition states and intermediates. They can also be used to assess the accuracy of different levels of calculations for a particular reaction. Although electronic structure calculations are currently limited to systems containing in the order of tens of atoms, developments in methodologies such as the divide-and-conquer approach [107, 108] promise to expand this range considerably. Semiempirical molecular orbital methods (e.g., AM1 [89] and PM3 [90]) are considerably less computer intensive than *ab initio* methods. As a consequence, they can be applied to much larger systems (on the order of hundreds of atoms). However, as a solvated enzyme system is likely to consist of tens of thousands of atoms, even with better scaling methods, full QM electronic structure investigations of enzyme reaction pathways (or molecular simulations of more than a few picoseconds) are too computationally demanding to be currently feasible. Therefore, other approaches have to be used.

Ideally, a method for the simulation of enzyme reactions should capture the essential details of the chemical reaction (which is fundamentally QM in nature) while treating at least a large part of the enzyme and the surrounding solvent. Hybrid approaches, which combine quantum chemistry with simpler molecular mechanics, are a powerful and increasingly important approach to modeling enzyme-catalyzed reactions. These combined methods are described as QM/MM techniques (as discussed in Section 3). QM/MM calculations combine a quantum chemical description of the groups directly involved in the reaction (at the enzyme active site) to calculate the electronic structure of the reacting system, with a simpler molecular mechanics treatment of the enzyme and the environment [137]. QM/MM calculations can be carried out at *ab initio*, semiempirical, or density-functional QM levels. Many research groups have carried out QM/MM investigations for a wide selection of enzymes [134, 145, 146, 216]. These studies have demonstrated that QM/MM modeling can provide biochemically useful and relevant insight into the mechanisms of enzymic reactions. Some interesting recent applications have been reviewed in this chapter (see Section 8).

It is clear that applications of biological molecules and systems in nanotechnology will continue to grow quickly in importance. Further developments and exploitation of this area will rely on the use of enzymes in synthesis, modification, signaling, recognition, and replication. Design and developments in these areas will require detailed understanding of enzyme catalytic mechanisms and predictive techniques for protein engineering. Molecular modeling and simulation of enzymes and their reactants will play a vital and growing role in the exciting emerging field of bionanotechnology.

ACKNOWLEDGMENTS

A. L. B. thanks EPSRC and BBSRC for current funding. A. J. M. thanks EPSRC, BBSRC, The Wolfson Trust, The Royal Society, and the IBM Life Sciences Outreach Programme for support. We thank Dr. C. Murray for assistance with the figures.

REFERENCES

1. A. Fersht, "Structure and Mechanism in Protein Science. A Guide to Enzyme Catalysis and Protein Folding." Freeman, New York, 1999.
2. L. Stryer, "Biochemistry." 4th edn. W.H. Freeman, New York, 1995.
3. U. S. Department of Energy, "Pacific Northwest National Laboratory Institutional Plan FY 2001–2005." Pacific Northwest National Laboratory, Richland, WA, 2000.
4. M. E. Bruins, A. E. M. Janssen, and R. M. Boom, *Appl. Biochem. Biotech.* 90, 155 (2001).
5. F. B. Perler, S. Kumar, and H. Kong, *Adv. Protein Chem.* 48, 377 (1996).
6. U. T. Bornscheuer, J. Altenbuchner, and H. H. Meyer, *Bioorg. Med. Chem.* 7, 2169 (1999).
7. E. Henke and U. T. Bornscheuer, *Biol. Chem.* 380, 1029 (1999).
8. O. Lockridge, R. M. Blong, P. Masson, M.-T. Froment, C. B. Millard, and C. A. Broomfield, *Biochem.* 36, 786 (1997).
9. A. E. Nixon and S. M. Firestine, *IUBMB Life* 49, 181 (2000).
10. A. Skandalis, L. P. Eneell, and L. A. Loeb, *Chem. Biol.* 4, 889 (1997).
11. W. P. C. Stemmer, *Proc. Natl. Acad. Sci.* 91, 10747 (1994).
12. W. P. C. Stemmer, *Nature* 370, 389 (1994).
13. J. C. Moore, H. M. Jin, O. Kuchner, and F. H. Arnold, *J. Mol. Biol.* 272, 336 (1997).

14. S. Nishikawa, K. Homma, Y. Komori, M. Iwaki, T. Wazawa, A. H. Iwane, J. Saito, R. Ikebe, E. Katayama, T. Yanagida, and M. Ikebe, *Biochem. Bioph. Res. Co.* 290, 311 (2002).
15. S. Diez, C. Reuther, C. Dinu, R. Seidel, M. Mertig, W. Pompe, and J. Howard, *Nano Lett.* 3, 1251 (2003).
16. A. Ikai, *Superlattice Microst.* 31, 43 (2002).
17. M. L. Merlau, M. del Pilar Mejia, S. T. Nguyen, and J. T. Hupp, *Angew. Chem.* 113, 4369 (2001).
18. G. Gilardi and A. Fantuzzi, *Trends Biotechnol.* 19, 429 (2001).
19. G. Gilardi, Y. T. Meharenna, G. E. Tsotsou, S. J. Sadeghi, M. Fairhead, and S. Giannini, *Biosens. Bioelectron.* 17, 119 (2002).
20. C. Branden and J. Tooze, "Introduction to Protein Structure." Garland, New York, 1998.
21. S. Remington, G. Wiegand, and R. Huber, *J. Mol. Biol.* 158, 111 (1982).
22. F. Todone, R. O. J. Weinzierl, P. Brick, and S. Onesti, *Proc. Natl. Acad. Sci.* 97, 6306 (2000).
23. "Sybyl 6.9." Tripos, St. Louis, MO, 2002.
24. "MOE." Chemical Computing Group, Montreal, Canada.
25. E. Martz, *Trends Biochem. Sci.* 27, 107 (2002).
26. W. Humphrey, A. Dalke, and K. Schulten, *J. Mol. Graphics* 14, 33 (1996).
27. "QUANTA." Accelrys, San Diego.
28. "Insight II." Accelrys, San Diego.
29. M. L. Connolly, *J. Appl. Crystallogr.* 16, 548 (1983).
30. M. L. Connolly, *Science* 221, 709 (1983).
31. C. L. Brooks III, M. Karplus, and B. M. Pettitt, "Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics." Wiley, New York, 1990.
32. D. E. Koshland, *Proc. Natl. Acad. Sci.* 44, 98 (1958).
33. R. B. Von Dreele, *J. Appl. Crystallogr.* 32, 1084 (1999).
34. R. B. Von Dreele, P. W. Stephens, G. D. Smith, and R. H. Blessing, *Acta Crystallogr. D* 56, 1549 (2000).
35. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, *Nucleic Acids Res.* 28, 235 (2000).
36. M. Karplus, J. D. Evanseck, D. Joseph, P. A. Bash, and M. J. Field, *Faraday Discuss.* 93, 239 (1992).
37. E. B. Nickbarg, R. C. Davenport, G. A. Petsko, and J. R. Knowles, *Biochemistry* 27, 5948 (1988).
38. A. J. Mulholland, G. H. Grant, and W. G. Richards, *Protein Eng.* 6, 133 (1993).
39. A. J. Mulholland and M. Karplus, *Biochem. Soc. T* 24, 247 (1996).
40. A. J. Mulholland and W. G. Richards, *J. Phys. Chem. B* 102, 6635 (1998).
41. L. Ridder, A. J. Mulholland, I. M. C. M. Rietjens, and J. Vervoort, *J. Mol. Grap. Modell.* 17, 163 (1999).
42. F. Jensen, "Introduction to Computational Chemistry." Wiley, Chichester, 1999.
43. A. R. Leach, "Molecular Modelling: Principles and Applications." Prentice Hall, Harlow, 2001.
44. C. J. Cramer, "Essentials of Computational Chemistry: Theories and Models." Wiley, Chichester, 2002.
45. A. D. MacKerell, D. Bashford, R. L. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus, *J. Phys. Chem. B* 102, 3586 (1998).
46. W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, *J. Am. Chem. Soc.* 117, 5179 (1995).
47. W. R. P. Scott, P. H. Hunenberger, I. G. Tironi, A. E. Mark, S. R. Billeter, J. Fennen, A. E. Torda, T. Huber, P. Kruger, and W. F. van Gunsteren, *J. Phys. Chem. A* 103, 3596 (1999).
48. S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner, *J. Am. Chem. Soc.* 106, 765 (1982).
49. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, *J. Comp. Chem.* 4, 187 (1983).
50. X. Daura, A. E. Mark, and W. F. Van Gunsteren, *J. Comp. Chem.* 19, 535 (1998).
51. L. D. Schuler, X. Daura, and W. F. van Gunsteren, *J. Comp. Chem.* 22, 1205 (2001).
52. T. A. Halgren, *J. Comp. Chem.* 17, 490 (1996).
53. T. A. Halgren and R. B. Nachbar, *J. Comp. Chem.* 17, 587 (1996).
54. U. Burkert and N. L. Allinger, "Molecular Mechanics." American Chemical Society, Washington, DC, 1982.
55. J. W. Ponder, "TINKER 4.1." Washington University School of Medicine, St. Louis, MO (2003).
56. T. Darden, D. York, and L. Pedersen, *J. Chem. Phys.* 98, 11089 (1993).
57. M. Born and J. R. Oppenheimer, *Ann. Physik* 84, 457 (1927).
58. W. J. Hehre, L. Radom, P. V. Schleyer, and J. Pople, "*Ab Initio* Molecular Orbital Theory." Wiley, New York, 1986.
59. W. Z. Pauli, *Z. Physik* 31, 765 (1925).
60. J. C. Slater, *Phys. Rev.* 35, 509 (1930).
61. J. C. Slater, *Phys. Rev.* 34, 1293 (1929).
62. C. C. J. Roothaan, *Rev. Mod. Phys.* 23, 69 (1951).
63. G. G. Hall, *Proc. R. Soc. Lon. Ser-A* 205, 541 (1951).
64. J. Almlöf, K. Faegri, and K. Korsell, *J. Comp. Chem.* 3, 385 (1982).
65. M. C. Strain, G. E. Scuseria, and M. J. Frisch, *Science* 271, 51 (1996).
66. M. Challacombe and E. Schwegler, *J. Chem. Phys.* 106, 5526 (1997).
67. C. Ochsenfeld, C. A. White, and M. Head-Gordon, *J. Chem. Phys.* 109, 1663 (1998).
68. J. M. Millam and G. E. Scuseria, *J. Chem. Phys.* 106, 5569 (1997).

69. W. J. Hehre, R. F. Stewart, and J. A. Pople, *J. Chem. Phys.* 51, 2657 (1969).
70. W. J. Hehre, R. Ditchfield, and J. A. Pople, *J. Chem. Phys.* 56, 2257 (1972).
71. J. Chandrasekhar, J. G. Andrade, and P. von Rague Schleyer, *J. Am. Chem. Soc.* 103, 5609 (1981).
72. D. Feller, *J. Chem. Phys.* 96, 6104 (1992).
73. J. A. Pople and R. K. Nesbet, *J. Chem. Phys.* 22, 571 (1954).
74. G. H. Grant and W. G. Richards, "Computational Chemistry," Oxford Univeristy Press, Oxford, 1995.
75. C. Møller and M. S. Plesset, *Phys. Rev.* 46, 618 (1934).
76. R. G. Parr and W. Yang, "Density-Functional Theory of Atoms and Molecules," Oxford University Press, Oxford, 1989.
77. T. Ziegler, *Chem. Rev.* 91, 651 (1991).
78. P. Hohenberg and W. Kohn, *Phys. Rev.* 136, 864 (1964).
79. W. Kohn and L. J. Sham, *Phys. Rev.* 140, 1133 (1965).
80. S. H. Vosko, L. Wilk, and M. Nusair, *Can. J. Phys.* 58, 1200 (1980).
81. A. D. Becke, *Phys. Rev. A* 38, 3098 (1988).
82. C. T. Lee, W. T. Yang, and R. G. Parr, *Phys. Rev. B* 37, 785 (1988).
83. A. D. Becke, *J. Chem. Phys.* 98, 5648 (1993).
84. R. E. Stratmann, G. E. Scuseria, and M. J. Frisch, *Chem. Phys. Lett.* 257, 213 (1996).
85. M. Elstner, T. Frauenheim, E. Kaxiras, G. Seifert, and S. Suhai, *Phys. Status Solidi B* 217, 357 (2000).
86. M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, and G. Seifert, *Phys. Rev. B* 58, 7260 (1998).
87. H. G. Bohr, K. J. Jalkanen, M. Elstner, K. Frimand, and S. Suhai, *Chem. Phys.* 246, 13 (1999).
88. M. J. S. Dewar and W. Thiel, *J. Am. Chem. Soc.* 99, 4899 (1977).
89. M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, and J. J. P. Stewart, *J. Am. Chem. Soc.* 107, 3902 (1985).
90. J. J. P. Stewart, *J. Comp. Chem.* 10, 209 (1989).
91. J. A. Pople, D. P. Santry, and G. A. Segal, *J. Chem. Phys.* 43, S129 (1965).
92. R. C. Bingham, M. J. S. Dewar, and D. H. Lo, *J. Am. Chem. Soc.* 97, 1294 (1975).
93. W. Thiel and A. A. Voityuk, *J. Mol. Struc. (Theochem)* 119, 141 (1994).
94. S. Scheiner, *Theor. Chim. Acta* 57, 71 (1980).
95. J. J. P. Stewart, *J. Comput. Aided Mol. Des.* 4, 1 (1990).
96. J. J. P. Stewart, "MOPAC2002," Fujitsu Ltd., Tokyo, 1999.
97. H. S. Rzepa and M. Y. Yi, *J. Chem. Soc. Perkin Trans. 2* 6, 943 (1990).
98. Y. Zheng and K. M. Merz, *J. Comput. Chem.* 13, 1151 (1992).
99. M. W. Jurema and G. C. Shields, *J. Comp. Chem.* 14, 89 (1993).
100. S. Fischer, S. Michnick, and M. Karplus, *Biochem.* 32, 13830 (1993).
101. J. J. P. Stewart, *J. Computer-Aided Mol. Design* 4, 1, (1990).
102. I. Rossi and D. G. Truhlar, *Chem. Phys. Lett.* 233, 231 (1995).
103. Q. Cui and M. Karplus, *J. Phys. Chem. B* 106, 1768 (2002).
104. G. Tresadern, H. Wang, P. F. Faulder, N. A. Burton, and I. H. Hillier, *Mol. Phys.* 101, 2775 (2003).
105. L. Ridder, I. M. C. M. Rietjens, J. Vervoort, and A. J. Mulholland, *J. Am. Chem. Soc.* 124, 9926 (2002).
106. T. D. Poulsen, M. Garcia-Viloca, J. L. Gao, and D. G. Truhlar, *J. Phys. Chem. B* 107, 9567 (2003).
107. W. T. Yang and T. S. Lee, *J. Chem. Phys.* 103, 5674 (1995).
108. S. L. Dixon and K. M. Merz, *J. Chem. Phys.* 107, 879 (1997).
109. A. Klamt and G. Schuurmann, *J. Chem. Soc. Perkin Trans.* 2, 799 (1993).
110. S. L. Dixon, A. Van der Vaart, V. Gogonea, J. J. Vincent, E. N. Brothers, D. Suarez, L. M. Westerhoff, and K. M. Merz, "Divcon99," The Pennyslvania State University, University Park, Pennyslvania, 1999.
111. V. Gogonea and K. M. Merz, *J. Phys. Chem. A* 103, 5171 (1999).
112. J. J. Vincent, S. L. Dixon, and K. M. Merz, *Theor. Chem. Acc.* 99, 220 (1998).
113. G. Nadig, L. C. Van Zant, S. L. Dixon, and K. M. Merz, *J. Am. Chem. Soc.* 120, 5593 (1998).
114. N. Diaz, D. Suarez, T. L. Sordo, and K. M. Merz, *J. Phys. Chem. B* 105, 11302 (2001).
115. J. J. P. Stewart, *J. Mol. Struc. (Theochem)* 401, 195 (1997).
116. S. J. Titmuss, P. L. Cummins, A. A. Bliznyuk, A. P. Rendell, and J. E. Gready, *Chem. Phys. Lett.* 320, 169 (2000).
117. S. P. Greatbanks, J. E. Gready, A. C. Limaye, and A. P. Rendell, *J. Comp. Chem.* 21, 788 (2000).
118. W. Yang and D. G. Drueckhammer, *J. Phys. Chem. B* 107, 5986 (2003).
119. A. Warshel, "Computer Modeling of Chemical Reactions in Enzymes and Solutions," Wiley, New York, 1997.
120. J. Villà and A. Warshel, *J. Phys. Chem. B* 105, 7887 (2001).
121. J. K. Hwang, G. King, S. Creighton, and A. Warshel, *J. Am. Chem. Soc.* 110, 5297 (1988).
122. A. Warshel and R. M. Weiss, *J. Am. Chem. Soc.* 102, 6218 (1980).
123. J. Åqvist and A. Warshel, *Chem. Rev.* 93, 2523 (1993).
124. J. Bentzien, R. P. Muller, J. Florian, and A. Warshel, *J. Phys. Chem. B* 102, 2293 (1998).
125. A. Warshel, *Annu. Rev. Biophys. Biomol. Struct.* 32, 425 (2003).
126. H. J. Kim and J. T. Hynes, *J. Am. Chem. Soc.* 114, 10508 (1992).
127. U. W. Schmitt and G. A. Voth, *J. Phys. Chem. B* 102, 5547 (1998).
128. R. Vuilleumier and D. Borgis, *Chem. Phys. Lett.* 284, 71 (1998).
129. J. Florian, M. F. Goodman, and A. Warshel, *J. Am. Chem. Soc.* 125, 8163 (2003).
130. P. Varnai and A. Warshel, *J. Am. Chem. Soc.* 122, 3849 (2000).
131. A. Warshel and M. Levitt, *J. Mol. Biol.* 103, 227 (1976).

132. A. J. Mulholland, in "The QM/MM Approach to Enzymatic Reactions" (L. Eriksson, Ed.), p. 597. Elsevier, Amsterdam, 2001.
133. J. Gao and M. A. Thompson, "Combined Quantum Mechanical and Molecular Mechanical Methods." American Chemical Society, Washington, DC, 1998.
134. D. Bakowies and W. Thiel, *J. Phys. Chem.* 100, 10580 (1996).
135. M. Svensson, S. Humbel, R. D. J. Froese, T. Matsubara, S. Sieber, and K. Morokuma, *J. Phys. Chem.* 100, 19357 (1996).
136. A. Lewandowicz, J. Rudzinski, L. Tronstad, M. Widersten, P. Ryberg, O. Matsson, and P. Paneth, *J. Am. Chem. Soc.* 123, 4550 (2001).
137. M. J. Field, P. A. Bash, and M. Karplus, *J. Comput. Chem.* 11, 700 (1990).
138. U. Röthlisberger, in "Combined Quantum Mechanical and Molecular Mechanical Methods" (J. Gao and M. A. Thompson, Eds.), p. 264. American Chemical Society, Washington, DC, 1998.
139. G. G. Ferenczy, P. J. Winn, and C. A. Reynolds, *J. Phys. Chem. A* 101, 5446 (1997).
140. M. A. Thompson, *J. Phys. Chem.* 100, 14492 (1996).
141. S. R. Gooding, P. J. Winn, R. I. Maurer, G. G. Ferenczy, J. R. Miller, J. E. Harris, D. V. Griffiths, and C. A. Reynolds, *J. Comp. Chem.* 21, 478 (2000).
142. J. L. Gao, *J. Comp. Chem.* 18, 1061 (1997).
143. I. Antes and W. Thiel, *J. Phys. Chem. A* 103, 9290 (1999).
144. S. P. Greatbanks, J. E. Gready, A. C. Limaye, and A. P. Rendell, *Proteins* 37, 157 (1999).
145. J. Gao, *Acc. Chem. Res.* 29, 298 (1996).
146. P. D. Lyne, M. Hodoscek, and M. Karplus, *J. Phys. Chem. A* 103, 3462 (1999).
147. J. Gao and C. Alhambra, *J. Chem. Phys.* 107, 1212 (1997).
148. L. L. Ho, A. D. MacKerell, and P. A. Bash, *J. Phys. Chem.* 100, 4466 (1996).
149. M. Freindorf and J. L. Gao, *J. Comp. Chem.* 17, 386 (1996).
150. G. Monard, M. Loos, V. Théry, K. Baka, and J.-L. Rivail, *Int. J. Quantum Chem.* 58, 153 (1996).
151. X. Assfeld and J.-L. Rivail, *Chem. Phys. Lett.* 263, 100 (1996).
152. J. Gao, P. Amara, C. Alhambra, and M. J. Field, *J. Phys. Chem. A* 102, 4714 (1998).
153. V. Théry, D. Rinaldi, J.-L. Rivail, B. Maigret, and G. G. Ferenczy, *J. Comp. Chem.* 15, 269 (1994).
154. G. Monard, X. Prat-Resina, A. Gonzalez-Lafont, and J. M. Lluch, *Int. J. Quantum Chem.* 93, 229 (2003).
155. T. A. Wesolowski and A. Warshel, *J. Phys. Chem.* 97, 8050 (1993).
156. P. Amara and M. J. Field, *Theor. Chem. Acc.* 109, 43 (2003).
157. "HyperChem Users Manual," HyperCube, Inc., Waterloo, Ontario, 2002.
158. N. Reuter, A. Dejaegere, B. Maigret, and M. Karplus, *J. Phys. Chem. A* 104, 1720 (2000).
159. Y. Zhang, T.-S. Lee, and W. Yang, *J. Chem. Phys.* 110, 46 (1999).
160. L. Ridder, A. J. Mulholland, J. Vervoort, and I. M. C. M. Rietjens, *J. Am. Chem. Soc.* 120, 7641 (1998).
161. D. J. Vocadlo, G. J. Davies, R. Laine, and S. G. Withers, *Nature* 412, 835 (2001).
162. A. J. Mulholland and W. G. Richards, *Proteins* 27, 9, (1997).
163. J. Antosiewicz, J. M. Briggs, A. H. Elcock, M. K. Gilson, and J. A. McCammon, *J. Comp. Chem.* 17, 1633 (1996).
164. A. J. Mulholland and W. G. Richards, *J. Mol. Struc. (Theochem)* 429, 13 (1998).
165. A. J. Mulholland and W. G. Richards, *Int. J. Quantum Chem.* 51, 161 (1994).
166. F. Himo and P. E. M. Siegbahn, *Chem. Rev.* 103, 2421 (2003).
167. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery, R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, A. G. Baboul, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, J. L. Andres, C. Gonzalez, M. Head-Gordon, E. S. Replogle, and J. A. Pople, "Gaussian 98, Revision A.7." Gaussian, Inc., Pittsburgh, PA, 1998.
168. "Jaguar 4.2." Schrondinger Inc., Oregon, 2000.
169. K. Fukui, *Acc. Chem. Res.* 14, 363 (1981).
170. P. Derreumaux, G. H. Zhang, T. Schlick, and B. Brooks, *J. Comp. Chem.* 15, 532 (1994).
171. O. Farkas and H. B. Schlegel, *J. Chem. Phys.* 111, 10806 (1999).
172. S. Patchkovskii and W. Thiel, *J. Comp. Chem.* 17, 1318 (1996).
173. W. Thiel, *J. Mol. Struc. (Theochem)* 398, 1 (1997).
174. F. Maseras and K. Morokuma, *J. Comp. Chem.* 16, 1170 (1995).
175. A. J. Turner, V. Moliner, and I. H. Williams, *Phys. Chem. Chem. Phys.* 1, 1323 (1999).
176. V. Moliner, A. J. Turner, and I. H. Williams, *Chem. Commun.* 14, 1271 (1997).
177. S. Marti, J. Andres, V. Moliner, E. Silla, I. Tunon, and J. Bertran, *Theor. Chem. Acc.* 105, 207 (2001).
178. R. Castillo, J. Andres, and V. Moliner, *J. Am. Chem. Soc.* 121, 12140 (1999).
179. R. Elber and M. Karplus, *Chem. Phys. Lett.* 139, 375 (1987).
180. R. Elber and M. Karplus, *Chem. Phys. Lett.* 311, 335 (1999).
181. R. Elber and D. Shalloway, *J. Chem. Phys.* 112, 5539 (2000).
182. O. S. Smart, *Chem. Phys. Lett.* 222, 503 (1994).
183. O. S. Smart and J. M. Goodfellow, *Mol. Simulat.* 14, 291 (1995).

184. K. E. Ranaghan, L. Ridder, B. Szefczyk, W. A. Sokalski, J. C. Hermann, and A. J. Mulholland, *Mol. Phys.* **101**, 2695 (2003).
185. S. Fischer and M. Karplus, *Chem. Phys. Lett.* 194, 252 (1992).
186. M. Karplus and G. A. Petsko, *Nature* 347, 631 (1990).
187. A. T. Hadfield and A. J. Mulholland, *Int. J. Quantum Chem.* 73, 137 (1999).
188. H. A. Carlson and W. L. Jorgensen, *J. Am. Chem. Soc.* 118, 8475 (1996).
189. C. L. Brooks III and M. Karplus, *J. Mol. Biol.* 208, 159 (1989).
190. C. L. Brooks III, A. T. Brunger, and M. Karplus, *Biopolymers* 24, 843 (1985).
191. G. M. Torrie and J. P. Valleau, *Chem. Phys. Lett.* 28, 578 (1974).
192. F. Proust-De Martin, R. Dumas, and M. J. Field, *J. Am. Chem. Soc.* 122, 7688 (2000).
193. S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, *J. Comput. Chem.* 13, 1011 (1992).
194. J. Chandrasekhar, S. F. Smith, and W. L. Jorgensen, *J. Am. Chem. Soc.* 107, 154 (1985).
195. J. Chandrasekhar and W. L. Jorgensen, *J. Am. Chem. Soc.* 107, 2974 (1985).
196. W. L. Jorgensen, *Acc Chem. Res.* 22, 184 (1989).
197. R. V. Stanton, M. Perakyla, D. Bakowies, and P. A. Kollman, *J. Am. Chem. Soc.* 120, 3448 (1998).
198. G. Alagona, C. Ghio, and P. A. Kollman, *J. Comp. Chem.* 24, 46 (2003).
199. O. Donini, T. Darden, and P. A. Kollman, *J. Am. Chem. Soc.* 122, 12270 (2000).
200. B. Kuhn and P. A. Kollman, *J. Am. Chem. Soc.* 122, 2586 (2000).
201. Y. K. Zhang, H. Y. Liu, and W. T. Yang, *J. Chem. Phys.* 112, 3483 (2000).
202. R. Car and M. Parrinello, *Phys. Rev. Lett.* 55, 2471 (1985).
203. D. K. Remler and P. A. Madden, *Mol. Phys.* 70, 921 (1990).
204. F. Alber and P. Carloni, *Protein Sci.* 9, 2535 (2000).
205. M. Eichinger, P. Tavan, J. Hutter, and M. Parrinello, *J. Chem. Phys.* 110, 10452 (1999).
206. U. Röthlisberger and P. Carloni, *Int. J. Quantum Chem.* 73, 209 (1999).
207. U. Röthlisberger, P. Carloni, K. Doclo, and M. Parrinello, *J. Biol. Inorg. Chem.* 5, 236 (2000).
208. L. Ridder, J. N. Harvey, I. M. C. M. Rietjens, J. Vervoort, and A. J. Mulholland, *J. Phys. Chem. B* **107**, 2118 (2003).
209. B. Entsch, D. P. Ballou, and V. Massey, *J. Biol. Chem.* 251, 2550 (1976).
210. B. Entsch, B. A. Palfey, D. P. Ballou, and V. Massey, *J. Biol. Chem.* 266, 17341 (1991).
211. K. Eschrich, F. T. van der Bolt, A. de Kok, and W. J. H. van Berkel, *Eur. J. Biochem.* 216, 137 (1993).
212. J. Vervoort, I. M. C. M. Rietjens, W. J. H. van Berkel, and C. Veeger, *Eur. J. Biochem.* 206, 479 (1992).
213. K. Maeda-Yorita and V. Massey, *J. Biol. Chem.* 268, 4134 (1993).
214. R. K. Wierenga, R. J. Dejong, K. H. Kalk, W. G. J. Hol, and J. Drenth, *J. Mol. Biol.* 131, 55 (1979).
215. H. A. Schreuder, P. A. J. Prick, R. K. Wierenga, G. Vriend, K. S. Wilson, W. G. J. Hol, and J. Drenth, *J. Mol. Biol.* 208, 679 (1989).
216. L. Ridder, A. J. Mulholland, I. M. C. M. Rietjens, and J. Vervoort, *J. Am. Chem. Soc.* 122, 8728 (2000).
217. L. Ridder, B. A. Palfey, J. Vervoort, and I. M. C. M. Rietjens, *FEBS lett.* 478, 197 (2000).
218. S. Peelen, I. M. C. M. Rietjens, M. G. Boersma, and J. Vervoort, *Eur. J. Biochem.* 227, 284 (1995).
219. M. A. Cunningham, L. L. Ho, D. T. Nguyen, R. E. Gillilan, and P. A. Bash, *Biochem.* 36, 4800 (1997).
220. P. A. Bash, M. J. Field, R. C. Davenport, G. A. Petsko, D. Ringe, and M. Karplus, *Biochem.* 30, 5826 (1991).
221. M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. J. Su, T. L. Windus, M. Dupuis, and J. A. Montgomery, *J. Comp. Chem.* 14, 1347 (1993).
222. S. R. Billeter, C. F. W. Hanser, T. Z. Mordasini, M. Scholten, W. Thiel, and W. F. van Gunsteren, *Phys. Chem. Chem. Phys.* 3, 688 (2001).
223. W. F. van Gunsteren and H. J. C. Berendsen, "Groningen Molecular Simulation (GROMOS) Library Manual." Biomos, Groningen, 1987.
224. W. F. van Gunsteren, P. K. Weiner, and A. J. Wilkinson, "Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications," Vol. 2. Escom Science, Leiden, 1994.
225. P. C. Babbitt, *Proc. Natl. Acad. Sci.* 97, 10298 (2000).
226. M. J. S. Dewar and Y.-C. Yuan, *Inorg. Chem.* 29, 3881 (1990).
227. P. A. Bash, L. L. Ho, A. D. MacKerell, D. Levine, and P. Hallstrom, *Proc. Natl. Acad. Sci.* 93, 3698 (1996).
228. C. Bartels and M. Karplus, *J. Comput. Chem.* 18, 1450 (1997).
229. W. J. Chen, G. F. Graminski, and R. N. Armstrong, *Biochem.* 27, 647 (1988).
230. S. W. Huskey, W. P. Huskey, and A. Y. H. Lu, *J. Am. Chem. Soc.* 113, 2283 (1991).
231. P. Zhang, S. Liu, S. O. Shan, X. L. G. L. Gilliland, and R. N. Armstrong, *Biochem.* 31, 10185 (1992).
232. P. D. Lyne, A. J. Mulholland, and W. G. Richards, *J. Am. Chem. Soc.* 117, 11345 (1995).
233. R. J. Hall, S. A. Hindle, N. A. Burton, and I. H. Hillier, *J. Comp. Chem.* 21, 1433 (2000).
234. S. E. Worthington, A. E. Roitberg, and M. Krauss, *J. Phys. Chem. B* 105, 7087 (2001).
235. W. J. Stevens, H. Basch, and M. Krauss, *J. Chem. Phys.* 81, 6026 (1984).
236. O. Wiest and K. N. Houk, *J. Am. Chem. Soc.* 117, 11628 (1995).
237. M. M. Davidson, I. R. Gould, and I. H. Hillier, *J. Chem. Soc. Perkin Trans. 2* 4, 525 (1996).
238. S. Marti, J. Andres, V. Moliner, E. Silla, I. Tunon, and J. Bertran, *J. Phys. Chem. B* 104, 11308 (2000).
239. S. Hur and T. C. Bruice, *J. Am. Chem. Soc.* 125, 1472 (2003).
240. L. Pauling, *Am. Sci.* 36, 51 (1948).
241. J. B. S. Haldane, "Enzymes." Longman, Green & Co., London, 1930.

242. A. T. Hadfield, D. J. Harvey, D. B. Archer, D. A. MacKenzie, D. J. Jeenes, S. E. Radford, G. Lowe, C. M. Dobson, and I. N. Johnson, *J. Mol. Biol.* 243, 856 (1994).
243. K. E. Ranaghan, L. Ridder, B. Szefczyk, W. A. Sokalski, J. C. Hermann, and A. J. Mulholland, *Org. Biomol. Chem.* 2, 968 (2004).
244. A. Crespo, D. A. Scherlis, M. A. Marti, P. Ordejon, A. E. Roitberg, and D. A. Estrin, *J. Phys. Chem. B* 107, 13728 (2003).
245. Y. Cha, C. J. Murray, and J. P. Klinman, *Science* 243, 1325 (1989).
246. T. Jonsson, M. H. Glickman, S. J. Sun, and J. P. Klinman, *J. Am. Chem. Soc.* 118, 10919 (1996).
247. J. Basran, M. J. Sutcliffe, and N. S. Scrutton, *Biochem.* 38, 3218 (1999).
248. J. Basran, S. Patel, M. J. Sutcliffe, and N. S. Scrutton, *J. Biol. Chem.* 276, 6234 (2001).
249. S. C. Tsai and J. P. Klinman, *Biochem.* 40, 2303 (2001).
250. Y.-L. Hyun and V. L. Davidson, *BBA-Protein Struct. M* 1251, 198 (1995).
251. H. B. Brooks, L. H. Jones, and V. L. Davidson, *Biochem.* 32, 2725 (1993).
252. D. G. Truhlar and B. C. Garrett, *Acc Chem. Res.* 13, 440 (1980).
253. C. Alhambra, M. L. Sanchez, J. C. Corchado, J. Gao, and D. G. Truhlar, *Chem. Phys. Lett.* 355, 388 (2002).
254. Y. P. Liu, G. C. Lynch, T. N. Truong, D. H. Lu, D. G. Truhlar, and B. C. Garrett, *J. Am. Chem. Soc.* 115, 2408 (1993).
255. M. Garcia-Viloca, C. Alhambra, D. G. Truhlar, and J. Gao, *J. Chem. Phys.* 114, 9953 (2001).
256. C. Alhambra, J. Corchado, M. L. Sanchez, M. Garcia-Viloca, J. Gao, and D. G. Truhlar, *J. Phys. Chem. B* 105, 11326 (2001).
257. P. R. Ortiz de Montellano, Ed., "Cytochrome P-450: Structure, Mechanism and Biochemistry," 2nd edn. Plenum, New York, 1995.
258. J. T. Groves and G. A. McClusky, *J. Am. Chem. Soc.* 98, 859 (1976).
259. J. T. Groves and M. Van der Puy, *J. Am. Chem. Soc.* 98, 5290 (1976).
260. J. T. Groves and D. V. Adhyam, *J. Am. Chem. Soc.* 106, 2177 (1984).
261. F. Ogliaro, N. Harris, S. Cohen, M. Filatov, S. P. de Visser, and S. Shaik, *J. Am. Chem. Soc.* 122, 8977 (2000).
262. N. Harris, S. Cohen, M. Filatov, F. Ogliaro, and S. Shaik, *Angew. Chem. Int. Ed.* 39, 2003 (2000).
263. K. Yoshizawa, T. Kamachi, and Y. Shiota, *J. Am. Chem. Soc.* 123, 9806 (2001).
264. T. Kamachi and K. Yoshizawa, *J. Am. Chem. Soc.* 125, 4652 (2003).
265. P. H. Toy, M. Newcomb, and P. F. Hollenberg, *J. Am. Chem. Soc.* 120, 7719 (1998).
266. S. P. de Visser, F. Ogliaro, N. Harris, and S. Shaik, *J. Am. Chem. Soc.* 123, 3037 (2001).
267. S. P. de Visser, F. Ogliaro, and S. Sason, *Chem. Commun.* 22, 2322 (2001).
268. J. T. Groves, Z. Gross, and M. K. Stern, *Inorg. Chem.* 33, 5065 (1994).
269. Z. Gross, S. Nimri, C. M. Barzilay, and L. Simkhovich, *J. Biol. Inorg. Chem.* 2, 492 (1997).
270. K. R. Korzekwa, D. C. Swinney, and W. F. Trager, *Biochem.* 28, 9019 (1989).
271. C. M. Bathelt, L. Ridder, A. J. Mulholland, and J. N. Harvey, *J. Am. Chem. Soc.* 125, 15004 (2003).
272. S. Dincturk and R. A. Jackson, *J. Chem. Soc. Perkin Trans.* 2, 1127 (1981).
273. R. Taylor, "Electrophilic Aromatic Substitution." Wiley, New York, 1989.
274. J. Andres, V. S. Safont, J. B. L. Martins, A. Beltran, and V. Moliner, *J. Mol. Struc. (Theochem)* 330, 411 (1995).
275. O. Tapia, R. Cardenas, J. Andres, and F. Colonna-Cesari, *J. Am. Chem. Soc.* 110, 4046 (1988).
276. P. L. Cummins and J. E. Gready, *J. Comp. Chem.* 19, 977 (1998).
277. P. L. Cummins, S. P. Greatbanks, A. P. Rendell, and J. E. Gready, *J. Phys. Chem. B* 106, 9934 (2002).
278. S. Ferrer, E. Silla, I. Tunon, S. Marti, and V. Moliner, *J. Phys. Chem. B* 107, 14036 (2003).
279. M. J. Rothman and L. L. Lohr, *Chem. Phys. Lett.* 70, 405 (1980).
280. J. B. Watney, P. K. Agarwal, and S. Hammes-Schiffer, *J. Am. Chem. Soc.* 125, 3745 (2003).
281. P. K. Agarwal, S. R. Billeter, P. T. R. Rajagopalan, S. J. Benkovic, and S. Hammes-Schiffer, *Proc. Natl. Acad. Sci.* 99, 2794 (2002).
282. I. F. Thorpe and C. L. Brooks III, *J. Phys. Chem. B* 107, 14042 (2003).
283. J. A. Gerlt, M. M. Kreevoy, W. W. Cleland, and P. A. Frey, *Chem. Biol.* 4, 259 (1997).
284. A. J. Mulholland and W. G. Richards, *Proteins* 27, 9 (1997).
285. A. J. Mulholland, P. D. Lyne, and M. Karplus, *J. Am. Chem. Soc.* 122, 534 (2000).
286. A. J. Mulholland and W. G. Richards, *J. Mol. Struc. (Theochem)* 427, 175 (1998).
287. H. Y. Liu, F. Muller-Plathe, and W. F. van Gunsteren, *J. Mol. Biol.* 261, 454 (1996).
288. D. Joseph, G. A. Petsko, and M. Karplus, *Science* 249, 1425 (1990).
289. G. Alagona, C. Ghio, and P. A. Kollman, *J. Am. Chem. Soc.* 117, 9855 (1995).
290. J. Åqvist and M. Fothergill, *J. Biol. Chem.* 271, 10010 (1996).
291. T. K. Harris, R. N. Cole, F. I. Comer, and A. S. Mildvan, *Biochem.* 37, 16828 (1998).
292. T. K. Harris, C. Abeygunawardana, and A. S. Mildvan, *Biochem.* 36, 14661 (1997).
293. Q. Cui, M. Elstner, E. Kaxiras, T. Frauenheim, and M. Karplus, *J. Phys. Chem. B* 105, 569 (2001).

CHAPTER 8

Modeling of Biologically Motivated Soft Matter Systems

Ilpo Vattulainen

Laboratory of Physics and Helsinki Institute of Physics, Helsinki University of Technology, Helsinki, Finland

Mikko Karttunen

Biophysics and Statistical Mechanics Group, Laboratory of Computational Engineering, Helsinki University of Technology, Helsinki, Finland

CONTENTS

ISBN: 1-58883-048-9

Handbook of Theoretical and Computational Nanotechnology
Edited by Michael Rieth and Wolfram Schommers
Volume 6: Pages (361–417)

1. INTRODUCTION

Up to about 10 years ago, computational physics was considered to be more or less a fringe discipline having very little fundamental value. This is clearly demonstrated by the extremely small number of universities that offered a proper curriculum in computational physics at that time. This attitude has changed drastically, however, and today computational physics has become part of the mainstream. It has been particularly successful in fields such as materials science and soft matter, in which the complexity of the systems poses formidable challenges for the understanding of their behavior.

The negative attitude may have had its origin in the desire to be able to solve physical problems either by exact analytical means or by well-established experimental techniques. As any computer simulation only provides numerical solutions to models (as compared to exact solutions), it is likely that simulation was not considered to have the same seriousness and fundamental value as other methods. A related dilemma is known by most of those who teach computational methods in interdisciplinary programs: before they learn what computational modeling is about, there are always a few opportunistic students who have the idea that computational modeling is an easy way out without having to know much of the underlying theory or having to deal with complicated equipment. Yet in many ways computational modelers are like experimentalists. Instead of microscopes or accelerators, however, they use computers, and instead of designing experiments, they try to build accurate predictive models—for that they have to have solid knowledge of both the underlying theory and possible experiments.

Today, there is no doubt about the value of computer simulations. It has become very clear that advances in theory, experiments, and computational modeling go hand in hand. This is particularly so in interdisciplinary fields such as soft matter and biophysics [1–4], in which one needs to combine both knowledge and methodologies from very different fields. Computer simulations have also become, or are becoming, standard tools in particle physics, materials science, drug design, and biology, and the list could be easily continued to include almost any field in science and engineering.

In this review we focus on recent developments and advances in modeling soft matter and biological systems. Because of both fundamental and technological reasons, there is strong desire to understand the overall behavior of complex macromolecular and biological soft matter systems. From a theoretical point of view, this task is highly nontrivial because processes in these systems take place over a huge range of length and timescales, whereas current modeling and analytical techniques are feasible over relatively limited scales only.

These systems often belong to the class of so-called complex fluids, which are characterized by their multicomponent nature and simultaneous presence of several, often unseparable, time- and length scales. The presence of many time- and length scales is the crux of soft matter and biological modeling (see Fig. 1).

The time- and length scale problem is best illustrated by an example: water is always present in biological systems. The size of a water molecule is approximately 10^{-10} m. The relevant timescale is defined by molecular vibrations having characteristic times on the order of 10^{-15} s. However, the biologically relevant problem of protein folding can take anything from 1 μs up to about 1000 s, depending on the size of the protein. As for length scales, proteins are long polyampholytes (i.e., polymers carrying both positive and negative charges, and possessing a whole hierarchy of spatial ordering in different interdependent scales; see e.g., Ref. [5]). The other "classical" problem in biophysics involves DNA. The double stranded DNA has a length of 1 m and a persistence length of approximately 50–100 nm (having some sequence dependence). To make things even more complicated, one should remember that proteins, DNA, and lipids operate in cells, which from a biological point of view are the functional entities of interest. Furthermore, cells consist of a large variety of different components (proteins, different lipids, etc.) and have typical sizes around a few tens of micrometers. As if the above would not be problematic enough, in living matter they are actually dissipative structures operating under nonequilibrium conditions. Deciphering the physical mechanisms of living matter is guaranteed to provide excitement and challenges for generations of scientists.

Before moving on to more specific issues of computational modeling, another example is provided by polymers. Although many biologically relevant molecules such as proteins and DNA can be considered as biopolymers, there is also an enormously large field of polymer science not related to living matter. Technologically, polymers are one of the central materials in modern-day society, with applications ranging from sophisticated medical materials to the plastic bags we encounter while grocery shopping. From the modeling point of view, it is really the field of polymers in which the ideas of linking many time- and length scales have developed the fastest. This is easy to understand through the following simple example: The timescales associated with bond vibrations are roughly 10^{-15} s, whereas conformational transitions associated with individual bonds occur typically in timescales of 10^{-11} s. The related changes taking place along the chain take orders of magnitude longer than these

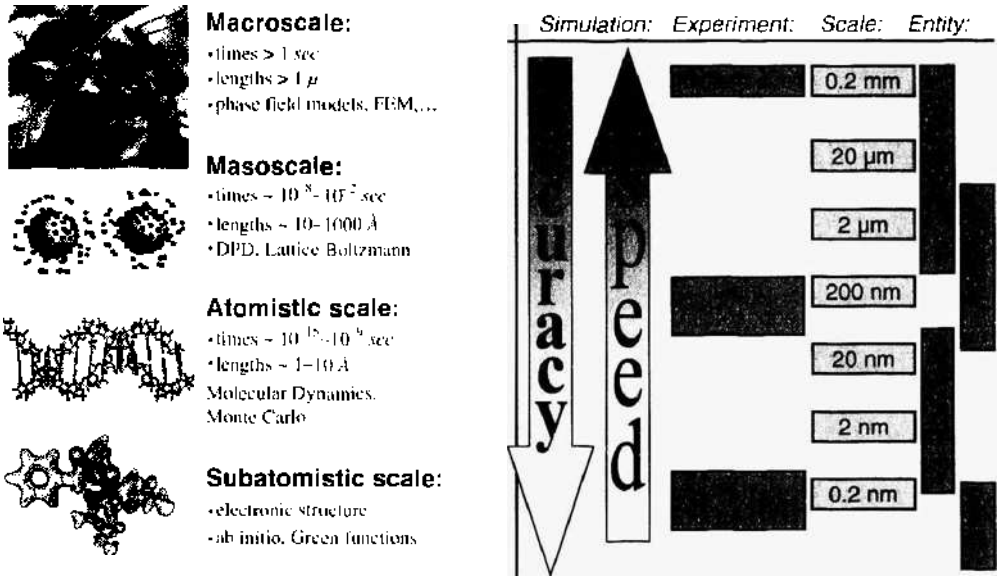


Figure 1. Different time- and length scales, typical computational methods used to study them, and biological entities related to various length scales. The dilemma between speed and accuracy is always present in simulations: it is always a trade-off between the two. In analogy to various simulation methods at different time- and length scales, it is not possible to use single experimental methods to cover all properties. The right-hand side of the figure is adapted from Ref. [4].

timescales. Furthermore, industrially important processes such as spinodal decomposition, or phase separation in general, have characteristic times of at least seconds.

What does the above mean from the point of view of computer simulations? There is a whole plethora of different methods, all having both advantages and limitations. In *ab initio* simulations, quantum mechanical details are taken into account, limiting the obtainable system sizes to a few hundred atoms as a result to the large number of electronic degrees of freedom—this is obviously not enough for polymeric or biological systems, is it? One can make an estimate that using the most powerful central processing units (CPUs) available at the moment for simulating the folding of the smallest proteins by using accurate *ab initio* methods, it would take about 10000 CPU-years—and this is a fairly optimistic scenario.

As a consequence, quantum mechanical approaches are appropriate only for issues in which electronic degrees of freedom cannot be neglected. In addition, they can be used to develop force fields for classical molecular dynamics (MD) simulations that are able to reach time scales of the order of 100 ns and linear system sizes of some tens of nanometers. Classical MD thus provides a unique tool to study various properties related to many biologically relevant soft matter systems, including small biomolecules and systems composed of these systems. The MD technique is indeed the workhorse in the modeling of biophysical systems, and it will be described in detail in the following sections.

The above very clearly expresses the fact that the range of problems in which quantum mechanics is appropriate is very limited. Further, the realistic time- and length scales accessible through MD simulations are also rather modest and far below those important in, say, polymer dynamics or protein folding. What this suggests is that one should think of building a hierarchy of different methods, each method being valid and useful over a certain well-defined length and timescale. By a proper design, the different methods can be made, ideally at least, rigorously related to each other through systematic coarse-graining approaches in which the detailed atomic or molecular description is replaced with a more simple one. This is the idea behind multiscale modeling which has recently attracted a rapidly increasing amount of attention in computational materials research [6–16]. The approach where classical force fields are obtained from quantum MD and then used in classical MD can, as a matter of fact, be seen as an example of a coarse-graining procedure.

The multiscale approach sounds appealing, but it includes a fundamental problem: There is no unique way to perform coarse-graining. There are many aspects to that problem, but one can immediately get a grasp of it by considering a simple-minded spatial coarse-graining of a cholesterol molecule depicted in Fig. 2: How should one select the new "superatoms"? What are the new interactions in the coarse-grained description? Is the new model consistent with the microscopic one (i.e., does it provide correct static and dynamic properties)? Are there alternative methods? And so on.

Because of the above reasons, and despite the difficulties, coarse-graining has become an increasingly active field of research. A variety of different approaches, some more and others less rigorous, have been presented. We will discuss some of them in the following sections but, to give a flavor, we mention a few here. As already discussed, finding a

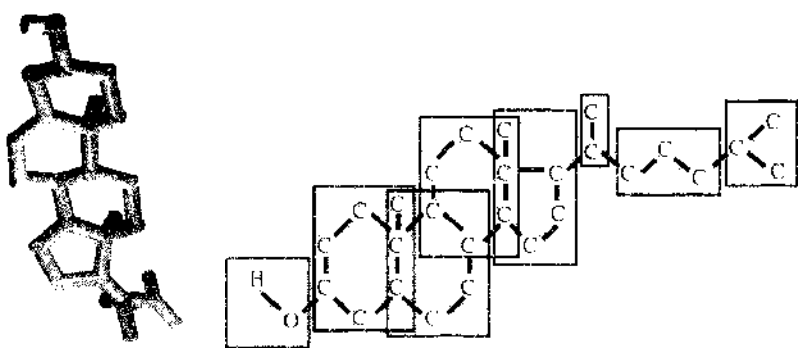


Figure 2. A model of cholesterol. Left: an atomic-level representation with the OH group shown with red and white at the top. Right: the same molecule with a possible definition of new coarse-grained interaction sites, or "superatoms," shown by the rectangular blocks.

systematic hierarchy is a major problem. Another important aspect concerns the fact that multiscale modeling involves two possible processes. First of all is coarse-graining or transforming detailed models to simplified descriptions with less degrees of freedom. This means averaging over some chosen properties of microscopic entities to form larger basic units for modeling. Second, fine-graining which is the opposite of coarse-graining. In this process, configurational properties are typically the key quantities as one maps a coarser model to a more detailed one.

In general, one tries to approach the problem via one of the following routes: by phenomenological methods based on (e.g., Ginzburg–Landau-type approach); by analytical approaches based on the projection operator formalism; by analyzing static properties and constructing the potential of mean force; by free-energy methods (i.e., configurational analysis); or by analyzing the occurrence rates of different processes.

The phenomenological approaches include methods such as dissipative particle dynamics (DPD; see, e.g., [17]), possibly the most used mesoscale modeling method for soft matter systems. The main idea there is to use "soft" phenomenological interparticle potentials and a momentum-conserving thermostat to ensure hydrodynamics. The intrinsic timescale in DPD is typically 10–20 ps, and the length scale is around 1 nm. DPD was introduced about a decade ago [18–20], and since then various generalizations have been introduced, providing the method a more rigorous theoretical background [21]. We will discuss DPD in detail in the following sections.

A more macroscopic phenomenological approach is based on Ginzburg–Landau-type formalism by introducing order parameters and free-energy and density functionals for the system. This approach is based on finding the relevant (slow) variables and general symmetry principles, and as such is close to a macroscopic description. It is particularly fruitful in dealing with phenomena involving spinodal decomposition or nucleation processes [22]. The review of Kawakatsu et al. [23] puts these models in perspective with more microscopic approaches, and a very detailed discussion of the mathematical aspects can be found in Ref. [24].

The most successful analytical approach is the so-called GENERIC, or general equation for non-equilibrium reversible–irreversible coupling, introduced by Grmela and Öttinger [25–27]. The advantage of this method is that it provides a physically sound approach based on statistical mechanics. It uses projection operator formalism in building up hierarchies between different levels of description, thus ensuring self-consistency. Despite its rather mathematical formalism, it can be used in connection with actual numerical simulations. An excellent review of GENERIC and other analytical approaches is provided by Español in Ref. [28].

Using static properties to obtain potentials of mean force has been a common approach in polymer science. One of the most interesting efforts has been the activity of the Mainz/Bremen region on optimized methods for generating potentials from structural information. The particularly interesting part of that is the inverse procedure (i.e., fine-graining alongside with the coarse-graining efforts). The problems in polymer science are challenging, and it is easy to appreciate them if one remembers that for polymer melts of low molecular weight, the conformational relaxation time (caused by diffusion) scales as N^2 (Rouse scaling), where N is the number of monomers. This said, the scaling is much worse for high-molecular-weight melt. In real-world applications, the relaxation times are from milliseconds upward. It is obvious that coarse-grained methods are needed and that this is really the field in which there has been a lot of effort. For reviews, see Refs. [29–32].

As another example, let us note that there have been only a few published coarse-grained lipid simulations, and to the authors' knowledge, they are all based on interaction site definitions (i.e., selecting interaction sites on some physical principles and then, often phenomenologically, choosing the relevant inter- and intramolecular potentials). That approach is, to a degree, always ad hoc, but the results of the Philadelphia group on lipid membranes [15, 33, 34] and the Groningen group on vesicle dynamics [35] have been very encouraging.

Thus, there is a timely practical problem of finding generally applicable and computationally efficient methods to coarse-grain molecular systems and of establishing a systematic link between the microscopic and macroscopic regimes by the means of statistical mechanics.

We will discuss many of these issues in detail, but we would also like to refer the reader to some of the existing reviews on specialized topics: Krőger [36] discusses practical issues related to the rheology of polymers in detail, including aspects related to GENERIC and projection operator methods. Baschnagl et al. [29] provide an extensive review on coarse-graining in polymer science. In addition, the lecture notes of the summer school Soft-Simu2002 [28] and the lecture notes of the conference "Bridging the Time-Scale Gap" [37] also address issues not covered in this review at length, including hydrodynamics, lattice Boltzmann, granular materials, and quantum-classical models.

In general, multiscale modeling has been a well-adopted and accepted field in Europe, and in particular by the European Science Foundation SIMU Research Network [38]. This is also the case in Japan, where the efforts are highlighted by Masao Doi's ambitious seamless zooming project, which produced the OCTA package for multiscale simulations of soft matter systems [39, 40]. Hence, it is rather surprising that there has been very little activity in North America. It is not difficult to predict that this situation will change.

The rest of the article is organized as follows. In Section 2 we discuss the field of molecular simulations in general, focusing on modeling techniques from the microscopic point of view. That is followed in Section 3 by a discussion of some of the novel and most used mesoscopic simulation methods, including DPD and its extensions, phase field modeling, lattice Boltzmann, and the GENERIC hierarchy. In Section 4 we discuss a number of methodological issues that are crucial in molecular modeling, and we continue in Section 5 by presenting a few applications of atomic molecular dynamics simulations in the context of biophysical systems. Finally, we present an example of coarse graining and mesoscale simulations in Section 6 and close this work by a brief summary in Section 7.

2. MOLECULAR SIMULATIONS

The field of molecular simulations is extremely wide, covering a variety of different methods for different purposes in the quantum, atomic-scale classical, and mesoscopic and continuum regimes. In this section, we briefly discuss the main ideas of atomic-level molecular simulations, starting from a historical perspective and proceeding to a few atomic-scale simulation techniques. As the topic is very wide, we have decided to focus on methods that are most relevant for the applications discussed later in this chapter.

Because of their importance for studies of large-scale properties of soft matter systems, coarse-graining and mesoscale simulation methods are discussed separately in Section 3.

2.1. Historical Perspective from ENIAC to GRID

The beginning of the era of computational sciences can be traced back to the early 1950s, after which the field went through a rapid increase in activities. The progress has taken place hand in hand with the development of computers and computational methods that are the basis of computational modeling of physical systems. Although the former of these two fields is related to microelectronics and semiconductor physics, and the latter is closely associated with applied mathematics and computational sciences, they are both closely coupled to molecular simulations. As a matter of fact, the history of these two fields of science is closely related to each other, too. Since the development of the first electronic computer, ENIAC [41], many of the people who took part in the pioneering work of developing computers have also had a strong effect on the development of computational methods and algorithms. For example, N. Metropolis originally suggested [42] an obvious name for the Monte Carlo (MC) simulation method [43, 44], as the method is largely based on the use of random numbers. Also, with J. von Neumann, Metropolis studied the randomness of the decimals of π and e [45] and developed the first algorithm for generating pseudorandom numbers (the so-called midsquare method [46, 47]).

Ever since, development of computers and simulation methods have taken steps hand in hand. From the pioneering computer simulations of simple liquids on MANIAC, we have come to a situation in which molecular simulations of complex biophysical systems such as DNA condensation and channel proteins in lipid membranes are almost daily routine.

On one hand, this is a result of computing resources that at present are tremendous compared to those that were available in the 1950s. On the other hand, this is also a result of major progress in the development of computer algorithms and the methodology in general.

How have we arrived in the present situation? How the modeling of molecular systems started in the first place? One is tempted to think that the history of molecular models is related to the development of computers in the 1950s, but actually modeling of molecular systems started long before computers were invented. To our knowledge, the first structural models for molecules were suggested in the end of the nineteenth and in the beginning of the twentieth centuries. Despite their rather crude nature, these models allowed one to better understand the three-dimensional structure and related physicochemical properties of liquids. The actual pioneer, however, is probably van der Waals, whose work lead to the well-known equation of state that predicted a first-order gas–liquid phase transition.

Mathematical models in terms of computer simulations joined the game 50 years ago, when the first computer simulation of simple liquids was carried out by Metropolis et al. [44]. The MC method used at that time allowed one to consider structural properties of liquids, and although the advances in computing power have been rapid, the method itself is still alive and well. As a matter of fact, it is nowadays used with success on an ever-increasing number of problems. Yet, as the MC method is not particularly appropriate for studies of dynamical properties, other techniques have been designed for this purpose. The most appropriate approach in this regard is, of course, the MD technique. In MD, a system is described as an ensemble of interacting particles whose evolution in time is found through an integration of Newton's equations of motion. This approach was first used for a system of hard spheres by Alder and Wainwright [48, 49]. A more realistic description of a liquid in terms of Lennard–Jones particles was later accomplished by Rahman in the 1960s [50].

After this groundwork, the progress has been rapid. Barker and Watts suggested [51] the first model for molecular liquid water using MC, and Rahman and Stillinger [52] followed using MD. Nevertheless, despite major activities that have extended for over 30 years, water has remained as one of the most difficult problems in the field of molecular simulations. Equally difficult is the description of proteins that originated from the early work by McCammon in 1977 [53]. The progress ever since has been remarkable, and it is rather confusing to try to understand how tedious biophysical systems can be tackled today despite their highly complex nature. Although the small proteins studied by McCammon in 1977 were considered in vacuum without electrostatics for a timescale of picoseconds, today one is able to simulate transmembrane water channels and other integral proteins in fully hydrated lipid membranes with full electrostatic interactions over a timescale of several nanoseconds [54, 55].

Although the first computer simulations were based on the use of punch cards [56], our work today is based on highly advanced computing architectures together with modern programming languages. Also, starting from the first computers such as ENIAC and MANIAC, we have come to a situation in which computer simulations are carried out in massively parallel supercomputers composed of thousands of central processing units (CPUs). Most recently, initiatives have been made to establish highly parallel and connected networks of computer clusters (GRIDs) to obtain similar capacity across the Internet. Although it is easy to look into the past, the rapid development in the field of molecular simulations does not make it too easy to make any conclusive predictions about the future. Anyhow, in the present and in the next section, we will find that the advances in simulation techniques have been substantial during the last decades. On this basis, it is easy to predict that the future of molecular simulations is both challenging and very positive.

2.2. Main Principles of Molecular Modeling

What is molecular modeling? To clarify this obvious matter, let us first quote the view by Tamar Schlick [57]: "Molecular modeling is the science and art of studying molecular structure and function through model building and computation."

Model building can be as simple as the Lego bricks that the authors liked a lot some 25 years ago—and still do. This approach can provide one with important insight into the

structure of liquids and more complex molecules. The structure of DNA, for example, was first discovered using this approach (though, to our knowledge, not with Lego bricks). The discovery of C_{60} Buckminsterfullerene provides another example, as the first prototype of its structure made by Richard E. Smalley was made of paper sheets [58]. On the computational side, model building starts from building the topology and initial structures of molecules that typically need input from experiments such as nuclear magnetic resonance (NMR) and various scattering techniques. Having done this, one needs to write down the Hamiltonian operator $\mathcal{H}$, which includes all interactions present in the model system. The Hamiltonian operator (often called "force field" in classical simulations) plays the key role in molecular simulations.

In principle, if we were clever enough, we could simply use $\mathcal{H}$ to solve the structure and the dynamics of the system using just paper, pen, and our brains. However, as we are often dealing with a many-body problem having much more than just two interacting components, an analytical exact solution is rarely possible [59]. Alternatively, one can simplify the problem and try to find approximate descriptions of the system, but then the validity of the approximations remains elusive. This barrier can be crossed by computer simulations, where the aim is to study the dynamics of a given system with the corresponding Hamiltonian through computers. Computer simulations allow us to do theoretical experiments without any approximations, thus providing a bridge between theory and experiments.

In the field of molecular simulations, this underlying idea has lead to the introduction of a number of useful simulation techniques, including *ab initio*, molecular mechanics, classical molecular dynamics, MC, free-energy calculation techniques, dissipative particle dynamics, and many many more. Below, we discuss the key issues of the main approaches designed for the atomic regime.

The topology of the molecules is here taken for granted. Thus we concentrate on the Hamiltonian operator and on the related issues of finding a proper force field for a given system. For a more thorough discussion on these techniques, see Refs. [57, 60–62].

2.3. Molecular Modeling in the Atomic Regime: From Quantum Mechanics to Classical Molecular Dynamics

2.3.1. Quantum Mechanical Techniques

Quantum mechanical (QM) simulation methods are based on the solution of the Schrödinger equation. The Schrödinger equation

$$\mathcal{H}\Psi_n = E_n\Psi_n \tag{1}$$

is written in terms of the Hamiltonian operator $\mathcal{H}$, which acts on the eigenfunctions Ψ_n whose corresponding energy eigenvalues (quantum states) are given by E_n. The Schrödinger equation therefore describes the spatial probability distributions of energy states in a stationary case.

As the system in this case is composed of both nuclear and electronic degrees of freedom, one often makes the assumption that the motions of electrons and nuclei can be separated from each other. This so-called Born–Oppenheimer approximation is almost without exceptions in use in present-day QM methods and is well justified because the nuclei are usually much heavier than the electrons, which implies that the positions of nuclei are essentially fixed on the timescale of electronic motion. This assumption reduces the computational load considerably compared to traditional electronic structure methods. Yet, further approximations are often crucial to allow studies of reasonable system sizes. In practice, there are two commonly adapted approaches that allow that.

Ab initio ("from the beginning") is based on the Born–Oppenheimer approximation of the time-independent Schrödinger equation, ignoring relativistic effects and using a linear combination of atomic orbitals that is specific to a chosen basis set. The energy of the system is calculated iteratively (self-consistently), which makes the calculation computationally expensive for reasonable system sizes. Another possibility is to use density-functional theory

(DFT) in which the method is constructed to be based on the use of the electron density function. Different DFT schemes [63] differ by their treatment of the exchange/correlation energy, though in general this class of methods is rather efficient and widely used. One popular technique, for example, is the approach suggested by Car and Parrinello. In this scheme, electronic DFT is used to calculate the energies and densities of the valence electrons "on the fly" [64]. Because the technique is basically MD combined with DFT in the adiabatic case using an extended Lagrangian approach, it is both efficient and highly useful in QM (as well as in classical) problems. Another issue that has attracted plenty of attention is the use of DFT for time-dependent problems. Because of recent progress, time-dependent DFT is nowadays the method of choice for calculating, for example, excitation energies of complex molecules, and it is becoming more and more popular for studies of spectroscopic properties of nanostructures.

Semiempirical QM methods provide an alternative approach. They are partly based on empirical information to replace a full QM description of parameters that define the forms and energies of atomic orbitals. As expected, this reduces the computational load considerably, possibly at the expense of accuracy.

In practice, QM approaches are applied in soft matter systems only to processes and phenomena for which quantum mechanics cannot be neglected. The action of enzymes close to their active site and other chemically activated processes are examples of such cases. Yet the computational burden is major and limits the timescales that can be studied with QM methods up to tens of picoseconds. As a consequence, recent work has focused on combining QM methods with classical approaches such that, for example, the active site of an enzyme is treated quantum mechanically, whereas the remaining part of the system is classical. The problem then is to treat the boundary between QM and classical regimes [65–68].

As we will not present applications based on QM approaches, we prefer not to discuss this issue in more depth. An interested reader is referred to Refs. [57, 62, 69, 70] for more information on QM techniques, and to Refs. [66, 67, 71–73] for recent applications.

2.3.2. Interactions in Classical Methods

Classical simulation techniques are obviously based on QM, but they describe a given system in a considerably more simplified fashion. In essence, in classical methods the quantum degrees of freedom are integrated over—"thrown under the carpet"—such that the QM effects are incorporated into the classical intramolecular and intermolecular interactions. For example, although electronic degrees of freedom are not included in classical molecular simulations, the exclusion rules of QM are taken into account in an effective sense: They are manifested in classical simulations through strongly repulsive interatomic interactions at short distances, which implies that two atoms cannot occupy the same space at the same time.

The absence of electronic degrees of freedom basically implies that classical simulation techniques are based on the assumption that the Born–Oppenheimer approximation is valid. Assuming that this assumption holds true (i.e., we can write the energy of the system as a function of nuclear coordinates, and the computational load of doing molecular simulations is reduced considerably). Instead of worrying about interactions in a QM level, one can then describe the energy of the system as a sum of contributions resulting from different kinds of processes, such as bond stretching and bond bending. The key word here is force field.

The force field is the heart of molecular computer simulations. It describes the internal energy of the system at any moment of time, and it obviously consists of a number of terms, each of which describes some physical process. Generally speaking, for a system of N particles whose coordinates are given by $\{\vec{r}_i\}$, one can write down the force field as follows:

$$V = \sum_i v_1(\vec{r}_i) + \sum_i \sum_{j>i} v_2(\vec{r}_i, \vec{r}_j) + \sum_i \sum_{j>i} \sum_{k>j>i} v_3(\vec{r}_i, \vec{r}_j, \vec{r}_k) + \cdots \tag{2}$$

The potential energy of the system is therefore divided into terms describing one-body, two-body, and three-body interactions, as well as other higher-order interactions that have not

been listed here. The one-body interaction (v_1) can be as simple as the action of a uniform electric field on a single atom. Two-body interactions (v_2) are obviously those between two interacting particles, and they are usually divided into bonded and nonbonded interactions. Bonded two-body interactions such as bond stretching are usually those between particles that are bonded by a covalent bond, and nonbonded interactions such as Coulombic interactions are between those atoms that are not bonded, such as atoms in different molecules. Higher-order interactions such as three- (v_3) and four-body interactions include bond bending, torsional interactions, and so forth.

In practice, a simple force field typically used for biologically relevant molecules can look like the following:

$$V = \sum_{\text{bonds}} \frac{k_i^b}{2}(l_i - l_i^{\text{ref}})^2 + \sum_{\text{angles}} \frac{k_i^\theta}{2}(\theta_i - \theta_i^{\text{ref}})^2 + \sum_{\text{torsions}} \frac{V_T}{2}[1 + \cos(n\omega - \gamma)] + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left\{ 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\} \tag{3}$$

Here the first term on the right-hand side describes the interaction resulting from bond stretching for a pair of bonded particles, whose reference distance is chosen to be l_i^{ref}. The interaction in this case is harmonic, and thus it is best justified for small fluctuations. As a remark, note that l_i^{ref} is not the equilibrium distance because that depends on thermodynamic conditions as well as the composition of the system. Rather, l_i^{ref} should be regarded as the distance that two particles will adapt to in the absence of any other interactions in a system. The second term in Eq. (3) is the (harmonic) bending interaction for three consecutive particles (A–B–C) in the same chain-like piece of a molecule, with the reference value of the valence angle being θ_i^{ref}. The torsional interaction presented third is a four-body term, whereas a short-range Lennard–Jones 6–12 potential is often used to describe the steep repulsion caused by the Pauli-exclusion principle as well as van der Waals interactions resulting from dispersion forces. Finally, there is an electrostatic term for long-range Coulombic interactions for a pair of charged particles whose distance from each other is r_{ij}.

For clarity, let us emphasize here that, to define a force field, one must specify not only the set of potential energy functions in a force field but also the force-field parameters (as well as other practical details such as cutoff distances used in truncating long-range interactions) used in the calculations. If one of the parameters is changed, then the force field is also changed.

Although Eq. (3) provides us with a simple but typical example of a force field used for biomolecules, it allows us to pose a number of relevant questions. How are the force fields determined? Where do the parameters in the force fields come from? How can one validate a force field?

Using Eq. (3) as an example, each of the terms has a clear physical meaning, and the parameters associated with a given term can be determined either from theory or from experiments. The theoretical approach is essentially based on quantum mechanical calculations and is often used to, for example, determine the distribution of partial charges within a molecule. Unfortunately, however, this approach is of limited use because the fitting is not easy to perform and depends critically on the quality of the QM computations, such as the basis set used in calculations. Thus, it is common to calibrate parameters in force fields by fitting them to empirical data.

The parameters in the bond stretching and bending interactions, for example, can be determined by spectroscopic techniques, and the parameters in the van der Waals potential can be derived by fitting parameters to crystal structures. Anyhow, even if the parameter fitting process is done very carefully, there is no point in claiming that the obtained force field is perfect. The number of different combinations of parameters is huge, and slight weaknesses in some of them can be compensated for by adjustments in others. Therefore, it should not be surprising to realize that there are many different force fields that are commonly employed in the field of biomolecular modeling and simulation. For one reason or another, each of them seems to have pros and cons that limits their use. To our knowledge,

CHARMM (chemistry at Harvard macromolecular mechanics) force fields, for example, serve well for simulations of a number of DNA structures [74]. The force fields used in Gromacs [75] (see: http://www.gromacs.org), in turn, are likely a better choice for modeling of lipids, whereas their performance in DNA simulations could be better. In addition to these, many other force fields are in common use. AMBER, OPLS, NAMD, and MMFF are just some of the many force fields available (see Ref. [57]).

Hence, we think it is fair to say that all force fields have certain deficiencies. They are largely based on intuition and the validity of QM and experimental results used to fit the force-field parameters. Further, all force fields are based on some set of potential energy functions, whose validity can be questioned. The functional forms of the potential energy functions can be very complicated, as is the case in the MM4 force field, allowing, for example, accurate predictions of molecular structures. In contrast, one can employ simpler functional forms, which makes it more feasible to model large molecules such as proteins or DNA. Typical force fields in this category are AMBER and CHARMM. There is an endless need for improvements of force fields as comparisons among their results with respect to experimental observations reveal deviations. In the meantime, we can use current force fields rather safely. Even if their predictions are not completely correct quantitatively, they can provide us with plenty of qualitative insight into both structural and dynamic properties of complex biological systems.

Although the above discussion is definitely short, an extensive discussion of the topic can be found from a number of textbooks and recent review articles (see Refs. [57, 62]).

2.3.3. Dynamics in Classical Techniques

Although a force field is a crucial part of any molecular model, we need some means to find the equilibrium structure of the system. This can be done in various ways. One can apply energy minimization techniques or the MC simulation technique [60, 61], for example. These methods use the force field as an input and eventually provide one with structures corresponding to equilibrium conditions. However, as they cannot tell much of the dynamics in a system, we shall not consider them any further in this chapter. Instead, we discuss an approach that is the one most commonly used in (classical) atomic-level studies of biomolecular systems: the MD technique [60, 61].

MD is highly popular because of its simplicity and physical appeal. In MD, one essentially calculates forces acting on particles and then uses Newton's equations of motion to update their positions and velocities.

For simplicity's sake, let us consider a system of N particles in the microcanonical NVE ensemble (constant volume and constant energy without dissipation). Assuming that we know the force field of the system, we can calculate the forces acting on the particles. Let the force acting on particle i at time t be $\vec{F}_i(t)$. Then, in the spirit of Newton, we know that $\vec{F}_i(t) = m_i\vec{a}_i(t)$, where m_i is the particle's mass and $\vec{a}_i(t)$ is the acceleration resulting from the force. In the simplest possible level, one can now calculate the new position and the velocity of the particle at time $t + \delta t$ using a Taylor expansion about time t:

$$\vec{v}_i(t + \delta t) = \vec{v}_i(t) + \delta t\vec{a}(t) \tag{4}$$

$$\vec{r}_i(t + \delta t) = \vec{r}_i(t) + \delta t\vec{v}(t) + \frac{1}{2}\delta t^2\vec{a}(t) \tag{5}$$

In principle, that is it. Once we have written down the force field, we can construct a system of, say, DNA attached to a lipid membrane surrounded by water and consider its dynamics through MD simulations by repeatedly solving the Newton's equations of motion.

However, as expected, there are a number of issues that might lead to problems. In addition to the trouble of finding a reliable force field for the system in question, one is bothered by various practical problems that are an inherent part of MD simulations. First of all is the integration of equations of motion. Instead of using Eq. (4), which is far too simple for practical purposes, one should apply an integration scheme that is both accurate and efficient. In practice, we need an accurate scheme that also allows the use of a large time

step δt. Although there is a wide variety of different integration schemes available, so-called symplectic integrators have turned out to be appropriate in many cases [70].

Second is how to deal with long-range interactions in a small (periodic) system. This question is very relevant because it has been shown that certain schemes for long-range interactions can lead to major artifacts in MD simulations of soft matter systems (see discussion in Section 4.1). At present, it seems like the Ewald-type summation schemes [60, 76] are most appropriate for this purpose.

Third, as experiments are typically done under conditions in which the temperature and some other thermodynamic variables are fixed, we have to ask ourselves how we can establish such conditions in MD simulations. This question with regard to different ensembles for different physical conditions and as regards ways to handle this issue in practice is discussed thoroughly (e.g., in Refs. [60, 61]). Further, as many soft matter systems are characterized by a wide range of different timescales such as the motion of bond stretching compared to the relative motion of two particles interacting through a long-range Coulombic potential, one should ask how to efficiently deal with this situation. The multiple-time step schemes discussed in Refs. [60, 77] provide one solution to this issue. Another and even more important issue is related to living systems in general, as they are basically never in equilibrium. If they are, they are dead. Thus, we need reliable ways to model biological soft matter systems under nonequilibrium conditions. The field of nonequilibrium MD aims to address related questions, and thanks to major activities in this challenging field [78], we are today at a stage at which many of the interesting research problems have been clarified, at least in part.

Processes in biologically relevant soft matter systems often take place over very large length and timescales, which puts some pressure on computational resources. Simulations in a parallel environment are therefore a rule rather than an exception, which leads to a number of practical questions, including, at the moment, how to deal with long-range interactions and random number generation in a parallel environment. Some of these issues will be discussed in more detail in Section 4.

For the time being, let us stress that the above issues can be a burden, but they can be (and in many cases have been) solved. Thus, classical atomic-level MD has reached a level at which it can be considered as a mature technique and can be applied to numerous intriguing problems in a wide range of fields. MD simulations of DNA [79, 80], protein systems [80, 81], protein folding [82, 83], and lipid membranes [84–88] are examples of this activity. Further examples will be discussed in Section 5.

3. COARSE-GRAINING: MOVING BETWEEN DIFFERENT TIME- AND LENGTH SCALES

The previous section summarized recent developments in molecular-level modeling of soft matter and biological systems. As mentioned in Section 1, the Holy Grail of modeling is to develop rigorous methods for linking different time- and length scales, such that both coarse-graining and fine-graining are well defined. In the following text, we will briefly review some of the new methods and provide a quick summary of their merits and problems. Of the specific methods, we provide a more detailed look into the dissipative particle dynamics in Section 3.2 and field theoretical methods in Section 3.3, as they represent two different philosophies (i.e., particle-based vs. continuum approaches; see Fig. 1).

As discussed earlier, classical MD simulations of biologically motivated systems are typically limited to system sizes not much larger than 10^4 atoms and timescales spanning up to around 100 ns in the best cases. The restrictions may be even worse if electrostatic interactions must be properly accounted for. In addition, for systems such as polymer melts, the computationally accessible time- and length scales are simply far too short for the system to be able to reach equilibrium as dynamic processes during equilibration occur under hydrodynamic conditions.

In Section 1, we mentioned that polymer research has been one of the leading fields in multiscale modeling in both practical and theoretical aspects. This is indeed the case, and there are a lot of different approaches that have been used. On the analytical side the projection operator-based formalism of Akkermans and Briels [6] provides an interesting

approach to study the emergence of fluctuating forces in a more coarse-grained description. The projection operator formalism has also been used successfully (although in a very different manner) in the GENERIC approach [25–27], which provides an analytically rigorous method for coarse-graining. In addition to the projection operator formalism that has proven to be fruitful, another commonly used analytical approach relies on the Ornstein–Zernike equation [89] and the hypernetted chain closure [90, 91]. We will not discuss those approaches here any further, but refer the reader to the above references and the references therein.

Inverting the radial distribution functions $g(r)$ to obtain pair potentials is another possible starting point. The theoretical basis for that was given by Henderson [92], who stipulated that under fixed conditions two pair potentials that give rise to the same $g(r)$ cannot differ by more than a constant. This constant is determined by the condition for the pair potential that

$$V(r \to \infty) \to 0, \tag{6}$$

where r is the interparticle distance. In this respect, this approach is analogous to the Hohenberg–Kohn theorem [93], which states that all ground-state properties are determined by the electron density. This theorem lies at the heart of modern DFT. It is important to notice that the radial distribution function obtained from a simulation includes effects from the many-body interactions. Furthermore, this way it is possible to define new interaction sites and to compute the radial distribution function between them, and thus readily obtain new coarse-grained models at different levels of description.

There are new simplex-algorithm based optimization procedures developed by Müller-Plathe et al. [94–96] that exploit the above described relation. For an overview of the current coarse-graining approaches in the field of polymer science, see, for example, Ref. [29, 32] and references therein.

As a conceptually simple approach, the DPD [18–20] have recently gained lot of popularity in soft matter simulations. We will discuss that approach in detail in the following text. Other coarse-grained approaches that take hydrodynamics properly into account include the lattice Boltzmann method (see, e.g., the article of Pagonabarraga in Ref. [28]) and a novel and very promising approach developed very recently by Malevanets and Kapral (MK) [12, 13]. The MK method couples a molecular-level description with a mesoscale treatment of solvent-conserving hydrodynamics. This approach is particularly appealing for studying dilute systems with hydrodynamics. Concurrently, it has been applied to a number of problems, including studies of dilute polymer systems [97], individual colloids under flow [13], diffusion of colloids [98], and crystalline clusters [99]. Malevanets and Yeomans have further developed a variant of the MK method and applied that to study structural and dynamical properties of individual polymer chains in a hydrodynamic medium [100, 101]. Recent developments of the technique are discussed in Refs. [102–107].

Before moving on to describe a few of the methods in detail, we would like to mention, as a detailed discussion is out of scope of this review, that there are also other types of developments. The coupling of quantum and classical systems is a new and exciting field and can open entirely new avenues in biomolecular simulations. We refer the reader to the article by Kalibaeva and Ciccotti in Ref. [28] for an in-depth discussion. Algorithmic developments are an important issue, and we refer the reader to one of the new textbooks in the field (e.g., the book by Leach [62]).

3.1. Effective Interactions: Inverse Monte Carlo

Conservative forces can consist of various kinds of forces [108] resulting from, for example, electrostatic interactions, as well as descriptions of detailed intermolecular interactions such as van der Waals forces. As suggested earlier, it is possible to obtain conservative forces by coarse-graining the pairwise interactions from an MD simulation using, for example, the Inverse MC (IMC) procedure [109] or the other procedures [31, 32] that are similar in spirit.

The IMC method bears a close resemblance to the Renormalization Group MC method (see, e.g., the book by Goldenfeld [110]), in which one iteratively solves the renormalization

constants. In IMC, one inverts the radial distribution functions, calculated in atomistic MD simulations, to obtain the effective potentials for a coarse-grained model with a fewer number of degrees of freedom. The essential features of the IMC procedure were introduced by Lyubartsev and Laaksonen in 1995; Ref. [109] contains the relevant computational details.

There are two aspects that we wish to point out. First, the effective potential includes corrections from many-body interactions to the well-defined potential of mean force (PMF) [89], which is defined as

$$V^{\mathrm{pmf}}(r) \equiv -k_{\mathrm{B}}T \ln g(r) \tag{7}$$

where r is the interparticle distance and $g(r)$ the pair correlation function. Therefore, the effective potential considered here is not the same as the PMF as defined by Eq. (7). Inclusion of the many-body corrections is the reason why the iterative IMC scheme is needed. Second, this procedure guarantees self-consistency; that is, the effective potentials, when used in mesoscale (say, DPD simulations), lead to the same pair correlation behavior as the underlying MD simulations. Thus, one can expect that certain response functions such as compressibility are properly described by the effective potentials. This method can be applied to studies of soft matter systems with full hydrodynamics.

3.1.1. The IMC: An Example of Inversion Methods

Next we will sketch the outline of the IMC procedure. As mentioned, there are other methods for using structural information to obtain interaction potentials. We have chosen to use the IMC because it will be used to discuss a specific example later in Section 6.

Let us assume a simple system consisting of identical particles with pairwise interactions only. Then we can write the Hamiltonian as

$$H = \sum_{i<j} V(r_{ij}) \tag{8}$$

where $V(r_{ij})$ is the pair potential and r_{ij} is the distance between the interaction sites i and j. Let us further assume that we do not know $V(r_{ij})$ but we know the radial distribution function $g(r_{ij})$, for example, from another simulation or experiments.

To obtain the interaction potential $V(r_{ij})$, we discretize the Hamiltonian such that the interval $[0, r_{\mathrm{cutoff}}]$ is divided into N grid points. We denote this discretized potential by V_α such that $\alpha = 1, \ldots, N$. With this discretization, we can approximate the Hamiltonian in Eq. (8) by

$$H_{\mathrm{discr}} = \sum_{\alpha} V_\alpha S_\alpha \tag{9}$$

Here, we have introduced S_α as the number of pairs that have their interparticle distance within the interval α. Recalling the definition of the radial distribution function immediately tells us that S_α is related to it through

$$\langle S_\alpha \rangle \sim 4\pi r^2 g(r_\alpha) N^2 \tag{10}$$

Because S_α is related to the radial distribution function, and thus also the discretized potential V_α, it is possible to write an expansion for S_α in terms of the discretized potential,

$$\Delta\langle S_\alpha \rangle = \sum_{\gamma} \frac{\partial \langle S_\alpha \rangle}{\partial V_\gamma} \Delta V_\gamma + \mathcal{O}(\Delta V^2) \tag{11}$$

Lyubartsev and Laaksonen [109] showed that the partial derivatives can be written in a compact and computationally tractable way, thus allowing one to use the above expansion for iteratively solving the discretized potential V_α.

To solve for V_α in practice, one has to perform regular MC simulations of the system to obtain $\langle S_\alpha \rangle$s. Then, solving V_α iteratively using Eq. (11), one obtains the interaction potential. The procedure resembles solving a multidimensional nonlinear equation using the Newton–Raphson method. The most natural starting point for the procedure is to use the potential

of mean force as defined above. It is also worth noticing that this procedure provides us with estimates of the uncertainty in $V(r)$. That can be formally done by eigenvalue analysis.

In Section 6, together with the DPD method, we will apply the IMC procedure to an aqueous sodium chloride solution.

3.2. Dissipative Particle Dynamics (DPD)

DPD was originally developed by Hoogerbrugge and Koelman to solve hydrodynamic flow problems in porous media. Since its introduction in 1992 [19, 20], DPD has become the most used coarse-grained simulation method in soft matter research, and several extensions and generalizations of DPD have been developed [21, 111–114]. Instead of porous media, DPD has gained most of its merits in simulations of polymers and surfactant systems [115–117] and, very recently, also in simulations of biological systems [10, 118–122]. To mention some examples, DPD has been applied to problems ranging from liquid-gas phase diagrams [123], to rupture of bacterial membranes [10], to self-assembly in Huntington's disease [122], and to packing of fillers in composites [124].

DPD is a coarse-grained method, and this is reflected in its interaction potentials, which are "soft" in contrast to the Lennard–Jones-type potentials used in classical MD. This is also where the strength and weakness of the standard (and the most used) version of DPD lies. By "softness" in potentials, we mean that the DPD potential has a finite value at zero particle separation (i.e., the Fermi exclusion principle is not accounted for). The standard DPD potential is shown in Fig. 3.

As can be seen from Fig. 3, the only parameters concerning the potential are the cutoff radius r_c and the amplitude of the potential. That this indeed is a reasonable approximation was shown, although in a slightly different context, by Forrest and Suter [9], who studied polymeric systems and explicitly averaged over fluctuations to obtain effective interaction potentials between monomers. They demonstrated that when averaged over sufficiently long timescales, the averaged, or coarse-grained, potentials approach a quadratic form: this is the form used virtually exclusively in DPD simulations. The underlying microscopic potentials were of the usual Lennard–Jones type. We will present the functional form of the DPD potential in the following subsection.

Despite its success, DPD does not come without problems and open questions. We will discuss them in the following sections. We also briefly review the latest developments concerning the inclusion of electrostatic interactions in DPD as well as novel and appealing coarse-graining procedures, providing a more solid basis for DPD. The latter are an example of true systematic coarse-graining linking MD to continuum representation.

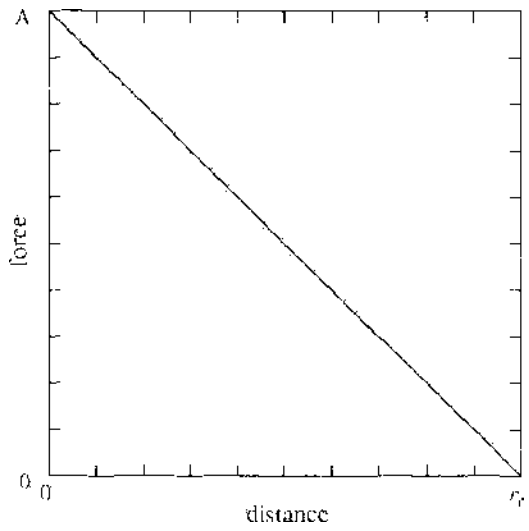


Figure 3. The conservative force in the "standard" DPD is linear with the cutoff distance r_c and the amplitudes of the pairwise interactions being the only parameters.

3.2.1. DPD Equations of Motion

Here we summarize the DPD equations of motion and discuss some aspects related to it. After that, we put the method in a more general context. For algorithmic details, a comprehensive discussion of the performance, implementation, and quality of various DPD integrators, we refer the reader to Refs. [125–127] and the references therein. For a more application-oriented and detailed review by one of the pioneers of the field, please see the article by Groot in Ref. [28].

One of the key ideas and motivations behind DPD is that it conserves hydrodynamics (i.e., all the interactions are pairwise conservative). In DPD the total pair force consists of three components; dissipative, random, and conservative. Thus, the force exerted on particle i by particle j can be written simply as

$$\vec{F}_{ij} = \vec{F}_{ij}^{D} + \vec{F}_{ij}^{R} + \vec{F}_{ij}^{C} \tag{12}$$

We start the discussion from the dissipative and random forces. In the standard notation, they are given as

$$\begin{aligned} \vec{F}_{ij}^{D} &= -\gamma\omega^{D}(r_{ij})(\vec{v}_{ij}\cdot\vec{e}_{ij})\vec{e}_{ij} \quad \text{and} \\ \vec{F}_{ij}^{R} &= \sigma\omega^{R}(r_{ij})\xi_{ij}\vec{e}_{ij}, \end{aligned} \tag{13}$$

where $\vec{r}_{ij} \equiv \vec{r}_i - \vec{r}_j$, $r_{ij} \equiv |\vec{r}_{ij}|$, $\vec{e}_{ij} \equiv \vec{r}_{ij}/r_{ij}$, and $\vec{v}_{ij} \equiv \vec{v}_i - \vec{v}_j$ for particles with positions $\vec{r}_i$ and velocities $\vec{v}_i$. The ξ_{ij} are symmetric Gaussian random variables with zero mean and unit variance and are independent for different pairs of particles and different times. The pairwise conserved fluctuations can be easily justified, as two particles being close enough to each other experience the same fluctuations.

Equations (13) contain two amplitudes γ and σ and two weight functions $\omega^{D}(r_{ij})$ and $\omega^{R}(r_{ij})$. The amplitudes define the strengths, and the weight functions the spatial extent, of the dissipative and random forces. The specific functional form of the weight functions is not specified, and any physically motivated choice can be used. In contrast to this freedom concerning the functional form, the two weight functions, as well as the amplitudes, are coupled via a fluctuation–dissipation relation.

The coupling of the dissipative and random forces $\vec{F}_{ij}^{D}$ and $\vec{F}_{ij}^{R}$ is the result of the fact that the thermal heat generated by the random force must be balanced locally by dissipation. The precise relationship between these two forces is determined by the fluctuation–dissipation theorem [18], which sets the following two conditions for both the weight functions and the amplitudes of the forces in Eqs. (13),

$$\omega^{D}(r_{ij}) = [\omega^{R}(r_{ij})]^2 \quad \text{and} \quad \sigma^2 = 2\gamma k_{\mathrm{B}}T^{*}, \tag{14}$$

where T^{*} is the canonical temperature of the system.

The most common choice for the weight functions ω^{D} and ω^{R} is the soft-repulsive form

$$\omega^{R}(r_{ij}) = \begin{cases} 1 - r_{ij}/r_c & \text{for } r_{ij} < r_c; \\ 0 & \text{for } r_{ij} > r_c, \end{cases} \tag{15}$$

where r_c is the cutoff distance and $\omega^{D}(r_{ij})$ is given by the fluctuation–dissipation relation above. It is somewhat surprising that, at least to the authors' knowledge, no other forms for the weight function have been used in published literature.

Thus far we have not said anything quantitative about the conservative force. To be precise, the DPD formulation does not specify its functional form. In principle, one is free to use any physically motivated choice, van der Waals forces, electrostatic interactions, or some coarse-grained interactions. The term DPD often refers to above formulation for random and dissipative forces, combined with the soft repulsive potential given as

$$\vec{F}_{ij}^{C} = F_{ij}^{\mathrm{soft}}(r_{ij})\vec{e}_{ij}, \tag{16}$$

where $\boldsymbol{F}_{ij}^{(C)}(r_{ij}) = A\omega(r_{ij})$. The parameter A is the amplitude of the force, and it can be defined via compressibilities or using Flory theory for polymers. For a detailed discussion, see the article by Groot in Ref. [28]. The force and the potential are plotted in Fig. 3.

What is usually referred to DPD is defined by Eqs. (13), (15), and (16) together with fluctuation–dissipation in Eq. (14). This definition is rather restrictive, as the dissipative and random forces are independent of the conservative force. This fact has been recognized by many authors, and it is appealing to think of DPD as a momentum-conserving thermostat [11, 111, 126, 128], as we will discuss in the next section.

The above formulation is the universally accepted and exclusively used form of DPD derived by Español and Warren [18] using a Fokker–Planck equation. It is also worth pointing out that the original DPD formulation by Hoogerbrugge and Koelman does not obey the fluctuation–dissipation theorem and hence does not relax to the canonical distribution, as required by thermodynamical equilibrium.

3.2.2. DPD as a Thermostat

As already mentioned in the previous section, DPD can be defined as a momentum-conserving thermostat for MD simulations. Earlier we defined DPD as a momentum-conserving coarse-grained MD using conservative potentials as given in Eq. (16). That is indeed the standard phenomenology, but it would be more appropriate to use the term DPD for a momentum thermostat defined by Eqs. (13) and (14) and the weight functions, for example, by Eq. (13).

These aspects have been recently studied in detail by Vattulainen et al. [126, 127] for general Lennard–Jones systems, and by Soddemann et al. [128], who showed the usefulness of the DPD thermostat also in nonequilibrium MD simulations.

Another problematic aspect with DPD has been the treatment of electrostatics. As it is easy to understand from the softness of the potentials, electrostatics needs special attention. This issue has been resolved by Groot [129, 130]. The idea is that the charges are smeared out and then the electrostatic field is solved in a grid. From the computational point of view, this approach is close to the multigrid method introduced by Sagui and Darden [131, 132]. Groot was also able to show that his approach is consistent with the Stillinger–Lovett sum rules, and thus it provides a solid basis for treating the electrostatics properly. It is likely that this will increase the usefulness of DPD in biomolecular simulations.

3.2.3. An Alternative: The Lowe–Andersen Thermostat

A different formulation for DPD was given by Lowe [133]. His approach is based on the thermostat developed in 1980 by Andersen [134]. The Andersen thermostat is different from other MD thermostats, as in this approach the velocities are periodically drawn from a Maxwell distribution. This formulation is computationally quite efficient, although it does not conserve momentum. Lowe [133] used the Andersen thermostat as a starting point and devised a momentum-conserving thermostat out of it. The idea here is that one chooses a parameter that describes the frequency (Γ in Lowe's original notation) for a particle to interact with the heat bath. Then, $\Gamma\delta t$, where δt is the time step, defines a probability for thermalization. This thermalization is done in a pairwise fashion, such that the method is momentum conserving, and it also produces the correct canonical distribution. One of the main advantages of this method is that it enhances viscosity and allows proper tuning of the Schmidt number, which is the ratio between kinematic viscosity and the diffusion coefficient.

Lowe's approach is appealing: There are no dissipative forces, and thus Lowe's approach is easier to use and performs well. The thermalization rate may be varied over a wide range, which implies that the dynamical properties of the system can be tuned in a controlled fashion. It is rather surprising, though, that in addition to Lowe's original article, there is only one other published work using this method [126].

3.2.4. A DPD Example: Liposome Formation

To give a brief illustration of a DPD simulation using the soft potentials as defined in Eq. (16), we have studied vesicle formation and dynamics [135]. The lipid molecules consist

of eight DPD particles, of which two belong to the headgroup, and the rest are divided equally between the two tails. The system illustrated in Fig. 4 has a total of 500 lipids and 188 000 waters (i.e, 228 000 DPD particles in total).

The amplitudes of the conservative force [i.e., the parameters a in Eq. (16)] between different species are shown in Table 1. In addition, the description of lipids included a harmonic freely jointed spring.

The simulation started with the formation of a stable bilayer. After that, the simulation was halted, allowing us to add a major amount of excess water around the bilayer. On this happening, the system was simulated under NVT conditions such the bilayer had to deform to a vesicular shape. This process and the formation of a spherical bilayer structure is clearly visible in Fig. 4. We are currently studying the dynamics of related vesicle systems.

3.2.5. Analytical Approaches: Systematic Derivation of DPD

The standard DPD as presented above is phenomenological. A very interesting formal approach has recently been presented by Flekkøy et al. [8, 21], who were able to formally link DPD to molecular level properties. They use a Voronoi tessellation–based technique to establish a systematic coarse-graining procedure that couples molecular dynamics to DPD and provides a systematic connection between continuum and particle representations. The great advantage that this method has is that it can be used to resolve different length scales simultaneously. The method is formally akin to the well-known renormalization group procedure extensively used in analytical treatment of critical phenomena. As such, this coarse-graining procedure provides a true systematic and tunable method. For details about the renormalization group procedure, see, for example, the book by Goldenfeld [110]. The mathematical details can be found from the original articles as cited above.

The approach of Flekkøy et al. provides a systematic and computational tractable, although not straightforward, method to coarse-grain molecular systems. From a computational point of view, their method is more demanding than particle-based methods such as the standard DPD, as the Voronoi tessellation procedure requires $N \log N$ operations and it comes with a fairly large additional prefactor.

Although not connected directly to DPD, we mention here another general method called GENERIC, introduced by Grmela and Öttinger [25–27]. The idea is that there is a general form for the time-evolution of nonequilibrium systems and that it can be written as

$$\frac{dx}{dt} = L(x)\frac{\delta E(x)}{\delta x} + M(x)\frac{\delta S(x)}{\delta x}$$

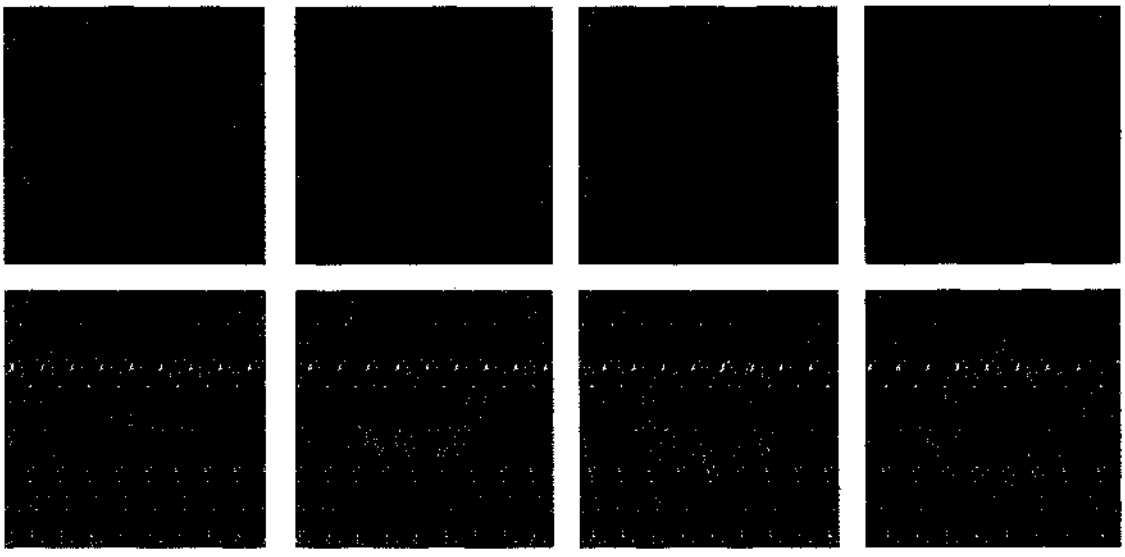


Figure 4. Formation of a vesicle in a DPD simulation [135]. The surrounding 188 000 water molecules have been removed for clarity. The lower panels show the corresponding cut through the middle of the vesicle during the formation process.

Table 1. Amplitudes of conservative interactions used in the present dissipative particle dynamics simulations for liposome formation.

	Head	Tail	Water
Head	100	160	100
Tail	160	100	160
Water	100	160	100

where x characterizes the state of the system, $L(x)$ is an antisymmetric matrix, and $M(x)$ is a symmetric and positive definite matrix. They are connected to the second and the first law of thermodynamics, respectively. The terms $E(x)$ and $S(x)$ are functionals for the total energy and entropy, respectively.

The equation above is the so-called GENERIC form, which uses the Poisson bracket and projection operator formalism to connect different levels of description. The basic ingredients of the theory are the energy and entropy functionals, which are the generators for reversible (energy) and irreversible (entropy) dynamics and the corresponding matrices.

The mathematical details are somewhat lengthy, and we refer the reader to the original articles [25–27] and the review by Español in Ref. [28], which discusses analytical approaches to coarse-graining at length.

3.3. Field Theoretical Methods

In contrast to the particle-based coarse-grained methods, the central idea of phase-field modeling is to provide a continuum description of a system using general symmetry principles, conservation laws, order parameters, and a free-energy functional. The name *phase-field* was coined by G. J. Fix [136] in 1982, but the approach itself is older [137]. This approach has several attractive features. First, it operates in meso- and macroscales, thus enabling studies of time- and length scales that would otherwise be difficult to reach. Second, it is suitable for describing out-of-equilibrium systems and interfaces between different phases that appear naturally in this approach. Third, it is possible to use projection operator techniques to study these systems using analytical theory.

Phase-field models have been used extensively in connection with problems involving nucleation and growth and spinodal decomposition (see, e.g., Ref. [138, 139]). In materials science, it has been used to study the properties of such diverse systems as directional solidification [140], morphological instabilities [141], eutectics [142], and charge-density waves [143], to mention a few examples. Surprisingly, the phase-field approach has received very little attention in soft matter problems, although it has great potential addressing questions related to macroscopic behavior. For details of phase-field models in a general context, we refer the reader to the recent article by Ala-Nissila et al. [22]. In the following text, we will describe the physical basis of this approach.

3.3.1. Landau Theory of Continuous Phase Transitions

The phase-field approach is based on Landau's theory of phase transition. Thus, the system is defined by an order parameter reflecting local order in the system. It can be, for example, a density difference between different liquids. From that it is clear that the order parameter is by definition a coarse-grained quantity. As such it does not include any information about the molecular details of the system. For the Ising model, one can perform a formal mapping to obtain the order parameter, which in that model is the magnetization of a block.

The Landau theory of continuous phase transitions is phenomenological and stresses the importance of overall general symmetry properties and analyticity over microscopic details in determining the macroscopic properties of a system; an excellent discussion is given in Ref. [110]. The validity of this approach is based on the following assumptions.

First, it is possible to define an order parameter Ψ, that characterizes the order in the system in the following way: $\Psi = 0$ in the disordered state (above the critical temperature T_c), and Ψ is small and finite in the ordered state (below T_c).

Second, it is possible to describe the system with a free energy. Close to T_c, it can be expanded in powers of the order parameter; that is,

$$F(\Psi) = \sum_{n=0}^{\infty} a_{2n}\Psi^{2n} \tag{17}$$

where F is the free energy and the expansion coefficients are phenomenological parameters that depend on temperature and the microscopic properties of the system under consideration. Therefore, at least in principle, the a_{2n} can be derived from first principles. The smallness of the order parameter guarantees that the expansion converges and can be truncated at a finite power.

Third, the expansion, Eq. (17), of the free energy must be consistent with the high temperature symmetry properties of the system under consideration. Because the Landau theory describes continuous transitions, no odd powers are allowed in the expansion of the free energy. When the system is cooled down to $T < T_c$, there is a spontaneous symmetry breakdown. This means that the ground state of the system has lower symmetry than the free energy shows: States $+\Psi_0$ and $-\Psi_0$ are equivalent, but the system must select one of them (see Fig. 5). Mathematically speaking, the Hamiltonian must commute with the symmetry group of the high-temperature phase.

Fourth, the free energy must be analytic. In addition to analyticity in Ψ, it is also required that the expansion coefficients be regular functions of the temperature. Because the second-order term is dominant, $a_2(T)$ must vary smoothly from $a_2(T) < 0$ for $T < T_c$, to $a_2(T) > 0$ when $T > T_c$, with $a_2(T = T_c) = 0$. This implies that $a_2(T) \sim (T - T_c)$. To see this, we first minimize the free energy with respect to Ψ,

$$\frac{\partial F}{\partial \Psi} = 2a_2\Psi + 4a_4\Psi^3 = 0 \tag{18}$$

The solutions are

$$\Psi_0 = 0 \qquad \text{and} \qquad \Psi_0 = \sqrt{-\frac{a_2}{2a_4}} \tag{19}$$

The first is the high-temperature solution, and the latter provides the minima for $T < T_c$. From Eq. (18), we can see that a_4 must be greater than zero, as otherwise $\Psi \to \infty$ would minimize the free energy, rendering the above construction useless. Next, we Taylor expand the expansion coefficients around T_c:

$$a_2(T) \approx a_2(T_c) + (T - T_c)\frac{\partial a_2(T)}{\partial T}\bigg|_{T=T_c} + \frac{1}{2!}(T - T_c)^2\frac{\partial^2 a_2(T)}{\partial T^2}\bigg|_{T=T_c} + \cdots \tag{20}$$

and

$$a_4(T) \approx a_4(T_c) + (T - T_c)\frac{\partial a_4(T)}{\partial T}\bigg|_{T=T_c} + \frac{1}{2!}(T - T_c)^2\frac{\partial^2 a_4(T)}{\partial T^2}\bigg|_{T=T_c} + \cdots \tag{21}$$

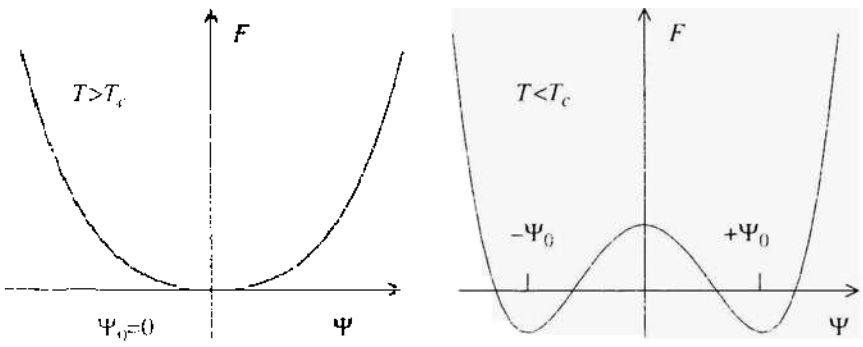


Figure 5. Schematic picture of the Landau free energy. Upon cooling below T_c the system undergoes a spontaneous symmetry breaking.

As argued above, $a_4(T) > 0$, and therefore we can approximate $a_4 = a_4(T_c) = \text{const.}$ This requirement is enough to guarantee the finiteness of the order parameter. Then, $a_2(T)$ must change its sign on cooling below T_c. Then,

$$\left.\frac{\partial a_2}{\partial T}\right|_{T=T_c} = \text{const.} > 0 \tag{22}$$

and $a_2(T) \sim (T - T_c)$.

The Landau theory is a mean-field theory because it does not take into account spatial inhomogeneities or thermal fluctuations. To account for inhomogeneities, we have to let the order parameter to become space dependent [i.e., $\Psi \equiv \Psi(\vec{x})$]. Conceptually, we should think of Ψ as a coarse-grained order parameter; that is, Ψ is defined only over a certain length scale. We must define a short wavelength (ultraviolet) cutoff for Ψ in such a way that Ψ varies smoothly in space; the order parameter cannot fluctuate on smaller length scales than the cutoff. In practice, the cutoff may often be thought of as the lattice spacing.

The response of the system to spatial inhomogeneities can be described as internal rigidity; Anderson [144] provides an in-depth discussion about the emergence and importance of rigidity in physical systems. Physically, it is clear that large differences between neighboring points are unfavorable. In other words, rigidity is a generic property of the system, and therefore there must be an energy cost associated with spatial inhomogeneities. Because the order parameter is a smooth and slowly varying function, we can take the rigidity of the system into account by making a gradient expansion and retaining only the lowest-order term compatible with the symmetry properties. The validity of the truncation of the gradient expansion depends on the smoothness and slowness of the order parameter variations.

Because the order parameter is now a local variable, the free energy in Eq. (17) becomes a functional of Ψ,

$$\mathcal{F}[\Psi] = \int d^d\vec{x}\left[F(\Psi) + \frac{1}{2}K(\nabla\Psi)^2\right] \tag{23}$$

where K is a phenomenological parameter describing the rigidity of the system and $F(\Psi)$ is the free energy defined in Eq. (17). Equation (23) is also known as the Ginzburg–Landau–Wilson free energy. Clearly, in Eq. (23), K must be positive for the free energy to be bounded from below.

3.3.2. Dynamical Models

The above description does not include time dependence. How time dependence can be added is to use linear response theory and assume relaxational dynamics (i.e., the system is dissipative and it is driven to equilibrium; see Fig. 5). The resulting models are often described as models A, B, C, and so on, depending on their symmetry properties, the dimensionality of the order parameter, and the conservation laws [137].

Being the simplest, model A serves as an example. It is a nonconserved system, and the equation of motion can be given as

$$\frac{\partial\Psi(\vec{x},t)}{\partial t} = -\Gamma\frac{\delta\mathcal{F}[\Psi(\vec{x},t)]}{\delta\Psi(\vec{x},t)} + \eta(\vec{x},t) \tag{24}$$

where we have included thermal fluctuations as Gaussian random noise, with the first and second moments defined as

$$\langle\eta(\vec{x},t)\rangle = 0 \tag{25}$$

and

$$\langle\eta(\vec{x},t)\,\eta(\vec{x}',t')\rangle = 2\Gamma k_B T\delta(\vec{x}-\vec{x}')\delta(t-t') \tag{26}$$

The angular brackets denote an average, and Γ is a kinetic coefficient describing the relaxation rate. An example of the evolution of a system described by model A is given in Fig. 6.

Figure 6. Phase separation in model A described by Eq. (24). The panels from left to right show the evolution of the system from random initial conditions towards equilibrium. Periodic boundary conditions are clearly visible.

Other models can be obtained by including conservation laws; for example, the dynamical equation of motion for a conserved order parameter is given by model B,

$$\frac{\partial \Psi(\vec{x}, t)}{\partial t} = \Gamma \nabla \left[\frac{\delta \mathcal{F}[\Psi(\vec{x}, t)]}{\delta \Psi(\vec{x}, t)} + \eta(\vec{x}, t) \right] \tag{27}$$

This can be derived using the continuity equation.

The system may have several order parameters that are coupled, thus leading to more complicated free energies and equations of motion; see Ala-Nissila et al. [22] and references therein.

It is rather surprising that this type of approach has, thus far, had only limited attention in soft matter and biological modeling. There are a few notable exceptions, though. Shore et al. [145, 146] coupled a phase-field model to the Navier–Stokes equation, including viscoelasticity to melt fracture in polymer extrusion experiments. The details of that model are beyond this review, but it suffices to note that the study of Shore et al. is a good example of the power of the phase-field approach, as it provides physical insight into (in this case, instabilities) both fundamentally and industrially important problems.

Other examples of the use of field theoretical models in soft matter include the OCTA software package (http://octa.jp), which in one of its parts uses self-consistent field theory from the Nagoya group [39, 40]. Self-consistent field theory has been used in the theoretical description of polymers for a long time. A very recent example of the developments in that field is the elegant study of tetrablock copolymers by Drolet and Fredrickson [147].

Another major effort to build modeling software around field-theoretical ideas is the MESODYN project of the Leiden group [24, 148]. They use the Ginzburg–Landau approach and DFT to build a systematic and computationally tractable system for polymer melts. A detailed description is again beyond this review, but with their approach it is possible to study phase separation even in three dimensional systems. This approach is closely related to model B of critical dynamics [137], as discussed above.

4. METHODOLOGICAL ISSUES IN MOLECULAR SIMULATIONS

One of the intriguing issues in science is indeed that there is the artistic side, too: The art of doing science. In the field of molecular simulations and computational sciences, this is largely related to the inventions and new ideas of seeing things done in an accurate but efficient manner. This brings us to the methodological side of doing molecular simulations. In this section, we consider a few aspects related to both atomistic and coarse-grained descriptions of molecular systems that illustrate the importance of developing the methodology.

We discuss three highly important methodological issues whose role for the reliability and accuracy of molecular simulations is particularly significant. First, especially for MD simulations of biologically relevant soft matter systems on atomic level, the treatment of long-range electrostatic interactions is a major issue. If this is not handled with care, the interpretation of simulation results may be very problematic because of the underlying artifacts caused by the mistreatment of electrostatics. Second, as all stochastic simulation methods are based on the use of noise produced by so-called pseudorandom number generators, the quality of pseudorandom number sequences is of prominent importance in all cases in which they are employed to generate the dynamics for the systems under study. Because pseudorandom

number generators produce deterministic rather than unpredictable sequences of numbers, the "randomness" of pseudorandom numbers is a very subtle issue and should never be taken for granted. Finally, we discuss an old issue that had been thought to be overcome several years ago: the artifacts resulting from integration schemes that yield the dynamics for systems governed by Newton's equations of motion. In contrast to classical MD, where this issue is well under control, in stochastic simulation techniques such as dissipative particle dynamics, the case is more subtle. Here we discuss most recent developments in this field and show how the problems can be overcome in an efficient and reliable fashion.

4.1. How to Handle Long-Range Electrostatic Interactions

Electrostatic interactions are an inherent part of all biologically relevant soft matter systems. For example, the behavior of water is to large extent driven by hydrogen bonds, in which electrostatics plays an important role. In the same spirit, the stability of proteins and DNA is largely based on hydrogen bonding. DNA, in particular, is a prominent example of the importance of electrostatics because it is probably the most charged molecule found in living systems. Other examples including the self-assembly and overcharging of DNA-lipid complexes [79, 80, 149–153] are therefore obvious and lead naturally to a wide class of biologically important systems composed of lipid molecules. The most important entities in this context are cell membranes that surround the cell. In addition to protecting it, cell membranes provide a heterogeneous, semipermeable, and fluidlike environment for numerous processes [154–158] facilitated by ion channels and other proteins embedded in or attached to membranes, thereby, for example, governing the transport of molecules and ions across a lipid membrane.

The importance of electrostatics in membrane systems can be demonstrated by a practically relevant example. Of the many techniques available for the introduction of genetic material into cells, many are based on nonviral delivery vectors composed of cationic liposomes. Understanding the properties of complexes composed of cationic membranes and DNA is therefore largely based on the understanding of the electrostatics associated with these systems. From the point of view of molecular simulations, it is hence crucial to treat long-range electrostatic interactions as accurately as possible.

Techniques available for this purpose are wide and include, for example, Ewald summation techniques [76] based on solving the Poisson equation for the electrostatic potential in a periodic system in a fashion in which all charged particles and their periodic images are accounted for systematically. The Particle–Mesh–Ewald (PME) technique [60, 159], specifically, has found increasing popularity in MD studies of soft matter systems. Other commonly used techniques include the fast multipole method [60, 160], the reaction-field technique, and truncation methods. The truncation method is particularly appealing because it leads to considerable reduction in the computational load. As a consequence, it is frequently used when computational requirements are substantial because of large system sizes or long timescales, which may be the case, for example, in studies of lipid–protein systems [161], self-assembly of lipids [162], and membrane fusion [163]. There is reason to emphasize, however, that the accuracy of electrostatics should not be sacrificed at the expense of speed.

The key point here is that the results based on MD simulations of biophysical systems depend on the scheme chosen for electrostatics. It has been shown that the truncation of electrostatic interactions may affect structural properties of systems like water [164–168], peptides [169, 170], proteins [171], and DNA [172, 173]. Here we demonstrate and discuss this issue in the context of lipid bilayers.

4.1.1. Model of a Dipalmitoylphosphatidylcholine Bilayer

We discuss this issue in a one-component lipid bilayer consisting of $N = 128$ dipalmitoylphosphatidylcholine (DPPC) molecules at 323 K (see Fig. 7); which is above the so-called main phase transition temperature T_m. The temperature T_m is characteristic to any lipid species and depends on various molecular details such as the headgroup and the length and saturation of lipid acyl chains. The main transition temperature characterizes the transition for the pure one-component lipid system between the high-temperature liquid-crystalline

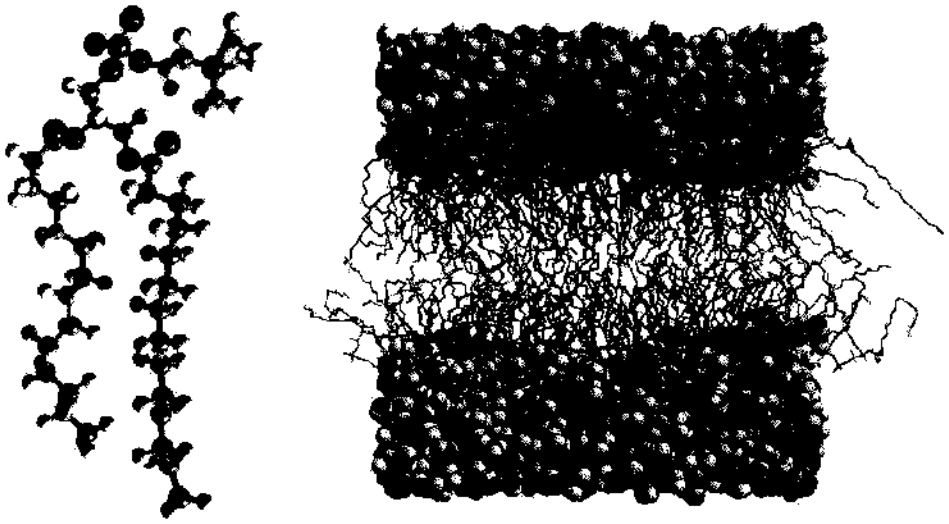


Figure 7. Representation of a saturated phospholipid molecule (left) and a lipid bilayer made of 128 dipalmitoylphosphatidylcholine and 3655 water molecules (right).

(also called liquid-disordered) phase and the low-temperature solid-ordered phase. The two phases are distinctly different in nature, as the high-temperature phase above T_m is characterized by the absence of any translational order in the plane of the membrane, as well as the absence of conformational ordering of the acyl chains. In contrast, the low-temperature phase below T_m is solidlike because of the translational order in the bilayer plane, in addition to which the acyl chains are highly ordered. In this work, in which we consider the DPPC bilayer at 323 K above $T_m = 315$ K, we are therefore in the high-temperature liquid–crystalline phase.

The bilayer is fully hydrated by 3655 water molecules, for which we used the single-point charge (SPC) model [174]. The choice of initial configuration and force fields is discussed in Ref. [175].

For electrostatic interactions, we discuss here the two most commonly used techniques. First, we consider group-based long-range electrostatic interactions handled by using an abrupt cutoff at the truncation distance r_{cut}. For r_{cut}, commonly used values in lipid membrane simulations range from 1.2 to 2.0 nm. Here we consider a number of values in the same range, with an objective to find possible trends in the limit where the truncation distance is as large as possible compared to the linear system size. To this end, we have chosen $r_{\text{cut}} = 1.8$, 2.0, and 2.5 nm.

In addition to truncation, we discuss the PME [60, 159] technique that takes the long-range interaction fully into account through Ewald summation rules. Although it has been shown that PME suffers from certain drawbacks as a result of the periodic nature of the simulation box, the PME method has been found to work well in various soft matter systems. Thus, it is used as a reference in this work.

It is worthwhile to emphasize that the present choice of interactions, time steps, thermostats and barostats, and the choice to use the twin-range scheme for long-range interactions, follows the practice made in numerous MD simulations of lipid bilayers. In particular, this approach follows the choice made for a DPPC bilayer in the pioneering works by Berger et al. [176] and Tieleman and Berendsen [177].

4.1.2. Area Per Molecule in a Lipid Bilayer

In lipid membrane systems, the area per molecule is perhaps the most central quantity. From the experimental point of view, it is one of the few quantities that can be determined rather accurately. In addition, the area per molecule is closely related to the ordering of lipid acyl chains in the hydrophobic part of the membrane, and area fluctuations in turn are related to a number of processes such as permeation through membranes, hole formation, and lateral diffusion of lipids and other molecules in the plane of the bilayer. To investigate

the influence of electrostatic interactions on the structural properties of lipid membranes, we therefore focus here on the area per lipid.

Results in Fig. 8 illustrate the time dependence of the area per DPPC molecule $A(t)$ over a timescale of 20 ns. As the equilibration of $A(t)$ takes about 10 ns, it is reasonable to discard this part of the trajectory and use only the last 10 ns for analysis. The simulations using PME then yield $\langle A \rangle = (0.655 \pm 0.010)$ nm^2. This result is consistent with recent experiments summarized by Nagle and Nagle [157], which for DPPC give $\langle A \rangle = 0.64$ nm^2. Truncation at 2.5 nm reduces the size of the bilayer by about 8% compared to the PME result, giving $\langle A \rangle = (0.604 \pm 0.009)$ nm^2. Further decrease of the cutoff distance to 2.0 nm leads to $\langle A \rangle = (0.582 \pm 0.027)$ nm^2. Finally, the smallest truncation distance of $r_{cut} = 1.8$ nm considered here yields $\langle A \rangle = (0.551 \pm 0.005)$ nm^2. Without doubt, this result deviates substantially from the PME result. Notably, if the scheme is changed from truncation at 1.8 nm to PME, the behavior of $\langle A \rangle$ rapidly evolves to the limit found for PME (see Ref. [175]). Thus, the deviations between the results for different schemes are solely the result of electrostatics, and not caused by initial conditions, for example.

Other simulation studies both in the low-temperature gel phase [178] and in the high-temperature fluid phase [179] are consistent with these findings. The study by Anézo et al. [179] is exceptionally revealing, as it covers a wide range of different electrostatic schemes and demonstrates the sensitivity of the area per lipid on the scheme chosen for electrostatic interaction.

The interactions used in the above systems have been completely similar in all cases, except for the scheme chosen for the long-range electrostatic interactions, yet the deviations between different model systems have been surprisingly large. To gain a better understanding of the reasons that lead to the above differences, we next focus on the most central quantity in all condensed and soft matter systems (i.e., the radial distribution function).

4.1.3. Radial Distribution Functions

The radial distribution function (RDF), $g(r)$, describes the probability of finding a pair of atoms a distance r apart, relative to the probability expected for a completely random distribution at the same density [61]. For the ideal gas characterized by the absence of conservative interactions, the RDF is therefore equal to 1 for all distances r. Interacting systems in turn are characterized by peaks and valleys that describe regions of high and low density, respectively. Hence, the RDF is an excellent means to gauge the ordering of any system.

Figure 9 shows the intermolecular RDF's between the nitrogen (N) atoms in the DPPC headgroups in a lipid bilayer. We find that the results for the PME technique are as expected, as in that case there is a strong main peak at about 0.85 nm and essentially no structure beyond 1.5 nm. This is consistent with the liquid–crystalline phase in which the lipids are in a fluid-like state, characterized by the lack of long-range ordering in the plane of the

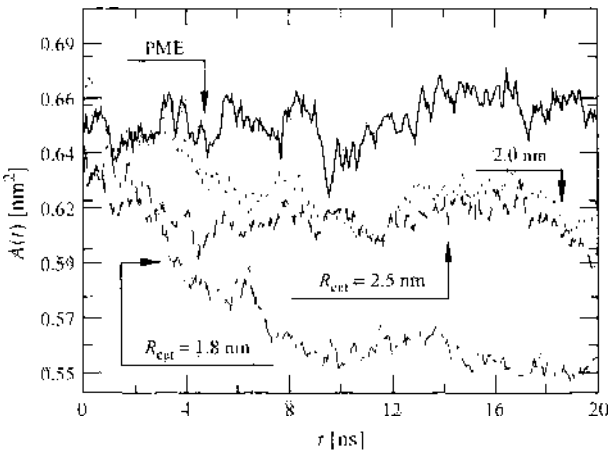


Figure 8. Temporal behavior of the area per molecule in a model system of a DPPC bilayer. Adapted from Ref. [175].

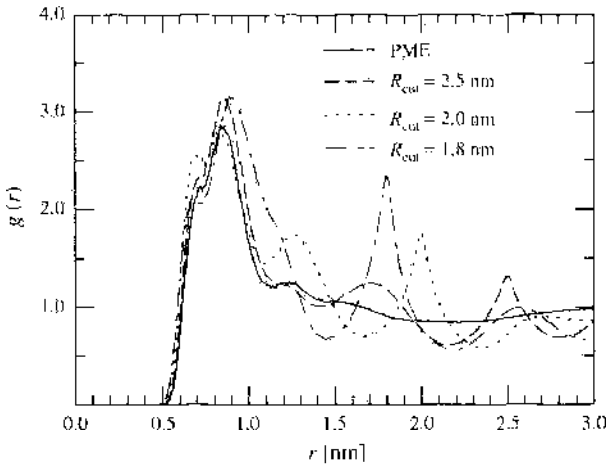


Figure 9. Radial distribution function for intermolecular N–N pairs in the headgroups of the dipalmitoylphosphatidylcholine molecules. Note the peaks exactly at the truncation distance. Adapted from Ref. [175].

membrane. The truncation techniques, in contrast, lead to pronounced peaks exactly at the truncation distance. For the truncation distance of 1.8 nm, the RDF has a prominent peak exactly at 1.8 nm. When the truncation distance is increased, the peak moves accordingly.

The above finding gives rise to major concern as it indicates that the truncation of electrostatic interactions may change the phase behavior of the system. As a matter of fact, this is the case, as was shown by Patra et al. in Ref. [175] for the RDF between the centers of mass positions of the lipids. The artificial ordering resulting from truncation leads to unexpected peaks that are entirely unphysical, thus changing the phase behavior of the membrane from a liquid–crystalline to some semi-fluid-like phase characterized by ordering at the length scale of the truncation distance. At the same time, the PME method does not yield any structure of the same kind.

Without any doubt, the above artifacts in structural quantities are the result of the fact that the truncation method does not account for the long-range component of electrostatic interactions. Apparently, the truncation of electrostatic interactions gives rise to artificial ordering in the bilayer plane, thus changing the phase behavior of the system. Although related artifacts have been observed in a number of different soft matter systems, the artifacts seem to be more pronounced in lipid membranes. Why so? To address this question, let us first recall that the dielectric constant of water is about 80. This implies that long-range interactions in water are strongly screened. Inside the membrane, however, the dielectric constant is about 2–4, and in the vicinity of the water-membrane interface it is about 20–40. Screening of long-range interactions in a membrane is therefore considerably weaker compared with pure water, thus enhancing the magnitude of artifacts compared to water-like systems. Second, it has been found that the energetic cost of changing the area per lipid in a membrane is very small, on the order of thermal energy [180], which implies that even minor changes in electrostatics may lead to substantial changes in the area of the membrane. As numerous both structural and dynamical quantities are influenced by the (excess) area in a membrane, it is therefore understandable how they are so strongly influenced by the choice of electrostatics in lipid membrane simulations. Clearly, this is a major issue and warrants particular attention in large-scale simulations.

4.1.4. Discussion on Ways to Handle Electrostatic Interactions

The above example demonstrates that the choice of a scheme for electrostatic interactions may dramatically affect the structural properties of lipids in a DPPC bilayer. Recent studies in the same lipid bilayer system [181] have further shown that dynamical properties such as lateral and rotational diffusion are equally sensitive to the choice of electrostatics. Now, of course, we can claim that the above findings are specific to this system rather than being generic and holding true in a variety of different lipid membrane systems. On one hand, the idea of the specific nature of our findings is true: The influence of the electrostatic

scheme on the structure and dynamics depends on the force field and other practical details chosen for the model. However, on the other hand, we may also conclude that the results discussed above are by and large generic. In particular, for the truncation methods, the location of artificial peaks found exactly at the truncation distance in the radial distribution functions is the most clear example to support this view. More generally, although this discussion has centred on lipid membranes, it is likely that similar kinds of findings would be faced in many other soft matter systems such as membrane-DNA complexes and systems composed of proteins and other solutes embedded in or attached to membranes. As we prefer to be on the safe side, we conclude that the truncation of electrostatic interactions should be avoided, if possible.

Having made this conclusion, let us ask ourselves why the truncation method has been used in the first place, and why it is still rather popular in large-scale simulations. The answer lies in the computational cost associated with handling electrostatic interactions. In this model system composed of a DPPC bilayer in water, the computational cost of using PME is larger by a factor of 1.8 compared with truncation at 1.8 nm. In large-scale simulations, where the computations may take several months (or even years) of CPU time, this is a major issue. In the absence of better techniques, the truncation distance is therefore still commonly used in systems where the computational load is excessive.

Recent studies may provide a better solution to this situation, however: The reaction-field technique [61]. In the reaction-field method, electrostatic interactions are explicitly accounted for up to some truncation distance (typically about 1.0–1.5 nm), beyond which the remaining long-range contribution is described by a reaction field correction using a single dielectric constant. Whereas the origin of this method goes back to the days of Onsager in the 1930s, it is still a reasonable and relatively useful choice for treating long-range interactions in an efficient and rather accurate manner. Of course, as the reaction-field method is typically implemented using a single dielectric constant through the whole system, it may be problematic in nonisotropic cases such as water-membrane systems and other interfaces. Further work would therefore be highly welcome for developing the RF approach in this respect. Nevertheless, we are confident that the reaction-field method offers a safer choice than the methods based on an abrupt truncation of electrostatics. This conclusion is supported by recent studies by Anézo et al. [179]. Recent investigations on a DPPC bilayer system provide further support for this idea [181, 182], showing that the artifacts in RDFs can be largely avoided by using the reaction-field technique instead of truncation. Yet, the computational load of the RF approach is comparable to the truncation technique.

Despite the above discussion, we wish to conclude that approaches such as the PME technique are currently the method of choice, if the computational load is not a limiting factor.

4.2. Role of Random Numbers in Parallel Stochastic Simulations

Perhaps surprisingly, random numbers are used extensively in numerous both scientific and practical applications. Stochastic simulations in physical sciences [60, 183, 184] characterize this fact very well. There, where the aim is to model complex systems over large length and timescales, atomistic approaches such as classical molecular dynamics are therefore no longer feasible. Examples of such stochastic techniques include the MC method, Brownian dynamics simulations, and DPD, all of which use pseudorandom numbers to generate the dynamics of model systems under study. Because of the deterministic nature of pseudorandom numbers, it is obvious that if there are any significant correlations within pseudorandom number sequences, then the dynamics of these model systems will be biased and the reliability of the whole approach may become questionable. To avoid such concerns, pseudorandom number generators should be thoroughly tested before extensive use in model simulations. This is particularly true as regards stochastic simulations in a parallel environment and in large-scale studies of soft matter systems, where the number of random numbers used is typically huge.

Below, we consider these two topics side by side. We first discuss recent ideas [185] on how to test the quality of random numbers in parallel applications in terms of random walks. As it turns out, this approach is both simple and efficient and is able to reveal correlation

effects in a number of commonly used pseudorandom number generators. Second, we study the effects of the quality of random numbers on DPD simulations [11, 17–19], which are based on solving Newton's equations of motion in the presence of a stochastic force component. The results indicate a few promising generators whose performance in the present tests is rather remarkable.

4.2.1. How to Test Parallel Random Number Generators by Random Walks

Parallel computing and parallel random number generation have been active fields of research for a relatively long time. With the exception of a few interesting studies [186–188], much less attention has been devoted to design tests, which are specifically suited for gauging correlations in parallel random number generators. This is partly the result of a great number of general test methods that have been developed and used during the last few decades [47]. In this general approach that is common to numerous standard tests, correlations are looked for from a very long random number sequence $\{x_i\}$. This approach works very well in serial computing, where the whole random number sequence is used on a single CPU. In parallel applications, however, the case is more complicated. Then it is more practical to consider relatively short subsequences $\{x_i\}^{(k)}$, $i = 1, \ldots, \Omega_k$, where subsequences $k = 1, \ldots, m$ do not overlap and are used on distinct CPUs during the calculation, as the task is distributed as subtasks between a number of different computing units. Obviously, the interest is now on cross-correlations between distinct random number sequences $\{x_i\}^{(1)}, \ldots, \{x_i\}^{(m)}$. To study such effects, one needs specific approaches that gauge long-range correlations between blocks of random numbers.

We consider this problem in terms of random walks, which are a common tool in a variety of disciplines, including physics, chemistry, biology, and economics. We use random walks to characterize the quality of random numbers in parallel calculations. The key idea is to consider a number of diffusing random walkers, each of which is governed by a distinct random number sequence. Through studies of their mutual correlations we are then able to characterize and quantify possible correlations between separate pseudorandom number sequences. To this end, we calculate quantities, such as intersection probabilities between different random walks, and compare their asymptotic behavior to known theoretical limits. The difference between simulation results and theoretical predictions serves as a measure of correlations in the random number generator under study.

In the following, we briefly describe three tests [185] that are based on this idea. They study both correlations within a single random number sequence $\{x_i\}^{(k)}$ and cross-correlations between distinct random number sequences $\{x_i\}^{(1)}, \ldots, \{x_i\}^{(m)}$. The sequences $\{x_i\}^{(k)}$ are generated by the pseudorandom number generator in question from $\{x_i\} = x_1, \ldots, x_\Omega, x_{\Omega+1}, \ldots, x_{2\Omega}, \ldots$ such that we obtain nonoverlapping sequences $\{x_i\}^{(1)} = x_1, \ldots, x_\Omega$, $\{x_i\}^{(2)} = x_{\Omega+1}, \ldots, x_{2\Omega}$, and so on. Here we consider the case where the sizes Ω_k of sequences $\{x_i\}^{(k)}$ are equal for all k. Random numbers x_i are uniformly distributed between zero and one.

4.2.2. Tests Based on Random Walks

In the height-correlation test, we consider the position y_i of a one-dimensional random walker versus the number of jumps made i. The position $y_i = \sum_{l=1}^{i} \delta y_l$ is a sum of displacements δy_i, which are random variables

$$\delta y_i = \begin{cases} +1, & \text{if } x_i \leq 1/3 \\ 0, & \text{if } 1/3 < x_i \leq 2/3 \\ -1, & \text{otherwise} \end{cases} \tag{28}$$

In this manner, we construct the paths $y_i^{(1)}$ and $y_i^{(2)}$ from the sequences $\{x_i\}^{(1)}$ and $\{x_i\}^{(2)}$, respectively. Using the initial condition $y_0^{(1)} = y_0^{(2)} = 0$, the height between the two random

walkers is defined as $h_t = y_t^{(1)} - y_t^{(2)}$. For a random process, the corresponding correlation function $H_t \equiv \langle |h_t - h_0| \rangle \sim t^{\phi}$ decays asymptotically as a power law with an exponent $\phi = 1/2$ [189].

The intersection test deals with two random walkers on a square lattice. The random walkers start from the origin at the same time and are allowed to jump independently on a lattice. Meanwhile, we consider the probability I_t that the two random walkers after t jumps have no intersection other than their common starting point. Note that the two random walks need not meet at the same site at the same time, but any common point in their paths is regarded as an intersection. For a random process, I_t behaves asymptotically like a power law $I_t \sim t^{-\alpha}$ with an exponent $\alpha = 5/8$ [190, 191].

In the S_N test, we consider N random walkers in one dimension and let them move simultaneously without any interaction such that, at any jump attempt, they can make a jump to the left or to the right with equal probability. After $t \gg 1$ jumps by all random walkers, the mean number of sites visited $S_{N,t}$ has an asymptotic form $S_{N,t} \sim f(N)t^{\gamma}$, where the scaling function $f(N) = (\ln N)^{1/2}$ and $\gamma = 1/2$ [192]. The value of γ observed serves as a measure of correlations.

Values of the parameters used in the tests discussed below are given in Table 2. There Ω is the number of jumps in a single random walk, and M is the number of independent runs in the test.

4.2.3. Results for Some Pseudorandom Number Generators

Here we discuss results of the three tests for a few carefully chosen pseudorandom number generators. The present results complement previous investigations discussed in more detail in Ref. [185].

The generators discussed in this work represent a variety of different generators often used within the physics community. R250 is an implementation of the generalized feedback shift-register (GFSR) algorithm [193] $x_n = x_{n-250} \oplus x_{n-103}$, where $\oplus$ is the bitwise exclusive OR operator. R89 is another example of GFSR generators and uses $x_n = x_{n-89} \oplus x_{n-38}$. RAN is a "minimal" linear congruential generator of Park and Miller (with multiplier 16807 and modulus $2^{31}-1$) combined with a Marsaglia shift sequence, and it has been proposed in a recent edition of *Numerical Recipes* [194]. RAN2 is based on the 32-bit combination generator first proposed by L'Ecuyer [195] and later published in *Numerical Recipes* [196], using shuffling. RANMAR [197, 198] combines a lagged Fibonacci generator with a simple arithmetic sequence and has been suggested as a good candidate when one aims towards a "universal generator" [198]. In addition, we test the Mersenne Twister [199], which has a huge period and good theoretical properties in view of recent studies by Matsumoto and Nishimura [199]. Finally, we consider the most "luxurious" version of RANLUX [200, 201], which is based on ideas of deterministic chaos. In RANLUX4, one generates 389 random numbers, delivers 24 of them, and throws the remaining 365 numbers away.

To determine the exponents ϕ, α, and γ, we considered their "running exponents" defined as

$$\phi_t \equiv \frac{\log(H_{t+\delta t}/H_t)}{\log[(t+\delta t)/t]} \tag{29}$$

where the present example is for the height correlation test. Running exponents of α and γ were determined in a similar fashion. The "time window" δt used in this work varied between 50 and 1000.

Demonstrative results for the running exponents are shown in Fig. 10. The initial behavior of some generators is clearly different and is discussed below in more detail. Nevertheless,

Table 2. Values of the parameters Ω and M used in the tests. The S_N test was carried out with $N = 2$.

	Height Correlation Test	Intersection Test	S_N Test
Ω	2000	4000	2000
M	10^7	10^8	10^7

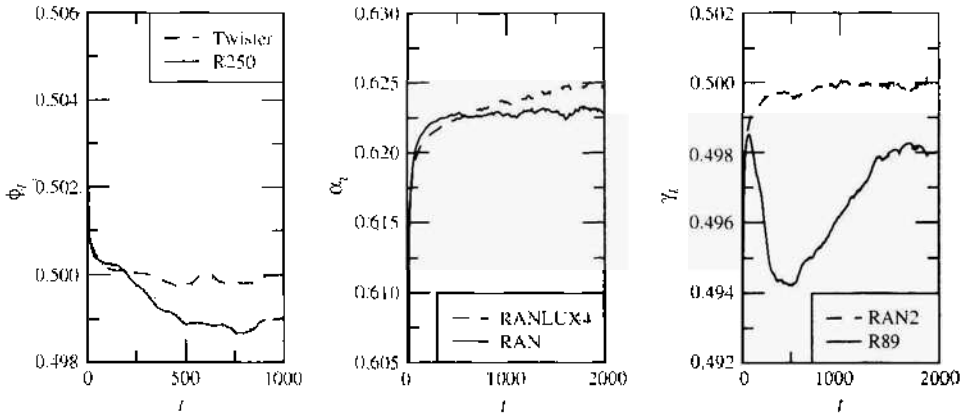


Figure 10. Demonstrative results of the running exponents (left) ϕ_t, (middle) α_t, and (right) γ_t in the three tests. The exponents shown in Table 3 have been extracted from the large-t limit. For clarity's sake, results of only a few generators are shown here.

asymptotically, we find the running exponents of all generators to converge to some limiting value at large t. This regime was therefore used to determine the exponents.

The results for the exponents ϕ, α, and γ are given in Table 3. Results of the GFSR generators R89 and R250 are not surprising, as they have recently failed in various random walk tests (see references in [185]). RANLUX4 and RANMAR, on the other hand, perform considerably better, in agreement with some previous studies [185, 198, 202–205]. Equally good performance is found for the Mersenne Twister and RAN2, whereas the results of RAN in the intersection test are not fully convincing. Then, the exponent α deviates from the theoretical limit, and this deviation seems to be systematic, as is demonstrated by Fig. 10 (middle). When the intersection test was repeated for RAN, we found $\alpha = 0.6239(4)$, which again fails the test.

The tests presented above focus on the asymptotic behavior of the corresponding correlation functions. This is because only the asymptotic behavior of the correlation functions is known theoretically. Yet correlations in pseudorandom number sequences are playing a role also at shorter scales, as is very evident from the running exponents in Fig. 10. For cases in which such short-range correlations are of interest, one can use a two-way analysis in which various pseudorandom number generators are judged against a reference generator. This approach was used in Ref. [185], which revealed further correlations in a few random number generators, including RANMAR.

At this stage, we conclude that even highly recommended and commonly used pseudorandom number generators may fail in tests that focus on their weak points. In general, one can safely say that there are no good pseudorandom number generators. All of them are bad. It is simply a question of finding those that are better and more reliable than the others. We will discuss this broad issue in more depth in Section 4.3.3. Meanwhile, to get further

Table 3. Results for the exponents of the three tests. The notation 0.4989(2) means 0.4989 ± 0.0002.

RNG	Height Correlation Test ϕ	Intersection Test α	S_N Test γ
R250	0.4989(2)	0.6265(5)	0.4984(1)
R89	0.4984(2)	0.6205(5)	0.4981(1)
RANLUX4	0.5001(1)	0.6257(5)	0.5000(1)
RANMAR	0.5000(1)	0.6250(4)	0.5001(1)
Twister	0.4999(1)	0.6242(5)	0.5000(1)
RAN	0.4999(1)	0.6224(4)	0.5000(1)
RAN2	0.5001(1)	0.6241(5)	0.5000(1)
Exact	1/2	$5/8 = 0.625$	1/2

Note: The exponents were extracted from the asymptotic tail of the corresponding correlation functions, and the exponents that deviate from the exact value by more than two error bars have been underlined.

insight into the use and effect of random numbers in stochastic simulations, we consider their influence on DPD simulations.

4.3. Role of Random Numbers in Dissipative Particle Dynamics Simulations

Above we have stressed the importance of developing techniques for multiscale modeling. Development of stochastic simulation techniques also plays a crucial role in this respect. This is simply because many important processes in soft matter systems take place at mesoscopic length and timescales (roughly 1–1000 nm and 1–1000 ns) that are beyond the limits of atomic-scale molecular dynamics. To overcome this problem, stochastic simulation methods such as DPD, discussed in Section 3.2, are hence needed.

The role of random numbers in DPD is crucial. This is because particles in DPD model systems are partly driven by stochastic noise, which is generated by pseudorandom numbers. Any correlations in pseudorandom number sequences can therefore lead to serious problems if they interfere with the true dynamics of the system.

To study how sensitive DPD actually is to such underlying correlations in pseudorandom number sequences, we consider a simple model fluid system described by N particles with masses m_i, coordinates $\vec{r}_i$, and velocities $\vec{v}_i$. Interparticle interactions are characterized by the pairwise conservative, dissipative, and random forces exerted on particle i by particle j, respectively, and are given by

$$\begin{aligned} \vec{F}_{ij}^{C} &= \alpha\omega(r_{ij})\vec{e}_{ij} \\ \vec{F}_{ij}^{D} &= -\gamma\omega^2(r_{ij})(\vec{v}_{ij}\cdot\vec{e}_{ij})\vec{e}_{ij} \\ \vec{F}_{ij}^{R} &= \sigma\omega(r_{ij})\xi_{ij}\vec{e}_{ij} \end{aligned} \tag{30}$$

where $\vec{r}_{ij} = \vec{r}_i - \vec{r}_j$, $r_{ij} = |\vec{r}_{ij}|$, $\vec{e}_{ij} = \vec{r}_{ij}/r_{ij}$, and $\vec{v}_{ij} = \vec{v}_i - \vec{v}_j$. The weight function for the different interaction terms has here been chosen to follow the same form. Random numbers appear in these equations by describing ξ_{ij}, which are symmetric random variables with zero mean and unit variance. The ξ_{ij} are expected to be uncorrelated for different pairs of particles and different times, and it is indeed random numbers whose task it is to satisfy this condition.

Remaining details are fixed by adopting a commonly made choice for the weight function

$$\omega(r_{ij}) = \begin{cases} 1 - r_{ij}/r_c & \text{for } r_{ij} < r_c \\ 0 & \text{for } r_{ij} > r_c \end{cases} \tag{31}$$

with a cutoff distance r_c [17] and $\omega^R(r_{ij}) = \omega(r_{ij})$. Then the equations of motion are given by the set of stochastic differential equations

$$d\vec{v}_i = \frac{1}{m_i}\left(\vec{F}_i^{C}\,dt + \vec{F}_i^{D}\,dt + \vec{F}_i^{R}\sqrt{dt}\right) \tag{32}$$

where $\vec{F}_i^{R} = \sum_{j\neq i}\vec{F}_{ij}^{R}$ is the total random force acting on particle i (with $\vec{F}_i^{C}$ and $\vec{F}_i^{D}$ defined correspondingly).

4.3.1. Model System

To maximize the role of random numbers, we study a simple model fluid system of identical particles ($m_i = m\ \forall i$) in the absence of conservative forces ($\alpha = 0$). This choice corresponds to an ideal gas, which provides us with some exact theoretical results to be compared with those of model simulations. The dynamics of the system then arise only from random noise and from a dissipative coupling between pairs of particles. The random force strength is chosen as $\sigma = 10$, and the dissipation strength is given by the fluctuation–dissipation relation $\sigma^2/\gamma = 2k_BT^*$ [18], where the desired thermal energy is chosen as $k_BT^* = 1$. The length scale is defined by $r_c = 1$, and time is given in units of $r_c\sqrt{m/k_BT}$.

In our simulations we use a three-dimensional box of size $10 \times 10 \times 10$ with periodic boundary conditions and a particle density $\rho = 4$. Equations of motion [Eq. (32)] are solved in practice using the recently suggested integration scheme known as DPD-VV [125–127],

and the time step used in this procedure was $\Delta t = 0.001$. With this choice, the average temperature during simulations $\langle k_B T \rangle$ remains at the desired value within 1%. The results shown below are for simulations of 10^6 time steps after equilibration, which takes about 200 CPU-hours each on a typical RISC workstation.

Because the results depend on how the random number sequences are used, we discuss this issue in some detail. Consider particles $i = 1, \ldots, N$. For every pair of particles at a given moment, we need to generate an independent and symmetric random variable ξ_{ij} with zero mean and unit variance, which here is done using uniformly distributed random numbers with unit variance [11]. Then using the particle "1" as an example, one finds all other particles "j" that interact with the particle "1" (i.e., particles for which $r_{1j} \leq r_c$). Suppose that the number of such particles $\{j_1, j_2, \ldots, j_K\}$ is K. Then one generates a random number sequence $\{x_i\} = x_1, \ldots, x_K$ such that x_1 determines ξ_{1j_1}, x_2 determines ξ_{1j_2}, and so forth. The same procedure is then performed for all particles interacting with the particles "2" through "$N - 1$," such that each pair of particles is considered exactly once.

The results below have not been published elsewhere, which possibly explains our interest in specifying this model in great detail.

4.3.2. Results for Some Pseudorandom Number Generators

In this discussion, we consider a small sample of the generators described in Section 4.2.3. We focus on the radial distribution function $g(r)$ [206] that is one of the most central observables in studies of liquids and solid systems. For the ideal gas considered here, $g(r) \equiv 1$ for all r, and therefore any deviation from one has to be interpreted as an artifact caused by the employed simulation scheme. This approach provides a simple and effective measure of correlation effects and serves our main interest of seeing how sensitive DPD is to the underlying correlations in pseudorandom number sequences.

Results shown in Fig. 11 (left) indicate that differences between the generators are small. For all generators tested, the radial distribution function is almost flat and has only minor deviations from the expected behavior. Did we expect something else that would have allowed us to conclude that some pseudorandom number generators lead to results that are simply nonsense? Maybe. At anyrate, we are definitely glad that we did not find major artifacts.

Aside from noise effects, the minor deviations found in Fig. 11 (left) may result from three different sources. First, the expected behavior of $g(r) = 1$ is rigorously true only in the limit $\Delta t \to 0$, whereas in practice the time step is always finite. Second, how Eq. (32) is integrated also affects physical quantities such as $g(r)$. Third, the deviations may result from correlations in random number sequences.

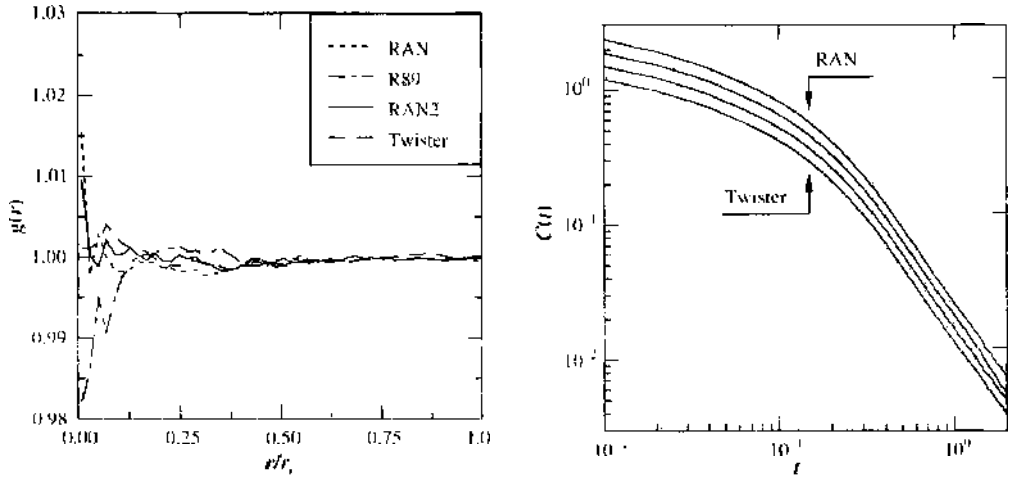


Figure 11. Left: Demonstrative results of the radial distribution function $g(r)$ for a few random number generators tested. Note that error bars are largest for small r, as the data was collected with a fixed bin size $dr = r_c/100$. Right: The decay of the velocity correlation function $C(t)$ for a few generators from top to bottom: RAN, R89, RAN2, and the Mersenne Twister. Results of R89, RAN2, and the Mersenne Twister have been shifted to clarify the presentation.

In this case, the time step Δt and the integration procedure are fixed, and therefore their effect is similar for all generators. Thus, the best we can do is to compare the four curves in Fig. 11 with one another. This comparison leads to a simple conclusion that all generators yield approximately similar results within error bars. Only R89 deviates from other generators at small r, and even in this case the fluctuations are rather weak.

To complement this test, we also calculated the isothermal compressibility κ_T defined as $\kappa_T \rho k_B T = 1 + 4\pi\rho \int_0^\infty dr r^2 [g(r) - 1]$, which is an example of a thermodynamic response function. Results of all generators were found to be very similar. Finally, we used the single-particle velocity correlation function $C(t) = \langle \vec{v}_i(t) \cdot \vec{v}_i(0) \rangle$ [206] to gauge possible problems with time-dependent quantities such as diffusion. These studies revealed that the results of all generators for $C(t)$ were essentially identical [see Fig. 11 (right)].

These results indicate that DPD model simulations are relatively insensitive to the underlying correlations in pseudorandom number sequences. This idea is supported by further calculations using rather poor generators (such as the GFSR generator $x_n = x_{n-31} \oplus x_{n-3}$), whose results were comparable to those discussed above. However, we cannot conclude that this finding is generic, as the test results depend on how random numbers are used within a simulation. In this case, we feel that the fluid-like nature of the system plays certain role. Namely, as was described in Section 4.3.1, the motion of the particle i is influenced by random forces acting between i and other neighboring particles nearby. Because the number of neighboring particles K is usually small and not conserved, the motion of the particle i is dictated by a short random number sequence $\{x_i\} = x_1, \ldots, x_K$, whose size K fluctuates in time. Furthermore, the random number sequences $\{x_i\}^{(1)} = x_1, \ldots, x_{K_1}$ and $\{x_i\}^{(2)} = x_{1+L}, \ldots, x_{K_2+L}$ used at consecutive moments for the same particle are separated by a sequence of size $L \sim \mathcal{O}(N\langle K \rangle)$. Thus, to conclude, it seems plausible that only strong, short-range correlations within $\{x_i\}$ influence the dynamics of simple fluid systems that have been considered in this work.

We found above that the quality of pseudorandom numbers is likely not crucial in DPD model simulations of simple fluid systems. This favorable idea should be taken with a grain of salt, however. A number of simulation studies have shown [185, 207, 208] that the underlying correlations in pseudorandom number sequences may interfere with the true dynamics of model systems. Therefore it is possible that the same problems are lurking behind the shadows and could be faced in DPD simulations, too. An example of such a situation is DPD simulations in a parallel environment, where pseudorandom number sequences can be used in various different ways (such as the leap-frog and blocking techniques, to name just two examples). This warrants particular care to be taken when pseudorandom number generators are being used in future large-scale applications. Although such tests have not been carried out yet (to our knowledge), they would be highly welcome to clarify this issue.

4.3.3. Discussion on the Future Aspects of Using Pseudorandom Numbers

The Buffon experiment [209] is probably one of the first and best known examples of the MC method. In this experiment, one throws needles of equal length at random over a plane marked with parallel and equidistant lines. By counting the number of intersections between lines and needles, one can estimate the value of π. Provided that the needles were thrown by a pseudorandom number generator, how precise and "universal" values for π would you expect?

If your answer is positive and in favor of using pseudorandom number generators without any concern, perhaps a few practical examples might change your mind. In the late 1980s, it was found [210] that the critical exponents in the three-dimensional Ising model depended on the random number generator used in the MC simulations. Later in the 1990s, Ferrenberg et al. [207] draw the same conclusion in the two-dimensional Ising model when cluster algorithms were employed for the dynamics of the system. Similar findings in other contexts are numerous and include, for example, studies of surface growth, deposition problems, and random walks [202, 208, 211].

The above examples demonstrate that in stochastic simulation techniques, there is an extra degree of freedom that should always be mentioned when the results are reported

in the literature: the pseudorandom number generator. This may sound sarcastic, but the fact is that the results of stochastic simulations are a function of the pseudorandom number generator used.

As a consequence, numerous test methods have been developed with an aim to finding (undesired) correlation effects from the pseudorandom number sequences. The standard tests that focus on general properties of random number sequences form the most significant test bench in this field [47]. They are complemented by more specific tests, often known as application-specific tests [202, 203] that are designed to mimic some particular application. For research topics such as surface growth or surface diffusion, using lattice-gas models, the Ising model is one of the most convenient starting points when one aims to develop an application-specific test for these purposes. In addition to these, theoretical tests such as the spectral test and discrepancy complement our knowledge of the properties of pseudorandom number algorithms. In all, there is a huge body of tests available. The key issue is to use them and, further, to develop new test methods for ever increasing requirements. When done carefully, one can avoid undesired artifacts resulting from the pseudorandom number generator used.

We have demonstrated above how correlations in some commonly used pseudorandom number generators may change the behavior of physical quantities such as time-dependent correlation functions. However, we have also found that some recently suggested generators such as the Mersenne Twister and RANLUX4 passed all the present tests. A pessimist could now say that all pseudorandom number generators have inherent weaknesses, and that eventually they fail anyway. Optimistically speaking, however, one can conclude that there are still many generators whose properties are reasonably good even for very challenging applications. To make sure that the situation will remain equally positive in the future as the computational power is increased, further work is definitely called for to develop more reliable pseudorandom number generators. When combined with novel test methods to challenge the quality of pseudorandom numbers in the future, we are on the right track.

4.4. How to Integrate Equations of Motion in DPD Simulations

In the above discussion, we have noted that there is an increasing demand for useful and efficient methods to deal with complex soft matter systems in the mesoscopic regime. DPD [11, 17–19] (see Section 3.2) is a particularly appealing technique in this regard. This is mainly because the "particles" of DPD correspond to coarse-grained entities, which allows the use of large time steps (compared to typical molecular dynamics simulations) as pair potentials for the coarse-grained description are considerably softer than for the underlying atomic description. Further, the pair potentials in DPD can be chosen such that the hydrodynamic modes are accounted for in a proper fashion.

In practice, to obtain full hydrodynamics, the choice of interactions is done through a pairwise coupling of particles through random and dissipative forces. As we will note below, this makes the integration of the equations of motion a nontrivial task. The main difficulty here arises from the dissipative force, which depends explicitly on the relative velocities of the particles, whereas the velocities in turn depend on the dissipative forces. An accurate description of the dynamics therefore requires a self-consistent solution, which in turn is computationally demanding.

Here we discuss how this problem can be overcome. In addition to the self-consistent approach, we discuss other complementary methods that account for the velocity dependence of dissipative forces in some approximate manner, allowing the integration to be carried out to a sufficient degree of computational efficiency.

4.4.1. Equations of Motion in DPD

The time evolution of particles in DPD can be described by the Newton's equations of motion

$$\begin{aligned} d\vec{r}_i &= \vec{v}_i\, dt \\ d\vec{v}_i &= \frac{1}{m_i}\left(\vec{F}_i^C dt + \vec{F}_i^D dt + \vec{F}_i^R \sqrt{dt}\right) \end{aligned} \tag{33}$$

Here $\vec{F}_i^C = \sum_{i \neq j} \vec{F}_{ij}^C$ is the total conservative force acting on particle i, and $\vec{F}_i^D$ and $\vec{F}_i^R$ are defined correspondingly for the dissipative and random forces, respectively.

The reason why we come back to this issue is the fact that Eq. (33) is at the heart of the problem as regards DPD. It shows that the velocities of the particles depend on the dissipative forces acting on them. At the same time, the dissipative forces in turn depend on the relative velocities of the DPD particles [see Eq. (30)]. The velocities and dissipative forces of all the particles are therefore intertwined. This is precisely the reason why the self-consistent solution is needed in the first place.

Equation (33) furthermore shows that the similarity of this case to classical MD simulations is close. Nevertheless, there is reason to point out that the random force term in Eq. (33) is coupled to the factor $\sqrt{dt}$ instead of dt. This can be justified by a Wiener process, as in stochastic differential equations. Here, it suffices to notice that physically the Wiener process models intrinsic (thermal) noise in the system and provides the simplest approach to modeling Brownian motion using stochastic processes (see Refs. [11, 18]).

4.4.2. Model Using DPD

Although the above continuous-time version of DPD satisfies detailed balance and describes the canonical NVT ensemble, in practice the time increments in Eq. (33) are finite and the equations of motion must be solved by some integration procedure. To study possible artifacts caused by the choice of the integration scheme, we consider a gaslike system of N identical particles. The model discussed here is essentially similar to that presented in Section 4.3.1.

In essence, we consider an ideal gas ("ideal DPD fluid") in three dimensions, where the random force strength is chosen to be $\sigma = 3$ in units of $k_B T^*$ and the particle number density is $\rho = 4$. The weight function $\omega^R(r_{ij}) = \omega(r_{ij})$, as well as other practical details have been chosen as in Section 4.3.1. Finally, the random numbers were generated by the Mersenne Twister.

The reason for using this model is that the ideal DPD fluid model provides us with some exact theoretical results that can be compared to results from model simulations. Hence, the dynamics of the present system arise only from thermal noise and dissipative coupling between pairs of particles.

4.4.3. Integration Schemes

The integration schemes discussed below have been suggested very recently. They complement each other in the sense that the velocity dependence of the dissipative forces is accounted for in all cases, but the approaches differ substantially from each other.

The standard velocity–Verlet (VV) [212] is typically the integrator in classical MD simulations. It is time-reversible and symplectic and works well especially at large time steps [61]. Here we use VV as a starting point for integrators designed specifically for DPD simulations.

The modified VV algorithm (DPD-VV) differs from the standard VV in the sense that it accounts for the velocity dependence of dissipative forces in an approximate fashion by updating the dissipative forces for a second time at the end of each integration step. When this idea is extended further, we obtain a self-consistent variant of DPD-VV. This SC-VV algorithm [125–127] determines the velocities and dissipative forces self-consistently through functional iteration, and the convergence of the iteration process is monitored by the instantaneous temperature $k_B T$. This approach is similar in spirit to the self-consistent leap-frog scheme introduced recently by Pagonabarraga et al. [213]. Other schemes considered here include the approach by den Otter and Clarke (OC) [214], in which the parameters of a leap-frog algorithm are adjusted such that the effects resulting from the velocity dependence of dissipative forces are reduced as much as possible, and the approach (SI) by Shardlow [215], in which the integration process is factorized such that the conservative forces are calculated separately from the dissipative and random terms.

In addition to the above integrators, we also discuss the performance of Lowe's integration scheme [133] based on combining the Andersen thermostat [134] with the VV scheme in a momentum conserving DPD-like fluid. This approach differs from the "traditional" DPD

in the sense that now the dissipative and random forces are not explicit. Rather, they are described by the Andersen thermostat that thermalizes the system for pairs of particles. In practice, one first integrates Newton's equations of motion with a time step Δt, and then for all pairs of interacting particles one decides with a probability $\Gamma\Delta t$ ($0 < \Gamma\Delta t \leq 1$) whether to take a new relative velocity from a Maxwell distribution. The dynamical properties of the system can be tuned by the parameter Γ [133], which is inversely proportional to the decay time for relative velocities.

A full description of the integrators is given in Ref. [126].

4.4.4. Demonstrative Results

As discussed above, this model is characterized by the absence of conservative forces, and thus any artifacts arising from the velocity-dependent forces are expected to be pronounced in this model. To study this possibility, we first discuss the deviations of the observed kinetic temperature

$$\langle k_B T \rangle = \frac{m}{3N-3}\left\langle \sum_{i=1}^{N} \tilde{v}_i^2 \right\rangle \tag{34}$$

from the desired temperature $k_B T^*$. As the simulations are employed in the canonical ensemble, the temperature conservation is one of the main conditions for a reliable performance of the model. The results for $\langle k_B T \rangle$ shown in Fig. 12 show that DPD-VV and SC-VV are reasonably good at small time steps, whereas larger time increments lead to major deviations from the desired temperature. For OC, $\langle k_B T \rangle$ decreases monotonically with Δt for $\Delta t \leq 0.15$. For larger time steps, we find the temperature to increase rapidly. Overall, we can conclude that the deviation in the case of OC is greater than for DPD-VV but considerably smaller than in the case of SC-VV. The Shardlow S1 integrator, in turn, has excellent temperature control, and the deviations remain less than 0.5% up to $\Delta t = 0.2$. The best temperature control is found for the method by Lowe, however, yielding $\langle k_B T \rangle = k_B T^*$ for all time steps Δt. Although Lowe's approach depends on the dynamic variable Γ, the conclusions with regard to temperature conservation have not been found to depend on the value chosen for this parameter [126].

Next, we examine the radial distribution function $g(r)$ [216], which is one of the most central observables in studies of structural properties of liquids and solids. For the ideal gas, the radial distribution function provides an excellent test for the integrators, as then $g(r) \equiv 1$ at the continuum limit. Therefore, any deviation from unity has to be interpreted as an artifact resulting from the integration scheme employed.

Results for $g(r)$ are shown in Fig. 13. We find that the deviations from the ideal gas limit $g(r) = 1$ are pronounced for OC, indicating that even for small time steps, this integration

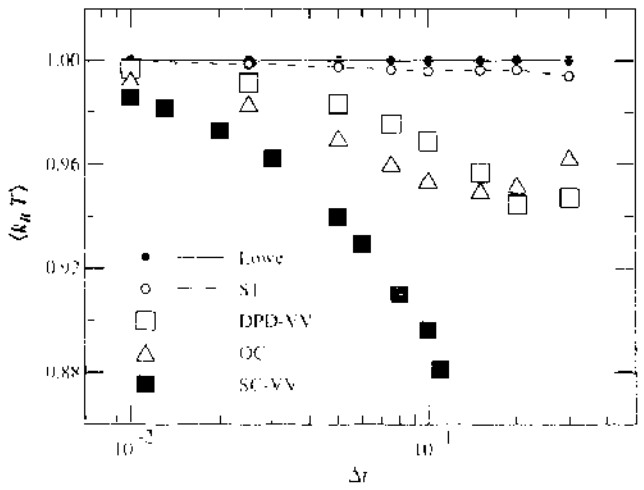


Figure 12. Results for the deviations of the observed temperature $\langle k_B T \rangle$ from the desired temperature $k_B T^* \equiv 1$ versus the size of the time step Δt [126]. The error in $\langle k_B T \rangle$ is of the order of 10^{-3}. Adapted from Ref. [126].

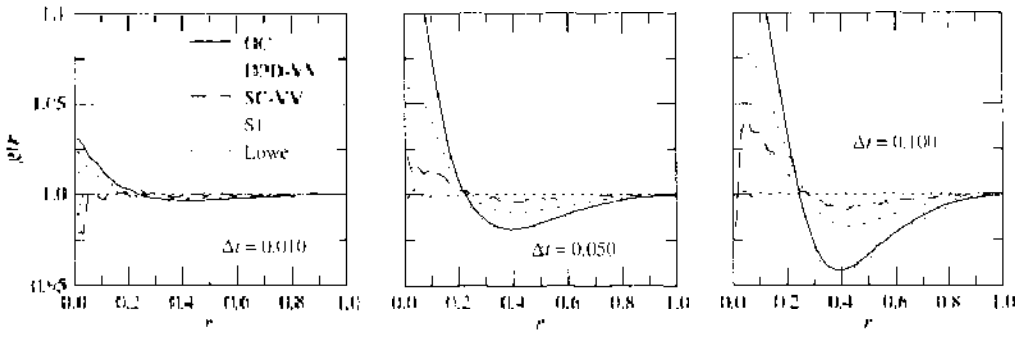


Figure 13. Radial distribution functions $g(r)$ with several values of time step Δt in model A: (a) $\Delta t = 0.01$, (b) $\Delta t = 0.05$, and (c) $\Delta t = 0.1$ [126]. The error in $g(r)$ is greatest at $r = 0.01$, where it takes the value of 0.01. Adapted from Ref. [126].

scheme gives rise to unphysical correlations. The performance of DPD-VV is considerably better, although artificial structures are yet rather pronounced, whereas SC-VV and S1 lead to a radial distribution function that is close to the theoretically predicted one. Completely structureless $g(r)$ is found only for the integrator by Lowe, however. Again, in this case, we have tested the behavior of $g(r)$ with various values of Γ, but the results remain the same. This confirms the expectation that Γ does not influence the equilibrium properties of the system.

The radial distribution function reflects equilibrium (time-independent) properties of the system. To complement the comparison of different integrators, we next consider the tracer diffusion coefficient

$$D_T = \lim_{t\to\infty} \frac{1}{6t} \langle [\vec{r}_i(t) - \vec{r}_i(0)]^2 \rangle \tag{35}$$

which can provide us with information of possible problems on the dynamics of the system. Here $\vec{r}_i(t)$ is the position of a tagged particle, and the mean-square displacement is then averaged over all particles in a system to get better statistics for D_T.

The results for the diffusion coefficient D_T in Fig. 14 are essentially consistent with the conclusions above. The integrator OC is not very useful in a system of this kind, since it seems to lead to substantial deviations from the expected behavior. The SC-VV and the integrator by Lowe perform much better, and the integrators S1 and DPD-VV are most stable in this case.

4.4.5. Discussion on Methodology of DPD

Above, we discussed the performance of various novel integration schemes that have been designed specifically for DPD simulations. On the basis of the ideal gas results, we can conclude that there are two approaches whose performance is above the others. Of these,

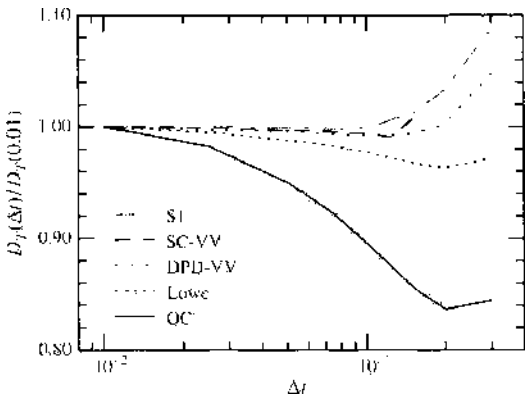


Figure 14. Results for the tracer diffusion coefficient $D_T(\Delta t)/D_T(0.01)$ versus the time step Δt [126]. The error in $D_T(\Delta t)/D_T(0.01)$ is on the order of 0.001. Adapted from Ref. [126].

Shardlow's integration scheme is based on splitting the equations of motion and can be applied to the usual DPD picture, whereas the approach by Lowe is distinctly different in nature and is related to the classical work by Andersen. Results for other, more strongly interacting models support this conclusion [126]. It is also noteworthy that both of these schemes are fast and easy to implement. Furthermore, and what is important when Lowe's method is compared to Shardlow's integration scheme, it provides an alternative and a very attractive description of dissipative particle dynamics. It is actually rather surprising that Lowe's approach has not attracted considerable attention yet. Perhaps the situation will improve shortly.

In addition to this issue, the methodology and the theoretical background of DPD have taken many significant steps very recently. From a simulation point of view, DPD has been shown to be a promising technique for studies of soft matter systems under non-equilibrium (flow) conditions [128]. On the theoretical side, a few extensions have been suggested recently [217, 218]. The Smoothed Particle Hydrodynamics framework suggested by Pep Español and his collaborators is particularly interesting as it lacks a number of limitations imposed by the traditional DPD picture [218]. We also would like to mention the recent work by Chris Lowe [133], whose alternative description for DPD is an attractive idea since it allows one to tune solvent properties in a reasonable and efficient way. Finally, as coarse graining is one of the grand challenges in the field of multiscale research, we would like to mention the recent work by Lyubartsev et al. [219]. Therein, the authors presented an approach to coarse-grain a MD description by the IMC technique such that the effective interactions found through this procedure can be applied to DPD simulations to study the large-scale properties of aqueous electrolyte solutions. This case will be discussed in more detail in Section 6.

We conclude that the development of methodology is an integral part of computational research. This is particularly true in soft matter research that deals with various fundamental questions, having a major effect on our understanding of biological systems. In particular, we hope that the case studies discussed above have highlighted some of the problems that may be hiding under the simulation protocols. Great care is therefore warranted to avoid artifacts resulting from the methods used, and more attention is called for to develop more accurate and efficient techniques to deal with complex, biologically relevant soft matter systems.

5. NANOSCIENCE OF BIOPHYSICAL SYSTEMS THROUGH ATOMIC-SCALE MOLECULAR DYNAMICS

At present, the length scales accessible to atomic-level molecular dynamics simulations are on the order of 5–50 nm. This may sound modest, but in practice it allows one to gain plenty of insight into the understanding of various biologically relevant molecules and molecular systems under hydrodynamic conditions. The above length scale further stresses the fact that MD is currently the method of choice for processes and phenomena that take place in the nanoscopic regime. In this section we therefore discuss a few examples that hopefully highlight the capability of molecular simulations, and MD in particular, in revealing the microscopic mechanisms and physical laws that govern biophysical systems on the atomic level. Because the field is enormous, we are bound to focus on one topic. We hence concentrate on the properties of lipid bilayers that are commonly adapted as basic models of cell membranes. In addition to these properties, we also, though very briefly, discuss recent computational modeling of other biomolecules related to membrane systems.

5.1. Lipid Membranes in a Nutshell

All cells are surrounded by cell membranes that are like thin elastic sheets. The thinness of the membranes is genuinely interesting, as whereas the diameter of a cell is typically 20–30 μm, the thickness of a cell membrane is just a fraction of this, being typically about 5 nm. Yet membranes are involved in essentially all processes related to the flow of nutrients and ions into and out of the cell. Membranes further provide an environment in which a

variety of biochemical processes occur, and in general the role of membranes in cells is crucial, as they function to organize biological processes by compartmentalizing them.

The main structural components of native biological membranes [220–222] are lipids (Greek: lipos, fat). As shown in Fig. 7, they have a polar headgroup and a nonpolar hydrocarbon tail. Because the polar "water-loving" headgroups are able to form hydrogen bonds, they favor contact with water, whereas the "water-hating" nonpolar tails try to avoid water as much as possible. This "schizophrenic" nature of lipid molecules is the underlying reason why they self-assemble as bilayer-like structures, where two lipid monolayers face each other (see Fig. 7).

Although this idea of describing lipid membranes simply as lipid bilayers is very convenient, it is yet somewhat misleading because biological membranes are not like single-component lipid bilayers but, rather, are mixtures of various types of lipids that differ in a number of ways [222]. Some differences may seem minor, such as the length of hydrocarbon tails or the position of a double bond in the tails of monounsaturated lipids. Yet there are differences that make some lipids distinctly different from each other, such as the chemical composition and the net charge of the polar headgroup. In all, there are hundreds (or even thousands) of different lipid species that are found in biological membranes. In addition to this, to make molecular modeling more challenging, one should account for the fact that lipid membranes include numerous kinds of proteins either embedded in (integral proteins, ion channels) or attached peripherally to the membrane. The proteins are also related to a dynamic rubberlike network known as the cytoskeleton [222], which is attached to the inner surface of the membrane, whereas the outer leaflet of the membrane is covered by a network that in this case is made of glycocalyx carbohydrates. All together, for a physicist modeling soft matter systems, biological membranes are a mess.

Life would be much easier if biological membranes could be described by simple theoretical tools such as models related to the Ising model. Perhaps surprisingly, this approach has indeed been applied to cell membranes, and the outcome has been very successful [223–228]. Sarcastically speaking, it has been said that life can be described by the Ising model, too: Of the two states available in the traditional Ising model, one is then "alive" and the other is "dead." However, even though the Ising model and related approaches can be very useful in some cases, it is clear that it is a very coarse-grained description for complex systems such as lipid membranes. We cannot expect that approaches of that kind can provide any insight into the understanding of membrane systems over microscopic scales. In this nanoscopic regime, where atomic and molecular degrees of freedom cannot be neglected, more detailed approaches are needed to resolve the microscopic mechanisms that dictate the properties of given systems.

5.2. Molecular Dynamics of One- and Two-Component Lipid Membranes

Molecular dynamics provides a unique means to consider lipid membranes as well as other biologically relevant soft matter systems from an atomic perspective (see recent review articles in Refs. [84–88]). In particular, it allows one to gain insight into the nature of atomic-scale phenomena and related issues with a level of detail missing in any experimental technique. Nevertheless, MD also allows one to analyze membrane properties that are of experimental interest, thus providing a connection to macroscopic properties of the system.

As MD simulations of native biological membranes are not feasible, a natural starting point is to consider simpler models that include only the most essential features of cell membranes. This line of thought has lead to numerous studies of single-component lipid bilayers. Even though this approach may sound too simplified, it can yield a substantial amount of knowledge of model membrane systems, and as a matter of fact, the same idea has been employed in experiments.

First, MD simulations of single-component lipid bilayers were carried out in the 1980s and early 1990s [229, 230]. The timescales and system sizes were obviously very limited, but the early studies were already able to demonstrate the usefulness of MD simulations in this field. This is particularly true as regards the structure of lipid bilayers, which was then the key

aspect because of the relatively short timescales accessible through MD simulations (on the order of 1 ns at that time). More recently, and in particular during the last few years, better computer resources have made it possible to examine some of the dynamical aspects, too.

Another direction of research that has attracted plenty of activity is the study of two-component lipid bilayers. This is simply a result of the fact that there is a rather limited number of lipid species that are abundant in cell membranes. As an example, in human erythrocytes the most abundant lipid species are [220] cholesterol (25%), phosphatidylcholine (PC) (19%), phosphatidylethanolamine (PE) (18%), sphingomyelin (SM) (17.5%), and phosphatidylserine (PS) (8.5%). In human myelin membranes the abundant lipid species are [222] cholesterol (27%), ganglioside (26%), PC (10%), PE (20%), SM (8.5%), and PS (8.5%). Whereas PC, PE, and SM molecules are neutral but zwitterionic, PS is charged (anionic). For the discussion below, it is remarkable to note that cholesterol, PCs, and SMs together make up about 62% of the composition of human erythrocytes, and about 45% of human myelin membranes. Other highly concentrated molecules playing a major role in cell membranes are, for example, glycolipids, sugar-containing lipid molecules whose polar headgroups are attached to oligosaccharides. Thus, they play a prominent role in cell–cell interactions.

The most relevant case for all eucaryotic plasma membranes that do contain large amounts of cholesterol is a mixture of PC and cholesterol molecules. The role of cholesterol is particularly important, as cholesterol is considered as one of the main regulators of the fluid-like nature of membranes that is the key to their various dynamic properties. The importance of cholesterol is further emphasized by the recent view by Simons and Ikonen [231], who proposed the so-called raft model for biological membranes. The raft model is in part based on the (rather old) idea of lateral organization of heterogeneities in cellular membranes, and second, for an assumption that sphingolipids and cholesterol segregate in membranes to form microdomains termed rafts. Cholesterol seems to be crucial in the formation of rafts, which have been reported to be rich in sphingolipids, cholesterol, and saturated lipids. Recent works have proposed rafts to be enriched in some integral proteins as well as in GPI-anchored (glycosylphosphatidylinositol-anchored) proteins in the outer monolayer of a plasma membrane. As a consequence, rafts have been thought to act as platforms for adhesion and signaling, as they confine proteins involved in a signal transduction event into the same microdomain. Anyhow, despite many studies, unambiguous evidence for the existence of rafts is currently lacking. For a more thorough account of this topic the reader is referred to reviews in Refs. [232–235]. Nevertheless, without doubt, the role of cholesterol in lipid membranes is clear: There is a variety of reasons to consider cholesterol as one of the most important (lipid) molecules in cells.

Having said this, let us consider a PC/cholesterol bilayer system as an example to demonstrate how MD can be used to probe complex biologically relevant systems. As the space is limited, we focus on a few structural quantities that are closely related to experiments.

Figure 15 shows a snapshot of a DPPC/cholesterol bilayer after a 100-ns MD simulation [236]. To better understand the structure of the bilayer, mass density profiles along the normal direction of the membrane are shown in Fig. 16 [237]. Experimentally, information of the same kind can be obtained by, for example, x-ray diffraction measurements that yield electron density profiles of given bilayer systems [157].

The center of the bilayer is at $z = 0$; thus we find that the thickness of the membrane in this case is about 5 nm. Water penetrates rather deeply into the polar headgroup region of the bilayer, where the phosphate and choline groups of DPPC molecules (as well as the hydroxyl groups of cholesterol molecules) reside, whereas water does not favor the hydrophobic region characterized by the DPPC tails and the steroid ring structure of cholesterol. Rather, the cholesterol molecule prefers to locate itself such that the polar OH-group is in the vicinity of the water-membrane interface and the steroid structure is in close contact with the acyl chains of DPPC molecules. The density at the center of the membrane is rather small, implying that the free volume and free voids are presumably largest at this part of the system. One might then assume that the intermonolayer friction is very small. We are not aware of related experiments, however; thus, this idea should be taken as suggestive. What is definitely clear, however, is that the headgroup-water interface is the most dense part of

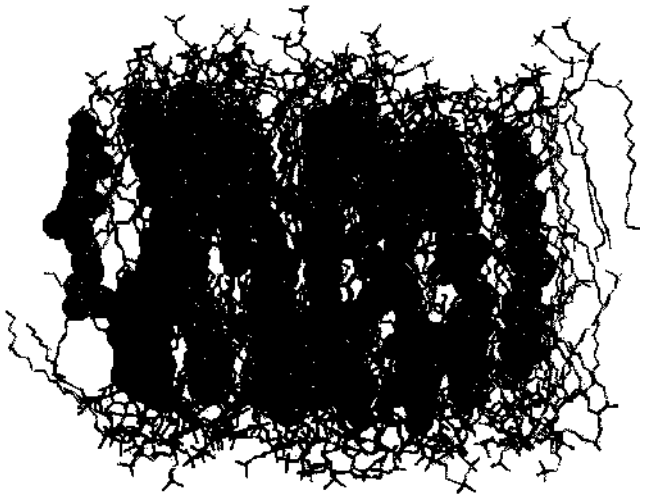


Figure 15. A snapshot of a lipid bilayer of dipalmitoylphosphatidylcholine and cholesterol molecules. The cholesterol mole fraction is 20%. For clarity, water is not shown.

the system. For one-component PC systems, the pressure across the membrane also seems to largest at the interface [180].

As the rigid steroid ring system of cholesterol is in close contact with the DPPC tails, one might expect that this has some influence on the ordering of lipid tails. This is precisely what happens as is depicted in Fig. 17. There, the order parameter S_n describes the ordering of the hydrocarbon tails in DPPC molecules following a definition based on the second order Legendre polynomial. In this case, $S_n = \frac{1}{2}\langle 3\cos^2\theta_n - 1\rangle$, where θ_n is the angle between the bilayer normal and the orientation of the vector along a C–H bond of the nth carbon atom of the hydrocarbon chain [86]. The same order parameter is commonly adapted to studies of liquid crystals. In membrane systems, S_n can be determined through nuclear magnetic resonance measurements. In this case it suffices to mention that $S_n = 0.5$ corresponds to full ordering (where acyl chains are full-trans—like a "zig-zag" conformation), whereas $S_n = 0$ corresponds to a disordered chain, in which case the hydrocarbon chain acts like a random walker. The main finding here is that for this system in the fluid-like state, above the so-called main phase transition temperature, cholesterol orders the acyl chains of

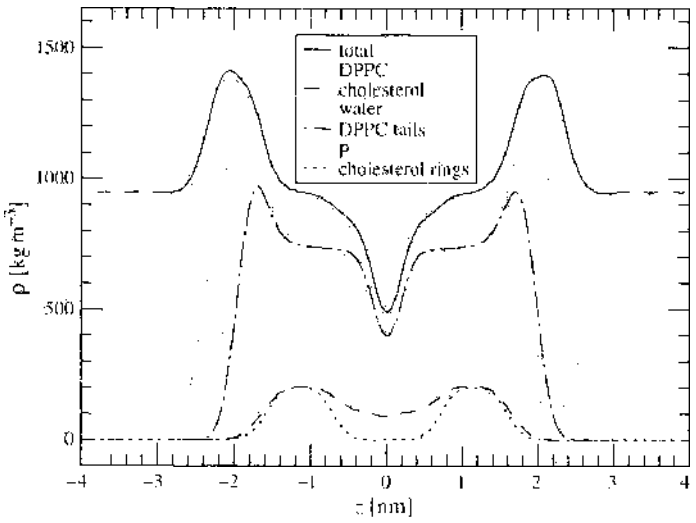


Figure 16. Mass density profiles of various structural components in a dipalmitoylphosphatidylcholine (DPPC)/cholesterol bilayer with a cholesterol mole fraction of 20% [237]. The results are shown across the bilayer in the z-direction of membrane normal. The case $z = 0$ corresponds to the center of the lipid bilayer. The different curves represent total mass density (full black), DPPC (full gray), cholesterol (dashed black), water (dashed gray), DPPC tails (dash-dotted black), phosphor atoms (P) (dash-dotted gray), and cholesterol rings (dotted black).

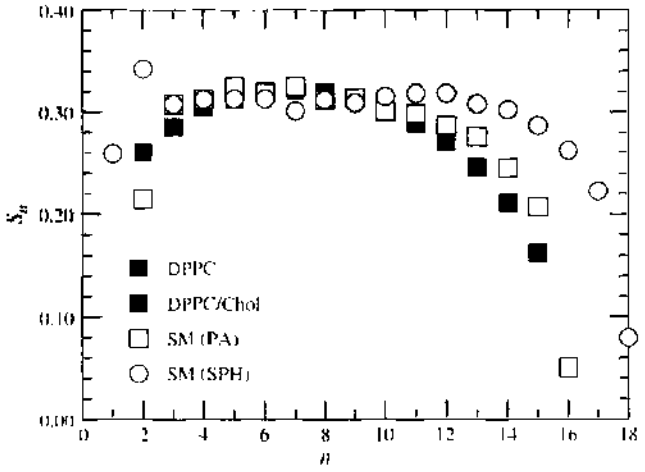


Figure 17. Order parameter S_n of the hydrocarbon chains of lipid molecules. Small n corresponds to carbons (methylene groups) close to the headgroup, whereas large n corresponds to carbons close to the center of the bilayer. In addition to the pure DPPC system [175], results are shown for a DPPC/cholesterol system at a cholesterol concentration of 20% [236] and for a pure SM bilayer (data taken from Ref. [248]). In the case of SM, the order parameters along both the sphingosine (SPH, 18:1) and palmitoyl (PA, 16:0) chains are given.

DPPC molecules. As recent MD simulations for lipid bilayer mixtures of glycerophospholipids and cholesterol have revealed, the enhancement of acyl chain ordering resulting from an increasing cholesterol concentration is related to a number of intriguing phenomena in membranes [236–246]. For example, as the cholesterol concentration is increased, the area per molecule in the plane of the bilayer decreases, the permeation of molecules across the bilayer is affected because of changes in the free energy profile, the likelyhood of gauche conformations of lipid acyl chains are affected considerably, the lateral diffusion of lipids and cholesterol molecules is reduced, and the bilayer in general becomes more rigid. These observations are in good agreement with experimental findings.

As far as other lipid membrane systems are concerned, the key issue is that MD simulations provide a versatile and essentially unique tool to analyze membrane properties in atomic detail. Basically, all possible issues (excluding electronic degrees of freedom) can be determined in full detail. In addition, as experimental techniques typically consider macroscopic aspects of the system, or the averages taken in experimental measurements are global ones, MD simulations allow one to gain a much deeper understanding of the system. In addition to global averages, one can calculate local properties, such as the influence of cholesterol on the ordering of phospholipids that are neighbors to the given cholesterol [244]. One can also follow the trajectories of individual molecules and signaling events in time, and thus, for example, resolve the diffusion mechanisms and diffusion pathways of molecules migrating inside lipid bilayers. The work by Söderhäll and Laaksonen of the diffusion of ubiquinone inside a DMPC bilayer provides an excellent recent example of this issue [247]. Now, as basically every atomic detail is within reach, does this imply that MD is the method of choice without any limits? It does not. There are two major issues that limit the use of molecular dynamics for simulations of biomolecular systems. First, the development of force fields together with rigorous testing to confirm their validity, and second, the resources. The first of these is a general issue with which one has to learn to live—it is and it will be part of our lives in all cases in which new models are being designed and tested. The second one will get better in time as the development of algorithms and the progress in computer technology permits computer simulations of more and more challenging problems.

Now, let us come back to the idea of lipid rafts. The second major component in rafts are sphingolipids, and SM in particular. It is therefore rather surprising that there are just a few studies that have explored SM systems through MD simulations [248–251]. In essence, one has found that SM is distinctly different from other lipids such as PCs in the sense that its hydrogen-bonding capacity is substantial. Unlike PCs, for example, SM molecules are able to act both as a donor and as an acceptor, thus allowing strong intramolecular as well as intermolecular hydrogen bonding. This leads to compact SM bilayers in which the area per

molecule in the plane of the membrane is much smaller than in PC systems, which in turn is expected to play a key role in various dynamical processes such as lateral and rotational diffusion as well as permeation of molecules across SM bilayers. Further, the small area per molecule and the concominant reduction of free volume inside SM bilayers is manifested in the ordering of SM hydrocarbon tails. As shown in Fig. 17, the order parameter $S_{\parallel}$ for a pure SM bilayer is about 50% larger than that for a pure DPPC bilayer, and approximately of the same magnitude as the order parameter for a DPPC/cholesterol mixture with a cholesterol content of 20%.

As there are no published computational studies of lipid bilayers with both cholesterol and SM (at the time of writing the present article), essentially all relevant questions with regard to the atomic perspective of lipid raft systems remain to be posed. We are looking forward to future work that hopefully will address these issues, together with many other questions related to many-component lipid membrane systems in general.

5.3. Recent Studies of Lipid Membranes Through Molecular Dynamics

Developments in the field of MD simulations of model lipid bilayer systems have been discussed in a number of rather recent review articles [84–88]. In this work, we therefore focus on the advances made in the last few years.

Most studies in the field of model membranes have concentrated on the "high-temperature case" characterized by the fluid-like L_α phase above the main phase transition temperature. Much less attention has been paid to the low-temperature gel phase characterized by ordering in the plane of the membrane, as well as conformational ordering of the lipid acyl chains. The studies by Venable et al. [178] and Snyder et al. [252] provide an exception to the rule. Although certain observations such as tilting of lipid acyl chains and reduced area per molecule can be made in agreement with experimental data, a detailed and solid analysis of the low-temperature phase is very difficult to perform because of the dynamics, which are very slow in this phase.

Rather extensive MD investigations over the years have lead to a good understanding of bilayers composed of saturated lipids, whereas the understanding of the properties of bilayers composed of unsaturated lipids has been much more limited. Recently, though, there has been increasing interest in investigating these systems [253–255]. This is simply because unsaturated lipids, and polyunsaturated lipids in particular, are an essential component of some cellular membranes. For example, the cell membrane of the nervous system is known to contain a large fraction of polyunsaturated docosahexaenoic, a fatty acid with six double bonds. Polyunsaturated lipids are further thought to play a key role in neurological dysfunctions such as Parkinson's disease and in diseases such as atherosclerosis and cancer. Although there is a long way to go until the properties of these systems are understood in detail, MD studies have already demonstrated that bilayers comprising polyunsaturated lipids are distinctly different from bilayers composed of saturated lipids. Among others, they have shown that polyunsaturated lipids in bilayers are highly flexible and characterized by a high degree of conformational disorder. As a consequence, as these findings imply that the influence of polyunsaturated lipids on the structure within a bilayer is considerable, it has been suggested that polyunsaturated lipids are particularly relevant for the proper functioning of membrane proteins [253].

Phospholipid (Langmuir) monolayers are rather commonly used as model systems of lipid bilayers. This approach is definitely attractive, as the monolayer, being half of a membrane, is a well-defined planar system that is easier to control than model vesicles and liposomes. As a consequence, it has been adopted to a wide range of studies, including intermolecular interactions between various molecules in monolayers, drug delivery issues, and penetration of peptides into lipid monolayers [256, 257]. The actual correspondence between quantities measured for bilayers and monolayers is not often clear, however, which may render interpretation of the data. In a recent study, Kaznessis et al. [258] employed MD simulations to investigate lipid monolayers at the water/air interface and found evidence that some structural quantities can indeed be different from their counterparts in lipid bilayers. In particular,

they observed that the order parameter S_n for a DPPC monolayer was essentially constant ($S_n \approx 0.2$) for carbons 2–14. In DPPC bilayers the behavior of S_n is markedly different (see Fig. 17). Further studies would be highly welcome to clarify the monolayer versus bilayer issue in more detail.

Biological membranes at physiological pH are surrounded by an aqueous buffer containing ions such as Na^+, Ca^{2+}, and Cl^-. At the same time, about 10% of the lipids in biomembranes are charged (typically anionic), which implies that their charge has to be compensated by counterions such as Na^+. Recent MD studies [259–261] have thus aimed to clarifing the effect of salt on lipid bilayers. As Boeckmann et al. demonstrated very clearly [259], this task is computationally very demanding. This is because relaxation times associated with the diffusion of ions are very large, which in turn implies that the timescales required for equilibration and gathering sufficient statistics are substantial (on the order of 100 ns or more). Nevertheless, recent attempts have shown that salt plays a major role in determining the structural and electrostatic properties of lipid bilayers and has a major effect on the lateral diffusion of lipids in a bilayer.

In addition to the above studies for pure lipid bilayers, we would like to mention recent work on cationic lipid bilayers [151, 262] because of their relevance in gene therapy, drug delivery, and specifically on DNA–membrane interactions. Some years ago, Bandyopadhyay et al. performed an MD study of a lipid mixture of neutral (zwitterionic) dimyristoylphosphatidylcholine (DMPC) and cationic dimyristoyltrimethylammonium propane (DMTAP) in the presence of a short DNA fragment [151]. It took several years until this very elegant piece of work was complemented by a second study focusing on cationic bilayers. Gurtovenko et al. [262] studied DMPC/DMTAP lipid bilayer mixtures over a wide concentration range of DMTAP and showed that there is a strong interplay between positively charged TAP headgroups and the zwitterionic PC heads. This interplay was shown to lead to considerable reorientation of PC headgroups for an increasing DMTAP concentration, in agreement with experimental data. It is likely that these observations play a significant role in the condensation of DNA onto the membrane surface.

Biomembranes are semipermeable barriers that control the flow of ions and other molecular species into and out of the cell. As a consequence, the permeation of small molecules such as water and ions has been studied rather extensively during the last decade. Recently, more ambitious attempts have been made to study the permeation and distribution of more complex molecules such as sugars, amphiphilic drugs, and anesthetic molecules in lipid bilayers [263–267]. It is evident that the influence of the compound on the structure of lipid bilayers depends on the compound in question. Thus, there is currently no point in drawing any general conclusions. Rather, we would like to emphasize the challenge associated with the modeling of these processes; the study of the distribution of compounds in membranes is in general hindered by the limited timescale accessible through MD simulations.

The same problem is encountered in studies of enzyme–membrane and protein–membrane systems because of large relaxation times associated with the conformational changes of the molecules. Nevertheless, some successful attempts have been made recently to better understand the action and function of different kinds of enzymes and protein molecules in membranes. For example, lipases are enzymes involved in a variety of processes, such as signal transduction and the hydrolysis of some bonds in phospholipids. MD simulations have been able to shed light on the influence of these molecules on the structure of lipid bilayers [268] and on the orientation and conformations of certain enzymes at the water–membrane interface [269]. As far as integral proteins are concerned, let us here mention just a few demonstrative examples of recent activities in the field. Sansom et al. [270, 271] have investigated potassium ion channels that enable rapid movement of K^+ ions passively across a membrane, and Baudry et al. [272] have featured recent progress in the modeling of light-driven proton pumps in the purple membrane. Perhaps the most topical issue, however, is the work based on MD studies of water channels [54, 55]. This example is related to the Nobel Prize in Chemistry 2003 and highlights the remarkable fact of how experiments and computer simulation studies can indeed complement each other.

5.4. From Atomistic to Coarse-Grained Molecular Descriptions of Lipid Membranes

In this chapter, we have in many places faced a fact that biological systems are characterized by a variety of different length and timescales. For the shortest ones, there are indeed techniques based on quantum mechanics and atomic-scale MD that describe systems in full atomic detail. As a consequence, they are obvious techniques for studies of microscopic properties. However, these approaches are not feasible for cases in which the length scales are larger than ~10 nm and the timescales are beyond ~100–1000 ns.

What we can do under these circumstances is to replace the atomic-level description of the system with a less detailed one. (See discussion in Section 3.) In practice, one still employs MD simulations for a system of interacting particles, using precisely the same framework as in atomic-level MD. The only difference is the molecular description and the choice of interactions. Although the above discussion has been based on an atomic view, in coarse-grained MD the molecular description does not account for atomic details. Rather, the particles describe clusters of atoms rather than individual ones.

One example of this approach is the recent work by Marrink et al. [273]. They used the Gromacs simulation package (first developed for atomic MD) to study membranes comprised of lipid molecules. The particles in this description did not represent atoms, however. Rather, they represented clusters of methyl groups or clusters of atoms in the polar headgroup, thus reducing the computational complexity of the model significantly. Goetz et al. [274, 275], Shelley et al. [276–278], and Kranenburg et al. [120, 121] have recently pursued ideas of a similar nature. Another approach is to calculate material properties from atomic-scale MD simulations and to then use them to parameterize mesoscale and macroscale models. Ayton et al. have recently applied this idea to study elastic properties of membrane systems beyond the atomic regime, including both mesoscopic [279] and macroscopic levels [280–283].

Overall, the key word in these simplified descriptions is coarse graining, which takes us higher in the hierarchy of molecular simulations and allows one to design and study mesoscopic, and possibly even macroscopic, models for biological membranes and other soft matter systems.

6. SIMULATIONS OF SOFT MATTER SYSTEMS OVER MESOSCOPIC SCALES: COARSE-GRAINING OF IONIC SOLUTIONS

Thus far we have discussed molecular level modeling in detail and have reviewed some of the recent developments in multiscale modeling in soft matter and biological systems. Further, we have discussed some practical aspects such as the importance of high-quality random numbers in numerical simulations and the use of MD simulations for microscopic studies of biophysical systems. Here, we take one step further and discuss how soft matter systems can be coarse grained and modeled over scales far larger than those accessible through atomic-level MD. More specifically, we use one particular coarse-graining method, namely the IMC method (see Section 3.1) to combine micro- and mesoscale approaches to demonstrate coarse-graining in practice. We discuss the merits and problems of this method and put it into a more general context. The method we have chosen is applied to an ionic solution. Despite its apparent simplicity, this system, consisting of water and NaCl only, is nontrivial from both a practical and a theoretical point of view.

6.1. Coarse-Graining Strategy

The approach we present here is conceptually simple and systematic, combining classical molecular dynamics (MD) and a momentum conserving DPD thermostat (i.e., we obtain the conservative potentials through a coarse-graining procedure from MD simulations and employ DPD as a thermostat in a manner discussed in Section 3.2. Figure 18 shows an illustration of the original and coarse-grained systems.

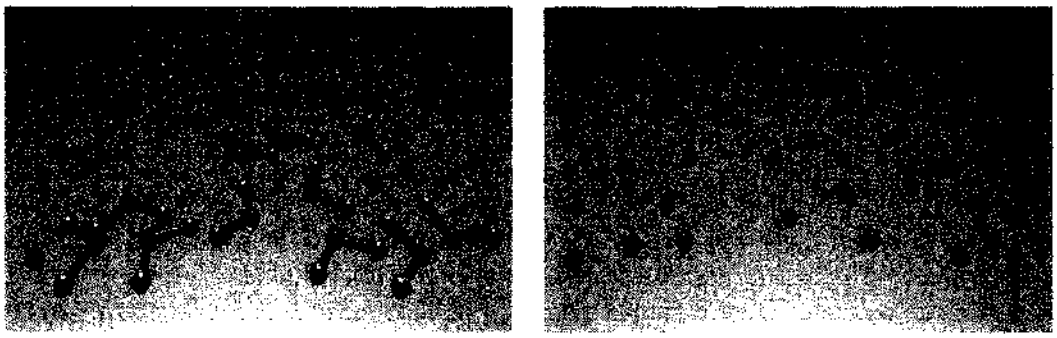


Figure 18. The coarse-grained and the underlying molecular system. Left: aqueous NaCl with atomic detail. Right: the coarse-grained model in which waters have been replaced with just one particle.

This method has several advantages, as will be discussed below, and it consists of three independent parts: first, MD simulations are carried out to obtain radial distribution functions $g(r)$ between different atoms, molecules, or molecular groups; second, the IMC procedure (see Section 3.1 for details) is applied to invert the radial distribution functions to obtain effective interaction potentials $V^{\text{eff}}(r)$ between the new interaction sites; and finally, the effective potentials are used within the DPD algorithm (see Section 3.2 for details) to study the long time- and length scale properties with full hydrodynamics for the new coarse-grained system.

One of the main advantages of this approach is that it allows easy tuning of the level of coarse-graining while preserving the essential molecular information. It also provides a well-defined and systematic coupling between the microscopic and coarse-grained systems in the sense that the pair correlation properties of the two descriptions are consistent. Although we have not checked that, this indicates that the coarse-graining procedure proposed here also preserves the essential information of phase behavior. Furthermore, this method speeds up simulation times considerably, thus making it possible to examine large-scale properties of complex molecular systems with full hydrodynamics. Finally, it allows one to study the limits of coarse-graining (i.e., how coarse-graining depends on physically relevant variables such as the density, temperature, pressure, and salt concentration).

6.2. Obtaining the Interaction Potentials

Let us first validate the method against the underlying microscopic MD simulations. To be self-consistent, the coarse-grained potentials should produce the same pair correlation functions. It is important to note that coarse-graining of even simple liquids is not a simple feat, especially if one wants to reproduce both static and dynamic properties. This is where the approach and results presented here are particularly valuable.

For the underlying microscopic MD simulations, we used the flexible SPC water model [284] and the so-called Smith–Dang parameters for Na^+ and Cl^- ions [285]: that is, the Lennard–Jones parameters for the sodium ions were $\sigma = 2.35$ Å and $\varepsilon = 0.544$ kJ/mol, and for chloride $\sigma = 4.4$ Å and $\varepsilon = 0.419$ kJ/mol. The temperature was set to 300 K, the simulations were done in the NVT ensemble, and we used a salt concentration of 0.87 M. Electrostatic interactions were computed using the Ewald summation. The full set of simulation parameters are given in [109, 286]. As the central quantity, the radial distribution functions between all different pairs of particles were computed from the MD simulations. They were used as an input in the IMC procedure (described in Section 3.1). The new coarse-grained interaction potentials are shown in Fig. 19. The potentials are not as "soft" as the DPD potentials in Fig. 3 but have extra features and a relatively hard core. That simply reflects the degree of coarse-graining and the fact that this method preserves the identity of the coarse-grained entities.

As a little digression a short comment regarding the model should be said. Whereas there are several well established and extensively studied models for water (see, e.g., Ref. [287] and references therein), flexible SPC being one of them, that is not the case for NaCl. The chosen Smith–Dang parameterization is hand optimized for NaCl and reproduces some of the

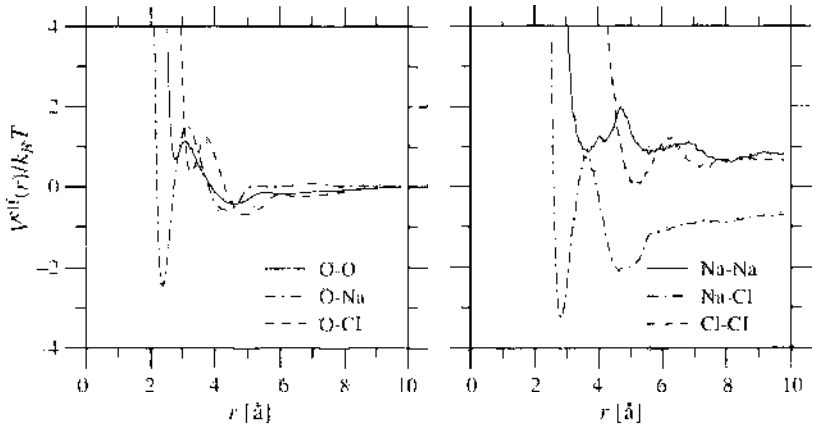


Figure 19. The effective interaction $V^{eff}(r)$ versus the radial distance r for all pairs of particles as obtained from the radial distribution functions of MD simulations at 0.87 M salt concentration. "O" refers to the oxygen atom in a water molecule.

properties well. In a recent study [288], we compared a number of force fields for simulations of NaCl in aqueous solutions and found that there are some important differences between them. We are pointing this out because one should always validate the coarse-grained results against the particular model used to obtain the new potentials.

A comparison of the effective interactions to the potentials of mean force revealed [288] that the overall structures of the potentials of mean force were roughly similar to those of the effective interactions—as expected—but there were significant differences in the form and amplitudes of some of the peaks. Furthermore, it is important to notice that the long-range tails of $V^{eff}(r)$ include a contribution from Coulombic forces, whereas the tails of the potentials of mean force do not (i.e., they decay rapidly to 0 for small r, resembling the behavior typical for the commonly used Derjaguin–Landau–Verwey–Overbeek [DLVO] theory for colloids; see, e.g., Ref. [289]). From the comparison with the results in Ref. [288], we conclude that the potential of mean force serves as a reasonable approximation of $V^{eff}(r)$ but is distinctly different quantitatively. This difference becomes more pronounced for charged particles for which even the qualitative behavior may be different.

At this point, we would like to stress the following six features:

1. All DPD calculations that will be discussed below were carried out with the $V^{eff}(r)$ obtained from the MD simulations at 0.87 M salt concentration. As was noticed previously [286], the effective potentials are rather weakly affected by a change of salt concentration (i.e., it is not necessary to recompute the effective interactions for all different salt concentrations).
2. We used a cutoff $r_c = 9.6$ Å in all the cases. This was done to enhance the efficiency of the model, and it did not give rise to artifacts in, for example, $g(r)$, as explicit tests confirmed. If larger cutoffs are desired, one can either compute $V^{eff}(r)$ over a larger interparticle distance or approximate the long-distance part by a Coulombic tail [109].
3. The electrostatic interactions that were explicitly accounted for in the MD simulation are now implicitly included in the effective interactions. This is reflected in the shape of the potentials.
4. DPD simulations were carried out with coarse-grained water molecules, each of them represented by a single spherical particle. The dynamics resulting from hydrogen bonds are implicitly described by the effective interactions.
5. The $V^{eff}(r)$ have a relatively soft core: that is, the potentials shown in Fig. 19 were extrapolated to a finite value as $r \to 0$.
6. The weight functions were chosen as to be of the standard form $\omega^R(r) = [1 - r/r_c]$ (see Section 3.2). The dissipation strength γ [see Eq. (13)] was determined such that the decay rate of the single-particle velocity correlation function at early times was approximately identical in MD and DPD simulations. That yielded $\gamma = 0.72$, which was used in simulations. Note that γ is the only fitting parameter in this model.

Whereas the first five items are straightforward, the last item is the critical step that connects the DPD thermostat to the MD simulations through the γ parameter. It is important to understand that there is no unique way to fix the strength of the dissipative force γ. Here, we adjusted the rate of the dynamics by finding an optimal γ in the sense that the initial decay rate of the single-particle velocity correlation function,

$$\phi(t) \equiv \langle \vec{v}_i(t) \cdot \vec{v}_i(0) \rangle \tag{36}$$

(separately for Na^+, Cl^-, and H_2O), was approximately similar in MD and DPD calculations at 0.87 M. This way, γ does not fix the long-time decay of $\phi(t)$s, and hence does not fix the diffusion coefficients. This scaling is in agreement with the pair collision theory [290]. The single-particle velocity autocorrelation functions are shown in Fig. 20.

6.3. Comparison Between Molecular Dynamics and Coarse-Grained Simulations

To check the consistency of our coarse-grained approach with the MD simulations, we first compared the radial distribution functions $g(r)$ of the two approaches. As can be seen in Fig. 21, the results for the $g(r)$s are in excellent quantitative agreement. In addition to this, the coordination numbers [219] were found to be in perfect agreement with those measured in the MD simulations. This proves that the method is self-consistent, at least in terms of static properties.

Even when static properties are well reproduced, there is no reason for the dynamics of the coarse-graing system to resemble the true dynamics of the system [32]. To test the dynamical properties of the system, we measured the tracer diffusion coefficient D using the Einstein relation for mean-square displacements,

$$D = \lim_{t \to \infty} \frac{1}{6t} \langle [\vec{r}_i(t) - \vec{r}_i(0)]^2 \rangle \tag{37}$$

which can provide us with information of possible problems on the dynamics of the system. Here $\vec{r}_i(t)$ is the position of a tagged particle at time t. To study dynamics even further, we have also computed the shear viscosity by a Green–Kubo relation [219]. Dynamical properties are a different matter, and there is no reason to expect, *a priori*, for them to match. This is particularly so in systems in which water and charges, or polarizable particles, are present, as hydrogen bonds are directional.

The results for tracer diffusion are presented in Fig. 22. It should be noted that in all of the coarse-grained DPD, we used the potentials obtained at 0.87 M salt concentration and shown in Fig. 19. Thus, this also tests the applicability of the potentials.

From Fig. 22, it is clear that within the statistical error the qualitative behavior of the diffusion coefficients versus the salt concentration is virtually identical in MD and DPD

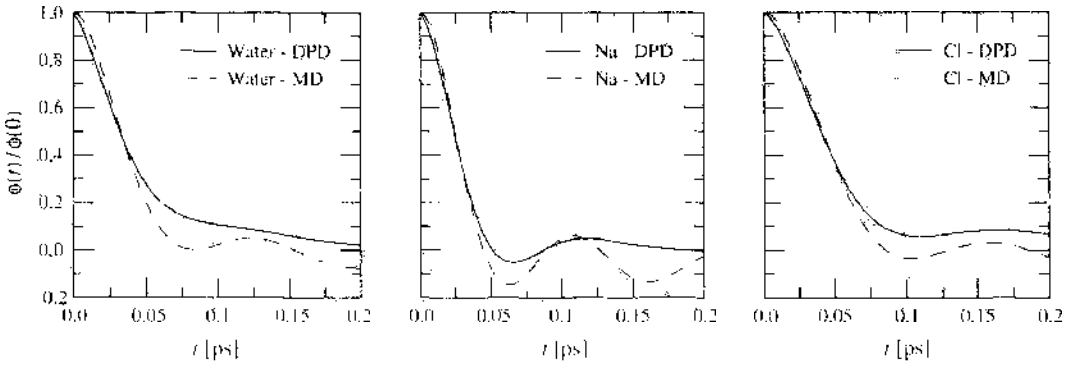


Figure 20. The decay of the single-particle velocity autocorrelation function $\phi(t)$ at early times. Results shown here are for water, Na^+, and Cl^- ions. As the data illustrates, the early-time decay of $\phi(t)$ is essentially identical between molecular dynamics and dissipative particle dynamics (DPD) simulations for $\gamma = 0.72$ used in the DPD simulations. Adapted from Ref. [219].

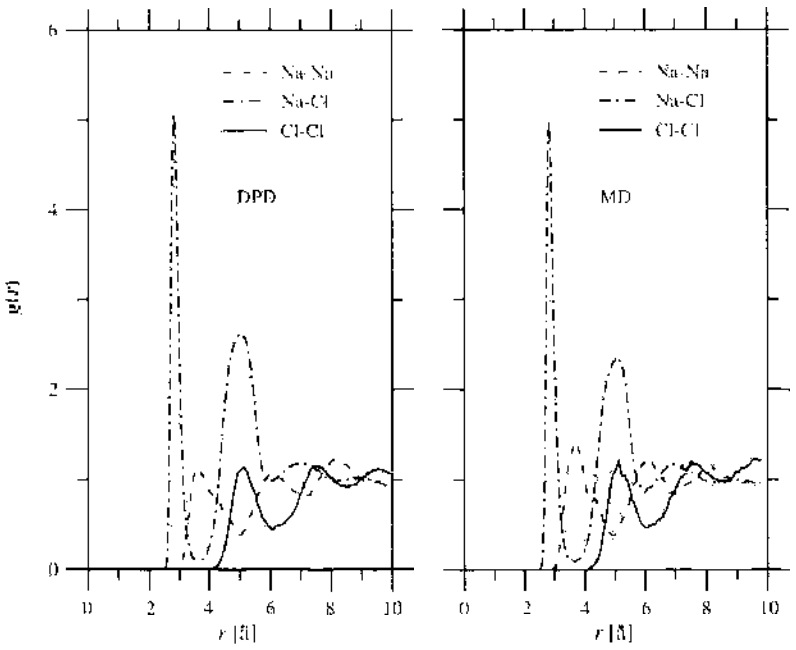


Figure 21. The radial distribution functions obtained from the coarse-grained simulation using the DPD thermostat with the effective potentials (left), and the $g(r)$ obtained from the microscopic MD simulations (right). The agreement between the two approaches is remarkably good.

simulations. Even at lower molarities, the deviations are small and the respective tracer diffusion coefficients for Na^+, Cl^-, and H_2O have the same overall behavior in both cases. The increasing deviations at low molarities are the result of the growing screening length. Furthermore, the diffusion coefficients are in good agreement with Langevin simulations (using the generalized Langevin equation formalism) and experimental results [291]. The agreement with experiments is mainly the result of MD simulations. The true test for our DPD-based approach is the comparison against the MD simulations that were used to obtain the effective potentials. As the comparison clearly shows, the coarse-grained method performs very well indeed.

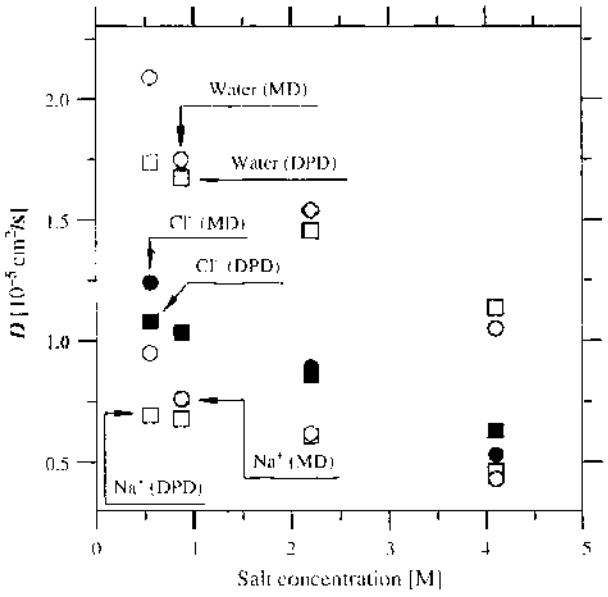


Figure 22. Comparison of the tracer diffusion coefficients obtained from molecular dynamics and dissipative particle dynamics (DPD) simulations. The DPD simulations are based on effective interactions extracted from the MD studies at 0.87 M. These effective interactions were used in all consequent DPD simulations with a varying salt concentration. DPD data has been scaled to allow comparison. Error bars are about the size of the symbols.

Our initial studies for the temperature dependence of D through DPD simulations further support the consistency of this approach, as we found that the effective diffusion barriers, obtained through an Arrhenius analysis of D around 300 K at 1 M, of H_2O and Cl^- were essentially identical, whereas the barrier of Na^+ was about 15% larger, in agreement with MD simulations. In our previous study [219], we also studied the behavior of the shear viscosity coefficient. It was found that it increases monotonously for an increasing salt concentration, in agreement with experiments [292]. The shear viscosity coefficient was not determined from MD simulations, as it would have been too time consuming.

6.4. Computational Aspects

In addition to the above physical aspects, there are important computational gains as well. The coarse-graining procedure, even at this very modest level, led to a speed-up of the order of 20, as the CPU times for MD and DPD simulations were almost identical but the particle numbers studied by MD and DPD were about 300 and 7000, respectively (at the same densities and simulation times of about 2 ns). As a result of the coarse-graining procedure, the fast internal and orientational degrees of freedom related to water have been integrated out leading to the speed-up.

We expect that in more complex systems where the molecular description can be coarse-grained more than here, leading to a molecular description with a smaller number of particles and softer interparticle interactions, the speedup will be significantly larger. A further important aspect is that once the MD simulation for obtaining the potentials has been done, the stored equilibrium configurations can be reused to obtain new effective potentials; that is, between different groupings of molecules. This provides a straightforward approach to study the effect of the level of coarse-graining on the properties of the system. In the case of lipid molecules, for example, such studies may reveal the essential interaction centers (such as molecular groups with a finite charge) and the key features of molecular structure that are crucial for the behavior of the system. Obviously, these features should then be accounted for in the coarse-grained model. Further, it is noteworthy that this methodology can be applied to a wide range of scales starting from *ab initio* simulations.

6.5. Discussion and Relation to Other Methods

In this section we have discussed in detail the linking of classical molecular dynamics simulations to a higher-level description via the IMC method [109]. We chose to have only a very modest level of coarse-graining to be able to better validate our results against the microscopic MD simulations.

The agreement between the present coarse-grained approach with the MD results and experiments clearly demonstrates that the method itself has a sound physical foundation as well as potential for further applications. In particular, despite the coarse-graining of both the particles and the interactions, the results show that the method retains the characteristics of different molecules, or molecular groups, and the essential aspects of interparticle interactions. Furthermore, although the level of coarse-graining used here has been quite modest, the results do demonstrate that the method can reduce the computational load significantly as compared to detailed molecular models.

Here, the computational gain was a factor of 20. This may sound modest but one should keep in mind that, unlike the MD simulations, the mesoscopic simulations were performed using an unoptimized simulation code and that the factor of 20 alone would be able to push us to the microsecond scale to enable studies of, for example, dynamic properties. Thus, this approach allows one to study soft matter systems over mesoscopic scales beyond the atomistic regime. It is also noteworthy that as DPD satisfies momentum conservation, this approach allows studies of a given system with full hydrodynamics that is essential in various soft matter processes. Thus, we feel that the approach presented here provides a good starting point and a platform for further studies of macromolecular systems and of the general question of bridging different time and length scales.

An important and general question regards the limitations of coarse-graining, and how much does it depend on physically relevant variables such as density, temperature, pressure,

and salt concentration. Although we lack general knowledge, the results presented here indicate that it is not necessary to obtain the effective potentials for each condition studied. This idea is supported by further ongoing studies of lipid systems (DMPC) and various solutions (LiCl, $CaCl_2$), in which case we have studied the effects of varying density and temperature. The results from these simulations are in favor of our conclusions about the generality of this approach. One of the main issues in regard to charge carrying systems is to check that the coarse-grained systems actually obey the Stillinger–Lovett sum rules [89] and thus charge conservation. Work is in progress to address that question.

We are currently applying this method to more complex biomolecular systems of amphiphilic lipid molecules in explicit water, with a stronger focus on the coarse-graining. Macromolecular systems pose new conceptual problems, and lipids, because of their amphiphilic nature, are particularly challenging. For this reason, at the moment no systematic method for coarse-graining lipids exists. We are currently extending the hierarchical approach presented here to lipids as well.

Finally, we would like to note that the method discussed here bears a very close resemblance to the approach by Müller–Plathe and coworkers in the context of polymeric systems. Elegant reviews on coarse-graining of polymeric systems are provided in Refs. [29, 32]. As far as the coarse-graining of lipid systems is concerned, the work of Shelley et al. presents the current state-of-the-art concerning this issue [276].

7. CONCLUDING REMARKS

Liquid crystals, paints, polymers, clay, foam, complex fluids, and surfactants share many things in common. Most important, these materials are distinctly different from liquids and solid condensed matter systems for which reason an idea has emerged that they constitute a new class of materials called *soft matter*. What is common to soft matter systems is that they are all characterized by weak interactions of the order of thermal energy. Because $(k_B T) \approx 26$ meV, the energy scale associated with soft matter is well below typical interaction energies in condensed matter systems, where they are on the order of 1 eV. Second, what is peculiar to soft matter is a multitude of fascinating phenomena, many of which deal with biological issues and are therefore related to living matter and life sciences. This point is truly challenging because biological systems in particular are characterized by extremely wide length and timescales, ranging from femtoseconds to several days and from nanometers to several meters.

Without doubt, our understanding of soft matter and biological systems is largely based on experiments. Nevertheless, it is a plain fact that experiments cannot live on their own. First of all, the resolution of experimental approaches in time and space is often limited, which implies that there are always issues that cannot be probed by experiments. The interpretation of experimental results is further always based on some model, thus stressing the importance of theoretical work that can provide one with an understanding of the laws of nature that govern these systems. Analytical theory, however, is also unable to live on its own. In addition to the models whose underlying assumptions can be questioned, the analytical approach is troubled by the fact that analytical calculations are often based on approximations whose validity may be difficult to assess. To resolve these problems, and concurrently to bridge theory and experiments more closely together, it has become common to carry out computer simulations based on well-defined theoretical models.

The models used in computer simulations can be quantum mechanical, and thus be able to address questions where electronic degrees of freedom are important. However, if electronic degrees of freedom are not crucial, a somewhat simpler approach using classical molecular dynamics in atomic level is perhaps more appropriate. For phenomena that take place over much larger scales, even more coarse-grained approaches are needed, leading us to mesoscale and macroscale simulation approaches. Thus, the fact that we are dealing with a variety of different time- and length scales in biological systems implies that one needs to develop and employ multiscale modeling techniques, an approach in which different kinds of computer simulation methods are bridged together through coarse-graining approaches.

In this review, our aim has been to present an overview of the essential underlying ideas related to molecular modeling, molecular dynamics simulations in atomic detail, and mesoscopic simulation methods probing systems beyond the atomic regime. In addition, we have discussed coarse-graining techniques whose aim is to reduce the number of degrees of freedom and consequently to yield less detailed models for studies of large-scale properties. Although this overview is focused on some techniques that we have found most appropriate for the present case, we hope that it has been able to highlight the challenges and future prospects related to the field.

We consider it a fact that molecular modeling in general is a unique tool that can provide a great deal of insight into both the microscopic and the large-scale properties of biologically relevant soft matter systems. The numerous applications in biosciences as well as in technologically related soft matter fields are some of the reasons that justify and motivate this work. Yet the challenge related to understanding such extremely rich problems, many of which are of fundamental nature, provides another and even a more important reason.

ACKNOWLEDGMENTS

First and foremost, we wish to thank all our collaborators, and in particular Alexander Lyubartsev and Aatto Laaksonen for fruitful discussions over the years. In addition, we would like to acknowledge financial support from the Academy of Finland (Grant Nos. 54113 [MK] and 80246 [IV]), the Academy of Finland Center of Excellence Program (IV), and the Finnish IT Center for Science and the HorseShoe (DCSC) supercluster computing facility at the University of Southern Denmark for computer resources. Finally, we are grateful to Emma Falck, Petri Nikunen, Michael Patra, and Tuomas Torsti for their help in preparing many of the figures shown in this article.

REFERENCES

1. M. Daoud and C. E. Williams, Eds., "Soft Matter Physics." Springer, Berlin, 1999.
2. H. Flyvjberg, J. Hertz, M. H. Jensen, O. G. Mouritsen, and K. Sneppen, Eds., "Physics of Biological Systems: From Molecules to Species." Springer, Berlin, 1997.
3. M. E. Cates and M. R. Evans, Eds., "Soft and Fragile Matter: Nonequilibrium Dynamics, Metastability and Flow." Institute of Physics Publishing, Bristol, 2000.
4. B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson, "Molecular Biology of the Cell," 3rd ed. Garland, New York, 1994.
5. C. Branden and J. Tooze, "Introduction to Protein Structure," 2nd ed., Garland, New York, 1999.
6. R. L. C. Akkermans and W. J. Briels, *J. Chem. Phys.* 113, 6409 (2000).
7. P. Español and F. Vázquez, *Phil. Trans. R. Soc. Lond. A* 360, 1 (2002).
8. E. G. Flekkøy, P. V. Coveney, and G. De Fabritiis, *Phys. Rev. E* 62, 2140 (2000).
9. B. M. Forrest and U. W. Suter, *J. Chem. Phys.* 102, 7256 (1995).
10. R. D. Groot and K. L. Rabone, *Biophys. J.* 81, 725 (2001).
11. R. D. Groot and P. B. Warren, *J. Chem. Phys.* 107, 4423 (1997).
12. A. Malevanets and R. Kapral, *J. Chem. Phys.* 110, 8605 (1999).
13. A. Malevanets and R. Kapral, *J. Chem. Phys.* 112, 7260 (2000).
14. M. Murat and K. Kremer, *J. Chem. Phys.* 108, 4340 (1998).
15. J. C. Shelley, M. Y. Shelley, R. C. Reeder, S. Bandyopadhyay, and M. L. Klein, *J. Phys. Chem. B* 105, 4464 (2001).
16. W. Tschöp, K. Kremer, J. Batoulis, T. Bürger, and O. Hahn, *Acta Polymer.* 49, 61 (1998).
17. P. B. Warren, *Curr. Opin. Colloid Interface Sci.* 3, 620 (1998).
18. P. Español and P. Warren, *Europhys. Lett.* 30, 191 (1995).
19. P. J. Hoogerbrugge and J. M. V. A. Koelman, *Europhys. Lett.* 19, 155 (1992).
20. J. M. V. A. Koelman and P. J. Hoogerbrugge, *Europhys. Lett.* 21, 363 (1993).
21. E. G. Flekkøy and P. V. Coveney, *Phys. Rev. Lett.* 83, 1775 (1999).
22. T. Ala-Nissila, S. Majaniemi, and K. Elder, in "Novel Methods in Soft Matter Simulations," (M. Karttunen, I. Vattulainen, and A. Lukkarinen, Eds.), pp. 357–388. Springer, Berlin, 2004.
23. T. Kawakatsu, K. Kawasaki, M. Furusaka, H. Okabayashi, and T. Kanaya, *J. Phys. Condens. Matter* 6, 6385 (1994).
24. N. M. Maurits, Mathematical Modeling of Complex Systems, PhD thesis, Groningen University, 1998.
25. M. Grmela and H. C. Öttinger, *Phys. Rev. E* 56, 6620 (1997).
26. M. Grmela and H. C. Öttinger, *Phys. Rev. E* 56, 6633 (1997).
27. H. C. Öttinger, *Phys. Rev. E* 57, 1416 (1998).

28. M. Karttunen, I. Vattulainen, and A. Lukkarinen, Eds., "Novel Methods in Soft Matter Simulations," Springer, Berlin, 2004.
29. J. Baschnagl, K. Binder, P. Doruker, A. A. Gusev, O. Hahn, K. Kremer, W. L. Mattice, F. Müller-Plathe, M. Murat, W. Pauli, S. Santos, U. W. Suter, and V. Tries, *Adv. Polym. Sci.* 152, 41 (2000).
30. W. J. Briels and R. L. C. Akkermans, *Mol. Sim.* 28, 145 (2002).
31. F. Müller-Plathe, *Chem. Phys. Chem.* 3, 754 (2002).
32. F. Müller-Plathe, *Soft Materials* 1, 1 (2003).
33. J. C. Shelley and M. Y. Shelley, *Curr. Opin. Colloid Interface Sci.* 5, 101 (2000).
34. J. C. Shelley, M. Y. Shelley, R. C. Reeder, S. Bandyopadhyay, P. B. Moore, and M. L. Klein, *J. Phys. Chem B* 10, 9785 (2001).
35. S. J. Marrink and A. E. Mark, *J. Am. Chem. Soc.* 125, 11144 (2003).
36. M. Kröger, *Phys. Rep.* 390, 453 (2004).
37. P. Nielaba, M. Mareschal, and G. Ciccotti, "Bridging Time Scales: Molecular Simulations for the Next Decade," Springer, Berlin, 2002.
38. ESF Network Simu—Challenges in Molecular Simulations; http://simu.ulb.ac.be.
39. M. Doi, *J. Comp. Appl. Math.* 149, 13 (2002).
40. M. Doi, *Pure Appl. Chem.* 75, 1395 (2003).
41. H. H. Goldstine and A. Goldstine, *Math. Tables Other Aids Comput.* 2, 97 (1946).
42. N. Metropolis, *Los Alamos Sci.* (15, Special Issue), 125 (1987).
43. N. Metropolis and S. Ulam, *J. Am. Stat. Assoc.* 44, 335 (1949).
44. N. Metropolis, A. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *J. Chem. Phys.* 21, 1087 (1953).
45. N. C. Metropolis, G. Reitwiesner, and J. von Neumann, *Math. Tables Other Aids Comput.* 4, 109 (1950).
46. T. E. Hull and A. R. Dobell, *SIAM Rev.* 4, 230 (1962).
47. D. E. Knuth, "The Art of Computer Programming: Seminumerical Algorithms," 3rd ed., Vol. 2. Addison-Wesley, Reading, MA, 1998.
48. B. J. Alder and T. E. Wainwright, *J. Chem. Phys.* 27, 1208 (1957).
49. B. J. Alder and T. E. Wainwright, *J. Chem. Phys.* 31, 459 (1959).
50. A. Rahman, *Phys. Rev.* 136A, 405 (1964).
51. J. A. Barker and R. O. Watts, *Chem. Phys. Lett.* 3, 144 (1969).
52. A. Rahman and F. H. Stillinger, *J. Chem. Phys.* 55, 3336 (1971).
53. J. A. McCammon, B. R. Gelin, and M. Karplus, *Nature* 267, 585 (1977).
54. E. Tajkhorshid, P. Nollert, M. Ø. Jensen, L. J. W. Miercke, J. O'Connell, R. M. Stroud, and K. Schulten, *Science* 296, 525 (2002).
55. M. Ø. Jensen, E. Tajkhorshid, and K. Schulten, *Biophys. J.* 85, 2884 (2003).
56. D. Mac Kernan, Interview of Michele Parrinello, Simu News.; available at simu.ulb.ac.be, pp. 5–17.
57. T. Schlick, "Molecular Modeling and Simulation," Springer, New York, 2002.
58. H. W. Kroto, Nobel lecture in December 7, 1996; http://www.nobel.se.
59. R. J. Baxter, "Exactly Solved Models in Statistical Mechanics," Academic Press, London, 1982.
60. D. Frenkel and B. Smit, "Understanding Molecular Simulation: From Algorithms to Applications," 2nd ed. Academic Press, San Diego, 2002.
61. M. P. Allen and D. J. Tildesley, "Computer Simulation of Liquids," Clarendon Press, Oxford, 1993.
62. A. R. Leach, "Molecular Modelling: Principles and Applications," 2nd ed. Prentice Hall, Harlow, 2001.
63. R. G. Parr and W. Yang, *Ann. Rev. Phys. Chem.* 46, 701 (1995).
64. R. Car and M. Parrinello, *Phys. Rev. Lett.* 55, 2471 (1985).
65. M. Eichinger, P. Tavan, J. Hutter, and P. Parrinello, *J. Chem. Phys.* 110, 10452 (1999).
66. W. Andreoni, A. Curioni, and T. Mordasini, *IBM J. Res. Dev.* 45, 397 (2001).
67. G. J. Martyna and M. E. Tuckerman, in "Bridging Time Scales: Molecular Simulations for the Next Decade" (P. Nielaba, M. Mareschal, and G. Ciccotti, Eds.), pp. 381–412. Springer, Berlin, 2002.
68. R. Kapral and G. Ciccotti, in "Bridging Time Scales: Molecular Simulations for the Next Decade" (P. Nielaba, M. Mareschal, and G. Ciccotti, Eds.), pp. 445–472. Springer, Berlin, 2002.
69. F. Jensen, "Introduction to Computational Chemistry," Wiley, Chichester, 1999.
70. J. M. Thijssen, "Computational Physics," Cambridge University Press, Cambridge, 1999.
71. R. Bonneau and D. Baker, *Annu. Rev. Biophys. Biomol. Struct.* 30, 173 (2001).
72. P. Carloni, U. Rothlisberger, and M. Parrinello, *Acc. Chem. Res.* 35, 455 (2002).
73. U. Rothlisberger, M. Sprik, and J. Hutter, in "Bridging Time Scales: Molecular Simulations for the Next Decade" (P. Nielaba, M. Mareschal, and G. Ciccotti, Eds.), pp. 413–444. Springer, Berlin, 2002.
74. S. Y. Reddy, F. Leclerc, and M. Karplus, *Biophys. J.* 84, 1421 (2003).
75. E. Lindahl, B. Hess, and D. van der Spoel, *J. Mol. Model.* 7, 306 (2001).
76. C. Sagui and T. A. Darden, *Annu. Rev. Biophys. Biomol. Struct.* 28, 155 (1999).
77. G. Martyna, M. E. Tuckerman, D. J. Tobias, and M. L. Klein, *Mol. Phys.* 87, 1117 (1996).
78. B. J. Berne, G. Ciccotti, and D. F. Coker, Eds., "Classical and Quantum Dynamics in Condensed Phase Simulations," World Scientific, Singapore, 1998.
79. E. Giudice and R. Lavery, *Acc. Chem. Res.* 35, 350 (2002).
80. V. Makarov and B. M. Pettitt, *Acc. Chem. Res.* 35, 376 (2002).
81. B. Israelewitz, M. Gao, and K. Schulten, *Curr. Opin. Struct. Biol.* 11, 224 (2001).
82. C. L. Brooks III, *Acc. Chem. Res.* 35, 447 (2002).
83. V. Daggett, *Acc. Chem. Res.* 35, 422 (2002).

84. L. Saiz, S. Bandyopadhyay, and M. L. Klein, *Bioscience Rep.* 22, 151 (2002).
85. L. Saiz and M. L. Klein, *Acc. Chem. Res.* 35, 482 (2002).
86. D. P. Tieleman, S. J. Marrink, and H. J. C. Berendsen, *Biochim. Biophys. Acta* 1331, 235 (1997).
87. S. E. Feller, *Curr. Opin. Colloid Interface Sci.* 5, 217 (2000).
88. H. L. Scott, *Curr. Opin. Struct. Biol.* 12, 495 (2002).
89. J. P. Hansen and I. R. McDonald, "Theory of Simple Liquids," Academic Press, London, 1986.
90. P. G. Bolhuis, A. A. Louis, J.-P. Hansen, and E. J. Meijer, *J. Chem. Phys.* 114, 4296 (2001).
91. A. A. Louis, P. G. Bolhuis, J.-P. Hansen, and E. J. Meijer, *Phys. Rev. Lett.* 82, 2522 (2000).
92. R. L. Henderson, *Phys. Lett.* 49A, 197 (1974).
93. P. Hohenberg and W. Kohn, *Phys. Rev. B* 136, 864 (1964).
94. R. Faller, H. Schmitz, O. Biermann, and F. Müller-Plathe, *J. Comp. Chem.* 20, 1009 (1999).
95. S. Girard and F. Müller-Plathe, *Mol. Phys.* 101, 779 (2003).
96. D. Reith, H. Meyer, and F. Müller-Plathe, *Macromolecules* 34, 2235 (2001).
97. E. Falck, O. Punkkinen, I. Vattulainen, and T. Ala-Nissila, *Phys. Rev. E* 68, 050102(R) (2003).
98. E. Falck, J. Lahtinen, I. Vattulainen, and T. Ala-Nissila, *Eur. Phys. J. E* 13, 267 (2004).
99. S. H. Lee and R. Kapral, *Physica A* 298, 56 (2001).
100. A. Malevanets and J. M. Yeomans, *Europhys. Lett.* 52, 231 (2000).
101. N. Kikuchi, A. Gent, and J. M. Yeomans, *Eur. Phys. J. E* 9, 63 (2002).
102. T. Ihle and D. M. Kroll, *Phys. Rev. E* 63, 020201 (2001).
103. T. Ihle and D. M. Kroll, *Phys. Rev. E* 67, 066705 (2003).
104. T. Ihle and D. M. Kroll, *Phys. Rev. E* 67, 066706 (2003).
105. E. Tüzel, M. Strauss, T. Ihle, and D. M. Kroll, *Phys. Rev. E* 68, 036701 (2003).
106. A. Lamura and G. Gompper, *Eur. Phys. J. E* 9, 477 (2002).
107. N. Kikuchi, C. M. Pooley, J. F. Ryder, and J. M. Yeomans, *J. Chem. Phys.* 119, 6388 (2003).
108. I. Pagonabarraga and D. Frenkel, *J. Chem. Phys.* 115, 5015 (2001).
109. A. P. Lyubartsev and A. Laaksonen, *Phys. Rev. E* 52, 3730 (1995).
110. N. Goldenfeld, "Lectures on Phase Transitions and the Renormalization Group," Addison-Wesley, Reading, MA, 1992.
111. W. Dzwinel and D. A. Yuen, *J. Coll. Interf. Sci.* 225, 179 (2000).
112. P. Español, *Europhys. Lett.* 39, 605 (1997).
113. P. Español, *Europhys. Lett.* 40, 631 (1997).
114. P. Español, M. Serrano, and I. Zuñiga, *Int. J. Mod. Phys. C* 8, 899 (1997).
115. R. D. Groot and T. J. Madden, *J. Chem. Phys.* 108, 8713 (1998).
116. R. D. Groot, T. J. Madden, and D. J. Tildesley, *J. Chem. Phys.* 110, 9739 (1999).
117. S. Jury, P. Bladon, M. Cates, S. Krishnan, M. Hagen, N. Ruddock, and P. Warren, *Phys. Chem. Chem. Phys.* 1, 2051 (1999).
118. S. Yamamoto, Y. Maruyama, and S. Hyodo, *J. Chem. Phys.* 116, 5842 (2002).
119. S. Yamamoto and S. Hyodo, *J. Chem. Phys.* 118, 7937 (2003).
120. M. Kranenburg, M. Venturoli, and B. Smit, *Phys. Rev. E* 67, 060901(R) (2003).
121. M. Kranenburg, M. Venturoli, and B. Smit, *J. Phys. Chem. B* 107, 11491 (2003).
122. M. G. Burke, R. Wocholski, and S. N. Yaliraki, *Proc. Natl. Acad. Sci.* 100, 13928 (2003).
123. P. B. Warren, *Phys. Rev. E* 68, 066702 (2003).
124. J. A. Elliot and A. H. Windle, *J. Chem. Phys.* 113, 10367 (2000).
125. G. Besold, I. Vattulainen, M. Karttunen, and J. M. Polson, *Phys. Rev. E* 62, R7611 (2000).
126. P. Nikunen, M. Karttunen, and I. Vattulainen, *Comput. Phys. Commun.* 153, 407 (2003).
127. I. Vattulainen, M. Karttunen, G. Besold, and J. M. Polson, *J. Chem. Phys.* 116, 3967 (2002).
128. T. Soddemann, B. Dunweg, and K. Kremer, *Phys. Rev. E* 68, 046702 (2003).
129. R. D. Groot, *J. Chem. Phys.* 118, 11265 (2003).
130. R. D. Groot, *J. Chem. Phys.* 119, 10454 (2003).
131. C. Sagui and T. Darden, *J. Chem. Phys.* 115, 6578 (2001).
132. C. Sagui, L. G. Pedersen, and T. Darden, *J. Chem. Phys.* 120, 73 (2004).
133. C. P. Lowe, *Europhys. Lett.* 47, 145 (1999).
134. H. C. Andersen, *J. Chem. Phys.* 72, 2384 (1980).
135. P. Nikunen, I. Vattulainen, and M. Karttunen, (2004) (unpublished).
136. G. J. Fix, in "The Mathematics of Finite Elements and Applications IV" (J. R. Whiteman, Ed.), p. 265., Academic Press, London, 1982.
137. P. C. Hohenberg and B. I. Halperin, *Rev. Mod. Phys.* 49, 435 (1977).
138. J. D. Gunton and M. Droz, "Introduction to the Theory of Metastable and Unstable States," Springer, Berlin., 1983.
139. J. D. Gunton, M. San Miguel, and P. S. Sahni, in "Phase Transitions and Critical Phenomena" (C. Domb and M. S. Green, Eds.), Vol. 8, p. 267, Academic Press, London, 1983.
140. N. Provatas, Q. Wang, M. Haataja, and M. Grant, *Phys. Rev. Lett.* 91, 155502 (2003).
141. J. Müller and M. Grant, *Phys. Rev. Lett.* 82, 1736 (1999).
142. F. Drolet, K. R. Elder, M. Grant, and J. M. Kosterlitz, *Phys. Rev. E* 61, 6705 (2000).
143. M. Karttunen, M. Haataja, K. R. Elder, and M. Grant, *Phys. Rev. Lett.* 83, 3518 (1999).
144. P. W. Anderson, "Basic Notions of Condensed Matter Physics," Addison-Wesley, Redwood City, 1984.
145. J. D. Shore, D. Ronis, L. Piche, and M. Grant, *Phys. Rev. Lett.* 77, 655 (1996).

146. J. D. Shore, D. Ronis, L. Piche, and M. Grant, *Phys. Rev. E* 55, 2976 (1997).
147. F. Drolet and G. H. Fredrickson, *Phys. Rev. Lett.* 83, 4317 (1999).
148. J. G. E. M. Fraaije, B. A. C. van Vlimmeren, N. M. Maurits, M. Postma, O. A. Evers, C. Hoffmann, P. Altevogt, and G. Goldbeck-Wood, *J. Chem. Phys.* 106, 4260 (1997).
149. B. M. Ladanyi and M. S. Skaf, *Annu. Rev. Phys. Chem.* 44, 335 (1993).
150. I. Koltover, T. Salditt, J. O. Rädler, and C. R. Safinya, *Science* 281, 78 (1998).
151. S. Bandyopadhyay, M. Tarek, and M. L. Klein, *J. Phys. Chem. B* 103, 10075 (1999).
152. A. Y. Grosberg, T. T. Nguyen, and B. I. Shklovskii, *Rev. Mod. Phys.* 74, 329 (2002).
153. M. Patra, M. Patriarca, and M. Karttunen, *Phys. Rev. E* 67, 031402 (2003).
154. M. Bloom, E. Evans, and O. G. Mouritsen, *Q. Rev. Biophys.* 24, 293 (1991).
155. R. Lipowsky and E. Sackmann, Eds., "Structure and Dynamics of Membranes: From Cells to Vesicles," Elsevier, Amsterdam, 1995.
156. K. M. Merz, Jr. and B. Roux, Eds., "Biological Membranes: A Molecular Perspective from Computation and Experiment," Birkhäuser, Boston, 1996.
157. J. F. Nagle and S. Tristram-Nagle, *Biochim. Biophys. Acta.* 1469, 159 (2000).
158. J. Katsaras and T. Gutberlet, Eds., "Lipid Bilayers: Structure and Interactions," Springer, Berlin, 2001.
159. U. Essman, L. Perera, M. L. Berkowitz, H. L. T. Darden, and L. G. Pedersen, *J. Chem. Phys.* 103, 8577 (1995).
160. L. Greengard and V. Rokhlin, *J. Comput. Phys.* 73, 325 (1987).
161. R. Zangi, M. L. de Vocht, G. T. Robillard, and A. E. Mark, *Biophys. J.* 83, 112 (2002).
162. S. J. Marrink, E. Lindahl, O. Edholm, and A. E. Mark, *J. Am. Chem. Soc.* 123, 8638 (2001).
163. S. J. Marrink and D. P. Tieleman, *Biophys. J.* 83, 2386 (2002).
164. H. E. Alper and R. M. Levy, *J. Chem. Phys.* 91, 1242 (1989).
165. S. E. Feller, R. W. Pastor, A. Rojnuckarin, A. Bogusz, and B. R. Brooks, *J. Phys. Chem.* 100, 17011 (1996).
166. P. Mark and L. Nilsson, *J. Comput. Chem.* 23, 1211 (2002).
167. H. E. Alper, D. Bassolino, and T. R. Stouch, *J. Chem. Phys.* 98, 9798 (1993).
168. H. E. Alper, D. Bassolino-Klimas, and T. R. Stouch, *J. Chem. Phys.* 99, 5547 (1993).
169. P. E. Smith and B. M. Pettitt, *J. Chem. Phys.* 95, 8430 (1991).
170. H. Schreiber and O. Steinhauser, *Biochemistry* 31, 5856 (1992).
171. D. M. York, T. A. Darden, and L. G. Pedersen, *J. Chem. Phys.* 99, 8345 (1993).
172. D. M. York, W. Yang, H. Lee, T. Darden, and L. G. Pedersen, *J. Am. Chem. Soc.* 117, 5001 (1995).
173. J. Norberg and L. Nilsson, *Biophys. J.* 79, 1537 (2000).
174. H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans, in "Intermolecular Forces" (B. Pullman, Ed.), pp. 331–342. Reidel, Dordrecht, 1981.
175. M. Patra, M. Karttunen, M. T. Hyvönen, E. Falck, P. Lindqvist, and I. Vattulainen, *Biophys. J.* 84, 3636 (2003).
176. O. Berger, O. Edholm, and F. Jahnig, *Biophys. J.* 72, 2002 (1997).
177. D. P. Tieleman and H. J. C. Berendsen, *J. Chem. Phys.* 105, 4871 (1996).
178. R. M. Venable, B. R. Brooks, and R. W. Pastor, *J. Chem. Phys.* 112, 4822 (2000).
179. C. Anézo, A. H. de Vries, H.-D. Höltje, D. P. Tieleman, and S. J. Marrink, *J. Phys. Chem. B* 107, 9424 (2003).
180. E. Lindahl and O. Edholm, *J. Chem. Phys.* 113, 3882 (2000).
181. M. Patra, M. Karttunen, M. T. Hyvönen, E. Falck, and I. Vattulainen, *J. Phys. Chem. B* 108, 4485 (2004).
182. M. Patra, M. Karttunen, M. T. Hyvönen, E. Falck, and I. Vattulainen, (2004) (unpublished).
183. J. E. Gentle, "Random Number Generation and Monte Carlo Methods," 2nd ed. Springer, Berlin, 2003.
184. K. Binder and D. W. Heermann, "Monte Carlo Simulation in Statistical Physics," 4th ed. Springer, New York, 2002.
185. I. Vattulainen, *Phys. Rev. E* 59, 7200 (1999).
186. M. Mascagni, *Parallel Comput.* 24, 923 (1998).
187. K. Entacher, *ACM Trans. Model. Comput. Simul.* 8, 61 (1998).
188. A. Srinivasan, M. Mascagni, and D. Ceperley, *Parallel Comput.* 29, 69 (2003).
189. J. Krug, *Adv. Phys.* 46, 139 (1997).
190. B. Duplantier and K.-H. Kwon, *Phys. Rev. Lett.* 61, 2514 (1988).
191. K. Burdzy and G. F. Lawler, *Ann. Prob.* 18, 981 (1990).
192. H. Larralde, P. Trunfio, S. Havlin, H. E. Stanley, and G. H. Weiss, *Phys. Rev. A* 45, 7128 (1992).
193. T. G. Lewis and W. H. Payne, *J. Assoc. Comput. Mach.* 20, 456 (1973).
194. W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, "Numerical Recipes in Fortran 90, The Art of Parallel Scientific Computing," 2nd ed., pp. 1141–1142. Cambridge University Press, Cambridge, 1996.
195. P. L'Ecuyer, *Commun. ACM* 31, 742 (1988).
196. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, "Numerical Recipes in Fortran, The Art of Scientific Computing," 2nd ed., pp. 271–273. Cambridge University Press, Cambridge, 1992.
197. F. James, *Comput. Phys. Commun.* 60, 329 (1990).
198. G. Marsaglia, A. Zaman, and W.-W. Tsang, *Stat. Prob. Lett.* 8, 35 (1990).
199. M. Matsumoto and T. Nishimura, *ACM Trans. Mod. Comp. Simul.* 8, 3 (1998).
200. M. Lüscher, *Comput. Phys. Commun.* 79, 100 (1994).
201. F. James, *Comput. Phys. Commun.* 79, 111 (1994).
202. I. Vattulainen, T. Ala-Nissila, and K. Kankaala, *Phys. Rev. Lett.* 73, 2513 (1994).
203. I. Vattulainen, T. Ala-Nissila, and K. Kankaala, *Phys. Rev. E* 52, 3205 (1995).
204. I. Vattulainen, K. Kankaala, J. Saarinen, and T. Ala-Nissila, *Comput. Phys. Commun.* 86, 209 (1995).
205. L. N. Shchur and P. Butera, *Int. J. Mod. Phys. C* 9, 607 (1998).

206. N. H. March and M. P. Tosi, "Atomic Dynamics in Liquids." Dover, New York, 1976.
207. A. M. Ferrenberg, D. P. Landau, and Y. J. Wong, *Phys. Rev. Lett.* 69, 3382 (1992).
208. R. M. D'Souza, Y. Bar-Yam, and M. Kardar, *Phys. Rev. E* 57, 5044 (1998).
209. J. M. Hammersley and D. C. Handscomb, "Monte Carlo Methods." Methuen & Co., London, 1964.
210. G. Parisi and F. Rapuano, *Phys. Lett.* 157B, 301 (1985).
211. F. J. Resende and B. V. Costa, *Phys. Rev. E* 58, 5183 (1998).
212. L. Verlet, *Phys. Rev.* 159, 98 (1967).
213. I. Pagonabarraga, M. H. J. Hagen, and D. Frenkel, *Europhys. Lett.* 42, 377 (1998).
214. W. K. den Otter and J. H. R. Clarke, *Europhys. Lett.* 53, 426 (2001).
215. T. Shardlow, *SIAM J. Sci. Comp.* 24, 1267 (2003).
216. J. P. Boon and S. Yip, "Molecular Hydrodynamics." Dover, New York, 1980.
217. C. J. Cotter and S. Reich, *Europhys. Lett.* 64, 723 (2003).
218. P. Español, in "Novel Methods in Soft Matter Simulations" (M. Karttunen, I. Vattulainen, and A. Lukkarinen, Eds.), pp. 69–115. Springer, Berlin, 2004.
219. A. P. Lyubartsev, M. Karttunen, I. Vattulainen, and A. Laaksonen, *Soft Materials* 1, 121 (2003).
220. D. Voet and J. G. Voet, "Biochemistry," 2nd ed. Wiley, New York, 1995.
221. G. Cevc and D. Marsh, "Phospholipid Bilayers: Physical Principles and Models." John Wiley & Sons, New York, 1987.
222. R. B. Gennis, "Biomembranes: Molecular Structure and Function." Springer, New York, 1989.
223. M. Nielsen, L. Miao, J. H. Ipsen, O. G. Mouritsen, and M. J. Zuckermann, *Phys. Rev. E* 54, 6889 (1996).
224. M. Nielsen, L. Miao, J. H. Ipsen, M. J. Zuckermann, and O. G. Mouritsen, *Phys. Rev. E* 59, 5790 (1999).
225. M. Nielsen, L. M. J. H. Ipsen, O. G. Mouritsen, M. J. Zuckermann, J. Thewalt, and M. Bloom, *Europhys. Lett.* 52, 368 (2000).
226. J. M. Polson, I. Vattulainen, H. Zhu, and M. J. Zuckermann, *Eur. Phys. J. E* 5, 485 (2001).
227. L. Miao, M. Nielsen, J. Thewalt, J. H. Ipsen, M. Bloom, M. J. Zuckermann, and O. G. Mouritsen, *Biophys. J.* 82, 1429 (2002).
228. M. J. Zuckermann, M. Bloom, J. H. Ipsen, L. Miao, O. G. Mouritsen, M. Nielsen, J. Polson, J. Thewalt, I. Vattulainen, and H. Zhu, *Methods Enzym.* 383, 198 (2004).
229. P. van der Ploeg and H. J. C. Berendsen, *J. Chem. Phys.* 76, 3271 (1982).
230. R. M. Venable, Y. Zhang, B. J. Hardy, and R. W. Pastor, *Science* 262, 223 (1993).
231. K. Simons and E. Ikonen, *Nature* 387, 569 (1997).
232. D. A. Brown and E. London, *Annu. Rev. Cell. Dev. Biol.* 14, 111 (1998).
233. M. Edidin, *Annu. Rev. Biophys. Biomol. Struct.* 32, 257 (2003).
234. J. R. Silvius, *Biochim. Biophys. Acta* 1610, 174 (2003).
235. M. Dykstra, A. Cherukuri, H. W. Sohn, S.-J. Tzeng, and S. K. Pierce, *Annu. Rev. Immunol.* 21, 457 (2003).
236. E. Falck, M. Patra, M. Karttunen, M. T. Hyvönen, and I. Vattulainen, *Biophys. J.* 87, 1076 (2004).
237. E. Falck, M. Patra, M. Karttunen, M. T. Hyvönen, and I. Vattulainen, (2004) (submitted).
238. A. M. Smondyrev and M. L. Berkowitz, *Biophys. J.* 80, 1649 (2001).
239. S. W. Chiu, E. Jakobsson, and H. L. Scott, *J. Chem. Phys.* 114, 5435 (2001).
240. S. W. Chiu, E. Jakobsson, R. J. Mashl, and H. L. Scott, *Biophys. J.* 83, 1842 (2002).
241. M. Pasenkiewicz-Gierula, T. Róg, K. Kitamura, and A. Kusumi, *Biophys. J.* 78, 1376 (2000).
242. T. Róg and M. Pasenkiewicz-Gierula, *Biophys. J.* 81, 2190 (2001).
243. C. Hofsäß, E. Lindahl, and O. Edholm, *Biophys. J.* 84, 2192 (2003).
244. P. Jedlovszky and M. Mezei, *J. Phys. Chem. B* 107, 5311 (2003).
245. P. Jedlovszky and M. Mezei, *J. Phys. Chem. B* 107, 5322 (2003).
246. P. Jedlovszky and M. Mezei, *J. Phys. Chem. B* 108, 465 (2004).
247. J. A. Söderhäll and A. Laaksonen, *J. Phys. Chem. B* 105, 9308 (2001).
248. M. T. Hyvönen and P. T. Kovanen, *J. Phys. Chem. B* 107, 9102 (2003).
249. E. Mombelli, R. Morris, W. Taylor, and F. Fraternali, *Biophys. J.* 84, 1507 (2003).
250. S. W. Chiu, S. Vasudevan, E. Jakobsson, R. J. Mashl, and H. L. Scott, *Biophys. J.* 85, 3624 (2003).
251. P. Niemelä, M. T. Hyvönen, and I. Vattulainen, *Biophys. J.* in press (2004).
252. R. G. Snyder, K. Tu, M. L. Klein, R. Mendelssohn, H. L. Strauss, and W. Sun, *J. Phys. Chem. B* 106, 6273 (2002).
253. T. Huber, K. Rajamoorthi, V. F. Kurze, K. Beyer, and M. F. Brown, *J. Am. Chem. Soc.* 124, 298 (2002).
254. L. Saiz and M. L. Klein, *Biophys. J.* 81, 204 (2001).
255. S. E. Feller, K. Gawrisch, and A. D. MacKerell, Jr., *J. Am. Chem. Soc.* 124, 318 (2002).
256. H. Möhwald, in "Structure and Dynamics of Membranes: From Cells to Vesicles" (R. Lipowsky and E. Sackmann, Eds.), pp. 161–211. Elsevier, Amsterdam, 1995.
257. R. Maget-Dana, *Biochim. Biophys. Acta* 1462, 109 (1999).
258. Y. N. Kaznessis, S. Kim, and R. G. Larson, *Biophys. J.* 82, 1731 (2002).
259. R. Boeckmann, A. Bac, T. Heimburg, and H. Grubmuller, *Biophys. J.* 85, 1647 (2003).
260. S. A. Pandit, D. Bostick, and M. L. Berkowitz, *Biophys. J.* 84, 3743 (2003).
261. S. A. Pandit and M. L. Berkowitz, *Biophys. J.* 82, 1818 (2002).
262. A. Gurtovenko, M. Patra, M. Karttunen, and I. Vattulainen, *Biophys. J.* 86, 3461 (2004).
263. A. K. Sum and J. J. de Pablo, *Biophys. J.* 85, 3636 (2003).
264. A. K. Sum, R. Faller, and J. J. de Pablo, *Biophys. J.* 85, 2830 (2003).

265. M. Pasenkiewicz-Gierula, T. Róg, J. Grochowski, P. Serda, R. Czarnecki, T. Librowski, and S. Lochyński, *Biophys. J.* 85, 1248 (2003).
266. J. Ulander and A. D. J. Haymet, *Biophys. J.* 85, 3475 (2003).
267. P. Mukhopadhyay, H. J. Vogel, and D. P. Tieleman, *Biophys. J.* 86, 337 (2004).
268. M. T. Hyvönen, K. Öörni, P. T. Kovanen, and M. Ala-Korpela, *Biophys. J.* 80, 565 (2001).
269. M. Ø. Jensen, T. R. Jensen, K. Kjaer, T. Bjørnholm, O. G. Mouritsen, and G. H. Peters, *Biophys. J.* 83, 98 (2001).
270. I. H. Shrivastava and M. S. P. Sansom, *Biophys. J.* 78, 557 (2000).
271. C. E. Capener and M. S. P. Sansom, *J. Phys. Chem. B* 106, 4543 (2002).
272. J. Baudry, E. Tajkhorshid, F. Molnar, J. Phillips, and K. Schulten, *J. Phys. Chem. B* 105, 905 (2001).
273. S. J. Marrink, A. H. de Vries, and A. E. Mark, *J. Phys. Chem, B* 108, 750 (2004).
274. R. Goetz and R. Lipowsky, *J. Chem. Phys.* 108, 7397 (1998).
275. R. Goetz, G. Gompper, and R. Lipowsky, *Phys. Rev. Lett.* 82, 221 (1999).
276. J. C. Shelley, M. Y. Shelley, R. C. Reeder, S. Bandyopadhyay, and M. L. Klein, *J. Phys. Chem. B* 105, 4464 (2001).
277. J. C. Shelley, M. Y. Shelley, R. C. Reeder, S. Bandyopadhyay, and M. L. Klein, *J. Phys. Chem. B* 105, 9785 (2001).
278. S. O. Nielsen, C. F. Lopez, P. B. Moore, J. C. Shelley, and M. L. Klein, *J. Phys. Chem. B* 107, 13911 (2003).
279. G. Ayton and G. A. Voth, *Biophys. J.* 83, 3357 (2002).
280. G. Ayton, S. G. Bardenhagen, P. McMurtry, D. Sulsky, and G. A. Voth, *IBM J. Res. Dev.* 45, 417 (2001).
281. G. Ayton, S. G. Bardenhagen, P. McMurtry, D. Sulsky, and G. A. Voth, *J. Chem. Phys.* 114, 6913 (2001).
282. G. G. Ayton, A. M. Smondyrev, S. G. Bardenhagen, P. McMurtry, and G. A. Voth, *Biophys. J.* 82, 1226 (2002).
283. G. Ayton, A. M. Smondyrev, S. G. Bardenhagen, P. McMurtry, and G. A. Voth, *Biophys. J.* 83, 1026 (2002).
284. K. Toukan and A. Rahman, *Phys. Rev. B* 31, 2643 (1985).
285. D. E. Smith and L. X. Dang, *J. Chem. Phys.* 100, 3757 (1994).
286. A. P. Lyubartsev and A. Laaksonen, *Phys. Rev. E* 55, 5689 (1997).
287. P. Mark and L. Nilsson, *J. Phys. Chem. A* 105, 9954 (2001).
288. M. Patra and M. Karttunen, *J. Comput. Chem.* 25, 678 (2004).
289. D. F. Evans and H. Wennerström, "The Colloidal Domain: Where Physics, Chemistry, Biology, and Technology Meet," 2nd ed. Wiley, New York, 1999.
290. A. J. Masters and P. B. Warren, *Europhys. Lett.* 48, 1 (1997).
291. M. Canales and G. Sesé, *J. Chem. Phys.* 109, 6004 (1998).
292. D. R. Lide, Ed., "CRC Handbook of Chemistry and Physics," 82nd ed. CRC Press, Boca Raton, 2001.

CHAPTER 9

Theoretical and Computational Treatments of DNA and RNA Molecules

Haijun Zhou
Max-Planck-Institute of Colloids and Interfaces, Potsdam, Germany

Yang Zhang
Center of Excellence in Bioinformatics, University at Buffalo, Buffalo, New York, USA

Zhong-Can Ou-Yang
Institute of Theoretical Physics, The Chinese Academy of Sciences, Beijing, China

CONTENTS

ISBN: 1-58883-048-9

Handbook of Theoretical and Computational Nanotechnology
Edited by Michael Rieth and Wolfram Schommers
Volume 6: Pages (419–487)

1. INTRODUCTION

Since the first direct measurement of the elasticity of single DNA in 1992 [1], mechanical properties of individual DNA and RNA macromolecules and interactions between DNA and various proteins have extensively been investigated by single-molecule methods (see, for example, review articles [2–8]). These single-molecule manipulation techniques include atomic force microscopy, magnetic beads, micropipettes, optical tweezers, and many others. Many new mechanical characteristics of DNA and RNA molecules were revealed by applying these unprecedented instruments. Inspired by this huge experimental revolution, there have been many theoretical modeling efforts in recent years. These theoretical and computational studies, combined with single-molecule observations, have resulted in greatly improved understanding on the elastic and statistical mechanical properties of DNA and RNA polymers.

This chapter tries to give a comprehensive review of the methods used in these theoretical researches and the insights gained from them. In this introductory section, we first summarize some basic knowledge of DNA and RNA for those readers who are not yet familiar with these molecules. It is followed by a brief survey of recent single-molecule experiments. In Section 1.3, we outline the main topics covered in this chapter.

1.1. Basics of DNA and RNA

Molecular biology became the main focus of modern biological studies ever since the discovery of the double-helix structure of DNA in 1953 by Watson, Crick, Wilkins, Franklin and coworkers [9–11]. Deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and protein are the main themes of molecular biology: they constitute the components of the central dogma (see, for example, books [12, 13]).

DNA is a polymer made of deoxyribonucleotides, and RNA is a polymer of ribonucleotides. In biological cells, DNA is usually very long. For example, the DNA sequence of *Escherichia coli* bacterium contains more than 9 million nucleotides, and the total number of nucleotides in one cell of our body is of the order of 10 billion (10^{10}). Nucleotides in DNA polymer are organized into two polynucleotide linear chains of equal length (Fig. 1). A nucleotide monomer is an asymmetric molecule, with a 3′-end and a 5′-end. In each of

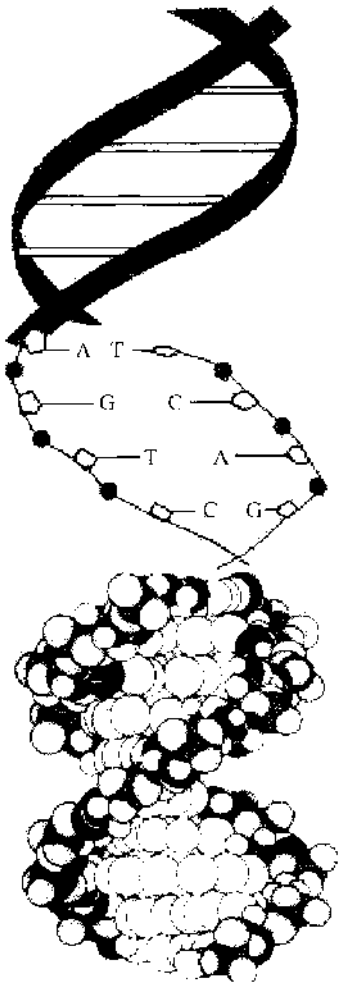


Figure 1. Schematic representation of DNA. The two strands of DNA run in opposite directions (indicated by the arrows along the phosphate–sugar backbones) and are bound together by many G–C and A–T base pairs. The two strands fold into a double-helix because of base-pair stacking interactions. In the relaxed state, each pitch of the double-helix contains 10.5 base pairs and the vertical distance between two neighboring base pairs is 0.34 nm. Adapted from http://ns2.d20.co.edu/kadets/lundberg/dna.html with permission from D. Lundberg.

the two strands of DNA, the 3′-end of one nucleotide is connected to the 5′-end of another nucleotide by a phosphate-diester covalent bond [14]. A polynucleotide chain therefore has a direction defined from its 5′-end to its 3′-end. There are four different types of nucleotide residues (bases); namely adenine (A), guanine (G), cytosine (C), and thymine (T). For most organisms, DNA is the molecule where the hereditary information is stored and maintained. The genetic information of an organism is recorded by the particular order of appearances of the four types of bases A, C, G, and T along each strand of DNA.

In DNA, the two poly(deoxyribon)nucleotide strands are antiparallel to each other and are twisted around a common central axis to form a right-handed double-helix structure, with the genetic information (the nucleotide residues) buried in the interior of the cylindrical molecule and the highly negatively charged sugar–phosphate backbones exposed outside. Such a structure has the benefit of prohibiting genetic information from being lost or destroyed, but it causes severe constraints on the access of genetic information and the transfer of it from one generation to another. During billions of years of evolution and selection, biological organisms have developed complicated mechanisms to facilitate gene duplication and transcription [12].

In physiological conditions, the double-helix structure of DNA is very stable. This stability is the result of the following three contributions.

First, the two strands of DNA are bound together by many hydrogen-bonded base pairs. The two polynucleotide strands of DNA are complementary to each other. Suppose each

strand has n nucleotides: complementarity means that, if at position i of one strand (counting from its 5′-end) the nucleotide is A, then at position $n - i$ of the other strand the nucleotide is T (the same rule holds for nucleotides G and C). Adenine and thymine can form a planar A–T base pair through the formation of two hydrogen bonds between them, Cytosine and guanine could form a planar G–C base-pair through the formation of three hydrogen bonds between them [14]. The sugar–phosphate backbone length of DNA between two consecutive base pairs is about 0.68 nm.

Second, the folded double-helix is stabilized by base-pair stacking interactions between adjacent nucleotide base-pairs [14]. The base-pair stacking potential is a short-ranged attractive interaction, with a maximum strength when the two DNA base-pair planes are separated by about 0.34 nm. Therefore, in its natural conditions, DNA is a right-handed double-helix, each helical period of which contains about 10.5 base pairs; two neighboring base pairs are separated vertically about 0.34 nm; and the pitch length of the double-helix along the central axis is about 3.6 nm [15].

Third, the existence of high concentration of counter-ions in biological cells screens out the negative charges of DNA. Therefore, electrostatic repulsive interactions among negatively charged phosphate groups are weak and cannot destroy the compact double-helix structure.

DNA is transferred from a parent cell to its daughter cells through a replication process [12]. The parent DNA is duplicated into two identical copies. When the parent cell divides into two daughter cells, each of them contains a copy of the duplicated DNA molecule. During each cell cycle, the DNA sequence is read by RNA polymerases, and genetic information is transcribed into a message RNA (mRNA) linear chain. This process is called transcription. The mRNA chain is then transmitted to another part of the cell and read by a protein synthesis machine called the ribosome. This last step is a translation process, where each word (of three nucleotides) in the input mRNA corresponds to an amino acid in the product protein chain.

During the replication and transcription processes, the DNA molecule is severely deformed: the double-helix is locally bended, untwisted, stretched, compressed, and the base-pair patterns are locally destroyed. The ability of DNA to perform reversible structural transitions in response to various external conditions is very important for the biological function of DNA as a hereditary material. DNA has a bending persistence length of about 50 nm [16], which means that if the bending deformation occurs over a length scale less than 150 base pairs, a considerable bending energy must be overcome. DNA could also be twisted by applying a torsional stress on the molecule. The twist modulus of DNA is recently measured to be 440 ± 40 pN nm^2 [17] (pN means piconewton, or 10^{-12} N). The stretching modulus of DNA is about 1000 pN [18]. However, when DNA is stretched with a force larger than 70 pN, its end-to-end extension suddenly increases to 1.7 times its relaxed contour length [18, 19]. The double-helix structure of DNA can also be severely changed by external torque [20–23].

The biological role of RNA is more versatile compared to that of DNA. Besides mRNA, there are tRNA (transfer-RNA), which brings amino acids to the protein-assembly machine, and rRNA (ribosomal RNA), which resides in the ribosome and is involved in protein synthesis. Some RNA chains can also have catalytic functions. Furthermore, in some kinds of viruses, the hereditary information is stored in RNA rather than in DNA.

RNA exists as a single-stranded polymer. The nucleotide residues of RNA are also composed of four types as in the case of DNA, but with uracil (U) replacing thymine. Although RNA is a single-stranded polymer, it can contain local right-handed double-helix structures. To form a local double-helix segment, one part of the RNA chain must bend at certain position to form a hairpin loop, and the hairpin structure is then stabilized by base-pairing and base-pair stacking interactions (for an example, see Fig. 2). Besides G–C base pairs, an adenine could form a A–U base pair with a uracil, and G–U base pairs can also be formed.

The ability for some RNA chains (such as tRNA) to fold into stable three-dimensional structures are very important for their biological functions. Given a RNA sequence, a computational challenge is to predict its stable structures. On the other hand, for mRNA, it might be unfavorable to have stable three-dimensional configurations. If a mRNA forms a hairpin pattern with two distinct RNA segments bound to each other, this hairpin configuration may restrict the passage of mRNA through the ribosome molecule, leading to a

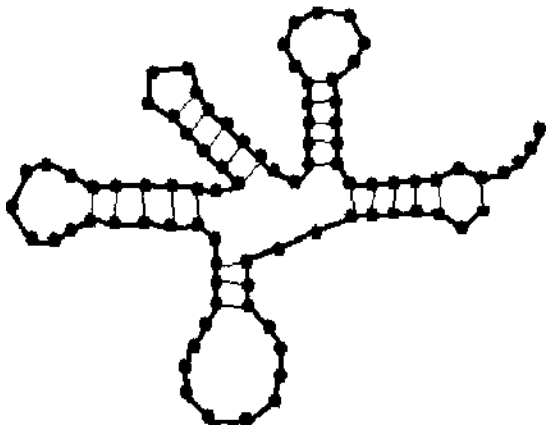


Figure 2. Clover leaf model of the base-pairing pattern of a tRNA molecule. Filled dots denote nucleotides and thin lines denote base pairs.

blockage of protein synthesis. For a mRNA sequence, it is possible to have many transient hairpin patterns. These hairpin patterns may not be stable, therefore they can be destroyed by the force generated by the ribosome on mRNA during translation. In eukaryotes, it is also conceivable that these hairpin patterns might play some role in RNA intron splicing and alternative gene expression. The links between RNA sequence and biology is still not quite clear.

1.2. A Brief Survey of Single-Molecule Experimental Studies on DNA and RNA

Prior to the advent of single-molecule methods, properties of polynucleotides were inferred from bulk thermodynamic measurements. An example along this line is measuring DNA denaturation curves based on light absorption or calorimetry (see, for example, review article [24] and book [25]). Many thermodynamic parameters of the DNA double-helix were estimated by these ways. However, bulk experiments with a certain concentration of macromolecules can only measure the polymer's average behavior. Fluctuations due to molecular individuality cannot be traced. There is another drawback: it is not always easy to separate the effects due to intramolecular interactions from those resulting from intermolecular interactions.

Single-molecule methods, which can work in the broad force range of 10^{-15} to 10^{-9} N, directly record the response of a macromolecule under the action of environmental stimuli. It enables researchers to gain detailed view of a macromolecule's mechanical properties. It also stimulates the need to perform theoretical modeling efforts on single-molecule systems with the aim of quantitatively reproducing experimental observations.

In 1992, the three authors S. B. Smith, L. Finzi, and C. Bustamante [1] fixed one end of a lambda-phage DNA to a glass slide and attached a magnetic bead to the other end. They then recorded the force-extension of DNA by applying a magnetic force on the attached bead. Very novel elastic response of DNA was observed [1]. This research has stimulated a lot of efforts in single-molecule investigation on polynucleotides. Here, we briefly mention some of these studies to give the reader an overview of the current experimental situations. For other recent experimental reviews, see Refs. [2–8].

1.2.1. Elasticity of Double-Stranded DNA

Smith and coworkers [1] found that the force-extension curves of lambda-phage DNA do not follow predictions based on standard Gaussian polymer model or freely-jointed-chain model. Later it was realized that semiflexible wormlike-chain polymer could reproduce the observation of Smith et al. to a large extent [16].

Bensimon and coworkers [26] used a receding meniscus to measure the elasticity of DNA. They found that DNA could be stretched up to 2.14 times its relaxed contour length before

breaking. Bustamante and coworkers also announced briefly in a conference abstract [27] that DNA could be stretched to 1.85 times its B-form length before breaking and that the overstretching was initiated suddenly at about 50 pN. In the experimental setup of Smith et al. [27], the ends of a lambda-DNA were attached to two microbeads and moved by micropipettes. Later, Bustamante and coworkers repeated this experiment by using optical tweezers [18]. The unzipping force was more precisely measured to be about 65 pN, and the overstretched DNA was found to be 1.7 times its B-form length. Caron and coworkers [19] performed a similar experiment but with an optical fiber as the force lever; their result was consistent with that of Smith et al. [18].

Croquette and coworkers [20] fixed the two strands of a lambda-DNA to a glass slide at one end and to a magnetic bead at the other end. They then applied a magnetic field to rotate and stretch the bead, thereby adding torsional as well as elongational stress in the DNA double-helix [20]. Very interesting behaviors of DNA were observed in this and subsequent experiments [20–23]. The elastic response of an undertwisted DNA is quite different from that of an overtwisted DNA. Overtwisting might cause B-form DNA to transit into a new P-form DNA with exposed nucleotide bases [23], whereas undertwisted DNA might be in a strand-separated form [21] or a left-handed Z-form [28–30].

The dynamics of single DNA macromolecules in elongational or shear flows were studied by Chu and coworkers [31–33].

1.2.2. Elasticity of Single-Stranded DNA

Smith and coworkers [18] studied the elastic response of single-stranded (ss) lambda-phage DNA at the action of an external force. They found that the elasticity of ssDNA could be well modeled by a freely-jointed-chain model with Kuhn length 1.5 nm and stretching modulus 800 pN, or by a wormlike-chain model. Croquette and coworkers [34] also studied the elastic response of ssDNA using magnetic bead and magnetic force. At high salt conditions, both the modified freely-jointed-chain model and the wormlike-chain model could not reproduce the force-extension curve of ssDNA when the external force is less than 5 pN. Bustamante and coauthors [18, 35, 36] also observed similar behavior. This discrepancy between theory and experiment was attributed to the formation of hairpins in ssDNA chain.

At low salt conditions and low external forces, Bustamante et al. [2] also observed that the force-extension curve of ssDNA is much deviated from that of a freely-jointed chain. This was caused by strong electrostatic repulsive interaction between negatively charged nucleotide monomers.

1.2.3. DNA Unzipping and Rupture

Heslot and coworkers [37] separated the two strands of a lambda-phage DNA by fixing the 5′-end of one strand to a glass slide and the 3′-end of the other strand to a soft force lever and moving the force lever at constant velocity. They found that the resulting force-extension is directly related to DNA sequence [37–40]. DNA unzipping experiments were also performed by Danilowicz et al. [41] under the condition of fixed external force.

Lee and coauthors [42] fixed the 5′-end of one strand of a DNA oligomer of 12 base pairs to a surface and the 5′-end of the other strand to the tip of an atomic force microscope and then stretched the molecule. The rupture force between the two strands was measured to be 410 pN in 0.1 M NaCl [42, 43]. Similar experiments on DNA oligomers of various lengths and sequences were undertaken by Noy and coworkers [44], Strunz and coworkers [45], and Pope and coworkers [46] under different ionic conditions. Very surprisingly, the rupture forces for DNA oligomers are typically much higher than the melting forces of long DNA polymers. Rief and coworkers [47, 48] measured the melting force for lambda-phage DNA at room temperature to be about 150 pN. This dramatic difference may be due to the following fact: In lambda-phage DNA, there are many single-stranded breaks, and these nicks enable DNA's torsional strain induced by stretching to be released. However, single-stranded breaks are absent in the DNA oligomers studied in Refs. [42–46].

The unzipping kinetics of short DNA chains was studied by Sauer-Budge et al. [49] by letting single-stranded but not double-stranded DNA segment to pass through a nanopore.

They found that the strand separation for perfectly matched DNA double-helices of tens of base pairs was an all-or-nothing process. However, there are unzipping intermediate states when a DNA oligomer with a single mismatch was forced to traverse the nanopore. The experimental setup of Ref. [49] is therefore capable of discriminating between minute differences in DNA sequences.

Even the unzipping forces between single nucleotide base pairs were measured by atomic force microscopy. Boland and Ratner [50] found that the force to separate a single A–T base pair is 54 pN. This value, however, is much higher than the unzipping force of 9–12 pN measured in Refs. [37, 47].

1.2.4. RNA Unfolding

The unfolding processes of several native RNA molecules were studied by many authors [51–54]. These nonequilibrium force-extension measurements were used to construct the energy landscape of RNA unfolding [52, 55, 56].

1.2.5. DNA–Protein Interactions

The interaction between DNA and RNA polymerases were studied by Block and coworkers [57–60] and also by Bustamante and coauthors [61]. These authors found that, at low external forces, RNA polymerases move along DNA template with an average velocity of 16 nucleotides per second. The stalling force for RNA polymerases is about 25 pN.

DNA replication dynamics was studied by Maier, Bensimon, and Croquette [34] and by Bustamante and coworkers [36] using single-molecule methods. It was found that the rate of DNA replication depends strongly on the applied force. When the external force exceeds 40 pN, DNA polymerase began to move back along the DNA template [36].

The working mechanisms of DNA topoisomerases were probed in Refs. [62].

The mechanical property of single chromatin fibers were studied by Cui and Bustamante [63]. One of the findings is that, at physiological ionic conditions, condensation-decondensation transition could be induced by an external force of 5 pN, suggesting the chromatin fiber to be in a dynamic structure because of thermal fluctuations. The dynamic structure of individual nucleosomes was also explored by Wang and coauthors [64] by stretching a tandem array of nucleosomes. The typical force needed to destroy the nucleosome structure was measured to be about 25 pN. The kinetics of chromatin assembly was recorded by Viovy and coauthors [65].

Poirier and Marko [66] demonstrated convincingly that in a mitotic chromosome, there is no contiguous protein scaffold. Covalent connection in a chromosome is therefore due to the DNA chain.

The interaction of RecA protein with double-stranded and single-stranded DNA were studied experimentally by Chatenay and coauthors [67]. These researchers found that structural fluctuations in double-stranded DNA are coupled to the RecA-DNA binding reaction. Related experiments were also undertaken by Libchaber and coworkers [68, 69] and by Bustamante and coauthors [35].

1.2.6. DNA Sequencing

Quake and coworkers [70] reported that sequence information of DNA could be obtained by performing experiments on just one single DNA molecule. The four types of nucleotides are labeled with four kinds of fluorescent agents. The incorporation of these nucleotides by a DNA polymerase that moves along the DNA chain is then recorded in real time. The authors [70] found that single-nucleotide resolution could be reached by this method.

It is anticipated that more and more single-molecule experimental observations on the interaction between protein and DNA will be reported in the near future.

1.3. Outline of This Chapter

On the theoretical side, analytical as well as computational methods have been developed in understanding single-molecule experimental observations. In Section 2 we outline some models on the melting phenomenon of double-stranded DNA. In Section 3, theoretical work

on the structural transitions in RNA secondary structures will be mentioned in detail. In Section 4, we review the entropic elasticity theory of DNA. In Section 5, we discuss DNA overstretching transition. In Section 6, we briefly review works done on the supercoiling property of DNA. Finally, Section 7 is reserved for conclusions.

There is a huge theoretical literature on the study of DNA and its interaction with proteins. The materials presented in this chapter are by no means complete. They are inevitably biased by our own expertise and taste. For example, theoretical investigations on the interaction between DNA and protein are completely neglected in this chapter. The reader is referred to other recent theoretical reviews (Refs. [71, 72]) for topics not covered in this chapter.

2. DNA DENATURATION AND UNZIPPING

As mentioned in Section 1.1, DNA is a double-stranded polymer, with the two strands bound together by hydrogen bonds. Compared to a covalent bond, a hydrogen bond is very weak. The energy needed to break a hydrogen bond is comparable to thermal energy $k_B T$ at room temperature of $T \simeq 300\ K$, where k_B is Boltzmann's constant. An A–T base pair contains two hydrogen bonds and a G–C pair contains three. Because of thermal fluctuations, at physiological temperatures a base pair may transit frequently between its paired state and the unpaired open state. Researchers are now quite interested in the biological significance of such local dynamical behavior of DNA double-helix (see, for example, Refs. [73–75]). Here we focus on the global equilibrium properties of DNA double-helix and review theoretical works on the denaturation transition in long DNA chains.

Experimentally, it has been known for many years that when a solution of DNA macromolecules is heated to about 80°C, the base pairs in the DNA double-helix break up cooperatively, and the two DNA strands dissociate from each other to form two separated random coils. This phenomenon is referred to as DNA *denaturation* or thermal DNA melting [14]. Even at relatively low temperature where thermal DNA melting does not occur, the two strands of DNA could be separated by applying an oppositely directed force on the two strands at the DNA terminal point [37, 41]. This phenomenon is called force-induced DNA melting, or *DNA unzipping*. DNA unzipping is also referred to as directional DNA melting [37], because the strand separation propagates from one end of the polymer chain to the other end.

The two strands of a DNA can also be separated by pulling its two opposite 5′ ends with a force (see, for example, Refs. [42–46]). This process is called DNA rupture. A theoretical investigation of DNA rupture can be found in Ref. [76]. In this section, we discuss only DNA unzipping and will not treat DNA rupture.

We investigate the statistical mechanics of thermal and directional DNA melting. In Section 2.1, we discuss the de Gennes–Peyrard–Bishop model. In Section 2.2, we introduce the Montanari–Mézard model and outline its main results. In Section 2.3, the renowned Poland–Scheraga model is recalled, followed by reviewing the calculations of Kafri and coworkers on the order of DNA denaturation transition. In Section 2.4, we discuss how the sequence information in DNA will be manifested in the force-extension curves in an equilibrium DNA unzipping experiment. In Section 2.5, we discuss kinetic issues related to DNA unzipping.

Earlier reviews on DNA thermal melting can be found in Refs. [24, 25, 77].

2.1. DNA Denaturation: de Gennes–Peyrard–Bishop Model

A simplified continuous model of DNA denaturation was investigated first by de Gennes [78] and later by Peyrard and Bishop [79]. This model regards that both base-pairing and base-pair stacking interaction are important to understand DNA denaturation. These two types of interactions are considered by the following Hamiltonian:

$$\mathcal{H} = \sum_{n=1}^{N} V(x_n) + \sum_{n=1}^{N-1} \frac{k}{2}(x_{n+1} - x_n)^2 \tag{1}$$

In Eq. (1), x_n denotes the bond length of the nth base pair, and $V(x)$ is the base-pairing potential. $V(x)$ is a short-ranged potential, and Peyrard and Bishop modeled it by a Morse potential for analytical convenience:

$$V(x) = D[e^{-ax} - 1]^2 \tag{2}$$

D is set to 5.29×10^{-22} J and $a^{-1} = 1.8$ Å, which corresponds to mean value for the hydrogen bonds in A–T and G–C base pairs.

The second term of Eq. (1) models couplings between two adjacent base pairs caused by base-pairing stacking. The coupling strength k is of the order 4.5×10^{-24} J/Å [79]. In general, the values of the parameters D, a, and k should all be sequence-dependent.

The partition function is

$$\mathcal{Z}(N) = \int_{-\infty}^{\infty} \prod_{n=1}^{N} dx_n \exp[-\beta \mathcal{H}(N)] \tag{3}$$

where $\beta = 1/k_B T$. Because base-pair stacking introduces only nearest-neighbor coupling in Eq. (1), the partition function Eq. (3) could be evaluated by transfer-matrix method [79]. Denote ϵ_i and $\phi_i(x)$ as, respectively, the ith eigenvalue and eigenfunction of the integral equation

$$\int_{-\infty}^{+\infty} dx' \exp\left[-\frac{k}{2}(x - x')^2 + V(x)\right]\phi_i(x') = e^{-\beta\epsilon_i}\phi_i(x) \tag{4}$$

then Eq. (3) is rewritten as

$$\mathcal{Z}(N) = \sum_{i=0}^{\infty} e^{-(N-1)\beta\epsilon_i} \int_{-\infty}^{+\infty} dx_N \, \phi_i(x_N) \int_{-\infty}^{-\infty} dx_1 \, e^{-\beta V(x_1)} \phi_i(x_1) \tag{5}$$

As long as the temperature T is low enough such that $d = (\beta/a)(2kD)^{1/2} > 1/2$, Eq. (4) has a confined ground-state eigenfunction of

$$\phi_0(x) = a^{1/2} \frac{(2d)^{d-1/2}}{[\Gamma(2d-1)]^{1/2}} \exp(-de^{-ax}) \exp\left[-\left(d - \frac{1}{2}\right)ax\right] \tag{6}$$

In the thermodynamic limit of $N \to \infty$, Eq. (4) reduces to $\mathcal{Z}(N) \propto \exp[-\beta N \epsilon_0]$. The average separation of a base pair is equation to

$$\langle x \rangle = \int_{-\infty}^{+\infty} x' \phi_0^2(x') \, dx' \tag{7}$$

Peyrard and Bishop [79] calculated the average separation between a DNA base pair based on Eq. (7). They found that when the temperature is increased, the hydrogen-bond stretching has a sigmoidal form and changes gradually from about 2 Å to about 10 Å over a temperature range of about 50 K. They also found that the transition temperature is greatly influenced by the coupling constant k in Eq. (1). The work of Peyrard and Bishop [79] suggested that base-pair stacking interactions as well as the short-ranged nature of base-pairing interactions are quite important to understand DNA denaturation. More refined models following Ref. [79] were also constructed and investigated in extensive detail [80–87].

The experimentally observed DNA denaturation is much more cooperative than that predicted in Ref. [79]. For this purpose, Dauxois and coauthors [81, 86] introduced nonharmonic nearest-neighbor stacking interactions. This was achieved by assuming the coupling k in Eq. (1) to be a separation-dependent quantity. Their calculations lead to a much sharper behavior than the original model [81, 86]. In Ref. [85], twisting degrees of freedom are also incorporated into the model to improve the agreement between theory and experiment.

The original calculations of Ref. [79] considered only the (localized) ground-state eigenvector of Eq. (4). For short DNA molecules where the thermodynamic limit does not hold, it is necessary to consider both the localized and extended eigenfunctions of Eq. (4). Extensive numerical work of Zhang and coauthors [84] demonstrated that, after this improvement,

actually the Peyrard–Bishop model could reproduce the melting curves of some naturally occurring short DNA chains with high precision.

Zhou discussed the phase-transition issue of DNA denaturation based on the de Gennes–Peyrard–Bishop model. In his calculation, the Morse potential Eq. (2) was replaced by an asymmetric δ potential for analytical convenience

$$V(x) = \begin{cases} +\infty & x < 0 \\ -\gamma\delta(x-a) & x \geq 0 \end{cases} \tag{8}$$

In addition, an energy term $-fx_N$ is added into Eq. (1) to mimic the effect of external unzipping. By taking into all the eigenstates of this model system, the Laplace transform of the partition function is exactly worked out [87]

$$G(z) = \sum_{N=0}^{\infty} \exp(-zN)\mathcal{Z}(N) = \frac{8\sqrt{z(2k\beta)^3}a[\exp(\beta fa/2) - \exp(-\sqrt{2zk\beta}a)]}{[8zk\beta - \beta^2 f^2]2\sqrt{2zk\beta}a - \tau[1 - \exp(-2\sqrt{2zk\beta}a)]} \tag{9}$$

where $\tau = 2ak\gamma\beta^2$. The largest solution of the equation $1/G(z) = 0$ corresponds to the linear free energy density of the polymer system.

Reference [87] showed that denaturation is a localization–delocalization transition process. When $T < T_m = \sqrt{2k\gamma a}/k_B$, there is one localized eigenstate, corresponding to base-paired DNA. The phase-transition occurs at T_m when the localized ground-state disappears. When there is no external force, the denaturation transition is second-order, provided that the hydrogen-bonding interaction $V(x)$ is asymmetric [for symmetric potential $V(x)$ there is no true phase-transition]. When the external force is non-zero, the denaturation is first-order. The phase-diagram based on this simplified model is shown in Fig. 3.

The de Gennes–Peyrard–Bishop model is essentially one-dimensional. The configurational fluctuations in DNA strands are all neglected. In the next section, we discuss a more realistic model of DNA denaturation that considers DNA as a three-dimensional object.

2.2. DNA Denaturation: Montanari–Mézard Model

Montanari and Mézard [88] constructed a three-dimensional polymer model to study the secondary structures in RNA polymers (see Section 3). However, as we show here, this model is also suitable to study DNA denaturation and unzipping.

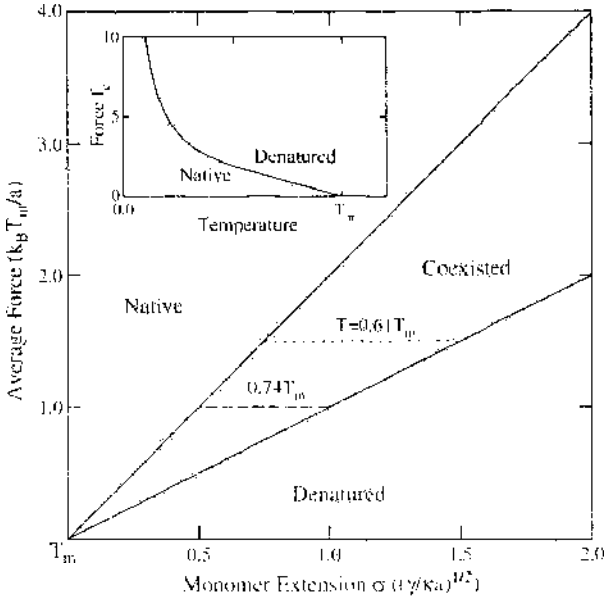


Figure 3. Phase diagram for a double-stranded biopolymer with an asymmetric potential. This system shows second-order thermal melting behavior at zero force. Inset shows how the threshold force for melting changes with temperature T.

In the Montanari–Mézard model, each DNA strand is regarded as a deformable freely jointed chain of N bonds and $N + 1$ beads (see Fig. 4). The $2(N + 1)$ beads along the two strands are located at spatial positions $\{\mathbf{r}_0^\alpha, \mathbf{r}_1^\alpha, \ldots, \mathbf{r}_N^\alpha\}$, with $\alpha = 1$ for the first strand and $\alpha = 2$ for the second strand. In each strand, the bond linking two consecutive beads i and $i + 1$ has the following bond length distribution [88]:

$$\mu(\mathbf{u}_i) \propto \exp\left(-\frac{(|\mathbf{u}_i| - b)^2}{2l_0^2}\right) \tag{10}$$

where $\mathbf{u}_i = \mathbf{r}_{i-1}^\alpha - \mathbf{r}_i^\alpha$ is the bond vector, b is the intrinsic length of the bond, and l_0 is the typical magnitude of bond length fluctuations. For single-stranded DNA, $b \simeq 1.7$ nm [2, 18] and $l_0 \sim 0.1$ nm.

A bead of the model is an effective interaction unit and corresponds to about three nucleotide bases of a real DNA strand. If the distance $\mathbf{r}_i = \mathbf{r}_i^1 - \mathbf{r}_i^2$ between the ith pair of beads is within some range a, there is an attractive interaction ϵ_i. The attractive interaction is base pair dependent, a G–C pair is stronger than an A–T pair. We are interested here in the general physics, so we assume all the pairing interactions to be sequence-independent: $\epsilon_i \equiv -\epsilon_0$, with ϵ_0 being the average pairing energy between two complementary DNA segments of length corresponding to one bead of the model. The effect of sequence heterogeneity to DNA unzipping will be discussed later. Cule and Hwa have also performed Monte Carlo simulation to study the changing of denaturation behavior due to sequence heterogeneity; see Ref. [83].

Apart from this kind of "on-site" base-pairing interactions, the stacking interaction between two consecutive base pairs is also considered. If both $|\mathbf{r}_i| \leq a$ and $|\mathbf{r}_{i-1}| \leq a$, then there is an additional stacking energy gain of $-\epsilon_s$. In principal, ϵ_s should also depend on sequence, but this complication is dropped here.

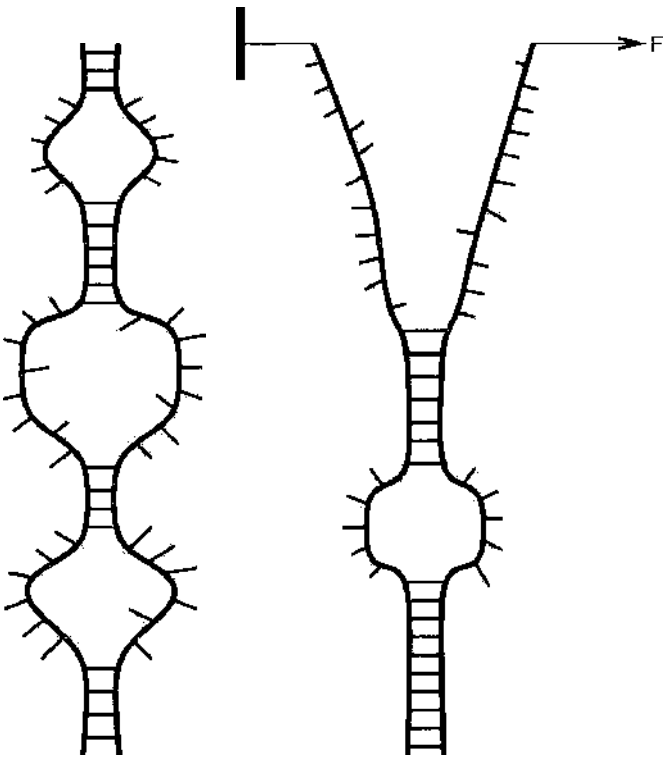


Figure 4. DNA denaturation and unzipping based on the Montanari–Mézard model. Each DNA strand is modeled as a deformable freely-jointed chain; there is on-site short-range attractive interaction between the ith pair of bases if their separation is less than distance a. (Left) DNA thermal melting (denaturation), with the formation of denaturation bubbles. (Right) An external force may act on the last base pairs to pull the two strands apart (unzipping).

In the following theoretical treatment, excluded volume effect is neglected, and the local helical structure of DNA is not considered. The model could be investigated by many means. Here we follow the graph-counting and generating-function method used in Refs. [88–93]. Section 2.2.1 investigates DNA denaturation and Section 2.2.2 treats unzipping.

2.2.1. DNA Thermal Melting

The partition function of a double-stranded polymer of $2n$ bonds is denoted by $Z_{ds}(n, \mathbf{r})$, where $\mathbf{r}$ is the separation between the last [$(n+1)$-th] pair of beads (see Fig. 4, left). We need to distinguish between two situations: (1) the last pair of beads are not base paired (i.e., $|\mathbf{r}| > a$); (2) the last pair of beads are base paired (i.e., $|\mathbf{r}| \le a$). Denote the partition function for the first case as $Z_{ds}^{(u)}(n, \mathbf{r})$ and that of the second case as $Z_{ds}^{(p)}(n, \mathbf{r})$. The total partition function is

$$Z_{ds}(n, \mathbf{r}) = \Theta(a - |\mathbf{r}|)Z_{ds}^{(p)}(n, \mathbf{r}) + [1 - \Theta(a - |\mathbf{r}|)]Z_{ds}^{(u)}(n, \mathbf{r}) \tag{11}$$

In Eq. (11), $\Theta(x)$ is the Heaviside step function: $\Theta(x) = 1$ for $x \ge 0$ and $\Theta(x) = 0$ if $x < 0$.

Based on the graph representation of Fig. 4 (left), the following recursive equations are written down for $n \ge 1$:

$$Z_{ds}^{(p)}(n, \mathbf{r}) = \Theta(a - |\mathbf{r}|)e^{\beta\epsilon_0} \int \left[\prod_{i=1}^{3} d\mathbf{u}_i\right] \mu(\mathbf{u}_1)\mu(\mathbf{u}_2)\delta\left(\sum_{i=1}^{3} \mathbf{u}_i - \mathbf{r}\right) \times \left[Z_{ds}^{(u)}(n-1, \mathbf{u}_3) + e^{\beta\epsilon_s} Z_{ds}^{(p)}(n-1, \mathbf{u}_3)\right] \tag{12}$$

$$Z_{ds}^{(u)}(n, \mathbf{r}) = [1 - \Theta(a - |\mathbf{r}|)] \int \left[\prod_{i=1}^{3} d\mathbf{u}_i\right] \mu(\mathbf{u}_1)\mu(\mathbf{u}_2) Z_{ds}(n-1, \mathbf{u}_3)\delta\left(\sum_{i=1}^{3} \mathbf{u}_i - \mathbf{r}\right) \tag{13}$$

Equations (12) and (13) are supplemented with the initial condition $Z_{ds}^{(p)}(0, \mathbf{r}) = \delta(\mathbf{r})$ and $Z_{ds}^{(u)}(0, \mathbf{r}) = 0$.

Hydrogen-bonding is a short-ranged interaction, with its range of interaction a much shorter than the Kuhn length b of a single-stranded DNA chain: $a \ll b$. Take into account this fact, a good approximation is to write $\Theta(a - |\mathbf{r}|) \simeq \frac{4\pi}{3}a^3\delta(\mathbf{r})$. This approximation makes our following analytical calculations much easier to perform.

The Fourier transform of a function $f(\mathbf{r})$ is indicated by an over-tilde and is defined as

$$\tilde{f}(\mathbf{p}) = \int d\mathbf{r} \exp(i\mathbf{p} \cdot \mathbf{r}) f(\mathbf{r}) \tag{14}$$

Equations (11), (12), and (13) are Fourier transformed to

$$\tilde{Z}_{ds}^{(p)}(n, \mathbf{p}) = \frac{a^3}{6\pi^2} e^{\beta\epsilon_0} \int d\mathbf{q} \tilde{\mu}^2(\mathbf{q}) \tilde{Z}_{ds}(n-1, \mathbf{q}) + \frac{a^3}{6\pi^2} e^{\beta\epsilon_0}(e^{\beta\epsilon_s} - 1) \int d\mathbf{q} \tilde{\mu}^2(\mathbf{q}) \tilde{Z}_{ds}^{(p)}(n-1, \mathbf{q}) \tag{15}$$

$$\tilde{Z}_{ds}^{(u)}(n, \mathbf{p}) = \tilde{\mu}^2(\mathbf{p}) \tilde{Z}_{ds}(n-1, \mathbf{p}) - \frac{a^3}{6\pi^2} \int d\mathbf{q} \tilde{\mu}^2(\mathbf{q}) \tilde{Z}_{ds}(n-1, \mathbf{q}) \tag{16}$$

$$\tilde{Z}_{ds}(n, \mathbf{p}) = \tilde{Z}_{ds}^{(p)}(n, \mathbf{p}) + \tilde{Z}_{ds}^{(u)}(n, \mathbf{p}) \tag{17}$$

In the above equations, $\tilde{\mu}$ is the Fourier transform of the bond vector distribution Eq. (10),

$$\tilde{\mu}(\mathbf{p}) = \int d\mathbf{r} \exp(i\mathbf{p} \cdot \mathbf{r}) \mu(\mathbf{r}) = \frac{\sin b|\mathbf{p}|}{b|\mathbf{p}|} \exp\left(-\frac{l_0^2|\mathbf{p}|^2}{2}\right) \tag{18}$$

The generating function $G(z, \mathbf{p})$ of $\tilde{Z}_{ds}(n, \mathbf{p})$ is defined by

$$G(z, \mathbf{p}) = \sum_{n=0}^{+\infty} \tilde{Z}_{ds}(n, \mathbf{p}) z^n \tag{19}$$

Based on Eqs. (15), (16), and (17), we arrive at the following expression:

$$G(z,\mathbf{p}) = \frac{1}{1 - z\tilde{\mu}^2(\mathbf{p})} \times \frac{1}{1 - \frac{za^3}{6\pi^2}e^{\beta\epsilon_{ij}}(e^{\beta\epsilon_s} - 1)\int d\mathbf{q}\,\tilde{\mu}^2(\mathbf{q})\left[1 + \frac{za^3}{6\pi^2}\int d\mathbf{q}\frac{\tilde{\mu}^2(\mathbf{q})}{1 - z\tilde{\mu}^2(\mathbf{q})}\right] - \frac{za^3}{6\pi^2}(e^{\beta\epsilon_{ij}} - 1)\int d\mathbf{q}\frac{\tilde{\mu}^2(\mathbf{q})}{1 - z\tilde{\mu}^2(\mathbf{q})}} \tag{20}$$

The total partition of a double-stranded polymer of $2n$ bonds is

$$Z_{ds}(n) = \int d\mathbf{r}\, Z_{ds}(n,\mathbf{r}) = \tilde{Z}_{ds}(n,\mathbf{0}) \tag{21}$$

The generation function of $Z_{ds}(n)$ is nothing but $G(z,\mathbf{0})$. For a linear polymer system with only short-range interactions, the total free energy is a linear function of system size n, plus correlation terms that scale logarithmically with n in leading order [89]. We expect to find the following asymptotic expression for the partition function $Z_{ds}(n)$ when $n \gg 1$:

$$Z^{(n)} \sim n^{\gamma} \exp\left(-\frac{ng}{k_B T}\right) \tag{22}$$

where g is the free energy linear density of the system and γ is a scaling constant. From Eq. (19) and Eq. (22), we know that $g = k_B T \ln z_c$, where z_c is the smallest positive singular point of the function $G(z,\mathbf{0})$.

Function $G(z,\mathbf{0})$ has a singular point at $z = z_{coil} = 1$ and another singular point at the point z_{native} determined by $y(z_{native}) = 0$, with

$$y(z) = 1 - \frac{za^3 e^{\beta\epsilon_{ij}}(e^{\beta\epsilon_s} - 1)}{6\pi^2}\int d\mathbf{q}\,\tilde{\mu}^2(\mathbf{q})\left[1 + \frac{za^3}{6\pi^2}\int d\mathbf{q}\frac{\tilde{\mu}^2(\mathbf{q})}{1 - z\tilde{\mu}^2(\mathbf{q})}\right] - \frac{za^3(e^{\beta\epsilon_{ij}} - 1)}{6\pi^2}\int d\mathbf{q}\frac{\tilde{\mu}^2(\mathbf{q})}{1 - z\tilde{\mu}^2(\mathbf{q})} \tag{23}$$

Function $y(z)$ monotonously decreases with z in the range $0 \le z \le 1$, with $y(0) = 1$. Because $1 - \tilde{\mu}^2(\mathbf{q}) \propto \mathbf{q}^2$ as $|\mathbf{q}| \to 0$, we know from Eq. (23) that, for three-dimensional systems of our interest, $y(1)$ is finite but $\frac{dy(z)}{dz}|_{z=1} = -\infty$.

When environment temperature T is low, $y(1) < 0$ and $y(0) > 0$. There must be a root of $y(z_{native}) = 0$ with $0 < z_{native} < 1$. The smallest positive singularity of $G(z,\mathbf{0})$ is at z_{native}. When T is high enough such that $T > T_m$ (the melting temperature), $y(1) > 0$ and $y(0) > 0$. There is no root of $y(z) = 0$ in the range $0 \le z \le 1$. The smallest singularity of $G(z,\mathbf{0})$ is at $z_{coil} = 1$. There is a phase transition at the melting temperature T_m when $y(1) = 0$.

At any temperature T, the fraction of hydrogen-bonded base pairs $n_{hp}(T)$ is determined by

$$n_{hp}(T) = -k_B T \frac{d\ln z_c}{d\epsilon_{ij}} = \begin{cases} -k_B T \frac{d\ln z_{native}}{d\epsilon_{ij}} & (T \le T_m) \\ 0 & (T > T_m) \end{cases} \tag{24}$$

In Fig. 5, the number of intact base pairs n_{hp} as a function of T is shown for a double-stranded DNA made of 50% G–C base pairs and 50% A–T base pairs (the sequence is assumed to be random). The melting temperature is set to $T_m = 363.82$ K. Different curves correspond to different $(\epsilon_{ij}, \epsilon_s)$ values, whereas the sum of $\epsilon_{ij} + \epsilon_s$ is fixed to about 11.0 kJ/mol.

When temperature $T < T_m$, DNA is in the native phase, and the fraction of intact base pairs is non-zero. As the temperature approaches T_m from below, the number of intact base pairs vanishes to zero; and when $T > T_m$, DNA is in the denatured phase, with very few hydrogen-bonded base pairs. Figure 5 shows that the cooperativity of this denaturation transition depends on the relative importance of base-pairing and base-pair stacking interactions. If the energy of double-stranded DNA is mainly contributed by base-pairing interactions, then the denature transition will be a gradual processes (the bottom curve of Fig. 5), whereas

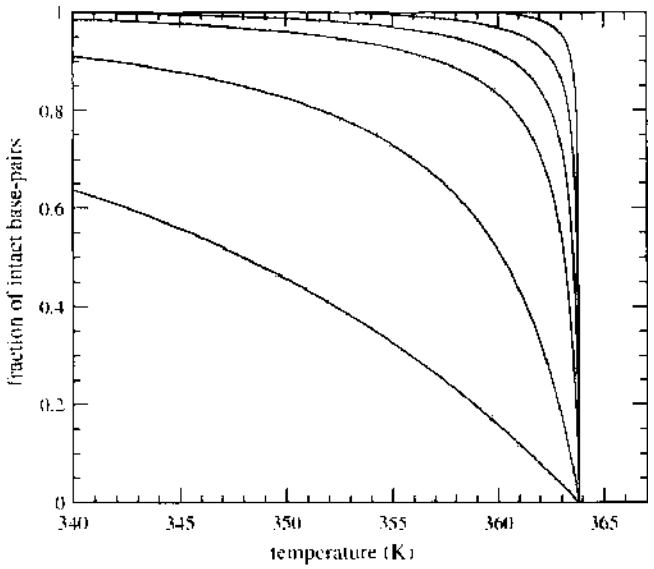


Figure 5. The melting curves of a hypothetical DNA made of 50% G-C base pairs and 50% A-T base pairs. The experimental melting temperature is about 90°C. The parameters used in the calculations are: $a=4.194$ Å, $b=1.7$ nm, $l_0=1.105$ Å. In these different curves, the base-pairing energy ϵ_0 and base-pair stacking energy ϵ_1 are set to $(\epsilon_0, \epsilon_1)=(1.247, 10.040)$, $(2.494, 8.768)$, $(3.742, 7.483)$, $(4.989, 6.178)$, $(7.483, 3.455)$, and $(9.977, 0.361)$ from top to bottom (here the energy is in units of kJ/mol).

if base-pair stacking energies contribute the main part of the energy, then the transition will be very cooperative (the top curve of Fig. 5). Experimentally, it was observed that the DNA thermal melting is a highly cooperative transition process [24, 94, 95], with a melting curve resembling the upper curves Fig. 5. When the temperature is slightly below T_m, most of the base pairs in DNA are intact, whereas when T is slightly above T_m, most of the base pairs are denatured. The Montanari–Mézard model with base-pair and base-pair stacking interactions is able to explain this highly cooperative phenomenon, if we assume that in DNA double-helix, base-pair stacking interactions are more important than base-pairing interactions.

One way to separate the contributions of base-pairing and base-pair stacking to the stability of DNA double-helix is to observe the melting curves of an ensemble of DNA molecules of the same nucleotide composition but different sequences. The total base-pairing interaction in these molecules are the same. Therefore, different melting behaviors must be due to differences in the base-pair stacking interactions. Recently, Montrichok and coworkers [96] combined UV spectroscopy and rapid quenching to demonstrate that the cooperativity of DNA denaturation is strongly sequence-dependent. For two DNA oligomers with similar A, T, G, and C compositions but with different sequence orders, one oligomer melts in a two-state way, whereas the other melted through many intermediates.

A question of academic interest is the order of the denature transition. Although the transition could be very cooperative as shown in Fig. 5, it is predicted to be *continuous* according to the model of this subsection. No latent heat is needed to complete the melting transition process. This is also consistent with the result of Ref. [87] based on the de Gennes–Peyrard–Bishop model of the preceding subsection. The transition is continuous because both the free energy and its first-order derivative with respect to temperature T are continuous at T_m. DNA thermal melting is driven by the competition between energy and entropy. A denaturation bubble (which is composed of two unpaired segments) could have many different configurations and hence are entropy-favored, whereas base-paired regions are energy-favored. When temperature is high enough, the gain in configurational entropy out-performs the loss in pairing energy and the two strands are therefore separated.

In Section 2.3, we will show that if self-avoidance effect is properly taken into account in the theoretical calculations, the melting transition could become first-order. Before doing this, however, let us first discuss force-induced DNA melting based on the present model.

2.2.2. Force-Induced DNA Melting

When there is an external force **F** pulling apart the double-stranded polymer (see Fig. 4, right), the total partition function Eq. (11) should be changed to $\exp(\beta \mathbf{F}\cdot\mathbf{r})Z_{ds}(n,\mathbf{r})$. The generating function of the new system has the same form as the expression in Eq. (20). The only difference is that **p** in Eq. (20) is replaced with $i\beta\mathbf{F}$.

For the new system, $z_{coil}=\tilde{\mu}^{-2}(i\beta F)<1$, and z_{native} is still given by the root of Eq. (23). When either the force or the temperature or both are large enough, the system is in the stretched coil state with two separated strands; when both the force and the temperature are small, the system is in the native paired state. The phase diagram in the force–temperature plan for a model DNA polymer made of 50% G–C pairs is shown in by the dashed line of Fig. 6. In this phase diagram, there is a critical force–temperature curve separating the native and the denatured phase. This phase diagram is determined by the equation

$$y(z_{coil})=1 \tag{25}$$

When the external force is nonzero, there exists a *discontinuous* first-order phase structural transition when the critical force–temperature curve is crossed. The phase transition is first-order because the free energy density is continuous, whereas its first-order derivative respect to temperature or force is discontinuous. The force-induced first-order phase transition is consistent with the de Gennes–Peyrard–Bishop model [87, 97]. Marenduzzo and coauthors [98] have studied an elegant three-dimensional directed lattice model of DNA force-induced unzipping. They also predicted a first-order phase transition.

According to the phase-boundary (the dashed line) of Fig. 6, one would anticipate that room temperature of $T=300$ K, an stretching force of about 4 pN is enough to pull the two strands of DNA apart. This is not the case. According to the experimental work of Essevaz-Roulet et al. [37] and Rief et al. [47], at room temperature, an external force in the range 9–20 pN is needed to separated the DNA double-helix.

There are many reasons that may related to this quantitative disagreement between theory and experiment. For example, in the model we mentioned earlier, there is only an energetic contribution to base-pair stacking; this is denoted by the constant energy parameter ϵ_s. However, in the actual case, for the two consecutive base pairs to stack onto each other, the DNA backbones must be twisted and folded. This confinement costs entropy, but this

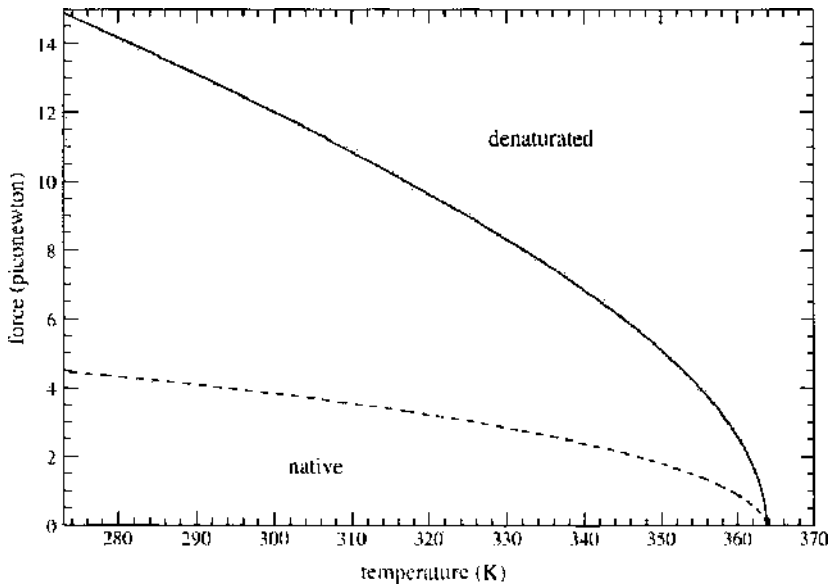


Figure 6. The phase-diagrams of a model DNA molecule made of 50% G–C base pairs and 50% A–T base pairs. In the calculation, $\epsilon_0=3.742$ kJ/mol, $\epsilon_s=7.483$ kJ/mol (for the dashed curve), and $\epsilon_s=\delta H_s-T\delta S_s$ (for the solid curve), where $\delta H_s=72.893$ kJ/mol and $\delta S_s=0.180$ kJ/(mol K). All the other parameters are the same as those listed in the caption of Fig. 5.

entropic effect is not considered in the model. Similarly, the free energy contribution to base-pairing is also assumed to be enthalpic in nature in our previous calculations. Unfortunately, to separate the enthalpic and entropic contributions to the free energy of pairing and stacking has not been an easy task and much work still needs to be done (see, for instance, Refs. [14, 25, 99, 100] and references cited therein). Here, to demonstrate the effect of taking into account of entropic contribution to base-pair stacking free energy, we separate the free energy of base-pair stacking into two parts; one part comes from the average stacking energy, and the other part comes from entropic penalties:

$$\epsilon_s = \delta H_s - T\delta S_s \tag{26}$$

In the above equation, δH_s is the enthalpic difference between a stacked and an unstacked pair of base pairs, and δS_s is the entropic difference between a stacked and unstacked pair of base pairs [14, 25]. Under this modification, the new phase boundary is shown by the solid line of Fig. 6, which is in better agreement with experimental observations. In general, one should also consider the temperature dependence of the pairing free energy ϵ_0.

A method of combining external force and temperature to measure various thermodynamic parameters of DNA could be suggested. One can fix the environment at various temperatures and measure the unzipping critical force of DNA. The resulting phase-diagram can then be fitted with theoretical results based on the model of this section to determine a set of free energy parameters that leads to a good fit to the experimental phase-diagram. We anticipate to see such progress in the near future.

2.3. Poland–Scheraga Model and Excluded-Volume Effect

It is quite difficult to consider excluded-volume interactions in the de Gennes–Peyrard–Bishop model and the Montanari–Mézard model. To incorporate the effect of excluded volume on the melting transition of DNA, we first mention a simplified two-state model, which is due to Poland and Scheraga [77, 90]. In the Poland–Scheraga model, each base pair exists either in a base-paired state or in an unpaired open state. A segment of DNA that contains only paired base pairs is called a double-helical segment, and a segment that is composed completely of opened base pairs is referred to as a denatured segment. DNA could be regarded as a linear chain of double-helical segments and denatured segments, with these two types of segments occur in alternative order. When two DNA segments come close together to each other, there are excluded volume interactions. Excluded volume interactions also exist between nucleotide bases in a single denatured DNA segment.

At first, let us discard any excluded volume effect and focus only on the base-pairing and base-pair stacking interactions. Without loss of generality, we assume the first and the last pair of bases are always in the paired state. Denote the partition function of a DNA double-helical segment of i ($i \geq 1$) base pairs by $Z_{\text{helix}}(i)$ and that of a denatured segment of j ($j \geq 1$) base pairs as $Z_{\text{bubble}}(j)$. The whole DNA polymer may be in the double-helix configuration, or there may be many double-helix segments and denaturation bubbles (see Fig. 7). The total partition function of a DNA of N base pairs is obtained by summarizing the contributions from all possible configurations:

$$Z(N) = Z_{\text{helix}}(N) + \sum_{s=1}\sum_{i_0=1}\sum_{j_1=i_0+1}\sum\sum\cdots\sum_{i_s=1}\sum_{j_s=1}\delta^{N}_{\sum_k i_k+\sum_l j_l}\prod_{m=1}^{s}[Z_{\text{helix}}(i_m)Z_{\text{bubble}}(j_m)]Z_{\text{helix}}(i_0) \tag{27}$$

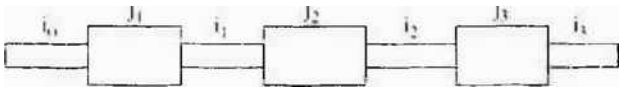


Figure 7. A schematic configuration of a DNA polymer in the Poland–Scheraga model. In this example, there are $s=3$ denaturation bubbles (length j_1, j_2, and j_3) and $s+1=4$ double-helical segments (length i_0, i_1, i_2, and i_3).

The generating function of Eq. (27) is

$$G(z)=\sum_{N=1}^{\infty} Z(N)z^N=\frac{G_{\text{helix}}(z)}{1-G_{\text{helix}}(z)G_{\text{bubble}}(z)} \tag{28}$$

where $G_{\text{helix}}(z)=\sum_{i=1}^{\infty} Z_{\text{helix}}(i)z^i$ and $G_{\text{bubble}}(z)=\sum_{j=1}^{\infty} Z_{\text{bubble}}(j)z^j$.

For a helical segment of i base pairs, the free energy is $-i\epsilon_0-(i-1)\epsilon_s$ due to base-pairing (ϵ_0) and base-pair stacking (ϵ_s). Therefore

$$G_{\text{helix}}(z)=\frac{ze^{\beta\epsilon_0}}{1-ze^{\beta(\epsilon_0+\epsilon_s)}} \tag{29}$$

The partition function $Z_{\text{bubble}}(j)$ is estimated as follows. As mentioned in Section 2.2, when neglecting all the possible interactions with a single strand, a single-stranded DNA could be modeled as a freely-jointed chain of bond length $b\simeq 1.7$ nm, with each bond corresponding to about three nucleotide bases. A DNA denaturation bubble of j base pairs therefore corresponds to a freely-jointed chain of $2(j+1)/3$ bonds, with the additional constraint that the two ends of the chain should be separated no more than $2a$ apart, where a is the base-pairing interaction range (see Section 2.2). Then

$$Z_{\text{bubble}}(j)=\Omega^{2(j+1)/3}\times \text{Probability}(|\mathbf{r}_1+\cdots+\mathbf{r}_{2(j+1)/3}|\le 2a) \tag{30}$$

$$\simeq c_0\frac{e^{j s_{\text{coil}}}}{(j+1)^{\gamma}}, \qquad \text{for } j\gg 1 \tag{31}$$

In Eq. (30), Ω denotes the total number of microscopic configurations of a bond; in Eq. (31), $\gamma=3/2$. $s_{\text{coil}}/2k_B$ is the configurational entropy of a single-stranded DNA base, and c_0 is a numerical constant. The scaling of the partition function of a denaturation bubble with bubble size, $j^{-3/2}$ is completely caused by entropic constraint of chain closure.

Based on Eq. (29) and Eq. (31), Eq. (28) is rewritten as

$$G(z)=\frac{ze^{\beta\epsilon_0}}{1-ze^{\beta(\epsilon_0+\epsilon_s)}-zc_0e^{\beta\epsilon_0}\sum_{j=1}^{\infty}\frac{(ze^{s_{\text{coil}}})^j}{j^{3/2}}} \tag{32}$$

According to Eq. (28), the free energy density of the system is

$$g_{\text{total}}=k_BT\ln z_s \tag{33}$$

where z_s is the smallest positive singular point of function $G(z)$.

$G(z)$ has two singular points: one is at $z=\exp(-s_{\text{coil}})$ and the other is at the point where $1/G(z)=0$. At low temperatures, the free energy of the whole system is determined by the second singular point. As temperature is increased, the second singular point approaches the first one, marking the transition from a native DNA into a denatured DNA. The transition is second-order, because at the transition point g_{total} is continuous and its first-order derivative with respect to temperature is also continuous [90].

2.3.1. Excluded-Volume Effect

How will the phase-diagram of DNA denaturation be changed if excluded-volume effect is taken into account? From our preceding discussion, it is obvious that the value of γ in Eq. (31) could influence the order of the denaturation transition. If $\gamma>2$, the transition is *first-order*, as Eq. (32) predicts.

Many analytical and computational efforts have been made to include excluded-volume effect in the Poland–Scheraga model. In 1966, Fisher [101] suggested a modified Poland–Scheraga model to include the excluded-volume effect between the two strands of each denaturation bubble. He found that $\gamma\simeq 1.75$ by this modification, so the transition is still second-order. Grassberger and coworkers [102] and Carlon and coworkers [103] performed numerical simulations on a lattice double-stranded polymer model. They have considered the self-avoidance effect in the simulation. The result of Refs. [102, 103] shows that, when

excluded volume effect is fully considered, the DNA denaturation transition is *first order*. The scaling exponent $\gamma \simeq 2.10$ as estimated by Carlon et al. [103]. Zhou [87] mapped the DNA melting problem to a quantum-mechanical problem and related the melting point with the vanishing of a confined state in the ground-state eigenvectors. However, because self-avoidance effect is only partially considered in the work of Ref. [87], the denaturation transition is still predicted to be continuous. In the model studied by Garel and coauthors [104], excluded-volume interaction between the two chains of DNA was considered, and intro-chain repulsions were neglected. This partially simplified model was found to predict a first-order denaturation transition in all dimensions [104] (but see comment by Bhattacharjee [105]). To further understand the results of Ref. [104], Baiesi and coauthors [106] considered the statistics of two mutually avoiding random walks and found that the reunion scaling exponent of such a pair of chains is $\gamma > 2$ both in two and three dimensions.

Here, we mainly review the work of Kafri, Mukamel, and Peliti [107, 108]. In the original Poland–Scheraga model, no excluded-volume effect is considered. When excluded-volume effect is taken into account, the available configuration of a DNA loop will be decreased. This decrease in the available configurations comes from two sources: first, the excluded-volume effect between the two strands of the loop [101]; second, the excluded-volume effect between the loop and the rest of the DNA chain. Fisher has considered the excluded-volume interaction between the two strands of the same loop, but it is not enough to account for the first-order phase transition. Kafri, Mukamel, and Peliti extended the work of Fisher to consider the interaction between one denaturation bubble and the rest of the DNA chain. They found that if this kind of interaction is considered, the scaling exponent $\gamma \geq 2.115$ [107], leading to a first-order phase transition. They have arrived at this conclusion taken using of the scaling properties of polymer networks developed by Duplantier and coworkers [109–111].

The arguments of Kafri and coauthors are as follows [108]. Consider the configuration of Fig. 8. When the polymer is considerably long, all the segments in Fig. 8 could be regarded as flexible. These segments form a polymer network of four chains and one loop. There are also two vertices (A and D) of degree unit and two vertices (B and C) of degree three. Generally, Duplantier and coworkers [109–111] showed that, for a polymer network made of M segments of length $l_1, l_2, \ldots, l_M$ (the total length is $L = \sum_i l_i$), the total partition function has the following scaling form:

$$Z_g = L^{\gamma_g - 1} z\left(\frac{l_1}{L}, \frac{l_2}{L}, \ldots, \frac{l_M}{L}\right) \tag{34}$$

where function $z(\cdots)$ is a scaling function and γ_g is obtained by

$$\gamma_g = 1 - 3\nu N_{\text{loop}} + \sum_{k=1} n_k \sigma_k \tag{35}$$

In Eq. (35), ν is the scaling exponent related with the radius of gyration of a self-avoiding chain in three dimensions, $\nu \simeq 0.588$ [111, 112]; N_{loop} is the total number of independent loops in the polymer network; n_k is the total number of vertices of degree k; and σ_k is a scaling exponent associated with a vertex of degree k.

Figure 8. A dsDNA of $2(N + n)$ bases, with a loop of $2n$ bases embedded in the middle. When both N and n are very large, the stem segments AB and CD and the two strands in the loop BCB could be regarded as flexible chains.

For the polymer network of Fig. 8, when $n \ll N$, Eq. (34) leads to

$$Z_g = (N+n)^{2\sigma_1+2\sigma_3-3\nu} z(n/N) \tag{36}$$

When $n/N \to 0$, Z_g should approach the partition function of a single polymer. Therefore, $g(n/N) \sim (n/N)^{2\sigma_3-3\nu}$. The total partition function then has the following scaling form:

$$Z_g = n^{2\sigma_3-3\nu} N^{2\sigma_1} \tag{37}$$

indicating that the total partition function is the partition function for a single loop, n^γ (with $\gamma = 3\nu - 2\sigma_3$), and that of the rest of the polymer, $N^{2\sigma_1}$.

It was known that $\sigma_3 = -0.175$ [111], so $\gamma \simeq 2.115$ in Eq. (31). Because it is larger than 2, according to the Poland–Scheraga model we mentioned earlier, the denaturation phase transition will be first-order.

Because the loop size distribution Eq. (31) is geometric with $2 < \gamma < 3$, the average loop size will be finite when the temperature approaches the melting temperature from below. However, the variance of the loop size becomes infinite as the melting temperature is approached.

It should be remarked that in the literature, there is still debate concerned with the excluded-volume effect between a denaturation loop and the remaining part of the DNA molecule (see, for example, Refs. [113, 114]). The theoretical reasoning of Refs. [107, 108] assume the polymer under study is extremely long, but cooperative denaturation behaviors were usually observed with DNA chains with about 10^2 to 10^3 base pairs. As our analysis in Section 2.2 (and also the discussion in the book of Poland and Scheraga [77]) demonstrated, DNA denaturation could be very cooperative when base-pair stacking interaction is considered. It might be suggested that for DNA chains with length of the order 10^3 base-pairs, base-pair stacking instead of excluded-volume interaction is the main contribution to the cooperativity of denaturation. More experimental investigations are certainly needed to resolve this issue.

2.4. Force-Induced Unzipping of Real DNA Molecules

In Sections 2.1, 2.2, and 2.3, we have discussed the general physical aspects of DNA denaturation and unzipping. In this and the following subsections, we discuss how real DNA molecules behave under the action of an external force. First, we consider the case of long DNA molecules. In Section 2.5, we will focus on the kinetics of structural transitions of DNA oligomers.

Essevaz-Roulet, Bockelmann, and Heslot [37–39] performed the following unzipping experiment on single DNA molecules (Fig. 9). The two strands on one end of a phage-λ DNA (48,502 base pairs, total contour length 16.2 μm) are separately attached to a glass slide and a very soft force lever (a microneedle with elastic modulus about 1.7 pN/μm), while the other end of the DNA macromolecule was kept free (In Ref. [40], the soft microneedle was replaced with optical tweezer to further increase the precision). At room temperature $T = 300$ K and nearly physiological ionic conditions (150 mM NaCl), the two strands of DNA was separated by displacing the glass slide with constant velocity (in the range 20–800 nm/s). During the DNA unzipping process, the distances between the glass slide and the force lever were recorded in real-time, as well as the bending extent of the force lever.

Heslot and coworkers [37–39] found that the typical force for DNA unzipping is in the range 10–15 pN. The force needed to unzip a GC-rich DNA segment is larger than that to unzip an AT-rich one, consistent with the fact that a G–C pair has three hydrogen bonds whereas an A–T pair has only two. They found that the unzipping force is directly related to the GC contents: peaks of unzipping force correspond to peaks in the DNA GC contents. The experimental curves are highly reproducible, suggesting the unzipping experiment was probing the equilibrium unzipping properties of double-stranded DNA. The irregularities in the force–displacement curves in the experimental curves are not random noises. These irregularities contain information about the DNA sequence.

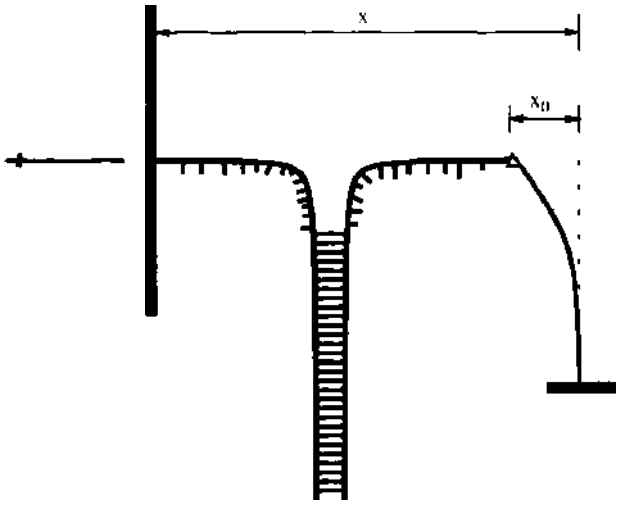


Figure 9. Schematic representation of the basic experimental setup of Essevaz-Roulet et al. [37–39]. The two strands on one end of a λ-DNA was attached to a movable glass slide and to a micro-needle. The strand separation process was followed by moving the glass slide with constant velocity and recording the time-dependent distance x between the glass slide and the micro-needle and the bending extent x_0 of the microneedle.

Trying to understand and reproduce the experimental unzipping force-extension profiles from theoretical calculations, an *equilibrium* slip-stick mechanism was suggested by Heslot and coworkers [38, 39]. Here we recall their model in some detail.

In the DNA–microneedle system, there exist the following three energy terms:

1. The energy cost of separating j base pairs, $E_{\mathrm{DNA}}(j)$. Bockelmann et al. [38, 39] suggested the following form

$$E_{\mathrm{DNA}}(j)=\sum_{\mu=1}^{j}E_{\mathrm{pair}}(\mu) \tag{38}$$

where $E_{\mathrm{pair}}(\mu)$ is the base-pairing energy of the μth base pair:

$$E_{\mathrm{pair}}(\mu)=\begin{cases}E_{\mathrm{G-C}}, & \text{for a G–C base pair}\\ E_{\mathrm{A-T}}, & \text{for an A–T base pair}\end{cases} \tag{39}$$

Therefore, E_{DNA} is determined by two phenomenological parameters $E_{\mathrm{G-C}}$ and $E_{\mathrm{A-T}}$, and sequence-dependent interactions between neighboring base pairs are averaged into these parameters. For the experiment with resolution in the range of about 100 base pairs [37], a two-parameter approximation is reasonable.

2. The elastic energy stored in the two single-stranded DNA segments. When the external force is F, the average extension of a single-stranded DNA segment of j base pairs is

$$x(F,j)=jl_{\mathrm{SS}}\left[\coth\left(\frac{Fb}{k_BT}\right)-\frac{k_BT}{Fb}\right]\left(1+\frac{F}{S}\right) \tag{40}$$

where l_{SS} is the contour length of one nucleotide unit in the single-strarded DNA, and S is the stretching modulus of a single-stranded DNA [115]. In deriving Eq. (40), the single-stranded DNA chain is modeled as a freely-jointed chain of Kuhn length b and stretching modulus S [115, 116]. The force needed at keeping the single-stranded DNA's extension fixed at x could be obtained by inversion of Eq. (40). The elastic energy to keep the extension of a single-stranded DNA at x is then

$$E_{\mathrm{elastic}}(x,j)=\int_0^x F(x')\mathrm{d}x'=F(x)x-\int_0^{F(x)}x(F',j)\mathrm{d}F' \tag{41}$$

3. The bending energy stored in the force lever,

$$E_{\text{lever}}(x_0) = \frac{k_{\text{lever}}}{2} x_0^2 \tag{42}$$

where x_0 is the total deflection of the tip of the force lever (Fig. 9).

The total energy of the system is

$$E_{\text{total}}(x_0, x_1, x_2, j) = E_{\text{DNA}}(j) + E_{\text{elastic}}(x_1, j) + E_{\text{elastic}}(x_2, j) + E_{\text{lever}}(x_0) \tag{43}$$

with the total displacement of the system fixed at $x = x_0 + x_1 + x_2$. The average force at total displacement x is

$$\langle F \rangle = \frac{\sum_{j=0}^{N} \int dx_1 \int dx_2 k_{\text{lever}}(x - x_1 - x_2) \exp[-\beta E_{\text{total}}(x - x_1 - x_2, x_1, x_2, j)]}{\sum_{j=0}^{N} \int dx_1 \int dx_2 \exp[-\beta E_{\text{total}}(x - x_1 - x_2, x_1, x_2, j)]} \tag{44}$$

The theoretical force–displacement profile obtained with fitting parameter $E_{G-C} = 2.9\ k_B T$, $E_{A-T} = 1.3\ k_B T$, $b = 1.5$ nm, $S = 800$ pN, $Nl_{SS} = 30\ \mu$m, and $T = 300$ K is in good agreement with experimental data [38, 39]. The fine details of the experimental unzipping curves could be quantitatively reproduced to some extent by this simple model.

A detailed analysis of the end-to-end distance of λ-DNA revealed the existence of many sawtooth patterns [37]. Typically, when the stretching force increases, at first the distance between the two ends of DNA almost keeps constant (the stick phase). Then, as the increase in force is about 1–2 pN, suddenly the end-to-end distance of DNA is increased by about 100–200 nm (the slip phase). As the displacement of the whole system keeps increasing with a constant velocity, this stick-slip process repeats many times. Bockelmann et al. [38, 39] regarded this phenomenon as an equilibrium stick-slip process.

The stick-slip phenomenon in unzipping λ-phage DNA has its origin from jumping of the DNA "opening fork" between adjacent energy local minima. We explain this insight by an example. In Fig. 10, the minimum value of the total energy Eq. (43) as a function of the opening-fork position j is shown at two fixed displacements. When the total extension of the system is $x = 18.1\ \mu$m, the energy absolute minimum is located at $j_1 \simeq 10,050$. This energy absolute minimum is separated from a metastable state at $j_2 \simeq 10,150$ by an energy barrier of 10 $k_B T$ and another metastable state at $j_3 \simeq 10,260$ by a barrier of 20 $k_B T$. Between the

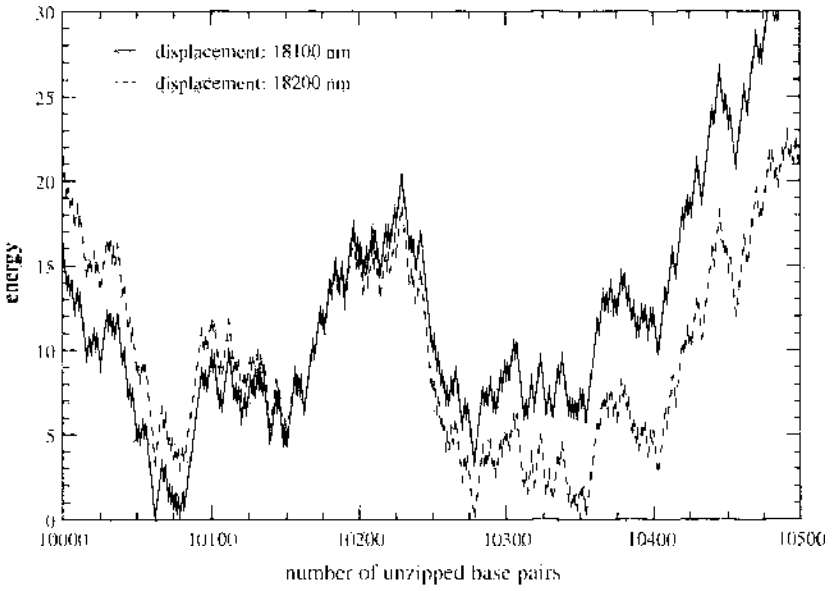


Figure 10. The energy landscape when the total displacement of the system is fixed at 18.1 μm (solid curve) and at 18.2 μm (dotted curve).

two metastable states at j_2 and j_3, there is also an energy barrier of 15 k_BT. At extension $x = 18.2$ μm, position j_3 now corresponds to the absolute energy minimum, and position j_1 becomes metastable. As the total extension changes from x_1 to x_2, the opening fork position will also tries to move from the old best position j_1 to the new best position j_3. This is achieved by two steps, from j_1 to j_2, and then from j_2 to j_3. Thermal activation causes a nucleotide base pair to switch frequently between the closed state and the opened state; the characteristic time scale is 3.6×10^{-6} s [75, 117]. However, to jump from j_1 to j_2, the thermal activation energy should be at least 10 k_BT, the probability of which is $e^{-10} \simeq 4.54 \times 10^{-5}$. Therefore, the typical time needed for the opening fork to jump from j_1 to j_2 is about 0.1 s. Similarly, the typical needed time to jump from j_2 to j_3 is about 1.0 s. In the experiment with displacement velocity 20 nm/s, it takes 5 s to move a distance of 100 nm. During this period, it is very possible that the jumping process $j_1 \to j_2 \to j_3$ has happened once. When the $j_1 \to j_2$ or the $j_2 \to j_3$ hopping happens, about 100 base pairs are unzipped cooperatively, leading to a rapidly drop in the pulling force.

For homogeneous DNA chains, the energy landscape is quite smooth, and the resulting force-extension profile will be featureless.

In Section 2.5, we will discuss more quantitatively why hopping processes such as that from position j_1 to position j_2 in Fig. 10 have a time scale of several seconds.

A practical issue is the possibility of using DNA unzipping to obtain sequence information of DNA. The analysis in this subsection suggests that, during DNA unzipping, the boundary between unzipped and base-paired DNA moves along DNA polymer by a hopping process: the hopping distance is of the order 10^1 to 10^2 base pairs. Therefore, one cannot attain single-nucleotide resolution by mechanical unzipping. On the other hand, as demonstrated by Quake and coworkers [70], combining fluorescence and moving of DNA polymerase along single-molecule DNA chain could be used to obtain sequence information with single-nucleotide resolution. We expect more experimental and theoretical investigations along this line.

2.4.1. Scaling Relationship in DNA Unzipping

The unzipping experiment of Essevaz-Roulet et al. [37] was performed by fixing the total extension of the DNA-microneedle system. It is also possible to pull the two strands of DNA apart with a constant force, as was done by Danilowicz et al. [41]. For an extremely long DNA under the action of constant external force, Section 2.2.2 showed that there exists a threshold force F_c. At given temperature, when the external force is lower than F_c, most of the base pairs of DNA are in the closed state, whereas when the external force is larger than F_c, the two strands of DNA are separated completely and there are no closed base pairs. It is interesting to study the scaling behavior of the number of opened base pairs as the external force approaches the transition force F_c from below.

First, let us consider the unzipping of a homogeneous DNA chain. Because of the first-order nature of the unzipping transition, the probability that n base pairs be opened is proportional to $\exp(-n\delta\epsilon_0)$, where $\delta\epsilon_0 \sim (F_c - F)$ is the free energy difference of a base pair in the paired state and in the opened state. Therefore, as $F \to F_c$, the average number of opened base pairs scales as

$$\langle n \rangle = \frac{\sum_{n=1}^{\infty} n \exp(-n\delta\epsilon_0)}{\sum_{n=1}^{\infty} \exp(-n\delta\epsilon_0)} \sim (F_c - F)^{-1} \tag{45}$$

When heterogeneity in a real DNA sequence (such as a λ-phage DNA) is considered, will the scaling relationship Eq. (45) still hold? Lubensky and Nelson [118] argued that the average number of opened base pairs should diverge even faster, with a scaling form of $\langle n \rangle \sim (F_c - F)^{-2}$! This is a surprising conclusion. Does the conclusion of Lubensky and Nelson [118] indicate that, as $F \nearrow F_c$, the free energy difference between a base pair in the closed state and that in the opened state is proportional to $(F_c - F)^2$? However, this should not be true, because the unzipping transition should be first order even in the presence of sequence heterogeneity.

Lubensky and Nelson [118] showed that for heterogeneous DNA, Eq. (45) is still true, but it is just one part of the contribution. The total number of opened base pairs has another contribution due to sequence heterogeneity. As $F_c - F$ is much smaller than a characteristic force determined by the sequence heterogeneity, this additional contribution, which scales as $(F_c - F)^{-2}$, becomes more important than Eq. (45).

Here, we explain the above-mentioned insight of Lubensky and Nelson in some detail. A more quantitative analysis could be found in the paper of Lubensky and Nelson [118].

When there is sequence randomness, the free energy increase caused by breaking n base pairs is

$$\delta E = n\delta\epsilon_0 + \sum_{i=1}^{n} \delta\epsilon_i \tag{46}$$

where $\delta\epsilon_0 \sim (F_c - F)$ is the average increase in the free energy caused by breaking one base pair, and $\delta\epsilon_i$ is the difference between the pairing energy of the ith base pair with respect to the average pairing energy. For simplicity, we model $\delta\epsilon_i$ as an independent and identically distributed Gaussian variable with standard deviation σ:

$$P(\delta\epsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\delta\epsilon_i)^2}{2\sigma^2}\right] \tag{47}$$

Figure 11 demonstrates the energy landscape of breaking n base pairs of a randomly generated DNA sequence according to Eqs. (46) and (47). The first term of Eq. (46) causes a constant slope of $\delta\epsilon_0$ to the free-energy profile, whereas the second term of Eq. (46) causes a deviation from this slope; the magnitude of this deviation scales as $\sqrt{n}\sigma$. Because of this random deviation, the free-energy minimum is not located at $n = 0$, but is shifted to n_{optimal} (in Fig. 11, $n_{\text{optimal}} \simeq 500$). This optimal number of base pairs opened is expressed by the following equation:

$$n_{\text{optimal}} \sim \frac{\sigma^2}{4(\delta\epsilon_0)^2} \tag{48}$$

Around the n_{optimal}, the free energy has a well of order of $-\frac{\sigma^2}{4\delta\epsilon_0}$, which is proportional to σ^2. When the number of broken base pairs goes beyond n_{optimal}, the first term of Eq. (46) dominates, and the free energy increases linearly with n. The total number of broken base pairs is then n_{optimal} plus Eq. (45):

$$\langle n \rangle \sim \frac{\sigma^2}{(\delta\epsilon_0)^2} + \frac{1}{\delta\epsilon_0} \tag{49}$$

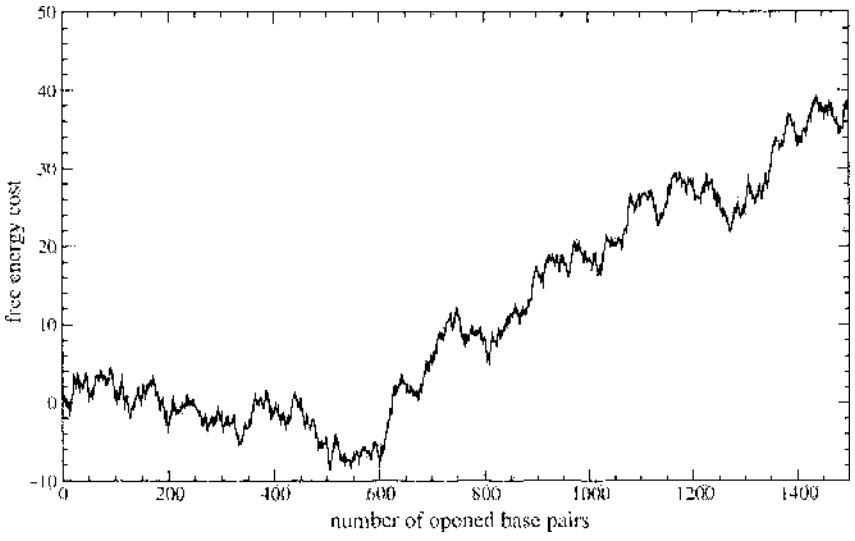


Figure 11. The free energy cost (in units of $k_B T$) due to unzipping n base pairs of a random-sequence DNA. In drawing the figure, $\delta\epsilon_0$ of Eq. (46) is set to 0.05 $k_B T$ and σ of Eq. (47) is set to 0.5 $k_B T$.

When the first term of Eq. (49) is less important than the second term, then Eq. (45) holds; however, when the first term outperforms the second term, then

$$\langle n \rangle \sim \frac{1}{(F_c - F)^2} \tag{50}$$

To observe the scaling form Eq. (50), one must have strong sequence heterogeneity [$\sigma \gg 0$ in Eq. (49)], or the external force is very close to the critical force F_c. Otherwise, the scaling form Eq. (45) will be observed. For the actual case of random sequence DNA, the binding energy of A–T and G–C pairs differ about 1 $k_B T$, which is considerable, so the scaling form Eq. (50) instead of Eq. (45) should be observed. Lubensky and Nelson have performed Monte Carlo simulations for several different heterogeneous DNA molecules [119] and their results confirmed the validity of Eq. (50).

2.5. Unzipping Kinetics of DNA/RNA Helix-Loops Under Constant Force

We mentioned that the hopping processes in unzipping λ-phage DNA occur with a characteristic time scale of *seconds*. Recently, single-molecule unzipping experiments were also performed on RNA [51, 52]. A helix-loop RNA of 49 bases, P5ab, has a native structure of a hairpin with 22 base pairs, plus a hairpin loop of four bases and a single-nucleotide bulge-loop. Liphardt et al. [51] revealed that, under a constant force of about 15 pN, P5ab switched between two well-defined "open" and "close" states on a time scale also of *seconds*.

On the other hand, experimental measurement of Altan-Bonnet and coauthors [75] reported that, at room temperature of $T = 300$ K, the closing–opening transition of a single nucleotide base pair takes place on a time scale of 10^{-6} s. In Section 2.4, we have briefly mentioned that this huge gap in time scales from microseconds to seconds is caused by the existence of high energy barriers. In this subsection, we look into this issue more quantitatively.

Coceo and coauthors [117] presented a semimicroscopic model to explain the above-mentioned huge gap in transition time scales. We review the calculations of Ref. [117] by focusing on the following model system: a hairpin-loop made of 12 G–C base pairs and a 4-base ATAT loop. The two ends of this single-stranded DNA are pulled apart by a constant force of magnitude f.

2.5.1. Model and Dynamic Rules

Under the pulling of a constant force f, the free energy associated with unzipping n bases is

$$G(n, f) = \sum_{i=1}^{n} \Delta g(i, f) \tag{51}$$

where $\Delta g(i, f)$ is the free-energy change due to opening the ith base pair at the boundary between the open and closed region.

$$\Delta g(i, f) = 2g_s(f) - g_0(i) \tag{52}$$

In Eq. (52), $g_0(i)$ is the free energy of opening base pair i. For the poly(GC) hairpin, the following values are used [117]:

$$g_0(i) = g_0 = -3.442\ k_B T, \quad (1 \le i \le 11)$$
$$g_0(12) = g_0 - g_{\text{loop}} = 3.062\ k_B T$$
$$g_{\text{loop}} = 6.504\ k_B T$$

$g_s(f)$ is the free energy of an opened DNA base. For simplicity, single-stranded DNA is regarded as a freely jointed chain, without considering its deformability:

$$g_s(f) = -k_B T \frac{l_{SS}}{b} \ln\left[\frac{\sinh \beta f b}{\beta f b}\right] \tag{53}$$

where $l_{SS} = 5.6$ Å, $b = 15$ Å [117].

The following semimicroscopic transition rates are assumed for each elementary step in the structural transition path. At time t, suppose the opening fork is at base-pair position n (the nth base pair is opened but the $n+1$th base pair is closed). The movement of this boundary is governed by an opening rate r_o and a closing rate r_c:

$$r_o(n) = re^{g_0(n)/k_BT} \tag{54}$$

$$r_c(f,n) = re^{2g_c(f,n)/k_BT} \tag{55}$$

In Eq. (54) and Eq. (55), r is essentially the microscopic rate for a base pair to move together or apart in the absence of tension or base-pairing interactions. Its value can be estimated by considering the inverse self-diffusion time for an object of nanometer scale [117]: $r = k_BT/2\pi\eta\ell^3 \approx 5\times10^6$ s^{-1}, with $\ell = 5$ nm, $\eta = 0.001$ Pa·s (the viscosity of water). Another way to estimate r is to fit the experimental data of Liphardt et al. [51] on P5ab unzipping. This fit gives $r = 3.6\times10^6$ s^{-1} [117].

The probability that at time t, the opening fork is at base pair n is $\rho_n(t)$. It is governed by the following Markovian equation

$$\begin{aligned}\frac{\partial\rho_n(t)}{\partial t} &= -[r_o(n+1)+r_c(f,n)]\rho_n(t)+r_o(n)\rho_{n-1}(t)+r_c(f,n+1)\rho_{n+1}(t)\\ &= -\sum_{m=0}^{N} T_{n,m}\rho_m(t)\end{aligned} \tag{56}$$

where the matrix $T_{n,m}$ $(0 \le n, m \le 12)$ is:

$$\begin{aligned}T_{n,n} &= r_o(n+1)+r_c(f,n)\\ T_{n,n+1} &= -r_c(f,n+1)\\ T_{n,n-1} &= -r_o(n)\end{aligned}$$

and all the other elements are zero.

2.5.2. Switching Kinetics

For the poly(GC) hairpin, the energy landscape as a function of opening fork position n is sawtooth shaped. At force $f = 15.95$ pN, the completely closed and the completely opened states have the same free energy. The highest energy state is at $n = 11$ base pairs, with a energy barrier of 10.17 k_BT.

The structural fluctuation of the poly(GC) hairpin is simulated based on the above-mentioned model. Figure 12 records the opening fork position as a function of time at fixed external force 15.95 pN. Figures 13 and 14 demonstrate the lifetime distribution for the polymer to be in the opened state and the closed state, respectively. Both of them obey exponential distribution. The mean lifetime of a closed state is $\bar{t}_{\text{mean close}} = 0.325$ s and that of an opened state is $\bar{t}_{\text{mean open}} = 0.208$ s; therefore, $\bar{t}_{\text{mean close}}/\bar{t}_{\text{mean open}} = 1.561$. These data are in close agreement with data obtained by equilibrium calculations mentioned below.

The eigenvalues of the matrix $T_{m,n}$ in Eq. (56) are listed below:

$$\begin{array}{ll}
1.\ 0 & 2.\ 2.233\times10^{-6}r\\
3.\ 1.748\times10^{-2}r & 4.\ 3.175\times10^{-2}r\\
5.\ 5.251\times10^{-2}r & 6.\ 7.790\times10^{-2}r\\
7.\ 1.060\times10^{-1}r & 8.\ 1.346\times10^{-1}r\\
9.\ 1.616\times10^{-1}r & 10.\ 1.850\times10^{-1}r\\
11.\ 2.032\times10^{-1}r & 12.\ 2.146\times10^{-1}r\\
13.\ 2.145\times10^{1}r &
\end{array} \tag{57}$$

Eigenvalue zero corresponds to equilibrium distribution of the opening fork position; the second eigenvalue corresponds to the transition between open and closed states, and it is

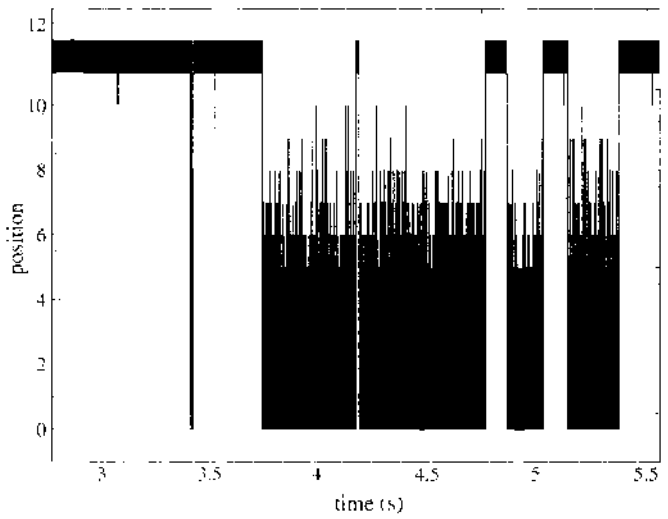


Figure 12. Typical structural fluctuations of a poly(GC) hairpin of 28 nucleotide bases. The external force is fixed at $f = 15.95$ pN. The position of the opening fork as a function of time is shown.

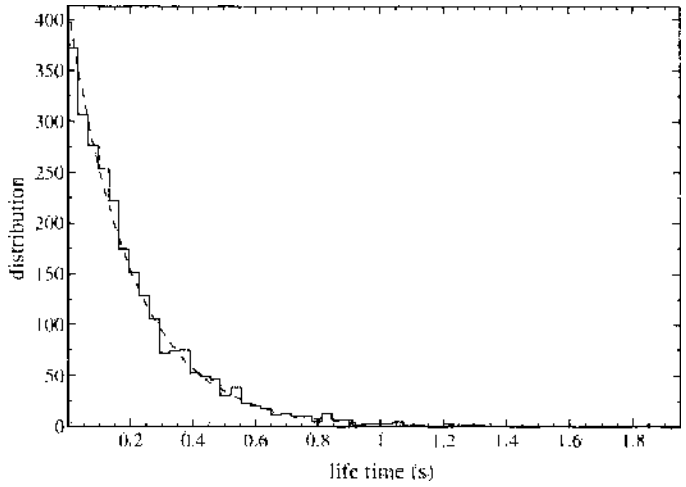


Figure 13. The lifetime distribution for a poly(GC) hairpin in the open configurations. The distribution could be well fitted by an exponential distribution (the dotted line) with mean reciprocal lifetime of $4.94 \pm 0.10\ s^{-1}$.

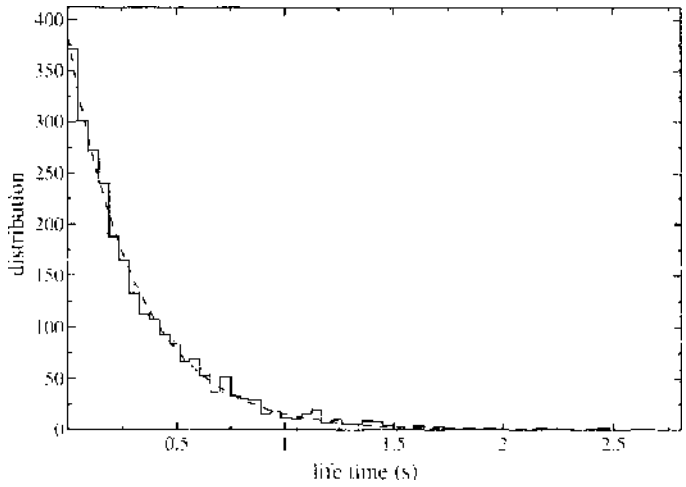


Figure 14. The lifetime distribution for a poly(GC) hairpin in the closed configurations. The distribution is fitted with an exponential distribution with mean reciprocal lifetime of $3.19 \pm 0.07\ s^{-1}$.

equal to 8.04 s^{-1}. All the other eigenvalues are of the order $10^{-1}r$ to $10^{1}r$ and correspond to internal structural transitions within the open and the closed states.

In the work of Cocco and coauthors [117], besides poly(GC) hairpins and P5ab RNA, some other real RNA molecules with more complex native structures were studied. The unzipping kinetic of λ-phage DNA under the action of a constant force was simulated [117]. In all these examples, unzipping at constant force proceeds by successive steps, each corresponding to the crossing of a free-energy barrier. This is similar with the conclusion of Section 2.4. Therefore, unzipping experiments performed at constant external force are also unlikely to access DNA sequence information with single-nucleotide resolution.

In Ref. [120], Cocco, Monasson, and Marko further considered the effect of torque contributions in the double-helix part and showed this torque could influence the unzipping and re-zipping kinetics when the force increase rate is above certain threshold that is comparable with the relaxation time of torques in double-stranded DNA. In Ref. [121], the force and kinetic barriers of initialization of DNA unzipping were studied detailedly.

In Ref. [122], Ritort et al. used a two-state kinetic model to study the unzipping dynamics of RNA and the relationship between work and load velocity. The dynamics of DNA unzipping was also studied by Marenduzzo et al. [123] by a lattice model. The relation between number of opened base pairs and time was obtained by simulation and was found to obey a scaling law.

In our above discussion, we have not considered the possibility of mismatched configurations. In RNA hairpins, such mismatched states are likely to occur. The folding–unfolding kinetics of RNA hairpins in the absence of external force was studied by Chen and Dill [124] and by Zhang and Chen [125]. It was found [124, 125] that when the environmental temperature is higher than a glass-transition temperature T_g, the folding–unfolding kinetics is two-state-like, similar to the result shown here. When the temperature is lower than T_g, the molecule is trapped in one of many low-free-energy intermediate states, and the folding–unfolding corresponds to trapping–detrapping transition.

A practical application of unzipping DNA hairpins maybe to measure the thermodynamic parameters of DNA base pairs. The experimental setup is relatively simple, and as demonstrated in this subsection, quantitative fitting of experimental data with theoretical calculations is easy to perform.

3. SECONDARY STRUCTURES OF RNA POLYMER

In biological bodies, RNA usually exists as a single-stranded biopolymer. Similar to DNA, it is also formed by four types of nucleotide bases, G, C, A, and U (uracil); and two types of canonical Watson–Crick base pairs, A–U and G–C, can be formed. The non-Watson–Crick base-pairing pattern G–U is also quite common in RNA structures.

Some kinds of RNA polymers, such as transfer RNA (tRNA), have specific stable three-dimensional native configurations. Such configurations are formed when the RNA chain folds up onto itself to form many base pairs between its complementary nucleotide bases. Such kind of RNA sequences should have been "designed" to make sure that there is only one native configuration, the free energy of which being much lower than those of other possible configurations. The sequences of tRNA molecules are obviously highly designed by natural selection. On the other hand, the tRNA sequences of different organisms also show some extent of variability. Investigation of this variability may shed some light on the evolutionary history of different organisms.

However, for many other RNA molecules including messenger RNA (mRNA) it is undesirable to have stable structures. Messenger RNA prefers to be in a random coil conformations with very few base-pairing interactions, so that its information can be easily read by the protein synthesis machine (the ribosome). The coding region of a genome must have been evolved to ensure that (1) its corresponding mRNA transcript contains very few mutually complementary segments, and (2) its encoding protein can fold into a stable three-dimensional native configuration. This is a very difficult combinatorial optimization problem in biology, and nature took an extremely long time to find the right answers. Today, identifying and interpreting correlations and anticorrelations in a DNA sequence becomes a

very important subfield of scientific research. To meet this challenge, a structural biology viewpoint will turn out to be essential.

In this section, we review recent works on the structural properties of RNA polymers. We focus on the stability of RNA secondary structures. By the secondary structure we mean the following picture [126–128]. The position of each nucleotide base along RNA chain with respect to the 5′-end of RNA is specified by an integer index i. Suppose in a RNA secondary structure, nucleotide bases at position i and position j (with $j > i$) form a base pair denoted as pair (i, j), and nucleotides base at position k and position l ($l > k$) form another base pair (k, l). Then, it is required that base pairs (i, j) and (k, l) are either *independent*, with $j < k$, or *nested*, with $i < k < l < j$.

In an actual RNA configuration, some base pairs may violate these conditions. These base pairs are classified into the higher-order tertiary structure of RNA. For example, in the three-dimensional yeast $tRNA^{Phe}$, there are base-pairing interactions between two RNA hairpin loops. These additional base pairs stabilize the twisted L conformation of tRNA [13]. However, for RNA, the main contribution to the free energy is from its secondary structures. Energy contributions from RNA tertiary structures could be regarded as perturbations to the free energy of the system. This separation of energy scales makes the study of RNA secondary structure of particular interest.

We discuss in Section 3.1, the structure of RNA polymers made of homogeneous nucleotide sequences. In Section 3.2, we show the effect of weak sequence design on RNA secondary structure. Glassy transition in RNA structures is discussed in Section 3.3, and the influence of electrostatic interactions to RNA structural stability is considered in Section 3.4. Section 3.5 focuses on the issue of predicting the most stable secondary structure of a given RNA sequence.

3.1. The Homogeneous-Sequence Approximation

As a starting point, this subsection neglects sequence heterogeneity and regards RNA as a homogeneous chain, with short-ranged attractive interactions between its constructing units.

When a RNA folds back onto itself and nucleotide bases form base pairs, many double-helical segments appear. For a RNA made of random sequences (referred to as a random-sequence RNA hereafter), because the correlation along the RNA chain is very weak, each such double-helical segment in general is very short (of several base pairs). According to recent experimental work of Ref. [75], at physiological temperature, double-helical segment of 2–10 base pairs is very easy to be disturbed by thermal noise. Therefore, secondary structures of a random-sequence RNA are transient, with typical lifetime of the order microseconds. The polymer could easily fluctuate from one secondary structure to another when the environmental temperature is not too low. Entropy plays an important role here.

Because mRNA could be well regarded as a random-sequence in some sense, the results of this subsection may have direct relevance to mRNA structural stability.

When one end of the RNA is fixed and the other end is pulled with an external force, it may be energetically unfavorable for the polymer to fold back and form double-helical segments. At some critical force, there is a so-called helix-coil structural transition, where the polymer is stretched considerably straight. It is of interest to estimate the magnitude of this critical force.

As a first step to understand the elastic property of a random-sequence RNA, we average over the sequence randomness and regard RNA as an effective chain with homogeneous sequence. Our treatment will be improved in Section 3.2 by considering the effect of weak sequence design; and in Section 3.3 we discuss how the homogeneity approximation of this subsection breaks down at low temperatures.

Montanari and Mézard [88] have studied the structural transitions in a random-sequence RNA by analytical means. In the treatment of Ref. [88], sequence randomness was smoothed out and replaced with a (weak) short-ranged average interaction between two nucleotide bases. The RNA chain was modeled as $n+1$ beads connected sequentially by n deformable chemical bonds. Each unit (a bead plus a bond) of the model corresponds to several nucleotide bases of a real RNA chain. Each bead can form a "base pair" with at most one

another bead at the same time. In the calculation of Ref. [88], the base-pairing interaction was the only interaction considered. The excluded-volume effect was neglected, and the possible base-pair stacking interactions were also neglected. It is believed that excluded-volume effect is irrelevant in high salt conditions but becomes important at low salt conditions (see Section 3.4). The pairing energy is short-ranged with range a and strength ϵ_0, as was assumed in Section 2.2. The deformable bond between two beads has the flowing bond length distribution [see Eq. (10)]:

$$\mu(\mathbf{u}) \propto \exp\left(-\frac{(|\mathbf{u}|-b)^2}{2l_0^2}\right) \tag{58}$$

where the bond length fluctuation l_0 is much shorter than the intrinsic bond length b.

Because the RNA chain is modeled as a homopolymer, the free energy and the partition function are sequence-independent and are functions only of the polymer length. Denote the partition function of a RNA chain of n bonds as $Z_0(n,\mathbf{r})$, where $\mathbf{r}$ is the end-to-end distance vector of the chain. According to Fig. 15, we have the following recursive equation [$Z_0(0,\mathbf{r})=\delta(\mathbf{r})$]:

$$\begin{aligned} Z_0(n,\mathbf{r}) = & \int d\mathbf{u}_1 d\mathbf{u}_2 \mu(\mathbf{u}_1) Z_0(n-1,\mathbf{u}_2)\delta(\mathbf{r}-\mathbf{u}_1-\mathbf{u}_2) \\ & + \Theta(a-|\mathbf{r}|)(e^{\beta\epsilon_0}-1)\int\left[\prod_{i=1}^{3} d\mathbf{u}_i\right]\mu(\mathbf{u}_1)\mu(\mathbf{u}_2)Z_0(n-2,\mathbf{u}_3)\delta\left(\mathbf{r}-\sum_{i=1}^{3}\mathbf{u}_i\right) \\ & + \sum_{l=0}^{n-3}\int\left[\prod_{i=1}^{3} d\mathbf{u}_i\right]Z_0(n-l-3,\mathbf{u}_1)\mu(\mathbf{u}_2) \\ & \times \Theta(a-|\mathbf{r}|)(e^{\beta\epsilon_0}-1)\int\left[\prod_{j=4}^{6} d\mathbf{u}_j\right]\mu(\mathbf{u}_4)\mu(\mathbf{u}_5)Z_0(l,\mathbf{u}_6)\delta\left(\mathbf{u}_3-\sum_{j=4}^{6}\mathbf{u}_j\right)s \end{aligned} \tag{59}$$

where $\Theta(x)$ is the Heaviside step function defined in Section 2.2.1.

With the same procedure as that in Section 2.2.1, we can perform a Fourier transform for the partition function $Z_0(n,\mathbf{r})$ and obtain an iterative expression for the Fourier transformed partition function $\tilde{Z}_0(n,\mathbf{p})$. The generating function of $\tilde{Z}_0(n,\mathbf{p})$ can then be calculated following the method of Section 2.2.1. The result reads [88]

$$G_0(z,\mathbf{p}) = \sum_{n=0}^{\infty}\tilde{Z}_0(n,\mathbf{p})z^n = \frac{1}{z}\frac{\omega(z)}{1-\tilde{\mu}(\mathbf{p})\omega(z)} \tag{60}$$

In Eq. (60), $\omega(z)$ is determined by the equation

$$\omega(z) = z + z^2 B(\omega) \tag{61}$$

with

$$B(\omega) = \frac{a^3}{6\pi^2}(e^{\beta\epsilon_0}-1)\int d\mathbf{q}\tilde{\mu}^2(\mathbf{q})\frac{\omega}{1-\tilde{\mu}(\mathbf{q})\omega} \tag{62}$$

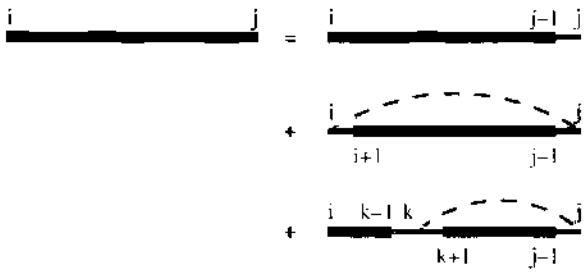


Figure 15. Hartree diagram on the secondary structure of RNA polymer. The thick lines such as (i,j) denote the partition function of the RNA segment from base i to base j, the thin segments denote a single nucleotide monomer, and the dotted lines denote base-pairing interactions.

and [see also Eq. (18)]

$$\tilde{\mu}(\mathbf{p}) = \frac{\sin b_0|\mathbf{p}|}{b_0|\mathbf{p}|} \exp\left(-\frac{l_0^2|\mathbf{p}|^2}{2}\right) \tag{63}$$

When an external force **F** is pulling the two ends of a RNA polymer apart, the generating function of the total partition function has the same form as Eq. (60), but with **p** replaced by $i\beta\mathbf{F}$ (see Section 2.2.2 for an explanation). A careful analysis [88] shows that in this case, Eq. (60) has two singular points: one corresponds to the point z_{coil} at which $\omega(z_{\text{coil}})\tilde{\mu}(i\beta\mathbf{F}) = 1$, and the other singular point is at z_{molten}, which is determined by $z_{\text{molten}} = \max_{\omega}[\frac{2\omega}{1+\sqrt{1+4\omega B(\omega)}}]$.

At given temperature T, when the external force is less than some threshold value $F_c(T)$, the polymer is in the so-called molten phase [91], with the configuration of RNA fluctuating frequently between many transient secondary structures. The average extension of the polymer is zero. When the force is larger than $F_c(T)$, the polymer is in the stretched coil phase, with few base-pairing interactions. The threshold force F_c at given temperature is determined by the equation

$$z_{\text{coil}}(F_c, T) = z_{\text{molten}}(T) \tag{64}$$

When $|\mathbf{F}| > F_c(T)$, the average extension of RNA in units of it total contour length Nb is calculated by

$$\text{Ex} = -\frac{1}{b} k_B T \frac{\partial \ln z_{\text{coil}}}{\partial |\mathbf{F}|} \tag{65}$$

The experimental and theoretical force-extension profile of a random-sequence RNA are shown in Fig. 16 (the solid line and symbols), where the solid line is obtained through Eq. (65).

In a random-sequence RNA, the main interaction is base-pairing interactions. The solid curve of Fig. 16 suggests that the Montanari–Mézard model could quantitatively explain the elastic response at high salt conditions of single-stranded charomid DNA [2, 34]. At room temperature $T \simeq 25°C$ and high ionic conditions (for example, 50 mM Na^+ and 5 mM Mg^{2+} buffer), when the external force is less than about 0.3 pN, the average end-to-end distance of RNA approaches zero, and the RNA molecule is in the collapsed molten phase.

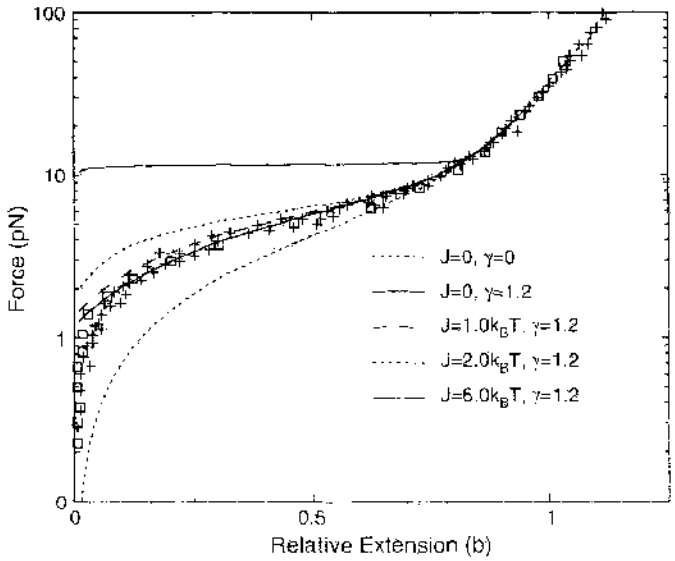


Figure 16. Force extension curve of random-sequence RNA. The experimental data come from Ref. [2] (pluses) and Ref. [34] (squares), and the ionic concentration is 5 mM Mg^{2+} and 50 mM Na^+. The theoretical parameters are $a = 4.194$ Å, $b = 1.7$ nm, $l_0 = 1.105$ Å, $T = 300$ K, $\epsilon_s = 7.482$ kJ/mol. In the figure, $\gamma = (4\pi a^3/3b^3)(e^{\beta\epsilon_0} - 1)$ and J characterize the average base-pair stacking interaction strength. Reprinted with permission from [93]. H. Zhou and Y. Zhang, *J. Chem. Phys.* 114, 8694 (2001), © 2001, American Institute of Physics.

When the external force is increased beyond this level, the secondary pairing structures in RNA is gradually pulled out and an increasing part of the polymer is in the stretched coil state. This gradual increase of coiled segment length with external force leads to the gradual force-extension profile of Fig. 16, without the appearance of a force plateau.

At the phase boundary $F_c(T)$, there is a continuous (*secondary-order*) molten-coil structural phase transition [88]. This transition is secondary-order, because at the transition point, the free energy and its first-order derivative are both continuous [77, 89, 90, 108]. The continuous nature of this phase transition can also be predicted by a simple two-state model [91]. The argument is as follows.

The RNA configuration could be divided into three parts: two stretched coil segments at the ends and a (complicated) hairpin segment in the middle. The partition function of a hairpin segment of m bases has the following scaling form:

$$Z_{11}^{\text{hairpin}}(m) \sim m^{-\frac{3}{2}} \exp\left(-n\frac{g_{\text{hairpin}}}{k_B T}\right) \tag{66}$$

where g_{hairpin} is the (model-dependent) free energy density of a RNA hairpin segment. Equation (66) will be derived in Section 3.3. Because of the scaling exponent of 3/2 in Eq. (66), we know from the Lifson argument [89] (see Section 2.3) that the transition between stretched coil state and hairpin state must be continuous. In both the continuous model [88] and the discrete model [91] mentioned here, self-avoidance effect is completely neglected. It is not yet completely clear whether the inclusion of this effect will qualitatively change the above-mentioned picture, as in the case of DNA thermal denaturation (see Section 2). However, the quantitative agreement between the theoretical result of the model of Ref. [88] and the experimental observations (see Fig. 16) suggests that self-avoidance effect may not be so important here, as the scaling exponent of 3/2 in Eq. (66) is not caused by the formation of loops but by the constraint of secondary structures.

3.1.1. Elasticity of Real Homogeneous RNA Chains

The homogeneous chain model studied by Montanari and Mézard has also real correspondence. For example, a chain with sequence poly(GC), poly(AU), or poly(CAG), and so forth, can be regarded as a homogeneous chain.

When such a "truly" homogeneous RNA molecule is stretched with an external force, the force-extension profile is much more sharper then the solid curve of Fig. 16. For example, Rief and coworkers [47] studied the elastic responses of single poly(AT) and poly(GC) nucleotide chains and found that the force-extension profile has a marked plateau at force 9 ± 3 pN for poly(AT) and 20 ± 3 pN for poly(GC). The existence of a force plateau suggests a discontinuous structural transition process. Pulling a poly(AT) or poly(GC) single-stranded DNA is just like pulling apart the two strands of a double-stranded DNA, indicating that in the absence of external force, the structure of a single-stranded poly(GC) or poly(AT) chain cannot fluctuate between many transient configurations. It forms a giant hairpin with an very long stem region and very few single-stranded loop regions. The entropic effect is suppressed.

However, based on the Montanari–Mézard model, it was found [93] that even when the pairing energy ϵ_0 is very high, the resulting force-extension curve is gradual and not abrupt. This suggests that the entropy is not suppressed by high base-pairing potentials and must be attributed to other effects including base-pair stacking interactions [92].

In the Montanari–Mézard model [88], base-pair stacking interaction is neglected. This may be a good approximation for a random-sequence RNA; but it may not be valid for a highly homogeneous polymer such as poly(GC), as the probability that two base pairs to be adjacent to each other in a poly(GC) will be much higher than in a random-sequence RNA. In Refs. [92, 93], base-pair stacking interactions were considered to explain the force-extension profile of a homogeneous RNA chain. When two formed base-pairs are adjacent to each other [i.e., a (i,j) base pair and a $(i+1,j-1)$ base pair], then there is an additional energy gain of magnitude ϵ_s (J in Fig. 16) [93].

The actual theoretical calculation is similar with that outlined in Section 2.2.1 and is detailed in Ref. [93]. Here we just mention the salient results. The effect of increasing

base-pair stacking interaction is evidently demonstrated in Fig. 16. When the base-pairing potential ϵ_0 is kept fixed and stacking potential ϵ_s (or J) is increased, the force-extension becomes more and more flat. When $\epsilon_s \simeq 6.0\ k_B T$ ($T = 300$ K) corresponding to an average stacking energy of $2\ k_B T$ between two adjacent nucleotide base pairs, a force plateau at 10 pN becomes very evident. Therefore, the cooperativity of hairpin-coil transition in a homogeneous RNA could be explained by base-pair stacking interactions. This conclusion reinforces the insight of Section 2.2, where we show that the cooperativity of DNA denaturation is also caused by base-pair stacking interactions.

Similar pictures can also be obtained by numerical Monte Carlo calculations including electrostatic interactions. The quantitative comparisons of the experimental data and the modeling calculations are displayed in Section 3.4 and its illustrations.

3.2. Structural Transitions in a Weakly Designed RNA Polymer

In the preceding subsection, the elastic property of a random-sequence RNA is studied. In this subsection, we study the elastic property of a long RNA chain that has a preferred native hairpin configuration [91]. The polymer is weakly designed in the following sense. First, generate a random RNA sequence of $2n+1$ nucleotide bases. Second, for index $i = 1$ till n, with probability $p \leq 1$, change the nucleotide base at position $2n+1-i$ to be complementary to the base at position i, and with probability $1-p$ keep the nucleotide base at position $2n+1-i$ unchanged. Energetically, the RNA polymer has a designed preferred configuration of to form a long hairpin of $2n+1$ bases. The designability is controlled by the parameter p. When p is much less than 1, the chain is only weakly designed. In this section, we mainly discuss this situation. The pairing energy is smoothed out along the polymer and the following pairing energy is assumed for the polymer chain:

$$\epsilon_{i,j} = -\epsilon_0 - (\epsilon_1 - \epsilon_0)\delta_{i+j}^{2n+1} \tag{67}$$

with $\epsilon_1 > \epsilon_0$. Of cause the difference $\epsilon_1 - \epsilon_0$ is controlled by the parameter p.

The thermal denaturation behavior of such system was first studied by Bundschuh and Hwa [91]. They concluded that when the temperature is high enough, the polymer will be in a molten phase with partition function governed by Eq. (66), and the polymer is not trapped by just one configuration. However, when the temperature is low, the polymer is in the designed hairpin configuration because of the energy preference of this state. There exists a second-order structural transition between the hairpin and the molten states at some critical temperature T_c.

In this subsection, the work of Ref. [91] is extended to consider also the effect of external force. The general picture is given in the phase diagram of Fig. 17. When both the temperature and the external force are small, the system is in the designed hairpin phase; when temperature is small but the external force is large, the system is in the stretched coil phase; when the temperature is high but the external force is small, the system is in the molten phase. The hairpin-coil transition is *first-order*; the molten-coil transition is *second-order*, and the hairpin-molten transition is also *second-order*.

Here we show the main steps of the analytical calculations that lead to the above-mentioned conclusions. We use the same continuous model as in the preceding subsection.

First, we define the following notations: (1) $Z_{\text{paired}}(n, \mathbf{r})$ is the total partition function of a designed RNA of $2n$ effective bonds, with the bases at the two ends forming a base pair of distance $\mathbf{r}$; (2) $Z_{\text{unpaired}}(n, \mathbf{r})$ is the total partition function of a designed RNA of $2n$ bonds, with the two bases at the ends separated with distance $\mathbf{r}$ ($|\mathbf{r}| > a$, the bonding distance); (3) $Z(n, \mathbf{r})$ is the total partition function of a designed RNA of $2n$ bonds.

Based on Fig. 18, we have the following expression for $Z_{\text{paired}}(n, \mathbf{r})$:

$$Z_{\text{paired}}(n, \mathbf{r}_1) = \sum_{m=1}^{n} \sum_{l_1=1} \cdots \sum_{l_m=1} \delta_{l_1+\cdots+l_m}^{n} \int d\mathbf{r}_2 \ldots d\mathbf{r}_{m+1} e^{-m\beta\epsilon_1}$$
$$\times \prod_{i=1}^{m} [\theta(a - |\mathbf{r}_i|) Z_{\text{unpaired bubble}}(l_i, \mathbf{r}_i, \mathbf{r}_{i-1})] \delta(\mathbf{r}_{m+1}) \tag{68}$$

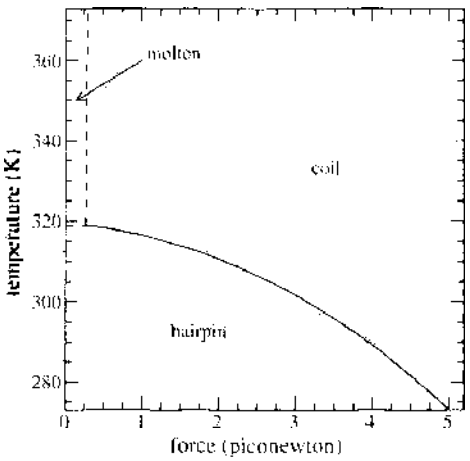


Figure 17. Phase-diagram of a weakly designed RNA polymer. The system has three phases: hairpin, molten, and stretched-coil. The hairpin-coil phase-transition is first-order (solid line), whereas hairpin-molten and molten-coil transitions are both second-order (dotted line and dashed line). The theoretical parameters are set to $a=0.4194$ nm, $b=1.7$ nm, $T_0=300$ K, $\epsilon_0=-3.0\ k_B T_0$, $\epsilon_1=3.6\ k_B T_0$.

where $Z_{\text{molten-bubble}}(l,\mathbf{r}_1,\mathbf{r}_2)$ is the partition function of a loop of $2l$ bonds. From Eq. (68), the partition function $Z_{\text{paired}}(n)=\int d\mathbf{r} Z_{\text{paired}}(n,\mathbf{r})$ can be obtained. Finally, we obtain the following generating function for this partition function:

$$G_{\text{paired}}(z)=\sum_{n=1}^{\infty} z^n Z_{\text{paired}}(n)=\frac{\frac{4\pi}{3}a^3 e^{\beta\epsilon_1} G_{\text{molten-bubble}}(z)}{1-\frac{4\pi}{3}a^3 e^{\beta\epsilon_1} G_{\text{molten-bubble}}(z)} \tag{69}$$

where $G_{\text{molten-bubble}}(z)=\sum_{l=1} z^l Z_{\text{molten-bubble}}(l,\mathbf{0},\mathbf{0})$. Clearly, $Z_{\text{molten-bubble}}(l,\mathbf{0},\mathbf{0})$ should be independent of the value of ϵ_1. To find the expression for $G_{\text{molten-bubble}}(z)$, we follow the idea of Ref. [91] to set ϵ_1 to be identical to ϵ_0 in Eq. (69) and calculate the partition function for a hairpinned configuration of such a "undesigned" polymer. This could be done following the same line as that in Section 3.1 and we find that

$$G_{\text{molten-bubble}}(z)=\frac{\frac{z}{(2\pi)^3}\int d\mathbf{q} G^0_{\text{even}}(z,\mathbf{q})}{1+\frac{4\pi}{3}a^3 e^{\beta\epsilon_0}\frac{z}{(2\pi)^3}\int d\mathbf{q}\tilde{\mu}^2(\mathbf{q})G^0_{\text{even}}(z,\mathbf{q})} \tag{70}$$

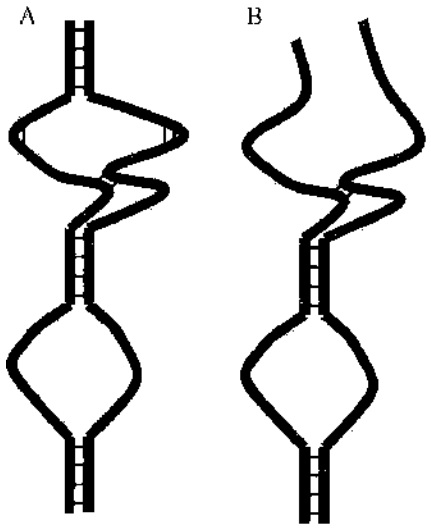


Figure 18. (A) Schematic representation of a weakly designed RNA with the last two bases forming a base pair. (B) Schematic representation of a weakly designed RNA where the last two bases do not form a base pair.

In Eq. (70), $G^0_{\text{even}}(z,\mathbf{p})=\sum_{n=0} z^n \tilde{Z}_0(2n,\mathbf{p})$, where $\tilde{Z}_0(2n,\mathbf{p})$ is the Fourier transform of the partition function of an undesigned polymer chain of $2n$ bonds (see Section 3.1). A careful analysis leads to the following expression

$$G^0_{\text{even}}(z,\mathbf{p})=\frac{1}{\sqrt{z}}\frac{1}{1-\tilde{\mu}(\mathbf{p})\omega(\sqrt{z})}\frac{\psi(\sqrt{z})}{1+\tilde{\mu}(\mathbf{p})[2\psi(\sqrt{z})-\omega(\sqrt{z})]} \tag{71}$$

where $\omega(\sqrt{z})$ is calculated according to Eq. (61), and

$$\psi(\sqrt{z})=\sqrt{z}\left[1+\frac{Az}{(2\pi)^3}\int d\mathbf{q}\tilde{\mu}^2(\mathbf{q})G^0_{\text{even}}(z,\mathbf{q})\right] \tag{72}$$

Similarly, the generating function for $Z_{\text{unpaired}}(n,\mathbf{r})$ can be obtained. Finally we arrive at the generating function for the whole system:

$$G(z,\mathbf{p})=\sum_{n=1}^{\infty} z^n Z(n,\mathbf{p})=\frac{1}{1-\frac{4\pi}{3}a^3(e^{\beta\epsilon_1}-e^{\beta\epsilon_{11}})\frac{z}{(2\pi)^3}\int d\mathbf{q}\tilde{\mu}^2(\mathbf{q})G^0_{\text{even}}(z,\mathbf{q})}\frac{G^0_{\text{even}}(z,\mathbf{p})}{}-1 \tag{73}$$

When an external force $\mathbf{F}$ is acting on the end of the polymer, the total generating function for the partition function has the same form as Eq. (72), but with $\mathbf{p}$ replaced by $i\beta\mathbf{F}$. From Eq. (72), we see that $G(z,i\beta\mathbf{F})$ has three singular points, corresponding respectively to the native hairpin phase, the denatured molten phase, and the denatured stretched coil phase. The phase diagram is shown in Fig. 17.

At given temperature and external force, the total relative extension of the polymer along the direction of the external force is obtained by the following expression:

$$\text{Ex}=-k_B T\frac{\partial \ln z}{\partial|\mathbf{F}|} \tag{74}$$

In Fig. 19, the force-extension profiles of a model designed RNA chain at different fixed temperatures are calculated based on Eq. (74). For the parameters chosen in this figure, when environmental temperature is higher than 318.88 K and the external stretching is lower than 0.3 pN, the model polymer is in the molten-globule state, and the force-extension curve above 0.3 pN is continuous. When temperature is lower than 318.88 K, the polymer is first in the designed hairpin state, and unzipping of this hairpin by external force leads to a force plateau.

It is of interest to measure the elastic behavior of weakly designed RNA polymers and compare the experimental results with theoretical calculations mentioned above.

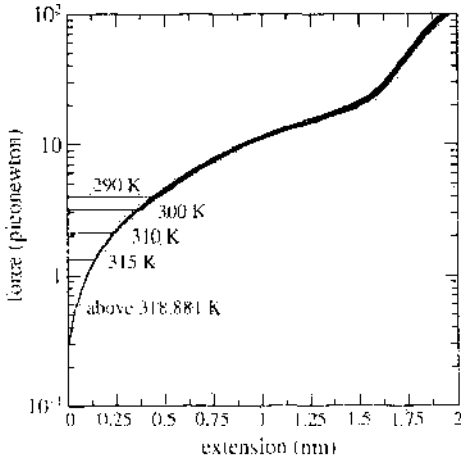


Figure 19. Force-extension relationship of a designed RNA polymer at different temperatures. When temperature is lower than $T_c = 318.881$ K, stretching RNA cause a first-order hairpin-stretched coil phase-transition. Therefore there is a force-plateau in each force curve. When temperature is higher than T_c, stretching cause a second-order The values of the theoretical parameters are listed in the caption of Fig. 17.

3.3. Glassy Transitions in RNA Sequences at Low Temperature

At low temperatures, a RNA polymer prefers to fold into a configuration that has the lowest possible energy. As discussed in the preceding subsection, if the RNA polymer under study has a designed native structure, it will fold to this structure at low temperatures. However, what will happen to a random-sequence RNA? Will many configurations of comparably low energies appear? Will the RNA be trapped in one of these low-energy configurations? These are very interesting questions that have attracted considerable theoretical thinking [127, 129–134]. Here we will only briefly discuss this issue, following the work of Bundschuh and Hwa [91, 131, 132].

In the analytical calculations of Refs. [91, 131, 132], each RNA monomer is either unpaired or is forming a base pair with another monomer. A given RNA secondary configuration S is a set of base pairs, any two of which are either mutually independent or are nested. The energy of a configuration S is assumed to be

$$E[S]=\sum_{(i,j)\in S}\epsilon_{ij} \tag{75}$$

where ϵ_{ij} is the base-pairing energy between base i and base j. In the analytical work of Bundschuh and Hwa [131], the pairing energies ϵ_{ij} is assumed to be an identically and independently distributed Gaussian random variable:

$$\rho(\epsilon)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(\epsilon-\epsilon_0)^2}{2\sigma^2}\right] \tag{76}$$

where ϵ_0 is the average pairing free energy and σ is the standard variation in the pairing free energy. The partition function of a RNA segment from base i to base j is denoted as $Z_{i,j}$:

$$Z_{i,j}=\sum_{S}\exp(-\beta E[S]) \tag{77}$$

This partition function satisfies the following recursive relationship:

$$Z_{i,j}=Z_{i,j-1}+e^{-\beta\epsilon_{kj}}Z_{i+1,j-1}+\sum_{k=i+1}^{j-2}Z_{i,k-1}e^{-\beta\epsilon_{kj}}Z_{k+1,j-1}+Z_{i,j-2}e^{-\beta\epsilon_{ij}} \tag{78}$$

with $Z_{k,k}\equiv 1$.

3.3.1. Homogeneous Sequence

First, let us consider the case of homogeneous RNA sequences [$\sigma=0$ in Eq. (76)]. In this case, the partition function $Z_{i,j}$ depends only on chain length $n=j-i$, so we denote it as $Z(n)$. Its generation function is

$$G(z)=\sum_{n=1}^{+\infty}Z(n)z^n \tag{79}$$

Combining Eq. (79) and Eq. (78) leads to the following expression [91]

$$G(z)=\frac{1}{2qz^2}\{1-z-2qz^2-\sqrt{(1-z)^2-4qz^2}\} \tag{80}$$

where $q=e^{-\beta\epsilon_{ij}}$.

The scaling form of the partition function ($n\gg 1$) could then be obtained:

$$\begin{aligned}Z(n)&=\frac{1}{2\pi i}\oint\frac{G(z)}{z^{n+1}}\mathrm{d}z=-\frac{1}{4\pi qi}\oint\frac{\sqrt{[1-(1+2\sqrt{q})z][1+(2\sqrt{q}-1)z]}}{z^{n+3}}\mathrm{d}z\\&=-\frac{(1+2\sqrt{q})^{n-2}}{4\pi qi}\oint\frac{\sqrt{[1-\zeta][1+\kappa\zeta]}}{\zeta^{n+3}}\mathrm{d}\zeta\\&=\text{const}\times\frac{(1+2\sqrt{q})^n}{n^{3/2}}\end{aligned} \tag{81}$$

where $\kappa = \frac{\sqrt{q}-1}{\sqrt{q}+1}$. The scaling exponent $\theta = \frac{3}{2}$ is explained in Ref. [131] using a "mountain" picture. This scaling exponent characterizes the RNA molten phase.

3.3.2. Heterogeneous Chain

For Heterogeneous RNA chain where the pairing free energy has the distribution Eq. (76), the quantity of interest is the average free energy over different realizations of the random sequence; that is,

$$[\mathcal{F}]_{\rm av} = -k_B T [\ln Z]_{\rm av} \tag{82}$$

However, it is extremely difficult to calculate the average of the logarithm of the partition function. A trick that is routinely used in studying spin-glass systems is to rewrite this average into another form:

$$[\ln Z]_{\rm av} = \lim_{n \to 0} \frac{[Z^n]_{\rm av} - 1}{n} \tag{83}$$

where Z^n means the total partition function of n RNA chains with the same sequence [135].

For our current system,

$$\begin{aligned} Z^n &= \sum_{S_1} \cdots \sum_{S_n} \exp\left[-\beta \left(\sum_{(i_1, j_1) \in S_1} \epsilon_{i_1 j_1} + \cdots + \sum_{(i_n, j_n) \in S_n} \epsilon_{i_n j_n} \right) \right] \\ &= \sum_{S_1} \cdots \sum_{S_n} \exp\left[-\beta \sum_{(i,j)} n_{ij} \epsilon_{ij} \right] \end{aligned} \tag{84}$$

where n_{ij} is the total number of (i, j) base pairs in the n replicas of the same system. Averaging over the random variables ϵ_{ij} in Eq. (84) leads to

$$[Z^n]_{\rm av} = \sum_{S_1} \cdots \sum_{S_n} \exp\left[-\beta \sum_{(i,j)} n_{ij} \epsilon_0 \right] \exp\left[\frac{\beta^2 \sigma^2}{2} \sum_{(i,j)} n_{ij}^2 \right]. \tag{85}$$

Another way to express Eq. (85) is as follows. Define $\theta_{ij}(k) = 1$ if in replica index k the two bases i and j forms a base pair (i, j), and $\theta_{ij}(k) = 0$ otherwise. Then $n_{ij} = \sum_k \theta_{ij}(k)$. One then has

$$[Z^n] = \sum_{S_1} \cdots \sum_{S_n} \prod_{k=1}^{n} u^{|S_k|} \prod_{1 \le k < l \le n} \tilde{u}^{|S_k \cap S_l|} \tag{86}$$

where

$$u = \exp\left[-\beta \epsilon_0 + \frac{1}{2} \beta^2 \sigma^2 \right] \tag{87}$$

$$\tilde{u} = \exp[\beta^2 \sigma^2] \tag{88}$$

Equation (86) is equivalent to Eq. (85), because $|S_k \cap S_l| = \sum_{(i,j)} \theta_{ij}(k) \theta_{ij}(l)$.

For general integer n, Eq. (86) is still very difficult to calculate. To make some progress, Bundschuh and Hwa focused on the case of $n = 2$ [131]. They found that Eq. (86) for this two-replica system is exactly solvable. When $N \gg 1$, the partition function has the following scaling form [131]

$$[Z^2] \sim N^{-\theta'} \zeta^N \tag{89}$$

where θ' is a scaling exponent and ζ is related to the average free energy density.

Bundschuh and Hwa [131] demonstrated that at given level of sequence randomness, when the temperature is decreased below some threshold, there is a phase transition in the two-replica system. The main conclusions are as follows.

1. There exists a critical temperature T_c determined by the equation $\tilde{u}=\tilde{u}_c$, where $\tilde{u}_c$ could be calculated by calculating the partition function of a double-stranded RNA system made of two independent RNA chains. When the sequence randomness is weak, that is, $\sigma \simeq 0$, this critical temperature T_c also approaches zero.
2. When the temperature $T > T_c$, the system is in the molten phase, with $\theta' = 3 = 2 \times 3/2$. This indicates that the two replicas are uncoupled in the thermodynamic limit. Therefore, the sequence randomness is irrelevant at this temperature range.
3. When $T < T_c$, the system is in a *weak* glassy phase, with $\theta' = 3/2$ in Eq. (89). The two replicas are locked together, and the two-replica system behaves as a single RNA chain. This glassy phase is weak, because the exciting energy scales *logarithmically* with the domain size.

The work of Bundschuh and Hwa [131, 132] therefore proves the existence of a glassy phase in random-sequence RNA chains at low-enough temperature. Related issues are also been discussed in the literature [127, 129, 130, 133, 134].

Complementary to these analytical investigations, in Ref. [125] the temperature-induced folding–unfolding kinetics of a hairpin-forming RNA chain was followed. Zhang and Chen [125] found that when the environment temperature T is higher than a glass-transition T_g, the folding–unfolding process is a two-state process, between a precquilibrated quasi-equilibrium macro-state and the unfolded coil macro-state. However, when $T < T_g$, the transition is not two-state-like but has many transition intermediates. This is consistent with the qualitative results of Bundschuh and Hwa [131]. At low temperature, the RNA chain is trapped in one of many free-energy local minima states, and there are many comparable time scales corresponding to transitions between these intermediate states.

Hyeon and Thirumalai [136] demonstrated that the typical energy scale ϵ characteristic of the roughness of the energy landscape could be measured by performing force-extension experiments on RNA over different ambient temperatures.

3.4. Single-Stranded DNA/RNA at Low Salt Solution

In our previous discussion of RNA or single-stranded (ss) DNA elastic properties, the intramolecular and polymer-solution electrostatic interactions are completely ignored. However, as native ssDNA/RNA molecules can achieve their biological activities only at their physiological salt milieu, it is of practical importance to understand the elasticity of ssDNA or RNA at different salt conditions. This subsection focuses on the influence of electrostatic interactions to ssDNA/RNA elasticity.

The electrostatic interactions of biopolymers in salt solution are usually theoretically approached by the Debye–Huckel method [137–140]. On the experimental side, during recent years, there are several experimental groups who have pulled ssDNA or RNA molecules and reported the force-extension data in a variety of salt environments [2, 18, 34, 36, 47, 141]. These data are crucial for checking these theoretical approaches to biopolymers in solution and meanwhile determine some important parameters of the molecules.

3.4.1. Debye–Huckel Theory of Biopolymers

In principle, the electrostatic potential $\psi(\mathbf{r})$ of an electrolyte solution at position $\mathbf{r}$ is governed by the Poisson–Boltzmann equation [142]

$$\nabla^2\psi(\mathbf{r}) = -\frac{4\pi}{D}\sum_{i=1}^{n} v_i e c_i \exp[-v_i e\psi(\mathbf{r})/k_B T] \tag{90}$$

Here, the solution is assumed to contain n different types of ions. The ith species has valence v_i, and the total number of ions of the ith type is N_i, with c_i ($=N_i/V$, V is the volume) being the bulk concentration of this species. D is the dielectric constant of the solution, and e denotes the protonic charge.

Equation (90) can be solved numerically under proper boundary conditions according to the series expansion method that was first used by Pierce [143]. In Fig. 20 are illustrative examples of the electrostatic potentials of an ssDNA/RNA cylinder in 2 mM NaCl and

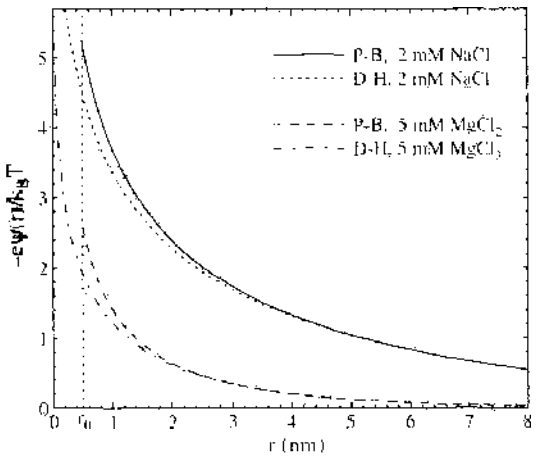


Figure 20. Electrostatic potential of ssDNA cylinder versus the radial distance from the cylinder axis in the solutions of 2 mM NaCl and 5 mM $MgCl_2$. The solid and dashed curves are the numerical solutions of Poisson–Boltzmann equation (P–B); the dotted and dashed-dotted denote corresponding Debye–Huckel approximations (G–H) with effective linear charge density ν along the axis listed in Table 1. Reprinted with permission from [140], Y. Zhang et al., *Biophys. J.* 81, 1133 (2001). © 2001, Biophysical Society of USA.

5 mM $MgCl_2$ solutions. Here, a cylinder radius of $r_0 = 0.5$ nm and surface charge density of $e/0.6$ nm are assumed, and potential function $\psi(\mathbf{r})$ is expanded up to the 17th and 14th order for NaCl and $MgCl_2$, respectively [140].

In case the electrostatic potential is small relative to $k_B T$, the terms higher than square in Pierce-expansion can be neglected. The electrostatic potential around a point charge q can be written implicitly in the Debye–Huckel form

$$\psi_{DH}(r) = \frac{q}{D|r|}\exp(-\kappa|r|) \tag{91}$$

where r is the radial distance from the charge, and κ is the inverse Debye length, equaling $(8\pi c_0 e^2/Dk_B T)^{1/2}$ (for NaCl solution) or $(24\pi c_0 e^2/Dk_B T)^{1/2}$ (for $MgCl_2$).

To count the influence of higher expansion terms of the Poisson–Boltzmann equation, one can phenomenologically change the amplitude of the Debye–Huckel potential of Eq. (91) to fit the precise solution of the Poisson–Boltzmann equation [144, 145]. According to Eq. (91), the electrostatic potential of a uniformly charged straight cylinder of infinite length can be written as

$$\psi_{DH}(r) = \int_{-\infty}^{\infty} \frac{\nu d\lambda}{D} \frac{\exp(-\kappa\sqrt{\lambda^2 + r^2})}{\sqrt{\lambda^2 + r^2}} = \frac{2\nu}{D} K_0(\kappa r) \tag{92}$$

where λ is arclength along the cylinder axis, ν is effective linear charge density, and K_0 is the first-order modified Bessel function [146]. By comparing Eq. (92) with the numerical Poisson–Boltzmann solution in the tail region as shown in Fig. 20, we can determine the effective linear charge density ν in different bulk ionic concentration c for both the NaCl and the $MgCl_2$ solution (see Table 1). In Table 1, we also show the effective charge density of double-stranded (ds) DNA, as well as ssDNA. All the data of ν can be very well fitted by a formula of

$$\nu = \exp(\alpha + \beta c^{2.5}) \tag{93}$$

with the fitting parameters α and β listed also in Table 1.

3.4.2. Elastic Freely-Jointed Chain Model and Interplay with Electrostatic Force

Because of high degree of flexibility, ssDNA/RNA molecules are usually modeled by a completely flexible "freely-jointed chain" (FJC) model: A chain consists of N rigid and independently oriented links, each of length b [147]. Because the contour length of the molecules

Table 1. The effective linear charge density ν (in unit of e/nm) of DNA molecules, calculated from the comparison of the Poisson–Boltzmann solution and the modified Bessel function (see text).[a]

Ionic concentration	ssDNA		dsDNA	
c_0 (M)	NaCl	$MgCl_2$	NaCl	$MgCl_2$
1.0	4.18	9.50	91.85	993.16
0.75	3.50	6.74	56.15	410.67
0.5	2.84	4.51	31.22	144.10
0.2	2.04	2.31	11.73	24.52
0.15	1.89	1.97	9.29	16.22
0.1	1.73	1.64	7.02	9.82
0.05	1.53	1.27	4.78	4.98
0.02	1.37	0.99	3.29	2.66
0.01	1.29	0.86	2.66	1.91
0.005	1.23	0.78	2.26	1.45
0.002	1.17	0.71	1.93	1.13
0.001	1.14	0.67	1.76	1.00
α	0.0338	−0.577	0.300	−0.505
β	1.36	2.80	4.18	7.33

[a] α and β are the parameters of Eq. (93) fitted to the data of ν. Reprinted with permission from [140], Y. Zhang et al., *Biophys. J.* 81, 1133 (2001). © 2001, Biophysical Society of USA.

can be elongated under high extended force, the modification of elastic links has been introduced into the FJC model [18]. Under the external force F, the free energy of this so-called elastic freely-jointed-chain model (eFJC) in salt solution can be written as

$$F_{\mathrm{eFJC}} = \frac{Y}{2}\sum_{i=1}^{N}(|\mathbf{r}_i - \mathbf{r}_{i-1}| - b)^2 + \frac{\nu^2}{D}\int \mathrm{d}\lambda_1\,\mathrm{d}\lambda_2 \frac{\exp(-\kappa|\mathbf{r}_i - \mathbf{r}_j|)}{|\mathbf{r}_i - \mathbf{r}_j|} - Fz_N \tag{94}$$

where $\mathbf{r}_i = (x_i, y_i, z_i)$ is the position of link i, and Young's modulus Y characterizes the stretch stiffness of backbone. The integration is performed over the molecule's curvilinear coordinates (with $i < j$).

In Fig. 21 is a comparison of force-extension data with Monte Carlo calculations of the eFJC model [140, 141]. In the experimental setup of Dessinges and coauthors [141], ssDNA molecules were treated physicochemically to suppress the base-pairing interaction. In this condition, ssDNA extension was observed to increase almost logarithmically with force,

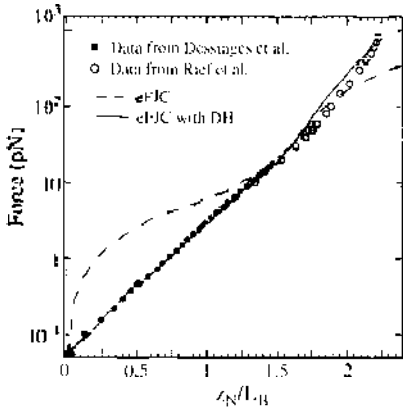


Figure 21. Extension of ssDNA molecules normalized by the contour length of an equivalent dsDNA molecule observed in 10 mM PB. Solid circles are from Dessinges et al. [141], where a physicochemical treatment has been employed on the molecule for the propose of suppressing the pairing interactions of the complementary bases; open circles are from Rief et al., which was pulled with an atomic force microscope at high force [47]. The continuous line is the result of a numerical simulation of the model described in Eq. (94) with $Y = 120\ k_BT/nm^2$. To fit the results at high forces ($F > 70$ pN), a nonlinear quartic elastic term with $Y_4 = 400\ k_BT/nm^4$ was added. The dashed line is the predictions of the eFJC model without electrostatic force included. Adapted with permission from [141], M.-N. Dessinges et al., *Phys. Rev. Lett.* 89, 248102 (2002). © 2002, American Physical Society.

which deviates strongly from the behavior of a pure elastic FJC (Fig. 21). According to the Monte Carlo simulation, the electrostatic interaction tends to swell the volume occupied by the chain and make the segments more easily aligned along the force direction. This is equivalent to enlarging the Kuhn length of the molecule. In principle, the lower the ionic concentration becomes, the larger the effective Kuhn length is and the more rigid the molecule looks. As shown in Fig. 21, good agreement with experimental data up to 70 pN can be obtained by the model of Eq. 94 without adjustable parameters. Beyond the external force of 70 pN, however, the extremely large deformation of the ssDNA backbone could not be described by the commonly used quadratic elastic approximation [88, 93, 140] but requires the introduction of a quartic elastic term with elastic module $Y_4 = 400\ k_B T/nm^4$ in the model [141]. This nonlinear elastic behavior may result from a structural transition in ssDNA as it is pulled above 70 pN [18].

The logarithmic increasing of extension with external stretching in ssDNA and RNA revealed in Fig. 21 is very striking. Recently, it was observed that similar force-extension behavior could happen in the stretching of spider-silk proteins [148]. It is yet to see whether this phenomenon is also caused by electrostatic interactions or not.

3.4.3. Base-Pairing Interactions of ssDNA/RNA in Solution

Long-range base pairs can be formed between the complementary bases of ssDNA/RNA, which tend to bend the molecule into specific secondary structures. Because hydrogen bonds of the base pairs themselves are short-ranged (~0.3–0.4 nm of donor–acceptor distance [149]), these secondary structures are strongly influenced by the excluded-volume interactions that can be dramatically enhanced by the electrostatic repulsion of phosphodiester backbone.

To include the self-pairing interaction in the eFJC model, we approximate the base-pairing potential by a node-pairing energy, i.e.

$$E_{\text{pair}} = \sum_{i=1}^{N_p} V_p \tag{95}$$

where N_p is the number of node pairs, and V_p is a sequence-dependent parameter denoting the intensity of a certain pairing mode.

As shown in Fig. 22, in low-salt solution, the electrostatic repulsive interaction along ssDNA backbone is weakly screened. For example, in 1 mM phosphate buffer (mainly NaH_2PO_4), the Debye length $l_D = 5.95$ nm, which is much larger than the hydrogen-bond interaction range. In this case, there are few base pairs being able to form because the dominant electrostatic repulsive potential excludes the bases from getting close into the

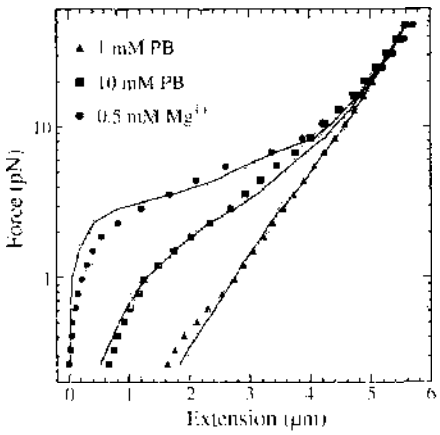


Figure 22. Stretching of a single ssDNA charomid of extension 5.7 μm in 1mM phosphate buffer (PB) (triangle), 10 mM PB (square), and 0.5 mM Mg^{++} (circle). Data are taken from Ref. [141]. The continuous curves are the results of Monte Carlo simulations of eFJC model [140] with a node-pairing strength of $V_p = 4.6\ k_B T$. Adapted with permission from [141]. M.-N. Dessinges et al., *Phys. Rev. Lett.* 89, 248102 (2002). © 2002, American Physical Society.

Watson–Crick base pair range. In high-salt solution, the electrostatic self-avoiding interaction is much weaker because of the strong screen effect. The complementary bases can therefore come spatially close and form base pairs. In this case, it need slightly larger force to elongate the molecule than a pure FJC chain to overcome additional base-pairing interaction in low force range.

3.4.4. Sequence Sensitivity of ssDNA/RNA Elasticity

There are three hydrogen bonds formed in a G–C base pair and two hydrogen-bonds in an A–T base pair. This difference of hydrogen-bonding strength leads to sequence sensitivity of ssDNA/RNA secondary structure and elasticity in high salt solutions.

In Fig. 23, two different sequences of charomid and pXΔII, which have 50% and 30% GC contents, respectively, are pulled under external forces [141]. The MC simulation of Eq. 94 plus pairing interaction Eq. 95 fits the data well with $V_p = 4.6\ k_BT$ for the charomid and $V_p = 3.8\ k_BT$ for the GC-poorer pXΔII construction. As expected, the higher the GC content, the higher the average pairing energy V_p, the more stable are the hairpins, and the higher the force required to stretch the ssDNA to a given length. This sensitivity to ionic conditions and nucleotide content disappears at force large enough to unzip DNA (>10 pN). As hairpins are less likely, the elasticity of ssDNA is less sensitive to variations of sequence and buffers. Furthermore, recent analytical results [88, 93] are in good agreement with experimental observations and Monte Carlo simulations.

When the pair-bases in RNA hairpin structures are neighboring along the sequence, there exists van der Waals interaction between consecutive base pairs. This stacking interaction most likely happens in homogeneous sequences, such as poly(G–C) and poly(A–T).

To include the stacking interaction in the eFJC model, we count the stacking energy as

$$E_{\text{stacking}} = \sum_{i=1}^{N_S} V_S \tag{96}$$

where N_S is the number of stacked node pairs, and V_S is the interactive potential between two neighboring node pairs. Two node pairs are considered as stacked only when they are nearest neighbors to each other.

As shown in Fig. 24, the stacking potential can dramatically change the conformation of ssDNA at low force. When the stacking potential is absent, the base pairs are formed quite randomly with formation of many interior loops and branched structures. So the dispersed hairpin structure can be easily pulled open in a medium force (left column of Fig. 24). When the stacking interaction exists, on the other hand, the base pairs are encouraged energetically to be neighboring and therefore lead to the formation of bulk hairpin structure (right column of Fig. 24). This kind of abrupt transition from hairpin to coil was indeed observed in recent

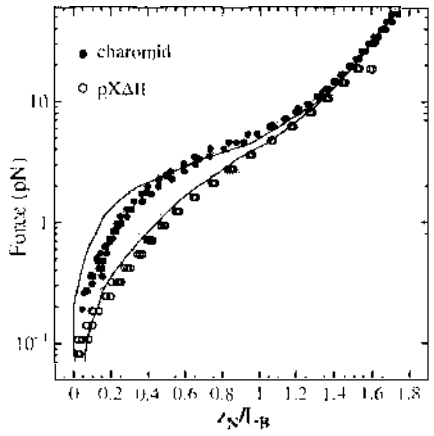


Figure 23. Stretching of two different ssDNA molecules: A 50% GC-rich charomid (solid circle) and a 30% GC-rich pXΔII (open circle) in 10 mM Tris buffer [141]. The continuous curves are the results of MC simulations of eFJC model plus a pair potential with $V_p = 4.6\ k_BT$ for the charomid and $V_p = 3.8\ k_BT$ for pXΔII. Adapted with permission from [141], M.-N. Dessinges et al., *Phys. Rev. Lett.* 89, 248102 (2002). © 2002, American Physical Society.

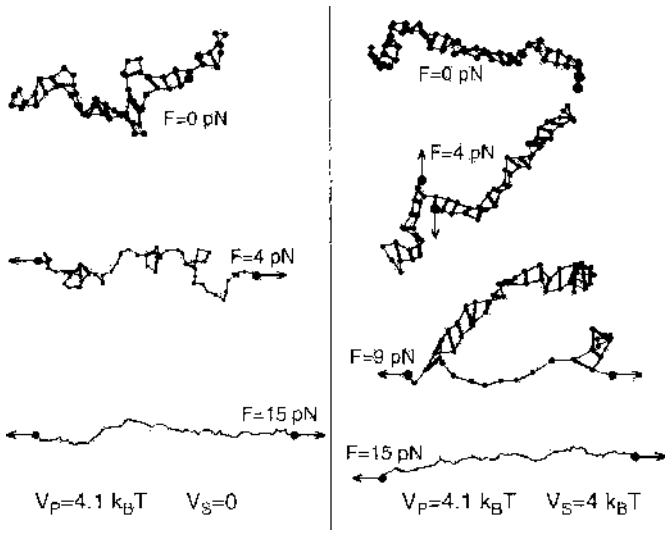


Figure 24. Typical conformations of ssDNA chains with random sequence (left) and designed poly(A–T) sequence (right) in 150 mM NaCl solution stretched at different external forces, which are produced in the Monte Carlo simulation of 60 nodes. The fictitious bases and backbone are expressed by green nodes and lines, with the nodes of two ends denoted by two bigger red globes. The hydrogen bonds of base pairs are denoted by magenta short lines. Reprinted with permission from [140]. Y. Zhang et al., *Biophys. J.* 81, 1133 (2001). © 2001. Biophysical Society of USA.

experiments by Rief and coauthors [47], in which two designed poly(A–T) and poly(G–C) ssDNA sequences are pulled (see Fig. 25). The value of threshold force, (i.e., the height of the plateau in the force-extension curve) is dependent on the pairing and stacking potentials of the nodes. The best fit of MC calculations [140] to the data are $V_P = 4.1\ k_BT$ and $V_S = 4\ k_BT$ in poly(A–T) sequence and $V_P = 10.4\ k_BT$ and $V_S = 6\ k_BT$ in poly(G–C) sequence. Bearing in mind that each Kuhn length (~1.6 nm) in our simulation contains about three nucleotide bases, we can infer that the pairing energy of each A–T base pair is $1.37\ k_BT$ and that for

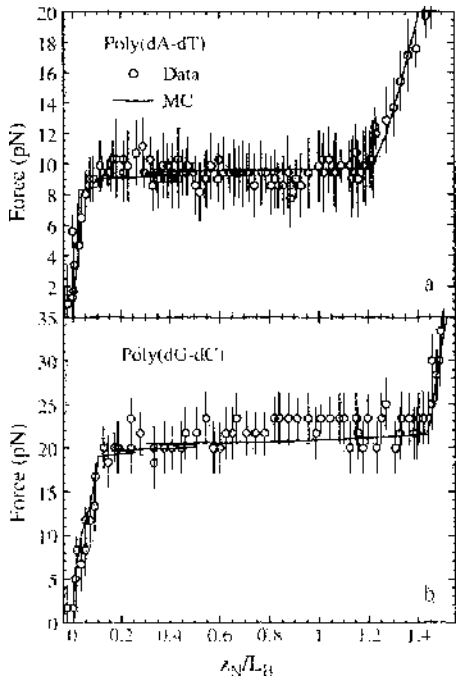


Figure 25. Stretching of designed poly(A–T) sequence (a) and poly(dG–dC) sequence (b). The data are taken from Ref. [47]; the lines are the Monte Carlo simulations of an eFJC model with base-pairing and base-stacking potentials. Reprinted with permission from [140]. Y. Zhang et al., *Biophys. J.* 81, 1133 (2001). © 2001. Biophysical Society of USA.

each G-C base pair is 3.47 k_BT. These values are comparable with the measurements of Bockelmann et al. [38] by unzipping the two strands of a dsDNA helix.

3.5. RNA Secondary Structure Prediction Through Mechanical Stretching

In the preceding several subsections, we have discussed the general physics of RNA secondary structures. Recently, the biological significance of RNA molecules have been acknowledged by more and more researchers. Besides the conventional functions of RNA, the mechanisms of small-RNA and RNA binding and RNA interference were discovered. Therefore, an issue of great biological interest is to predict faithfully the structure of a given RNA sequence. In this subsection, we mention some recent investigations on this respect.

RNA structure prediction is certainly not a new question. Many algorithms have already been existed for quite a long time; see, for example, Refs. [150–155] and references cited therein. These algorithms work mainly for the secondary structures of RNA. They try to identify the structures with the minimum free energy by calculating the partition function iteratively [see Eq. (78)] and by assigning different empirical statistical weights to different structural elements [150–152]. For short RNA sequences that contain hundreds of nucleotides, these algorithms usually predict a pairing pattern for a given RNA that is in reasonable agreement with experimental measurements.

However, these iterative algorithms are unreliable for very long RNA chains. Part of the reason is that these algorithms excluded the possibility of tertiary structures such as pseudo-knots. A pseudo-knot is formed when the nucleotide bases in the loop regions of two hairpin segments pairing with each other (see Fig. 26).

With the development of single-molecule manipulation techniques, the suggestion has been raised to detect the secondary structure of RNA by mechanical stretching. For long RNA sequences, however, experimental measurements [2, 34, 141] revealed that the force-extension curves of real RNA chains are very smooth and featureless. This situation is dramatically different from the unzipping curves of double-stranded DNA molecules [37].

To find out why the force-extension profiles of stretching RNA can not reveal its structure, Gerland and coworkers [156] performed numerical stretching simulations on several real RNA molecules with well-defined native structures. They also found that the equilibrium force-extension curves for natural or random RNA chains are considerably smooth and do not contain many features to enable the detection of the underlying secondary structure of RNA [Fig. 27(a)].

Two reasons may contributed to this observed smoothness. First is a compensation effect [156]. As demonstrated in Fig. 28(A), as the total extension of a RNA polymer is increased, new structural motifs may be opened, causing a transient decrease in the stretching force; however, some structural motifs that were opened previously may reform rapidly, leading to a reverse increase of the force. When pulling velocity is not so quick, the resulting force-extension will then be smooth. Lower the temperature or increasing the pulling rate may help improving the detection of RNA secondary structure.

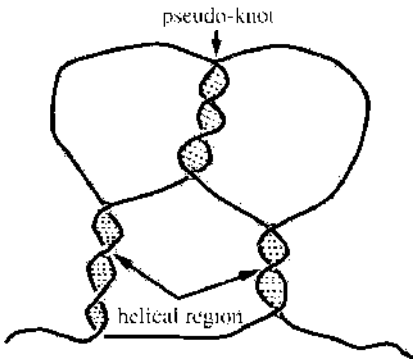


Figure 26. An schematic RNA configuration with two helical segments and one pseudo-knot.

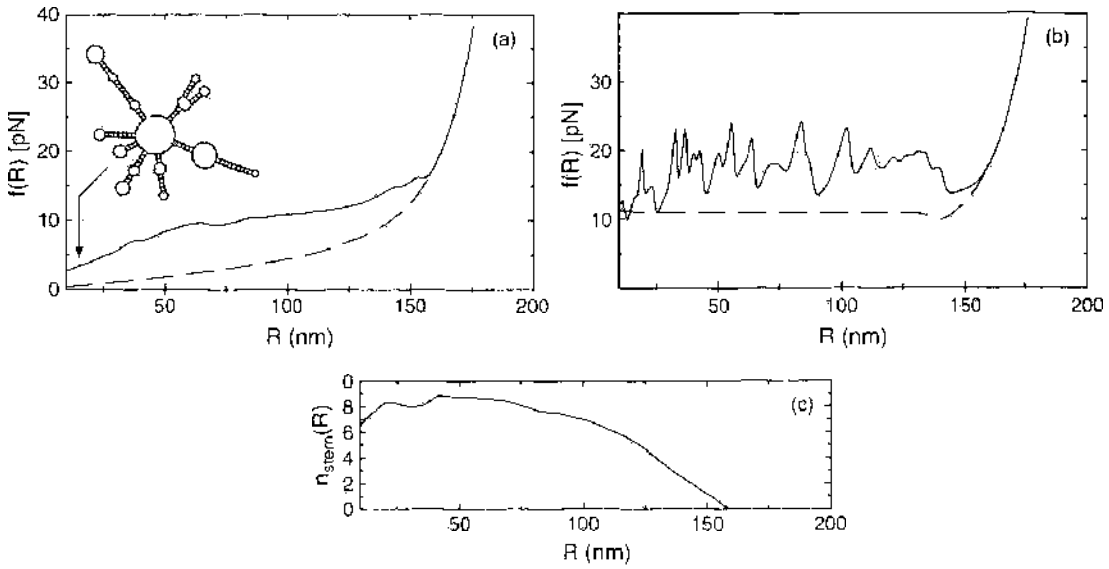


Figure 27. (a) Simulated force-extension curve for a group I intron (solid line) and that for a homogeneous RNA chain (dashed line). The native structure of the intron sequence is also shown. (b) The force-extension curve for a DNA hairpin of random sequence (solid line) and that of a homogeneous DNA hairpin (dashed line). (c) Mean number of external stems for the group I intron. Reprinted with permission from [156]. U. Gerland et al., *Biophys. J.* 81, 1324 (2001). © 2001, Biophysical Society of USA.

The second reason is the contribution of suboptimal structures [156]. According to the calculation in Section 3.1, fluctuations of RNA structures over different suboptimal configurations could result in a very smooth stretching curve.

To delete the compensation effect and to probe the native secondary structure of a RNA molecule by mechanical stretching, Gerland and coworkers [157] suggested an elegant way. The idea is to let the RNA molecule to pass through a nanopore [see Fig. 28(B)]. The radius of the nanopore is so small that only single-stranded RNA segments can pass through and double-stranded segments could not. By pulling the single-stranded part, the double-stranded hairpin segment nearest to the nanopore will be unzipped, whereas other double-stranded segments will not be destroyed. Therefore, the hairpin secondary structures in a RNA polymer will be pulled out in a discrete manner, starting from one end of the polymer. Gerland and coworkers performed simulations with this idea and their simulation results suggested that by this means, the native secondary structures of RNA can indeed be largely probed. Figure 29 shows the simulated force-extension curves of pulling a real RNA polymer through a nanopore. We see that sawtooth patterns, which signify the breaking of individual RNA

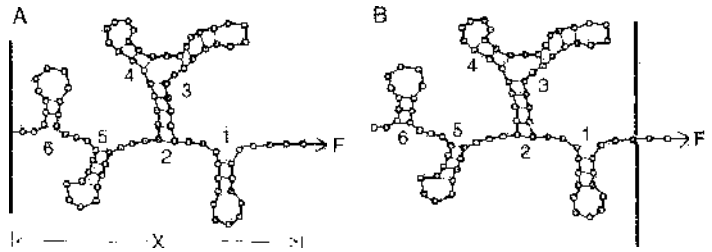


Figure 28. (A) Compensation effect in RNA unfolding. One end of a RNA chain is fixed and the other end is pulled with a stiff force lever. At extension X, hairpin with index 1 is opened; at a slightly increased extension $X+\delta X$, hairpin with index 5 is also destroyed, which leads to a transient drop in the pulling force. This drop in pulling force makes it favorable for hairpin 1 to be closed again, resulting in an increase in the pulling force. This compensation effect partially accounts for the smooth behavior of force-extension curves in RNA pulling experiment. (B) To use pulling experiment to explore the native structure of a RNA chain, Gerland and coworkers [157] suggested the use of a nano-pore. Only single-stranded RNA segment could pass through the nano-pore. When applied with a external force F, hairpin 1 will first be opened, followed by hairpin 2, and then hairpin 3 and so on. Because the hairpins break in a sequential order, from the recorded force-extension curve, the secondary structure of RNA could be inferred.

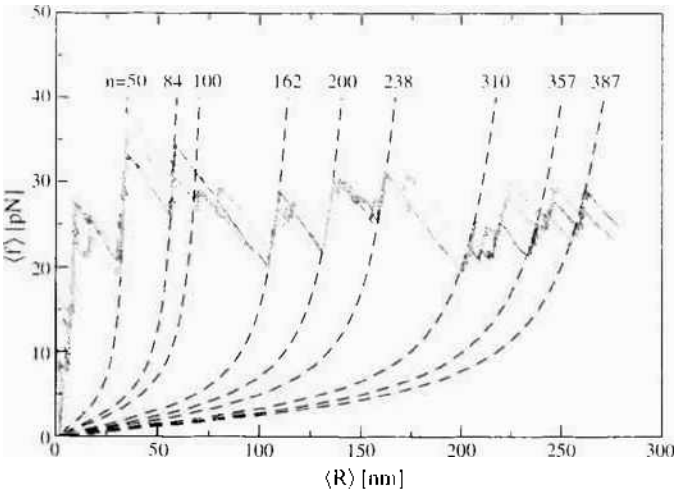


Figure 29. Typical simulated force-extension curves for pulling a RNA polymer through a nano-pore. Reprinted with permission from [157], U. Gerland et al. e-print: cond-mat/0306126 (2003).

structural elements, appear with considerable reproducibility. For more quantitative discussion of this issue, please refer to Ref. [157].

4. DNA ENTROPIC ELASTICITY

We have discussed in the preceding two sections structural transitions in DNA and RNA that are associated with breaking of nucleotide base pairs. In this and the next two sections, we focus on the mechanical property of DNA double-helix when the base-pairing pattern is kept intact.

Double-stranded DNA (dsDNA) is a long polymer of nucleotides. In solution its spatial configuration fluctuates over time because of thermal activations that are characterized by energy scale $k_B T$. What is the best model to describe the elastic behavior of solution dsDNA?

Here we first outline the main predictions of three polymer models: the Gaussian model, the freely-jointed-chain (FJC) model, and the wormlike-chain (WLC) model. The WLC model is found to be most suitable in describing the elastic response of dsDNA under the action of low external forces (less than 10 piconewton).

4.1. The Gaussian Model and the Freely-Jointed-Chain Model

In the Gaussian model, a polymer is assumed to be formed by tandemly connecting N bonds. The end-to-end extension is

$$\mathbf{R} = \sum_{i=1}^{N} \mathbf{r}_i \tag{97}$$

Each bond $\mathbf{r}_i$ has the following bond length distribution

$$P(\mathbf{r}_i) = \left(\frac{3}{2\pi b^2}\right)^{3/2} \exp\left(-\frac{3\mathbf{r}_i^2}{2b^2}\right) \tag{98}$$

The characteristic length scale b is called the Kuhn length [147], $\langle \mathbf{r}_i^2 \rangle = b^2$. Between any two bonds there is no other interactions.

Under the action of an external stretching force of magnitude f, the average extension of the polymer along the force direction is

$$X = \frac{Nb^2}{3k_B T} f \tag{99}$$

The average extension is proportional to force f.

The Gaussian model works well when the external force is small. However, under moderate or high stretching conditions, the assumption Eq. (98) becomes invalid, as each bond could have only finite extensibility. The freely-jointed-chain model replaces this assumption with that of a rigid bond,

$$P(\mathbf{r}_i) = (4\pi b^2)^{-1}\delta(|\mathbf{r}_i| - b) \tag{100}$$

However, the orientation of nearest-neighboring bonds are completely independent. Under this assumption, the force-induced total average extension is

$$X = Nb\left[\frac{\cosh(\beta fb)}{\sinh(\beta fb)} - \frac{1}{\beta fb}\right] = L\mathcal{L}(\beta fb) \tag{101}$$

where $L = Nb$ is the total contour length of the polymer and $\mathcal{L}(x) = \coth(x) - 1/x$ is known as the Langevin function.

When the force is small, Eq. (101) reduces to the linear relationship Eq. (99). However, for large force, Eq. (101) predicts that

$$1 - \frac{X}{L} = \frac{k_B T}{fb} \tag{102}$$

This scaling law characterizes the FJC model.

4.2. The Wormlike-Chain Model

In 1992, Smith, Finzi, and Bustamante [1] made the first measurement on the elastic response curve of λ-phage DNA (48,502 base pairs in length) under the action of an external force. These authors found that, neither the Gaussian model Eq. (99) nor the freely-jointed-chain model Eq. (101) could explain their experimental data [1]. Later it turned out that dsDNA is a semiflexible polymer, and to properly account the elasticity of dsDNA, one has to use the wormlike-chain model [16, 158].

In the WLC model, a polymer is regarded as an *inextensible* chain of contour length L and its configuration is described by the position vectors $\mathbf{r}(s)$ at each arclength point $0 \le s \le L$ along the chain. The chain is *semiflexible*, so there is an energy penalty associated with bending of the chain. The configurational energy is expressed as follows:

$$\mathcal{H}[\mathbf{r}] = \int_0^L ds \frac{k_B T \ell_p}{2}\left(\frac{d\mathbf{r}^2(s)}{ds^2}\right)^2 = \int_0^L ds \frac{k_B T \ell_p}{2}\left(\frac{d\mathbf{t}(s)}{ds}\right)^2 \tag{103}$$

In Eq. (103), $\mathbf{t}(s) = \frac{d\mathbf{r}(s)}{ds}$ is the unit tangent vector at arclength s. The physical meanig of the *bending persistence length* ℓ_p will be explained shortly.

The probability distribution of the tangent vector $\mathbf{t}$ at arclength s is determined by the following Green equation:

$$\frac{\partial P(\mathbf{t}, s)}{\partial s} = \frac{1}{2l_p}\nabla_{\mathbf{t}}^2 P(\mathbf{t}, s) \tag{104}$$

The boundary condition of Eq. (104) is $P(\mathbf{t}, 0) = \delta(\mathbf{t} - \mathbf{t}_0)$, where $\mathbf{t}_0$ is the tangent of the chain at arclength $s = 0$. The derivation of Eq. (104) is given in the Appendix.

Solution of Eq. (104) leads to

$$P(\mathbf{t}, s) = \sum_{l=0}^{+\infty} \exp\left[-\frac{l(l+1)}{2l_p}s\right]\frac{2l+1}{4\pi}P_l(\mathbf{t}\cdot\mathbf{t}_0) \tag{105}$$

where $P_l(\mathbf{t}\cdot\mathbf{t}_0)$ is l-th order Legendre function. Therefore, the tangent–tangent correlation is

$$\langle \mathbf{t}(s)\cdot\mathbf{t}_0 \rangle = \exp\left(-\frac{s}{l_p}\right) \tag{106}$$

From Eq. (106) we know that ℓ_p is the characteristic correlation length of the polymer's direction. For dsDNA, $l_p \simeq 53$ nm or 150 base pairs [16, 158].

Under the action of an external stretching force F, the total average end-to-end extension is determined by finding the ground-state eigenvalue and eigenvector of the following Green function (see Appendix):

$$\frac{\partial \Psi(\mathbf{t}, s)}{\partial s} = \left[\frac{1}{2l_p} \nabla_{\mathbf{t}}^2 + \beta F \mathbf{z}_0 \cdot \mathbf{t} \right] \Psi(\mathbf{t}, s) \tag{107}$$

where $\mathbf{z}_0$ denotes the direction of the external force. The ground-state eigenvalue and eigenvector of Eq. (107) could be obtained exactly (see, for example, Ref. [159]), and the resulting force-extension relation is shown in Fig. 30. Here, we also give an approximate variational solution, which turns out to be very accurate.

We choose the following variational ground-state eigenvector

$$\Psi_0 = \exp[-g_0 s + (\eta/2)\mathbf{t} \cdot \mathbf{t}_0] \tag{108}$$

and the ground-state "energy" is

$$g_0 = \frac{\int \Psi_0^* [-(2l_p)^{-1} \nabla_{\mathbf{t}}^2 - \beta F \mathbf{t} \cdot \mathbf{z}_0] \Psi_0 \, d\mathbf{t}}{\int d\mathbf{t} \, \Psi_0^* \Psi_0} = \left(\frac{\eta}{4l_p} - \beta F \right) \mathcal{L}(\eta) \tag{109}$$

The variational parameter η should be chosen such that g_0 attains its minimal value. Therefore, η is determined by

$$4\beta F l_p = \eta + \left(\frac{d \ln \mathcal{L}(\eta)}{d\eta} \right)^{-1} \tag{110}$$

and the average end-to-end extension along the force direction $\mathbf{z}_0$ is

$$X = -L k_B T \frac{\partial g_0}{\partial F} = L \mathcal{L}(\eta) \tag{111}$$

For many applications, the following famous interpolation formula [16, 158] for the force-extension relation is often used:

$$F = \frac{k_B T}{4\ell_p} \left[\frac{1}{(1 - \frac{X}{L})^2} - 1 + \frac{4X}{L} \right] \tag{112}$$

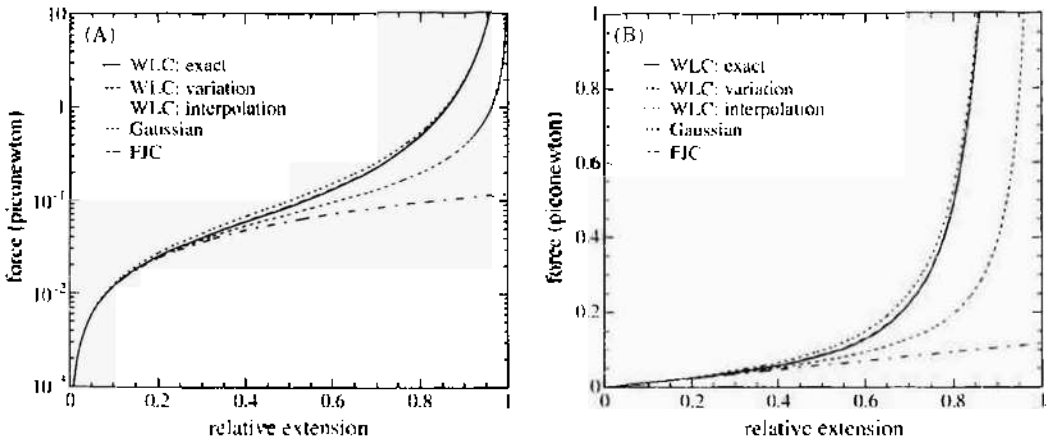


Figure 30. The force-extension curve for a wormlike-chain model and the comparison with those of the Gaussian model and the freely-jointed-chain model. The extension is in units of the total contour length (for the case of Gaussian polymer, it is in units of Nb, where N is the number of bonds). The model parameters is $T = 300$ K and $\ell_p = 53.0$ nm (correspondingly, $b = 106.0$ nm). The variational curve is calculated from Eq. (104) and the interpolation curve is calculated from Eq. (112). (A) linear-logarithmic scale; (B) linear-linear scale.

To understand Eq. (112), first we notice that when the external force is very weak, the average extension X is related to the force through the linear-response theorem

$$X = \frac{F}{k_B T}\left[\langle([\mathbf{r}(L)-\mathbf{r}(0)]\cdot\mathbf{z}_0)^2\rangle_{F=0} - \langle[\mathbf{r}(L)-\mathbf{r}(0)]\cdot\mathbf{z}_0\rangle_{F=0}^2\right] = \frac{2\ell_p L}{3k_B T}F \tag{113}$$

On the other hand, when the external force is very large, the configuration of the wormlike chain will be aligned along the direction of the external force; the tangent vector can only perform very small fluctuations around the direction of the external force. We can expand the tangent vector along the force direction to quadratic term and obtain the following equation

$$1 - \frac{X}{L} = \sqrt{\frac{k_B T}{4\ell_p F}} \tag{114}$$

Combining Eqs. (113) and (114) then leads to the interpolation formula Eq. (112).

Comparing Eq. (114) with Eq. (102), we know that as the external force F is large, the difference between a semiflexible WLC chain's contour length and its extension scales as $F^{-1/2}$ rather than F^{-1}. This $-1/2$ scaling law was observed in dsDNA [1, 16].

The results mentioned in this section are summarized in Fig. 30.

The most interesting behavior of a semiflexible polymer is when the total contour length L is of the same order as the bending persistence length ℓ_p. On such kind of systems, there has recently been much work (see, for example, Refs. [160–163]).

5. DNA OVERSTRETCHING TRANSITION

The wormlike-chain model of the preceding section regards DNA as an inextensible chain. When DNA is under the action of a relative large external force ≥ 10 pN, the molecule becomes extensible. In the force range 10 pN $\leq F \leq 70$ pN, the contour length of DNA linearly increases with external stretching, with a stretching modulus $\simeq 10^3$ pN [18]. Surprisingly, at external force $F \simeq 70$ pN, the contour length of DNA suddenly increases to 1.7 times its B-form length [18, 19, 30, 164]. This abrupt structural transition is referred to as DNA overstretching transition in the literature. When $F < 70$ pN, DNA is a double-helical molecule, with folded sugar–phosphate backbones; when $F > 70$ pN, DNA is overstretched, suggesting that the hidden length in the double-helix is pulled out. This section reviews recent efforts in understanding the nature of this DNA overstretching transition.

In the literature, there are two kinds of opinions concerning DNA overstretching transition. In one opinion, the transition is between B-form DNA and an overstretched S-form DNA. S-form DNA is also a double-helical molecule. In the other opinion, which was first suggested by Smith, Cui, and Bustamante [18] and later more thoroughly investigated by Bloomfield and coworkers [164–166], is that the overstretching transition is actually a force-induced DNA melting process.

5.1. Cooperative Overstretching Transition Viewed from a Discrete Model

Many theoretical and computational efforts have been made to understand the origin of DNA overstretching transition (see, for example, Refs. [19, 28, 167–173]). The first theoretical understanding of DNA overstretching was undertaken by Cluzel and coworkers [19] based on a discrete Zimm–Bragg model (and similar models are also considered in Refs. [168–170, 173]). Here we show the main points of this model.

Under the action of an external force f, each structural unit of DNA (i.e., a nucleotide pair) has two states: a B-form state with extension $l_B(f)$ and an overstretched S-form state with extension $l_S(f)$. There is an energy difference ΔE between these two states. Furthermore, there is an addition energy cost ω with each junction between an S-form segment and a B-form segment. A DNA chain of length N base pairs may have a configuration characterized by the following set of integers $\{n_0, n_1, m_1, \ldots, n_i, m_i, n_0\}$, where $n_0 \geq 0$ and $m_i \geq 1$

($i=1,\ldots,s$) are, respectively, the number of nucleotide pairs in the first, second,..., $s+1$th B-form segment; and $n_i \geq 1$ ($i=1,\ldots,s$) and $n_0 \geq 0$, are the corresponding S-form segments between any two consecutive B-form segments. The configuration energy is

$$\mathcal{H}[\{m_i, n_i\}] = -\sum_{i=0}^{s} m_i f l_B - \sum_{i=0}^{s} n_i (f l_S - \Delta E) + (2s+1)\omega - \omega\delta^0_{m_0} - \omega\delta^0_{n_0} \tag{115}$$

An application of the generating function method of Section 2.3 leads to the following expression for the average extension X of a nucleotide pair:

$$X = \frac{[\alpha_1 l_S + \alpha_2 l_B]\sqrt{(\alpha_1-\alpha_2)^2+4\alpha_3} + (\alpha_1-\alpha_2)[\alpha_2 l_B - \alpha_1 l_S] - 2\alpha_3[l_B + l_S]}{(\alpha_1+\alpha_2)\sqrt{(\alpha_1-\alpha_2)^2+4\alpha_3} - (\alpha_1-\alpha_2)^2 - 4\alpha_3} \tag{116}$$

where $\alpha_1 = \exp(\beta f l_B)$, $\alpha_2 = \exp(\beta f l_S - \beta\Delta E)$, and $\alpha_3 = \alpha_1\alpha_2\exp(-2\beta\omega)$.

When external force is larger than 10 pN, the B-form DNA could be well modeled by an extensible wormlike chain with bending persistence length $l_p = 53.0$ nm and stretching modulus $S_B = 1000$ pN [2]. Therefore,

$$l_B(f) = d_B \mathcal{L}(\eta_B)(1 + f/S_B) \tag{117}$$

where $d_B = 3.4$ Å is the distance between two nucleotide pairs in B-form DNA, and η_B is determined by Eq. (110) with the bending persistence length of B-form DNA. According to the work of Storm and Nelson [173], S-form DNA could also be regarded as a wormlike chain with bending persistence length $l'_p = 12.0$ nm. So we could write down

$$l_S(f) = d_S \mathcal{L}(\eta_S)(1 + f/S_S) \tag{118}$$

where d_S is set to $d_S = 5.8$ Å and η_S is also determined by Eq. (110) with the bending persistence length l'_p. The stretching modulus of S-DNA is set to $S_S = 2000$ pN.

Figure 31 demonstrated the predicted force-extension relationship for DNA according to this Zimm–Bragg model. We see that the threshold force 70 pN is determined by the energy difference ΔE between S-form DNA and B-form DNA, whereas the cooperativity of transition is determined by the junction energy cost ω. The higher the junction energy cost, the abrupt the B–S transition is. So this simple model is able to reproduce the experimental observations [18, 19].

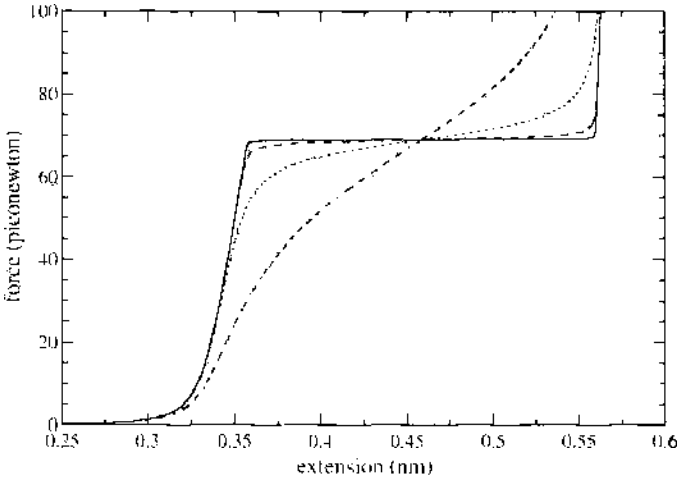


Figure 31. Theoretical force-extension profile of DNA based on Eq. (116). In the calculation, ΔE is set to 8.4 kJ/mol per nucleotide pair as in Ref. [19] and the other parameters are given in the main text. The cooperation factor ω is set to $\omega = 15.0$ kJ/mol (solid line), $\omega = 10.0$ kJ/mol (dashed line), $\omega = 5.0$ kJ/mol (dotted line), and $\omega = 1.0$ kJ/mol (dot-dashed line).

What is the physical origin of the cooperation factor ω? There are some possible alternative explanations. Cluzel et al. [19] performed molecular modeling of the DNA stretching process using the program JUMNA and suggested that in the overstretched DNA, the nucleotide bases are no longer perpendicular to the molecular axis. Therefore, it is possible that the discontinuities in inclination of the nucleotide bases in the B-form and S-form boundary will cause large energy penalty. However, the molecular modeling of Cluzel et al. [19] was performed by assuming the total twist of DNA to be fixed unchanged, which was in conflict with experimental situations. In the experimental situation, there were many single-stranded breaks (nicks) in the DNA backbone, which allow the twisting stress in DNA to be released by rotating the DNA molecule around such nicks. Another possibilities is that B–S transition was initiated only at these nicks. The existence of just a limited number of single-strand nicks can lead to a phenomenological energy penalty ω. However, this explanation is also not very convincing because (1) the same force-extension behavior was observed in many types of DNA molecules which differ in nucleotide composition and number of single-strand nicks, and (2) the overstretching transition is not sensitive to pulling velocity.

A semimicroscopic model of DNA will be introduced in the next subsection. This model suggests that the short-ranged nature of base-pair stacking interactions might be the main reason for the highly cooperative B–S transition.

5.2. Double-Stranded Polymer Model of DNA and the Overstretching Transition

This subsection investigates theoretically the possible effects of DNA short-ranged base-pair stacking interaction to the cooperativity of DNA overstretching transition. To incorporate base-pair stacking interactions, Zhou and coauthors [28, 159] introduced a double-stranded semiflexible polymer model for DNA and studied its property detailedly. Here we recall the main points of this model.

In the model of Zhou et al. [28], DNA is modeled as a double-stranded polymer with continuous degrees of freedom (Fig. 32). The two inextensible backbones of DNA [14] are characterized by the same bending rigidity $\kappa = k_B T \ell_p$, where $\ell_p \simeq 1.5$ nm is the bending persistence length of ssDNA. The position vector along ssDNA is $\mathbf{r}_i = \int^s \mathbf{t}_i(s')ds'$, where $\mathbf{t}_i$ $(i = 1, 2)$ is the unit tangential vector of the ith backbone, and s its arc length (Fig. 32). The nucleotide basepairs between the backbones are viewed as rigid planar structures with finite area and volume.

5.2.1. Bending Energy of DNA

First consider the bending energy of the backbones,

$$E_b = \int_0^L \frac{\kappa}{2}\left[\left(\frac{d\mathbf{t}_1}{ds}\right)^2 + \left(\frac{d\mathbf{t}_2}{ds}\right)^2\right]ds \tag{119}$$

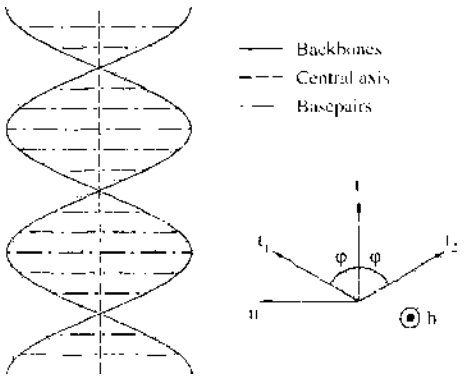


Figure 32. The double-stranded DNA model of Zhou et al. [28]. DNA is modeled as two inextensible wormlike chains winding along a common central axis. The two strands are bound together by many rigid rods that mimic base pairs. Reprinted with permission from [159]. H. Zhou et al., *Phys. Rev. E* 62, 1045 (2000). © 2000, American Physical Society.

where L is the contour length of each ssDNA backbone and s is its arc length parameter. To proceed, each base pair is regarded for the moment as a thin rigid rod of length $2R$, with a unit vector $\mathbf{b}$ pointing along it from $\mathbf{r}_1$ to $\mathbf{r}_2$, i.e., $\mathbf{r}_2(s)-\mathbf{r}_1(s)=2R\mathbf{b}(s)$. Relative sliding of the backbones is prohibited, the basepair planes are assumed to lie perpendicular to the DNA central axis and $\mathbf{b}\cdot\mathbf{t}_1=\mathbf{b}\cdot\mathbf{t}_2\equiv 0$. The central axis of dsDNA can be defined as $\mathbf{r}(s)=\mathbf{r}_1(s)+R\mathbf{b}(s)$, and its tangential vector is denoted by $\mathbf{t}$, with $\mathbf{t}\cdot\mathbf{b}=0$. Because both $\mathbf{t}_1$ and $\mathbf{t}_2$ lie on the same plane perpendicular to $\mathbf{b}$, we obtain that $\mathbf{t}_1=\mathbf{t}\cos\varphi+\mathbf{n}\sin\varphi$ and $\mathbf{t}_2=\mathbf{t}\cos\varphi-\mathbf{n}\sin\varphi$, where $\mathbf{n}=\mathbf{b}\times\mathbf{t}$ and φ is half the rotational angle from $\mathbf{t}_2$ to $\mathbf{t}_1$ ($\mathbf{b}$ being the rotational axis). We call φ the folding angle, it is in the range between $-\pi/2$ and $+\pi/2$ ($\varphi>0$ for right-handed rotations and <0 for left-handed ones). It is not difficult to verify that

$$\frac{d\mathbf{b}}{ds}=\frac{(\mathbf{t}_2-\mathbf{t}_1)}{2R}=-\mathbf{n}\frac{\sin\varphi}{R} \tag{120}$$

With Eq. (120) and the definition of $\mathbf{r}$, we know that

$$\frac{d\mathbf{r}}{ds}=\frac{(\mathbf{t}_1+\mathbf{t}_2)}{2}=\mathbf{t}\cos\varphi \tag{121}$$

With the help of Eqs. (120) and (121), the total bending energy of the backbones Eq. (119) can be rewritten as

$$E_b=\int_0^L\left[\kappa\left(\frac{d\mathbf{t}}{ds}\right)^2+\kappa\left(\frac{d\varphi}{ds}\right)^2+\frac{\kappa}{R^2}\sin^4\varphi\right]ds \tag{122}$$

The second and the third terms in Eq. (122) is deformation energy caused by folding of the backbones with respect to the central axis, and the first term, $\kappa(d\mathbf{t}/ds)^2$, is the bending energy of DNA central axis contributed by the backbone bending rigidity κ. So far, base pairs are viewed as thin rods, and their contribution to the bending rigidity of DNA chain is not considered. Because of steric effects caused by finite volume and area, base pairs will certainly increase the bending rigidity of DNA chain (the bending persistence length of double-stranded DNA is about 50 nm, quite larger than that of a single-stranded DNA chain). The simplest way to consider such effects is to replace κ in the first term of Eq. (122) with a phenomenological parameter κ^*, with $\kappa^*>\kappa$ [28].

5.2.2. Base–Pair Stacking Energy

Besides steric effects, nucleotide base pairs contribute also base-pair stacking energy. This energy mainly originates from noncovalent van der Waals interactions between adjacent base pairs [14]. Base-pair stacking interaction is short-ranged and is characterized by an attraction potential proportional to $1/d^6$ and a strong repulsion potential proportional to $1/d^{12}$ (here d is the axial distance between adjacent base pairs). In our continuous model, the line density of such Lennard–Jones type potential can be written as

$$\rho(\varphi)=\begin{cases}\dfrac{\epsilon}{r_0}\left[\left(\dfrac{\cos\varphi_0}{\cos\varphi}\right)^{12}-2\left(\dfrac{\cos\varphi_0}{\cos\varphi}\right)^{6}\right] & \text{for } (\varphi\geq 0)\\[2ex] \dfrac{\epsilon}{r_0}[\cos^{12}\varphi_0-2\cos^6\varphi_0] & \text{for } (\varphi<0)\end{cases} \tag{123}$$

In Eq. (123), r_0 is the backbone arc length between adjacent bases; φ_0 is a parameter related to the equilibrium distance between a DNA dimer; and ϵ is the base-pair stacking intensity, which is generally base-sequence specific. Here we focus on macroscopic properties of DNA, so we consider ϵ in the average sense by taking its averaged value over ten different DNA base-pair dimers (see Table 2) [14, 174]. In this work, ϵ is set to $14.0\,k_BT$, and $T=300\,K$.

The asymmetric base-pair stacking potential Eq. (123) ensures a relaxed DNA to take on a right-handed double-helix configuration with its folding angle $\varphi\sim\varphi_0$. However, if adjacent base pairs are pulled apart slightly from the equilibrium distance by external forces or thermal stretching fluctuations, the base-pair stacking interaction intensity quickly decreases

Table 2. The base-pair stacking energy of ten possible DNA base-pair stacking patterns based on quantum chemical calculations [174].[a]

DNA dimer	Stacking energy	DNA dimer	Stacking energy
5′ C·G G·C 3′	−24.09	5′ C·G | 5′ T·A A·T 3′ | G·C 3′	−17.35
5′ C·G 5′ A·T | T·A 3′ | G·C 3′	−16.20	5′ G·C | C·G 3′	−16.00
5′ G·C 5′ C·G | G·C 3′ C·G 3′	−13.64	5′ T·A | A·T 3′	−10.85
5′ G·C 5′ A·T | T·A 3′ C·G 3′	−10.85	5′ G·C | 5′ T·A | A·T 3′ | C·G 3′	−11.20
5′ A·T | 5′ T·A A·T 3′ | T·A 5′	−8.87	5′ A·T | T·A 3′	−6.31

[a]In the table, the arrows are pointing from the 5′-end to 3′-end of DNA. The stacking energy is in units of k_BT with $T = 300$ K.

because of its short-range nature. In other words, the base-pair stacking potential can endure only a limited pulling force.

The total base-pair stacking energy is

$$E_{LJ} = \int_0^L \rho(s)\,ds \tag{124}$$

5.2.3. External Stretching Energy

When an external force is pulling at one end of DNA, the DNA will try to align along the force direction $\mathbf{z}_0$. The energy associated with external force is

$$E_f = -\int_0^L f\mathbf{t}\cdot\mathbf{z}_0 \cos\varphi\, ds \tag{125}$$

The total energy of a particular configuration is the sum of the bending energy Eq. (122), the base-pair stacking energy Eq. (124), and the alignment energy Eq. (125):

$$E = E_{\text{bend}} + E_{LJ} + E_f \tag{126}$$

According to the derivation given in the Appendix, the Green function $G(\mathbf{t},\varphi;\mathbf{t}',\varphi';s)$ that determines the probability distribution of $\mathbf{t}$ and φ along DNA chain is governed by

$$\frac{\partial G}{\partial s} = \left[\frac{\partial^2}{4\ell_p^*\partial \mathbf{t}^2} + \frac{\hat{\sigma}^2}{4\ell_p\partial\varphi^2} - \frac{f\mathbf{t}\cdot\mathbf{z}_0}{k_BT}\cos\varphi - \frac{\rho(\psi)}{k_BT} - \frac{\ell_p\sin^4\varphi}{R^2}\right]G \tag{127}$$

where $\ell_p^* = \kappa^*/k_BT$. The spectrum of the above Green equation is discrete and hence for long chains, the average extension can be obtained either by differentiation of the ground-state eigenvalue, g, of Eq. (127) with respect to f:

$$\frac{X}{L} = (1/L)\int_0^L \langle \mathbf{t}\cdot\mathbf{z}_0\cos\varphi\rangle\, ds = k_BT\frac{\partial g}{\partial f} \tag{128}$$

or by a direct integration with the normalized ground-state eigenfunction, $\Phi(\mathbf{t},\varphi)$, of Eq. (127):

$$\frac{X}{L} = \iint |\Phi|^2 \mathbf{t}\cdot\mathbf{z}_0\cos\varphi\, d\mathbf{t}\, d\varphi \tag{129}$$

Both g and $\Phi(\mathbf{t},\varphi)$ can be obtained numerically through standard diagonalization methods, and identical results are obtained by Eqs. (128) and (129). The theoretical force-extension of dsDNA is calculated numerically and compared with experimental data of Ref. [19]. Figure 33 shows the result of this comparison. The theoretical curve in

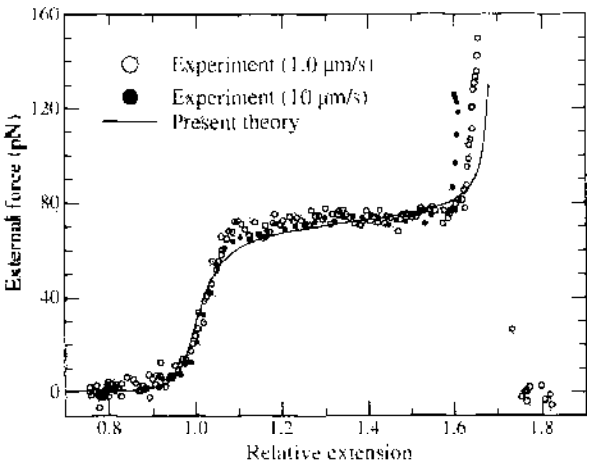


Figure 33. Force-extension relation of DNA. Experimental data is from Fig. 2(a) of Ref. [19]. Theoretical curve is obtained by the following considerations: (i) $\ell_p = 1.5$ nm and $\epsilon = 14.0\ k_BT$; (ii) $\ell_p^* = 53.0/2\langle\cos\varphi\rangle_{\ell=0}$ nm, $r_0 = 0.34/\langle\cos\varphi\rangle_{\ell=0}$ nm, and $R = (0.34 \times 10.5/2\pi)\langle\tan\varphi\rangle_{\ell=0}$ nm; (iii) adjust the value of φ_0 to fit the data. For each φ_0, the value of $\langle\cos\varphi\rangle_{\ell=0}$ is obtained self-consistently. The present curve is drawn with $\varphi_0 = 62.0°$ (in close consistence with the structural property of DNA), and $\langle\cos\varphi\rangle_{\ell=0}$ is determined to be 0.573840. DNA extension is scaled with its B-form contour length $L\langle\cos\varphi\rangle_{\ell=0}$. Reprinted with permission from [28], H. Zhou et al., *Phys. Rev. Lett.* 82, 4560 (1999). © 1999, American Physical Society.

Fig. 33 is obtained with just one adjustable parameter. The agreement with experiment is remarkable.

According to the continuous-degree model mentioned above, the onset of cooperative extension of DNA axial length at forces about 65–70 pN is mainly caused by the yielding of the short-range base-pair stacking interaction. After the collapse of base-pair stacking, the model assumes double-stranded DNA will take a planar ladder structure. In reality, more complex configurational transition might happen after the yielding of stacking interaction. For example, the unstacked base pair may become unstable and break into two unpaired bases. Alternatively, the vertically stretched base pairs may tilt with respect to the DNA central axis to gain some residual base-pair stacking. To take into account these and other kind of possibilities, one need to extend the basic model to include the possibility of base-pair breaking and tilting. Related investigations could be found in Refs. [175–180].

The continuous model of Zhou et al. [28] is also able to describe the elastic response of DNA at low external forces. Below the onset of cooperative elongation, DNA seems to be

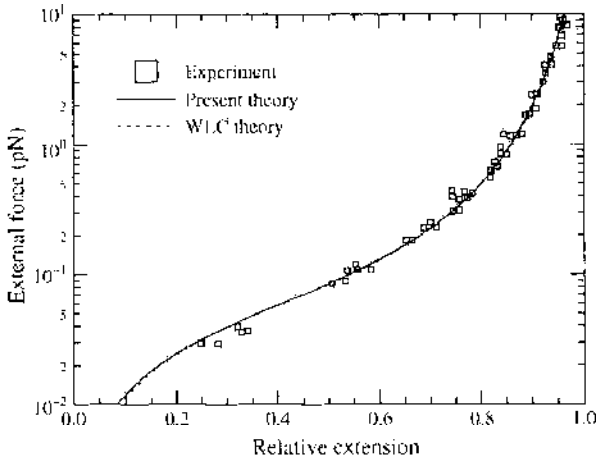


Figure 34. Low-force elastic behavior of DNA. Experimental data is from Fig. 5(B) of Ref. [1], the dotted curve is obtained for a wormlike chain with bending persistence length of 53.0 nm, and the parameters for the solid curve are the same as those in Fig. 33. Reprinted with permission from [28], H. Zhou et al., *Phys. Rev. Lett.* 82, 4560 (1999). © 1999, American Physical Society.

very stiff and calculations based on this model show that at $f = 50$ pN the total extension of DNA is only 4.1% longer than its B-form contour length, in close accordance with the value of 4.6% reported by Smith et al. [18]. This is related to the fact that the base-pair stacking intensity ϵ is very strong. At low forces ($f < 10$ pN), because the fluctuation of the folding angle φ is extremely small, it can just be neglected, and DNA elasticity is caused by thermal fluctuations of the axial direction t (entropic elasticity). The entropic elasticity (wormlike chain) model of Section 4.2 with contour length $L\langle\cos\varphi\rangle_{f=0}$ and persistence length $2\ell_p^*\langle\cos\varphi\rangle_{f=0}$ is an excellent approximation of the present theory (here $\langle\cos\varphi\rangle_{f=0}$ is the average of $\cos\varphi$ at zero force). This point is demonstrated in Fig. 34.

In Section 6.1, we will show that the model mentioned in this subsection can also quantitatively explain the supercoiling property of stretched dsDNA without any fitting parameter.

6. ELASTICITY OF SUPERCOILED DNA

The number of times the two strands of DNA double helix are interwound (i.e., the linking number Lk) is a topologically invariant quantity for closed DNA molecules [181]. It is also topologically invariant for a linear DNA polymer in case that the orientations of the two extremities of the linear polymer are fixed and no part of the polymer is allowed to go round the extremities of the polymer [20]. For an unstressed B-form DNA segment, it has one right-handed twist about per 3.4 nm along its length [14] (i.e., $Lk_0 = L_\beta/3.4$), where L_β is the DNA length. In real biological processes, however, DNA molecules are often overwound or underwound from its relaxed state in order to facilitate specific biological activities. For example, in human cells, the nucleus is just a few micrometers in diameter while it contains DNA with total length three meters within 23 chromosomes. If these chromosomes were in the relaxed coil form, it would be hard for them to fit inside the nucleus. It was observed via the reconstitution of nucleosome cores by urea/salt dialysis that a negatively twisted molecule can be packed more efficiently compared with linearized molecules [182]. The reciprocal relations between the twisting topology of chromatin and the genetic processes *in vivo* are also extensively acknowledged, and the twisting stress was shown to influence the efficiency of transcription [183] and to modulate the binding of trans-acting proteins to DNA elements [184, 185].

Conceptionally, in all cases when the linking number of DNA deviates from its relaxed value Lk_0, the DNA polymer is called "supercoiled." Because the same linking number deviation can induce different twisting stress in molecules of different size, it is more appropriate to measure extent of supercoiling by a properly normalized quantity. The supercoiling degree σ is measured by the normalized linking number deviation; that is,

$$\sigma = \frac{Lk - Lk_0}{Lk_0} \tag{130}$$

where as mentioned before, Lk_0 is the linking number of a relaxed B-form DNA of the same length. In living organisms, DNA molecules are usually negatively supercoiled, with supercoiling degree $\sigma \simeq -0.06$.

In this section, we investigate the elasticity of the supercoiled DNA molecules by analytical means and Monte Carlo simulations, based on the double-stranded model mentioned in Section 5.2. The main purpose is to understand the experimental data reported by Strick and coauthors [20, 22], who have pulled individual supercoiled DNA molecules and measured their elastic responses under various external forces and torques.

This section does not discuss the supercoiling property of circular DNA molecules, as studies on this topic have already well documented; see, for example Ref. [186].

6.1. Analytical Approximation

In the experiment of Ref. [20], the two strands of a λ-DNA at one end are bound to a glass cover slip and the two strands at the other end to a magnetic bead. The linking number of λ-DNA was then changed by rotating the magnetic bead with a magnetic field. At the same

time, an external force in the order of piconewton was applied to the DNA molecule to make it straight enough.

To model this situation theoretically, Zhou and coauthors [28, 159] introduced a torque energy in the double-stranded DNA model (Section 5.2). This torque energy is proportional to the total linking number of the DNA double-helix. Here we recall their calculations in some detail.

The total number of topological turns one DNA strand winds around the other, Lk, can be expressed as the sum of the twisting number, $Tw(\mathbf{r}_1, \mathbf{r})$, of backbone $\mathbf{r}_1$ around the central axis $\mathbf{r}$ and the writhing number, $Wr(\mathbf{r})$, of the central axis [181, 187, 188]:

$$Lk = Tw + Wr \tag{131}$$

Based on the model of Section 5.2, the twisting number is calculated by the following expression [181, 189, 190]

$$Tw(\mathbf{r}_1, \mathbf{r}) = \frac{1}{2\pi}\int_0^L \mathbf{t}\times\mathbf{b}\cdot\frac{\mathrm{d}\mathbf{b}}{\mathrm{d}s}\mathrm{d}s = \frac{1}{2\pi}\int_0^L \frac{\sin\varphi}{R}\,\mathrm{d}s \tag{132}$$

where L is the total contour length of a DNA strand and R is the cross-sectional radius of DNA double-helix. The writhing number of the central axis is much more difficult to calculate. It is expressed as the following Gauss integral over the central axis [187, 188]:

$$Wr(\mathbf{r}) = \frac{1}{4\pi}\iint \frac{\mathrm{d}\mathbf{r}\times\mathrm{d}\mathbf{r}'\cdot(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3} \tag{133}$$

In the case of linear chains, provided that some fixed direction (for example, the direction of the external force, $\mathbf{z}_0$) can be specified and that the tangent vector $\mathbf{t}$ never points to $-\mathbf{z}_0$, it was proved by Fuller [191] that the writhing number Eq. (133) can be calculated alternatively according to the following formula,

$$Wr(\mathbf{r}) = \frac{1}{2\pi}\int_0^L \frac{\mathbf{z}_0\times\mathbf{t}\cdot\mathrm{d}\mathbf{t}/\mathrm{d}s}{1+\mathbf{z}_0\cdot\mathbf{t}}\,\mathrm{d}s \tag{134}$$

Equation (134) can be further simplified for highly extended linear DNA chains whose tangent $\mathbf{t}$ fluctuates only slightly around $\mathbf{z}_0$. In this case, Eq. (134) is approximately expressed as

$$Wr(\mathbf{r}) \simeq \frac{1}{4\pi}\int_0^L \left[-t_y\frac{\mathrm{d}t_x}{\mathrm{d}s} + t_x\frac{\mathrm{d}t_y}{\mathrm{d}s}\right]\mathrm{d}s \tag{135}$$

where t_x and t_y are, respectively, the two components of $\mathbf{t}$ with respect to two arbitrarily chosen orthonormal directions ($\mathbf{x}_0$ and $\mathbf{y}_0$) on the plane perpendicular to $\mathbf{z}_0$.

The energy caused by the external torque of magnitude Γ is then equal to

$$E_{\text{torque}} = -2\pi\Gamma Lk = -2\pi\Gamma(Tw + Wr) \tag{136}$$

This energy term is added to Eq. (126) of the preceding section.

Based on Eq. (A7) of the Appendix, the Green equation for a supercoiled and highly stretched dsDNA is expressed as

$$\begin{aligned}\frac{\partial\Psi(\mathbf{t},\varphi;s)}{\partial s} = \Bigg[&\frac{\partial^2}{4\ell_p^*\partial\mathbf{t}^2} + \frac{\partial^2}{4\ell_p\partial\varphi^2} + \frac{f\cos\varphi}{k_BT}\mathbf{t}\cdot\mathbf{z}_0 - \frac{\rho(\varphi)}{k_BT} - \frac{\ell_p}{R^2}\sin^4\varphi \\ &+ \frac{\Gamma}{Rk_BT}\sin\varphi - \frac{\Gamma}{4k_BT\ell_p^*}\frac{\partial}{\partial\phi} + \frac{\Gamma^2}{16\ell_p^*(k_BT)^2}\sin^2\theta\Bigg]\Psi(\mathbf{t},\varphi;s)\end{aligned} \tag{137}$$

where (θ, ϕ) are the two directional angles of $\mathbf{t}$ [159]. Similar to what we have done in Section 5.2, we can now express the above Green equation in matrix form [159].

The ground-state eigenvalue and eigenfunction Φ_0 of Eq. (137) can be obtained numerically for given applied force and torque. The average extension is calculated through by the following formula

$$X = L\int \chi_0(\mathbf{t},\varphi)\mathbf{t}\cdot\mathbf{z}_0\cos\varphi\,\Phi_0(\mathbf{t},\varphi)\mathbf{dtd}\varphi \tag{138}$$

where $\chi_0(\mathbf{t},\varphi)$ is the ground-state left-eigenfunction of Eq. (137). The writhing number Eq. (135) is calculated according to Eq. (A16) as

$$\langle Wr\rangle = L\frac{\Gamma}{16\pi\ell_p^* k_B T}\int \chi_0(\mathbf{t},\varphi)\sin^2\theta\Phi_0(\mathbf{t},\varphi)\mathrm{dtd}\varphi \tag{139}$$

and average linking number is then expressed as

$$\langle Lk\rangle = \langle Tw\rangle + \langle Wr\rangle = \frac{L}{2\pi R}\int \chi_0\sin\varphi\,\Phi_0\,\mathrm{dt\,d}\varphi + \frac{L\Gamma}{16\pi\ell_p^* k_B T}\int \chi_0\sin^2\theta\,\Phi_0\,\mathrm{dt\,d}\varphi \tag{140}$$

The theoretical relationship between extension and supercoiling degree is shown in Fig. 35 and compared with the experimental result of Strick et al. [20]. In obtaining these curves, the values of the parameters are the same as those used in Fig. 33, and no adjustment has been done to fit the experimental data. We find that in the case of negatively supercoiled DNA, the theoretical and experimental results are in quantitative agreement, indicating that the present model is capable of explaining the elasticity of negatively supercoiled DNA; in the case of positively supercoiled DNA, the agreement between theory and experiment is not so good, especially when the external force is relatively large. In this simple model, we have not considered the possible deformations of the nucleotide base pairs. Though this assumption might be reasonable in the negatively supercoiled case, it may fail for positively supercoiled DNA chain, especially at large stretching forces. The work done by Allemand et al. [23] suggested that positive supercoiled and highly extended DNA molecule can take on Pauling-like configurations with exposed bases.

For negatively supercoiled DNA molecule, both theory and experiment reveal the following elastic aspects: (a) When external force is small, DNA molecule can shake off its torsional stress by writhing its central axis, which can lead to an increase in the negative writhing number and hence restore the local folding manner of DNA strands to that of B-form DNA; (b) However, writhing of the central axis causes shortening of DNA end-to-end extension, which becomes more and more unfavorable as the external force is increased. Therefore, at large forces, the torsional stress caused by negative torque (supercoiling degree) begins

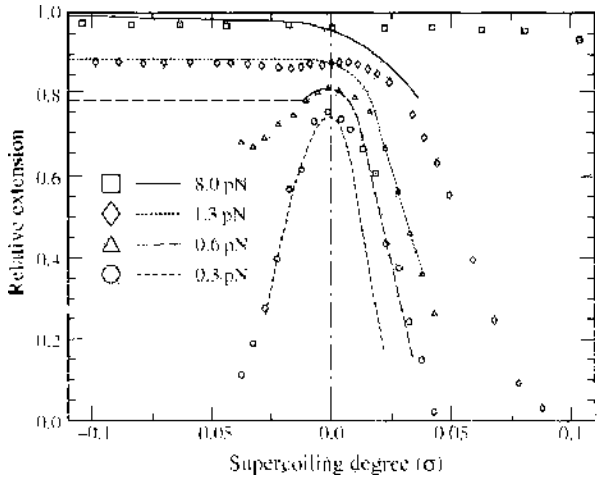


Figure 35. Extension versus supercoiling relations at fixed pulling forces for torsionally constrained DNA. The parameters for the curves are the same as Fig. 33, and experimental data is from Fig. 3 of Ref. [20] (symbols). Reprinted with permission from [28], H. Zhou et al., *Phys. Rev. Lett.* 82, 4560 (1999). © 1999, American Physical Society.

to unwind the B-form double-helix and triggers the transition of DNA internal structure, where a continuously increasing portion of DNA takes on some certain new configuration as supercoiling increases, while its total extension keeps almost invariant.

What is the new configuration? Strick and coauthors [20] suggested this new configuration corresponds to denatured DNA segments (i.e., negative torque leads to breakage of hydrogen bonds between the complementary DNA bases and consequently to strand-separation). To check this picture, Strick and coauthors [21] used short single-stranded homologous DNA segments to detect possible denaturation bubbles. They found that these homologous DNA probes indeed bind onto negatively supercoiled dsDNA molecules.

As seen in Fig. 35, although the simple theoretical model of this section does not taken into account the possibility of strand-separation, it can quantitatively explain the behavior of negatively supercoiled DNA. It is interesting to see what is the prediction of the new configuration according to this model. The calculations performed in Ref. [159] suggested that, in negatively supercoiled and highly extended DNA, the two strands of DNA might be in a left-handed double-helix configuration. The structural parameters of the left-handed configuration suggested by the theory are listed in Table 3 and compared with those of Z-form DNA [12]. The strong similarity in these parameters suggests that the torque-induced left-handed configurations may resemble the Z-form of double-stranded DNA [12, 30]. It may be possible that in this left-handed DNA configuration, the base-pairing is not stable. This can explain why short homologous DNA segments can bind to the DNA polymer [21].

The supercoiling property of stretched DNA were also studied theoretically using isotropic models by many others; see, for example, Refs. [170, 192–200].

The anisotropic twisting elasticity of double-stranded DNA was investigated by considering the coupling between DNA spontaneous curvature and twist; see, for example, Refs. [171, 201–204]. Garrivier and Fourcade considered the twist-bend coupling due to the intrinsic curvature in DNA, and very interesting numerical results were reported [202]. Panyukov and Rabin also demonstrated by analytical and numerical calculations that, under external stretching, a supercoiled helical ribbon polymer has anisotropic force-extension relationship [171, 203]. Theoretical work on this respect may have direct biological significance, as real DNA molecules has local sequence-dependent intrinsic curvature.

6.2. DNA Supercoiling Studied by Monte Carlo Simulation

The analytical calculations mentioned in Section 6.1 apply to highly stretched DNA. To go beyond this limit, a powerful method is to use Monte Carlo simulation. Vologodskii and Marko in Ref. [205] studied the supercoiling property of stretched DNA by using an achiral model of dsDNA. To better understand the experimental observations of Strick et al. [20], the chiral model of dsDNA mentioned in Section 5.2 and 6.1 was studied by Zhang and coauthors [206] following the procedure of Ref. [205]. Here we recall the work of Ref. [206] in some detail to demonstrate the main points in simulating the configurational property of topologically constrained DNA.

6.2.1. Discrete DNA Model

In the simulation, the double-stranded DNA molecule is modeled as a chain of discrete cylinders, or two discrete wormlike chains constrained by basepairs of fixed length $2R$ (Fig. 36).

Table 3. Structural parameters for the suggested torque-induced left-handed DNA configurations at different external force.[a]

Force (pN)	Torque (k_BT)	d (Å)	Pitch (Å)	Num
1.3	−5.0	3.59	41.20	11.48
1.0	−5.0	3.57	40.93	11.44
1.3	−4.0	3.83	46.76	12.19
1.0	−4.0	3.82	46.38	12.15
Z-form:		3.8	45.6	12

[a] d, the average rise per base pair; Num, the number of base pairs per turn of helix. The last row contains the corresponding values for Z-form DNA. Reprinted with permission from [159], H. Zhou et al., *Phys. Rev. E* 62, 1045 (2000). © 2000, American Physical Society.

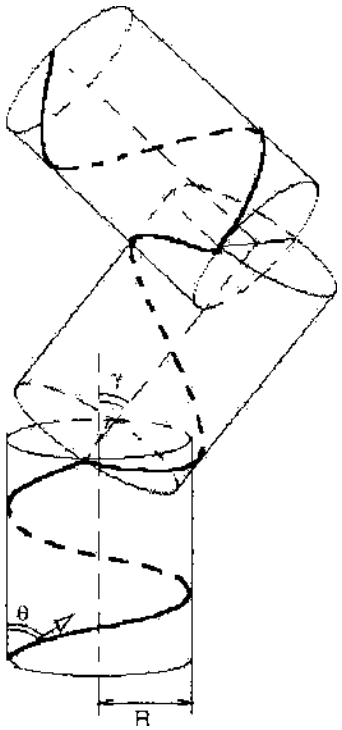


Figure 36. A three-segment configuration of discrete DNA chain in Monte Carlo simulation. Reprinted with permission from [206], Y. Zhang et al., *Biophys. J.* 78, 1979 (2000). © 2000, Biophysical Society of USA.

The conformation of the chain is specified by the space positions of vertices of its central axis, $\mathbf{r}_i=(x_i, y_i, z_i)$ in 3-D Cartesian coordinate system and the folding angle of the sugar-phosphate backbones around the central axis, θ_i, $i=1,2,\ldots,N$. Each segment is assigned the same amount of basepairs, n_{bp}, so that the length of the ith segment satisfies

$$\Delta s_i=|\mathbf{r}_i-\mathbf{r}_{i-1}|=0.34n_{\text{bp}}\frac{\cos\theta_i}{\langle\cos\theta\rangle_0} \tag{141}$$

where $\langle\cdots\rangle_0$ means the thermal average for a relaxed DNA molecule.

According to Eq. (126), the total energy of dsDNA molecule with N segments in our discrete computer model is expressed as

$$E=\alpha\sum_{i=1}^{N-1}\gamma_i^2+\alpha'\sum_{i=1}^{N-1}(\theta_{i+1}-\theta_i)^2+\frac{\kappa}{R^2}\sum_{i=1}^{N}\Delta s_i\sin^3\theta_i\tan\theta_i+\sum_{j=1}^{N_{\text{bp}}}U(\theta_j)-Fz_N \tag{142}$$

where γ_i is the bending angle between the $(i-1)$th and the ith segments, and z_N is the total extension of the DNA central axis along the direction of the external force F (assumed in the z-direction). Here, the bending rigidity constant α corresponds to the persistence length $p=53$ nm of dsDNA according to the direct discretization of Eq. (126); that is,

$$\alpha=\frac{p}{2b}k_BT \tag{143}$$

where b is the average length of segments.

In earlier approaches [205, 206], the bending rigidity constant α of a discrete chain was determined according to the Kuhn statistic length of wormlike chain, which is twice of the persistence length p. In fact, the Kuhn length l_{kuhn} of a discrete wormlike chain with rigidity α is written as (see, for example, Ref. [206])

$$l_{\text{kuhn}}=b\frac{1+\langle\cos\gamma\rangle}{1-\langle\cos\gamma\rangle} \tag{144}$$

where

$$\langle \cos\gamma \rangle = \frac{\int_0^{\pi} \cos\gamma \exp(-\alpha\gamma^2/k_B T)\sin\gamma \, d\gamma}{\int_0^{\pi} \exp(-\alpha\gamma^2/k_B T)\sin\gamma \, d\gamma} \tag{145}$$

The rigidity constant α defined in this way is only a function of $m = l_{kuhn}/b$, the number of links within one Kuhn length. The dependence of α versus m obtained from Eqs. (144) and (145) is shown numerically in Fig. 37. The rigidity constant α follows very well the linear dependence upon m, especially in the reasonable region of $m > 5$. As a comparison, we also show the line of Eq. (143) (i.e. $\alpha = (m/4)k_B T$). It is obvious that the rigidity constants of discrete chain are numerically equivalent in two algorithms. However, the processes of discretization of Eq. (126) are much more directly determined, especially to the models with complicated elasticity.

The constant α' in the second term of Eq. (142) should be associated with stiffness of the DNA single strand. As an approximation, we have taken here $\alpha' = \alpha$. The unpublished data show that the second term of Eq. (142) is quite small compared with the other four terms in small torsional deviations. The result of the simulation is not sensitive to α' in these cases.

6.2.2. Monte Carlo Simulation Procedure

The equilibrium sets of dsDNA conformations are constructed using the Metropolis MC procedure [207]. The conformational space is sampled through a Markov process. Three kinds of movements are considered [186, 205, 206]: (1) a randomly chosen segment is undertwisted or overtwisted by an angle λ_1 [Fig. 38(a)]; (2) a portion of the chain is rotated by λ_2 around the axis connecting the two ends of rotated chain [Fig. 38(b)]; (3) the segments from a randomly chosen vertex to the free end are rotated by λ_3 around an arbitrary orientation axis that passes the chosen vertex [Fig. 38(c)]. All these three types of movements satisfy the basic requirement of the Metropolis procedure of microscopic reversibility, and each of them is performed with a probability of 1/3. The value of λ_1, λ_2, λ_3 are uniformly distributed over interval $(-\lambda_1^0, \lambda_1^0)$, $(-\lambda_2^0, \lambda_2^0)$, and $(-\lambda_3^0, \lambda_3^0)$, respectively, and λ_1^0, λ_2^0, and λ_3^0 are chosen to guarantee an appropriate acceptance rate of trial movements.

A trial move from a conformation i to a conformation j is accepted on the basis of the probability $p_{i \to j} = \min(1, \rho_j/\rho_i)$, where ρ_i is the probability density of conformation i. Energetic importance sampling is realized in the Metropolis MC method by choosing the probability density ρ_i as the Boltzmann probability: $\rho_i = \exp(-E_i/k_B T)$, where E_i is the energy of conformation i calculated according to Eq. (142).

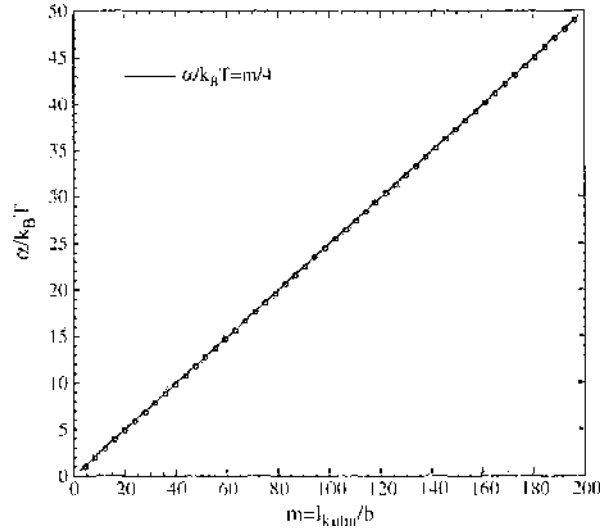


Figure 37. The bending rigidity constant α of a discrete wormlike chain as function of the number of segments within one Kuhn length $m = l_{kuhn}/b$. The circle points denote the result calculated from Eqs. (144) and (145); the solid line from direct discretization of Eq. (126).

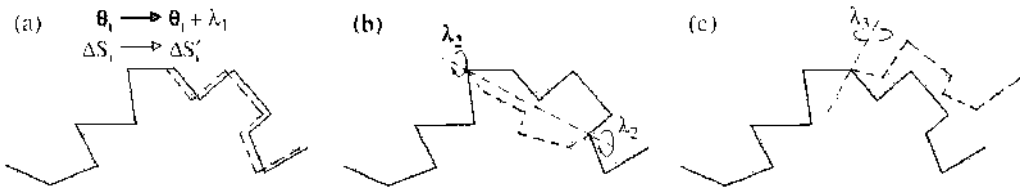


Figure 38. Movements of dsDNA chain during Monte Carlo simulations. The current conformation of the DNA central axis is shown by the solid lines and the trial conformations by dashed lines. (a) The folding angle θ_i in the ith segment is changed into $\theta_i + \lambda_1$. All segments between ith vertex and the free end are translated by the distance of $\Delta s_i - \Delta s_i'$. (b) A portion of the chain is rotated by an angle of λ_2 around the axis connecting the two ends of rotated chain. (c) The segments from a randomly chosen vertex to the free end are rotated by an angle λ_3 around an arbitrary orientation axis that passes the chosen vertex. Reprinted with permission from [206], Y. Zhang et al., *Biophys. J.* 78, 1979 (2000). © 2000. Biophysical Society of USA.

The starting conformation of the chain is unknotted. To avoid knotted configuration in the Markov process, we calculate the Alexander polynomial of each trial conformation [208, 209]. In a case where the trial movement knots the chain, the trial movement will be rejected. To incorporate the excluded-volume effect, for each trial conformation we calculate the distance from the points on the axis of a segment to the points on the axis of another nonadjacent segment. If the minimum distance for any two chosen segments is less than the DNA diameter $2R$, the energy of trial conformation is set to be infinite and the movement is then rejected.

During the configurational evolution of DNA, the supercoiling degree σ may distribute around all the possible values. In order to avoid the waste of computational samples and also for the propose of comparison with the supercoiling experiments [20, 22], we bound the supercoiling σ of DNA chain inside the region of $-0.12 \geq \sigma \geq 0.12$. When the torsion degree of trial conformation is beyond the chosen range, we simply neglect the movement and reproduce a new trial movement again.

The linking number Lk of each conformation is calculated according to Eq. (131), where the twist number Tw [Eq. (132)] can be re-expressed by

$$Tw = \frac{1}{2\pi R} \sum_{i=1}^{N} \Delta s_i \tan\theta_i \tag{146}$$

To enclose the linear DNA molecule without changing its linking number, we add three long flat ribbons to the two ends for each conformation and keep the ribbons in the same planar (Fig. 39). The writhing number Wr of the linear DNA is equal to that of the closed circular configuration and can be thereby calculated by Eq. (133).

6.2.3. Elasticity of Supercoiled dsDNA

To obtain equilibrium ensemble of DNA evolution, 10^7 elementary displacements are produced for each chosen applied force F. The relative extension and supercoiling degree σ of each accepted conformation of DNA chain are calculated. When the trial movement is rejected, the current conformation is counted up twice [207].

In order to see the dependence of mechanics property of dsDNA upon supercoiling degree, the whole sample is partitioned into 15 subsamples according to the value of the supercoiling degree σ. For each subsample, we calculate the averaged extension

$$\langle z_N \rangle_j = \frac{1}{N_j} \sum_{i=1}^{N_j} \frac{z_N(i)}{L_B} \qquad j = 1, \ldots, 15 \tag{147}$$

and the averaged torsion

$$\langle \sigma \rangle = \frac{1}{N_j} \sum_{i=1}^{N_j} \sigma_i \qquad j = 1, \ldots, 15 \tag{148}$$

where N_j is the number of movements of supercoiling that belong to jth subsample.

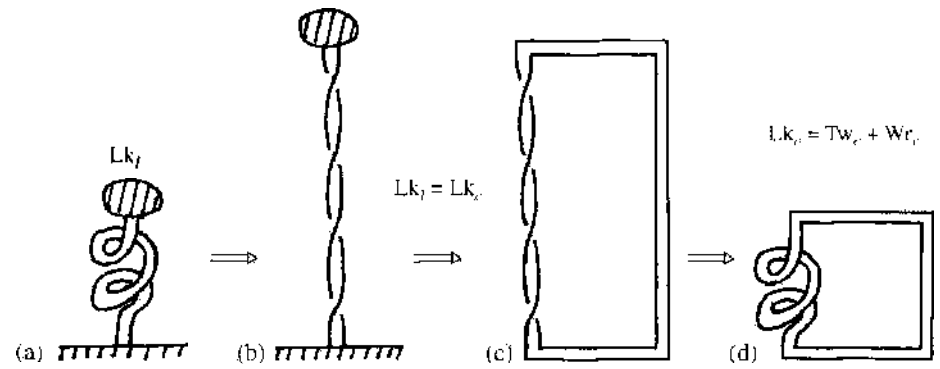


Figure 39. The schematic diagram to calculate link number in our simulations. (a) For a linear supercoiled DNA chain with one end attached to a microscope slide and with the other end attached to a magnetic bead, when the orientation of the bead is fixed and the DNA chain is forbidden to go round the bead, the number of times for two strands to interwind each other, the linking number of the linear DNA (Lk_l), is a topological constant. (b) The DNA double-helix is stretched to a fully extended form while the orientation of bead keeps unchanged. The link number of linear DNA chain is equal to the twist number (i.e. $Lk_l = Tw_l$). (c) Three long flat ribbons are connected to the two ends of the linear twisted DNA of (b). The link number of the new double-helix circle is equal to that of linear DNA chain, that is, $Lk_c = Tw_c = Tw_l = Lk_l$, because the writhe of the rectangle loop is 0. (d) The DNA circle in (c) can be deformed into a new circle, one part of which has the same steric structure as the linear supercoiled DNA chain in (a). So by adding three straight ribbons, the link number of linear double helix DNA can be obtained by calculating the link number of the new DNA circle (i.e. $Lk_l = Lk_c = Tw + Wr$). Reprinted with permission from [206], Y. Zhang et al., *Biophys. J.* 78, 1979 (2000). © 2000, Biophysical Society of USA.

We display in Fig. 40 the averaged extension as a function of supercoiling degree for three typical applied forces. In spite of quantitative difference between Monte Carlo results and experimental data, the qualitative coincidence is striking. Especially, three evident regimes exist in both experimental data and the Monte Carlo simulations:

(1) At a low force, the elastic behavior of DNA is symmetric under positive or negative supercoiling. This is understandable, as the DNA torsion is the cooperative result of hydrogen-bond constrained bending of DNA backbones and the base-pair stacking interaction in our model. At very low force, the contribution from applied force and

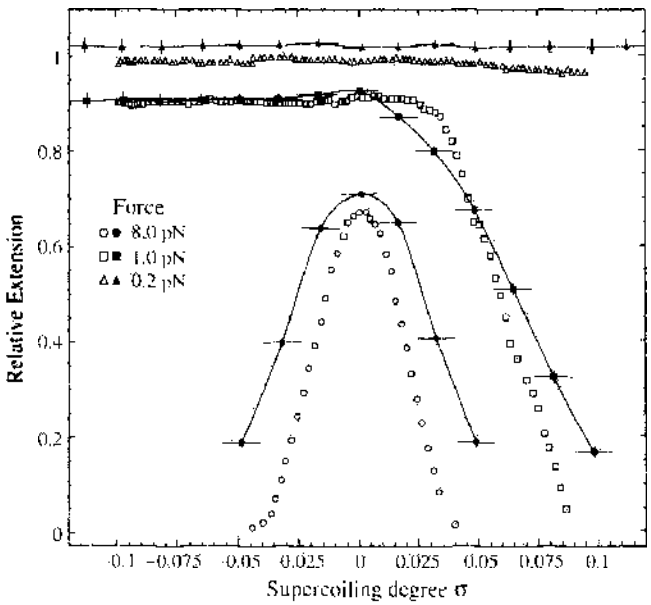


Figure 40. Relative extension versus supercoiling degree of DNA polymer for three typical stretch forces. Open points denote the experimental data [22] and solid points the results of Monte Carlo simulation of the discrete dsDNA model [206]. The vertical bars of the solid points signify the statistic error of the simulations, and the horizontal ones denote the bin-width that we partition the phase space of supercoiling degree. The solid lines connect the solid points to guide the eye. Reprinted with permission from [206], Y. Zhang et al., *Biophys. J.* 78, 1979 (2000). © 2000, Biophysical Society of USA.

the thermodynamic fluctuation perturbs the folding angle θ of base pair to derive very little from the equilibrium position, θ_0. Therefore, the DNA elasticity is achiral at this region. For a fixed applied force, the increasing torsion stress tends to produce plectonemic state that shortens the distance between the two ends, therefore, the relative extension of linear DNA polymer. These features can be also understood by the traditional approaches with harmonic twist and bending elasticity [196, 205].

(2) At intermediate force, the folding angle of basepairs are pulled slightly further away from equilibrium value θ_0 where van der Waals potential is not symmetric around θ_0. So the chiral nature of elasticity of the DNA molecule appears. In the negatively supercoiled region (i.e., $\theta < \theta_0$), the contribution of applied force to the total elastic energy dominates that of the stacking potential. So the extension is insensitive to negative supercoiling degree. On the other hand, the positive supercoiling still tends to contract the molecule.

(3) At higher force, the contribution of the applied force to the energy dominates that of van der Waals potential in both over- and underwound DNA. DNA is in the extended configuration, and its writhing number is small. The polymer is highly twisted.

The biological significance of DNA supercoiling has been discussed intensively in the literature (see, for example, Ref. [186]). Benham and coworkers [210, 211] quantitatively studied the base-pair opening probability as a function of supercoiling degree based on some simple models. The approach of Benham et al. [210, 211] was recently extended by Li and Ou-Yang [212] and very interesting predictions were obtained concerning the transcriptional initiation of several genes. It is anticipated that in coming years, the link between DNA supercoiling and gene transcription will be understood more thoroughly.

7. CONCLUSIONS

In this chapter, we have reviewed recent theoretical work on the statistical mechanics of single-molecule DNA and RNA polymers. The topics covered in this chapter include DNA double-helix denaturation, DNA unzipping, DNA entropic elasticity, DNA overstretching, DNA supercoiling, and RNA secondary structure formation and prediction. The theoretical methods used in these studies include generating functions, path integrals, Green equations, and Monte Carlo simulations. These methods were detailedly demonstrated. Here we summarize the main results of this chapter.

DNA thermal denaturation and force-induced unzipping were discussed in Section 2. Three different kinds of models were mentioned, namely the de Gennes–Peyrard–Bishop model, the Montanari–Mézard model, and the Poland–Scheraga model. One main conclusion is that the cooperativity of the DNA denaturation transition is largely due to the short-ranged base-pair stacking potential. It was also demonstrated that measuring the force-extension response curves of DNA unzipping is unlikely to probe DNA sequence information down to the single nucleotide resolution.

RNA secondary structure transition was studied in Section 3. The elastic properties of random-sequence RNA and weakly designed RNA were investigated, as well as the glassy phase transition in RNA secondary structures. It was demonstrated that the base-pair stacking interactions could stabilize the structure of RNA and make the RNA hairpin-coil transition a highly cooperative process.

Electrostatic interactions in RNA polymers were quantitatively accounted. The idea of using single-molecule force-extension measurement to identify the secondary structure of RNA was also reviewed in Section 3.

Following a short section (Section 4) on the entropic elasticity of DNA, Section 5 investigated the overstretching transition of double-stranded DNA. A double-stranded polymer model was mentioned. The highly cooperative overstretching transition in DNA is closely related to the short-ranged nature of DNA base-pair stacking interactions.

The last section (Section 6) focused on the behavior of a supercoiled and stretched DNA polymer. It was demonstrated that the double-stranded polymer model of Section 5 is able to quantitatively explain the asymmetric elasticity of negatively and positively supercoiled

DNA chains. It was also suggested that, under the action of an external force, a negatively supercoiled DNA might be in a left-handed Z-form configuration.

The mechanical property of DNA directly influences its biological functions. This chapter demonstrated that many of the elastic properties of DNA could be understood by taking into account the particular structure of DNA as a double-stranded, base-pair stacked, helical macromolecule. We should emphasize here that the short-ranged base-pair stacking interaction is very important to understand the mechanical property of DNA. On the one hand, the base-pair stacking interaction makes DNA double-helix considerably stable to small environmental variations; on the other hand, due to this short-ranged interaction, dramatic, abrupt, and reversible structural transitions of DNA could be achieved. This later property may be quite important for a biological organism to survive dramatic environmental changes.

The structure and elasticity of DNA also influences its interaction with proteins. Hopefully in the near future the interactions between DNA and various proteins will also be thoroughly interpreted.

ACKNOWLEDGMENTS

We are grateful to Dr. Ulrich Gerland for permitting us to reproduce Fig. 27 and Fig. 29, to Mr. Li Ming for reading and correcting an earlier version of the manuscript, and to Mr. Doug Lundberg for permitting us to reproduce Fig. 1. We thank Professor Reinhard Lipowsky and Professor Yu Lu for support. H. Z. acknowledges the hospitality of the Institute of Theoretical Physics (ITP) and the Interdisciplinary Center of Theoretical Studies (ICTS) in Beijing, where the main part of this manuscript was written.

APPENDIX: PATH INTEGRAL METHOD IN POLYMER STATISTICAL PHYSICS

In Section 4, we have used path integral and Green equation techniques in studying the elastic property of DNA. To make this review self-contained, in this appendix we review some basic ideas on the application of path integral method to the study of polymeric systems [147, 159, 213–215].

Consider a polymeric string, and suppose its total "arclength" is L, and along each arclength point s one can define an n-dimensional "vector" $\mathbf{r}(s)$ to describe the polymer's local state at this point. (For example, in the case of a flexible Gaussian chain, $\mathbf{r}$ is a three-dimensional position vector; in the case of a semiflexible chain such as the wormlike chain [158, 161], $\mathbf{r}$ is the unit tangent vector of the polymer and is two-dimensional.) We further assume that the energy density (per unit arclength) of the polymer can be written as the following general form:

$$\rho_e(\mathbf{r}, s) = \frac{m}{2}\left(\frac{d\mathbf{r}}{ds}\right)^2 + \mathbf{A}(\mathbf{r})\cdot\frac{d\mathbf{r}}{ds} + V(\mathbf{r}) \tag{A1}$$

where $V(\mathbf{r})$ is a scalar field and $\mathbf{A}(\mathbf{r})$ is a vectorial field. The total partition function of the system is expressed by the following integration:

$$\Xi(L) = \iint d\mathbf{r}_f \phi_f(\mathbf{r}_f) G(\mathbf{r}_f, L; \mathbf{r}_i, 0) \phi_i(\mathbf{r}_i) d\mathbf{r}_i \tag{A2}$$

where $\phi_i(\mathbf{r})$ and $\phi_f(\mathbf{r})$ are, respectively, the probability distributions of the vector $\mathbf{r}$ at the initial ($s=0$) and final ($s=L$) arclength point; $G(\mathbf{r}, s; \mathbf{r}', s')$ is called the Green function, and it is defined in the following way:

$$G(\mathbf{r}, s; \mathbf{r}', s') = \int_{\mathbf{r}'}^{\mathbf{r}} \mathcal{D}[\mathbf{r}''(s)] \exp\left[-\beta \int_{s'}^{s} ds'' \rho_e(\mathbf{r}'', s'')\right] \tag{A3}$$

where integration is carried over all possible configurations of $\mathbf{r}''$, and $\beta = 1/k_B T$ is the Boltzmann coefficient. It can be verified that the Green function defined above satisfies the

following relation:

$$G(\mathbf{r},s;\mathbf{r}',s')=\int d\mathbf{r}''G(\mathbf{r},s;\mathbf{r}'',s'')G(\mathbf{r}'',s'';\mathbf{r}',s') \qquad (s'<s''<s) \tag{A4}$$

The total free energy of the system is then expressed as

$$\mathcal{F}=-k_BT\ln\Xi \tag{A5}$$

To calculate the total partition function Ξ, we define an auxiliary function $\Psi(\mathbf{r},s)$ and call it the wave function because of its similarity with the true wave function of quantum systems. Suppose the value of Ψ at arclength point s is related to its value at s' through the following formula such that

$$\Psi(\mathbf{r},s)=\int d\mathbf{r}'G(\mathbf{r},s;\mathbf{r}',s')\Psi(\mathbf{r}',s') \qquad (s>s') \tag{A6}$$

[In fact, the choice of the wave function $\Psi(\mathbf{r},s)$ is not limited. Any function determined by an integration of the form of Eq. (A6) can be viewed as a wave function.]

We can derive from Eqs. (A1), (A3), and (A6) that:

$$\frac{\partial\Psi(\mathbf{r},s)}{\partial s}=\left[\frac{\nabla_r^2}{2m\beta}-\beta V(\mathbf{r})+\frac{\mathbf{A}(\mathbf{r})\cdot\nabla_r}{m}+\frac{\nabla_r\cdot\mathbf{A}(\mathbf{r})}{2m}+\frac{\beta\mathbf{A}^2(\mathbf{r})}{2m}\right]\Psi(\mathbf{r},s)=\hat{H}\Psi(\mathbf{r},s) \tag{A7}$$

Equation (A7) is called the Green equation; it is very similar to the Schrödinger equation of quantum mechanics. However, there is an important difference. In the case of $\mathbf{A}(\mathbf{r}\neq 0$, the operator $\hat{H}$ in Eq. (A7) is not Hermitian. Therefore, in this case, the matrix form o the operator $\hat{H}$ may not be diagonalized by unitary matrix.

Denote the eigenvalues and the right-eigenfunctions of Eq. (A7) as $-g_i$ and $|i\rangle=\phi_i(\mathbf{r})$ $(i=0,1,\ldots)$, respectively. Then it is easy to know, from the approach of quantum mechanics, that

$$\Psi(s)=\sum_i e^{-g_i(s-s')}|i(s)\rangle\langle i(s')|\Psi(s')\rangle \tag{A8}$$

where $\langle i|=\chi_i(\mathbf{r})$ $(i=0,1,\ldots)$ denote the left-eigenfunctions of Eq. (A7), which satisfy the following relation:

$$\langle i|i'\rangle=\int d\mathbf{r}\,\chi_i(\mathbf{r})\Phi_{i'}(\mathbf{r})=\delta_i^{i'}$$

In the case where $\hat{H}$ is Hermitian [i.e., $\mathbf{A}(\mathbf{r})=0$], then we can conclude that

$$\chi_i(\mathbf{r})=\Phi_i^*(\mathbf{r})$$

From Eqs. (A2), (A6), (A7), and (A8), we know that

$$\Xi(L)=\int d\mathbf{r}\int d\mathbf{r}'G(\mathbf{r},L;\mathbf{r}',s')\phi_f(\mathbf{r})\phi_i(\mathbf{r}')=\sum_i\langle\phi_f|i\rangle\langle i|\phi_i\rangle e^{-g_iL}$$

$$=e^{-g_0L}\langle\phi_f|0\rangle\langle 0|\phi_i\rangle \qquad [\text{for } L\gg 1/(g_1-g_0)] \tag{A9}$$

Consequently, for long polymer chains, the total free energy density is just expressed a

$$\mathcal{F}/L=k_BTg_0 \tag{A10}$$

and any quantity of interest can then be calculated by differentiation of $\mathcal{F}$. For example, the average extension of a polymer under external force field f can be calculated as $\langle z\rangle=\partial\mathcal{F}/\partial f=Lk_BT\partial g_0/\partial f$.

We continue to discuss another very important quantity, the distribution probability of $\mathbf{r}$ at arclength s, $P(\mathbf{r},s)$. This probability is calculated from the following expression:

$$P(\mathbf{r},s)=\frac{\int d\mathbf{r}_f \int d\mathbf{r}_i \phi_f(\mathbf{r}_f) G(\mathbf{r}_f, L; \mathbf{r}, s) G(\mathbf{r}, s; \mathbf{r}_i, 0) \phi_i(\mathbf{r}_i)}{\int d\mathbf{r}_f \int d\mathbf{r}_i \phi_f(\mathbf{r}_f) G(\mathbf{r}_f, L; \mathbf{r}_i, 0) \phi_i(\mathbf{r}_i)} \tag{A11}$$

Based on Eqs. (A6) and (A8), we can rewrite Eq. (A11) in the following form:

$$\begin{aligned} P(\mathbf{r},s) &= \frac{\int d\mathbf{r}_f \int d\mathbf{r}_i \int d\mathbf{r}' \phi_f(\mathbf{r}_f) G(\mathbf{r}_f, L; \mathbf{r}', s) \delta(\mathbf{r}'-\mathbf{r}) G(\mathbf{r}, s; \mathbf{r}_i, 0) \phi_i(\mathbf{r}_i)}{\int d\mathbf{r}_f \int d\mathbf{r}_i \phi_f(\mathbf{r}_f) G(\mathbf{r}_f, L; \mathbf{r}_i, 0) \phi_i(\mathbf{r}_i)} \\ &= \frac{\sum_m \sum_n \langle \phi_f | m \rangle \langle n | \phi_i \rangle \chi_m(\mathbf{r}) \Phi_n(\mathbf{r}) \exp[-g_m(L-s) - g_n s]}{\sum_m \langle \phi_f | m \rangle \langle m | \phi_i \rangle \exp(-g_m L)} \end{aligned} \tag{A12}$$

For the most important case of $0 \ll s \ll L$, Eq. (A12) then gives that the probability distribution of $\mathbf{r}$ is independent of arclength s; that is,

$$P(\mathbf{r},s) = \chi_0(\mathbf{r}) \Phi_0(\mathbf{r}) \qquad (\text{for } 0 \ll s \ll L) \tag{A13}$$

With the help of Eq. (A13), the average value of a quantity that is a function of $\mathbf{r}$ can be obtained. For example,

$$\langle Q(s) \rangle = \int d\mathbf{r}\, Q(\mathbf{r}) P(\mathbf{r},s) = \int d\mathbf{r}\, \chi_0(\mathbf{r}) Q(\mathbf{r}) \Phi_0(\mathbf{r}) = \langle 0 | Q | 0 \rangle \tag{A14}$$

and

$$\left\langle \int_0^L Q(\mathbf{r}(s)) ds \right\rangle = \int_0^L \langle Q(s) \rangle ds = L \langle 0 | Q | 0 \rangle \qquad [\text{for } L \gg 1/(g_1 - g_0)] \tag{A15}$$

Finally, we list the formula for calculating $\langle \mathbf{B}(\mathbf{r}) \cdot d\mathbf{r}/ds \rangle$; here, $\mathbf{B}(\mathbf{r})$ is a given vectorial field. The formula reads:

$$\left\langle \mathbf{B}(\mathbf{r}) \cdot \frac{d\mathbf{r}}{ds} \right\rangle = \frac{1}{2} \langle 0 | [\mathbf{r} \cdot (\hat{H}\mathbf{B}) - \mathbf{B} \cdot (\hat{H}\mathbf{r})] | 0 \rangle \qquad [\text{for } L \gg 1/(g_1 - g_0)] \tag{A16}$$

REFERENCES

1. S. B. Smith, L. Finzi, and C. Bustamante, *Science* 258, 1122 (1992).
2. C. Bustamante, S. B. Smith, J. Liphardt, and D. Smith, *Curr. Opin. Struct. Biol.* 10, 279 (2000).
3. H. Clausen-Schaumann, M. Seitz, R. Krautbauer, and H. Gaub, *Curr. Opin. Chem. Biol.* 4, 524 (2000).
4. T. Strick, J.-F. Allemand, V. Croquette, and D. Bensimon, *Physics Today* 54, 46 (2001).
5. M. C. Williams and I. Rouzina, *Curr. Opin. Struct. Biol.* 12, 330 (2002).
6. M. Hegner and W. Grange, *J. Muscle Res. Cell. Motility* 23, 367 (2002).
7. J. F. Allemand, D. Bensimon, and V. Croquette, *Curr. Opin. Struct. Biol.* 13, 266 (2003).
8. C. Bustamante, Z. Bryant, and S. B. Smith, *Nature* (*London*) 421, 423 (2003).
9. J. D. Watson and F. H. C. Crick, *Nature* (*London*) 171, 737 (1953).
10. M. H. F. Wilkins, A. R. Stokes, and H. R. Wilson, *Nature* (*London*) 171, 738 (1953).
11. R. E. Franklin and R. G. Gosling, *Nature* (*London*) 171, 740 (1953).
12. J. D. Watson, N. H. Hopkins, J. W. Roberts, J. A. Steitz, and A. M. Weiner, "Molecular Biology of the Gene," 4th ed. Benjamin/Cummings, Menlo Park, CA, 1987.
13. D. L. Nelson and M. M. Cox, "Lehninger Principles of Biochemistry," 3rd ed. Worth Publishers, New York, 2000.
14. W. Saenger, "Principles of Nucleic Acid Structure." Springer-Verlag, New York, 1984.
15. J. C. Wang, *Proc. Natl. Acad. Sci. U.S.A.* 76, 200 (1979).
16. C. Bustamante, J. F. Marko, E. D. Siggia, and S. Smith, *Science* 265, 1599 (1994).
17. Z. Bryant, M. D. Stone, J. Gore, S. B. Smith, N. R. Cozzarelli, and C. Bustamante, *Nature* (*London*) 424, 338 (2003).
18. S. B. Smith, Y. Cui, and C. Bustamante, *Science* 271, 795 (1996).
19. P. Cluzel, A. Lebrun, C. Heller, R. Lavery, J. L. Viovy, D. Chatenay, and F. Caron, *Science* 271, 792 (1996).
20. T. R. Strick, J. F. Allemand, D. Bensimon, A. Bensimon, and V. Croquette, *Science* 271, 1835 (1996).
21. T. R. Strick, V. Croquette, and D. Bensimon, *Proc. Natl. Acad. Sci. U.S.A.* 95, 10579 (1998).
22. T. R. Strick, J.-F. Allemand, D. Bensimon, and V. Croquette, *Biophys. J.* 74, 2016 (1998).
23. J. F. Allemand, D. Bensimon, R. Lavery, and V. Croquette, *Proc. Natl. Acad. Sci. U.S.A.* 95, 14152 (1998).

24. R. M. Wartell and A. S. Benight, *Phys. Rep.* 126, 67 (1985).
25. V. A. Bloomfield, D. M. Crothers, and I. Tinoco, Jr., "Physical Chemistry of Nucleic Acids." Harper & Row, New York, 1974.
26. D. Bensimon, A. J. Simon, V. Croquette, and A. Bensimon, *Phys. Rev. Lett.* 74, 4754 (1995).
27. S. B. Smith, Y. Cui, A. C. Hausrath, and C. Bustamante, *Biophys. J.* 68, A250 (1995).
28. H. Zhou, Y. Zhang, and Z.-C. Ou-Yang, *Phys. Rev. Lett.* 82, 4560 (1999).
29. H. Zhou and Z.-C. Ou-Yang, *Mod. Phys. Lett. B* 13, 999 (1999).
30. J. F. Leger, G. Romano, A. Sarkar, J. Robert, L. Bourdieu, D. Chatenay, and J. F. Marko, *Phys. Rev. Lett.* 83, 1066 (1999).
31. T. T. Perkins, D. E. Smith, and S. Chu, *Science* 276, 2016 (1997).
32. D. E. Smith and S. Chu, *Science* 281, 1335 (1998).
33. D. E. Smith, H. P. Babcock, and S. Chu, *Science* 283, 1724 (1999).
34. B. Maier, D. Bensimon, and V. Croquette, *Proc. Natl. Acad. Sci. U.S.A.* 97, 12002 (2000).
35. M. Hegner, S. B. Smith, and C. Bustamante, *Proc. Natl. Acad. Sci. U.S.A.* 96, 10109 (1999).
36. G. J. L. Wuite, S. B. Smith, M. Young, D. Keller, and C. Bustamante, *Nature (London)* 404, 103 (2000).
37. B. Essevaz-Roulet, U. Bockelmann, and F. Heslot, *Proc. Natl. Acad. Sci. U.S.A.* 94, 11935 (1997).
38. U. Bockelmann, B. Essevaz-Roulet, and F. Heslot, *Phys. Rev. Lett.* 79, 4489 (1997).
39. U. Bockelmann, B. Essevaz-Roulet, and F. Heslot, *Phys. Rev. E* 58, 2386 (1998).
40. U. Bockelmann, P. Thomen, B. Essevaz-Roulet, V. Viasnoff, and F. Heslot, *Biophys. J.* 82, 1537 (2002).
41. C. Danilowicz, V. W. Coljee, C. Bouzigues, D. K. Lubensky, D. R. Nelson, and M. Prentiss, *Proc. Natl. Acad. Sci. U.S.A.* 100, 1694 (2003).
42. G. U. Lee, L. A. Chrisey, and R. Colton, *Science* 266, 771 (1994).
43. A. D. MacKerrel and G. U. Lee, *Eur. Biophys. J.* 28, 415 (1999).
44. A. Noy, D. V. Vezenov, J. F. Kayyem, T. J. Meade, and C. M. Liebe, *Chem. Biol.* 4, 519 (1997).
45. T. Strunz, K. Oroszlan, R. Schäfer, and H.-J. Guntherodt, *Proc. Natl. Acad. Sci. U.S.A.* 96, 11277 (1999).
46. L. H. Pope, M. C. Davies, C. A. Laughton, C. J. Roberts, S. J. B. Tendler, and P. M. Williams, *Eur. Biophys. J.* 30, 53 (2001).
47. M. Rief, H. Clausen-Schaumann, and H. E. Gaub, *Nat. Struct. Biol.* 6, 346 (1999).
48. H. Clausen-Schaumann, M. Rief, C. Tolksdorf, and H. E. Gaub, *Biophys. J.* 78, 1997 (2000).
49. A. F. Sauer-Budge, J. A. Nyamwanda, D. K. Lubensky, and D. Branton, *Phys. Rev. Lett.* 90, 238101 (2003).
50. T. Boland and B. D. Ratner, *Proc. Natl. Acad. Sci. U.S.A.* 92, 5297 (1995).
51. J. Liphardt, B. Onoa, S. B. Smith, I. Tinoco, Jr., and C. Bustamante, *Science* 292, 733 (2001).
52. J. Liphardt, S. Dumont, S. B. Smith, I. Tinoco, Jr., and C. Bustamante, *Science* 296 1832 (2002).
53. B. Onoa, S. Dumont, J. Liphardt, S. B. Smith, I. Tinoco, Jr., and C. Bustamante, *Science* 299, 1892 (2003).
54. X. Zhuang, H. Kim, M. J. B. Pereira, H. P. Babcock, N. G. Walter, and S. Chu, *Science* 296, 1473 (2002).
55. C. Jarzynski, *Phys. Rev. Lett.* 78, 2690 (1997).
56. C. Jarzynski, *Phys. Rev. E* 56, 5018 (1997).
57. H. Yin, R. Landick, and J. Gelles, *Biophys. J.* 67, 2468 (1994).
58. H. Yin, M. D. Wang, K. Svoboda, R. Landick, S. M. Block, and J. Gelles, *Science* 270, 1653 (1995).
59. M. D. Wang, M. J. Schnitzer, H. Yin, R. Landick, J. Gelles, and S. M. Block, *Science* 282, 902 (1998).
60. H. Yin, I. Artsimovitch, R. Landick, and J. Gelles, *Proc. Natl. Acad. Sci. U.S.A.* 96, 13124 (1999).
61. J. R. Davenport, G. L. Wuite, R. Landick, and C. Bustamante, *Science* 287, 2497 (2000).
62. G. Charvin, D. Bensimon, and V. Croquette, *Proc. Natl. Acad. Sci. U.S.A.* 100, 9820 (2003).
63. Y. Cui and C. Bustamante, *Proc. Natl. Acad. Sci. U.S.A.* 97, 127 (2000).
64. B. D. Brower-Toland, C. L. Smith, R. C. Yeh, J. T. Lis, C. L. Peterson, and M. D. Wang, *Proc. Natl. Acad. Sci. U.S.A.* 99, 1960 (2002).
65. B. Ladoux, J. P. Quivy, P. Doyle, O. duRoure, G. Almouzni, and J.-L. viovy, *Proc. Natl. Acad. Sci. U.S.A.* 97, 14251 (2000).
66. M. G. Poirier and J. F. Marko, *Proc. Natl. Acad. Sci. U.S.A.* 99, 15393 (2002).
67. J. F. Leger, J. Robert, L. Bourdieu, D. Chatenay, and J. F. Marko, *Proc. Natl. Acad. Sci. U.S.A.* 95, 12295 (1998).
68. G. V. Shivashankar, M. Feingold, O. Krichevsky, and A. Libchaber, *Proc. Natl. Acad. Sci. U.S.A.* 96, 7916 (1999).
69. R. Bar-Ziv and A. Libchaber, *Proc. Natl. Acad. Sci. U.S.A.* 98, 9068 (2001).
70. I. Braslavsky, B. Hebert, E. Kartalov, and S. R. Quake, *Proc. Natl. Acad. Sci. U.S.A.* 100, 3960 (2003).
71. S. Cocco, J. F. Marko, and R. Monasson, *C. R. Physique* 3, 569 (2002).
72. J. F. Marko and S. Cocco, *Physics World* 16, 37 (2003).
73. E. Yeramian, *Gene* 255, 139 (2000).
74. T. Lipniacki, *Phys. Rev. E* 64, 051919 (2001).
75. G. Altan-Bonnet, A. Libchaber, and O. Krichevsky, *Phys. Rev. Lett.* 90, 138101 (2003).
76. P.-G. de Gennes, *C. R. Physique* 2, 1505 (2001).
77. D. Poland and H. A. Scheraga, "Theory of Helix-Coil Transitions in Biopolymers: Statistical Mechanical Mechanical Theory of Order-Disorder Transitions in Biological Macromolecules." Academic Press, New York, 1970.
78. P.-G. de Gennes, *Rep. Prog. Phys.* 32, 187 (1969).
79. M. Peyrard and A. R. Bishop, *Phys. Rev. Lett.* **62**, 2755 (1989).
80. T. Dauxois, M. Peyrard, and A. R. Bishop, *Phys. Rev. E* 47, 684 (1993).

81. T. Dauxois, M. Peyrard, and A. R. Bishop, *Phys. Rev. E* 47, R44 (1993).
82. J. Xiao, J. Lin, and B. Tian, *Phys. Rev. E* 50, 5039 (1994).
83. D. Cule and T. Hwa, *Phys. Rev. Lett.* 79, 2375 (1997).
84. Y. L. Zhang, W. M. Zheng, J. X. Liu, and Y. Z. Chen, *Phys. Rev. E* 56, 7100 (1997).
85. M. Barbi, S. Cocco, M. Peyrard, and S. Ruffo, *J. Biol. Phys.* 24, 97 (1999).
86. N. Theodorakopoulos, T. Dauxois, and M. Peyrard, *Phys. Rev. Lett.* 85, 6 (2000).
87. H. Zhou, "Force-Induced Melting and Thermal Melting of a Double-Stranded Biopolymer" (2000), e-print: cond-mat/0007015, http://de.arxiv.org/abs/cond-mat/007015.
88. A. Montanari and M. Mézard, *Phys. Rev. Lett.* 86, 2178 (2001).
89. S. Lifson, *J. Chem. Phys.* 40, 3705 (1964).
90. D. Poland and H. A. Scherage, *J. Chem. Phys.* 45, 1464 (1966).
91. R. Bundschuh and T. Hwa, *Phys. Rev. Lett.* 83, 1479 (1999).
92. H. Zhou, Y. Zhang, and Z.-C. Ou-Yang, *Phys. Rev. Lett.* 86, 356 (2001).
93. H. Zhou and Y. Zhang, *J. Chem. Phys.* 114, 8694 (2001).
94. V. M. Pavlov, J. L. Lyubchenko, A. S. Borovik, and Y. Lazurkin, *Nucl. Acids. Res.* 4, 4052 (1977).
95. A. S. Borovik, Y. A. Kalambet, Y. L. Lyubchenko, V. T. Shitov, and E. Golovanov, *Nucleic Acids Res.* 8, 4165 (1980).
96. A. Montrichok, G. Gruner, and G. Zocchi, *Europhys. Lett.* 62, 452 (2003).
97. S. M. Bhattacharjee, *J. Phys. A* 33, L423 (2000).
98. D. Marenduzzo, A. Trovato, and A. Maritan, *Phys. Rev. E* 64, 031901 (2001).
99. I. Rouzina and V. A. Bloomfield, *Biophys. J.* 77, 3242 (1999).
100. I. Rouzina and V. A. Bloomfield, *Biophys. J.* 77, 3252 (1999).
101. M. E. Fisher, *J. Chem. Phys.* 45, 1469 (1966).
102. M. S. Causo, B. Coluzzi, and P. Grassberger, *Phys. Rev. E* 62, 3958 (2000).
103. E. Carlon, E. Orlandini, and A. L. Stella, *Phys. Rev. Lett.* 88, 198101 (2002).
104. T. Garel, C. Monthus, and H. Orland, *Europhys. Lett.* 55, 132 (2001).
105. S. M. Bhattacharjee, *Europhys. Lett.* 57, 772 (2002).
106. M. Baiesi, E. Carlon, E. Orlandini, and A. L. Stella, *Eur. Phys. J. B* 29, 129 (2002).
107. Y. Kafri, D. Mukamel, and L. Peliti, *Phys. Rev. Lett.* 85, 4988 (2000).
108. Y. Kafri, D. Mukamel, and L. Peliti, *Eur. Phys. J. B* 27, 135 (2002).
109. B. Duplantier, *Phys. Rev. Lett.* 57, 941 (1986).
110. B. Duplantier, *J. Stat. Phys.* 54, 581 (1989).
111. L. Schäfer, C. von Ferber, U. Lehr, and B. Duplantier, *Nucl. Phys. B* 374, 473 (1992).
112. P.-G. de Gennes, "Scaling Concepts in Polymer Physics." Cornell University Press, Ithaca, NY, 1979.
113. A. Hanke and R. Metzler, *Phys. Rev. Lett.* 90, 159801 (2003).
114. Y. Kafri, D. Mukamel, and L. Peliti, *Phys. Rev. Lett.* 90, 159802 (2003).
115. T. Odijk, *Macromolecules* 28, 7016 (1995).
116. F. Bueche, "Physical Properties of Polymers." Interscience, New York, 1962.
117. S. Cocco, J. F. Marko, and R. Monasson, *Eur. Phys. J. E* 10, 153 (2003).
118. D. K. Lubensky, and D. R. Nelson, *Phys. Rev. Lett.* 85, 1572 (2000).
119. D. K. Lubensky and D. R. Nelson, *Phys. Rev. E* 65, 031917 (2002).
120. S. Cocco, R. Monasson, and J. F. Marko, *Phys. Rev. E* 66, 051914 (2002).
121. S. Cocco, R. Monasson, and J. F. Marko, *Phys. Rev. E* 65, 041907 (2002).
122. F. Ritort, C. Bustamante, and I. Tinoco, Jr., *Proc. Natl. Acad. Sci. U.S.A.* 99, 13544 (2002).
123. D. Marenduzzo, S. M. Bhattacharjee, A. Maritan, E. Orlandini, and F. Seno, *Phys. Rev. Lett.* 88, 028102 (2002).
124. S.-J. Chen and K. A. Dill, *Proc. Natl. Acad. Sci. U.S.A.* 97, 646 (2000).
125. W. Zhang and S.-J. Chen, *Proc. Natl. Acad. Sci. U.S.A.* 99, 1931 (2002).
126. Z. Wang and K. Zhang, in "Current Topics in Computational Molecular Biology" (T. Jiang, Y. Xu, and M. Q. Zhang, Eds.), Chap. 14, pp. 345–363. Tsinghua University Press, Beijing, China, 2002.
127. P. G. Higgs, *Phys. Rev. Lett.* 76, 704 (1996).
128. I. Tinoco, Jr., and C. Bustamante, *J. Mol. Biol.* 293, 271 (1999).
129. F. Krzakala, M. Mézard, and M. Müller, *Europhys. Lett.* 57, 752 (2002).
130. A. Pagnani, G. Parisi, and F. Ricci-Tersenghi, *Phys. Rev. Lett.* 84, 2026 (2000).
131. R. Bundschuh and T. Hwa, *Phys. Rev. E* 65, 031903 (2002).
132. R. Bundschuh and T. Hwa, *Europhys. Lett.* 59, 903 (2002).
133. E. Marinari, A. Pagnani, and F. Ricci-Tersenghi, *Phys. Rev. E* 65, 041919 (2002).
134. M. Müller, F. Krzakala, and M. Mézard, *Eur. Phys. J. E* 9, 67 (2002).
135. M. Mézard, G. Parisi, and M. A. Virasoro, "Spin Glass Theory and Beyond." World Scientific, Singapore, 1987.
136. C. Hyeon and D. Thirumalai, *Proc. Natl. Acad. Sci. U.S.A.* 100, 10249 (2003).
137. S. Sugai and K. Nitta, *Biopolymers* 12, 1363 (1973).
138. J. J. Delrow, J. A. Gebe, and J. M. Schurr, *Biopolymers* 42, 455 (1997).
139. K. Klenin, H. Merlitz, and J. Langowski, *Biophys. J.* 74, 780 (1998).
140. Y. Zhang, H. Zhou, and Z.-C. Ou-Yang, *Biophys. J.* 81, 1133 (2001).
141. M. N. Dessinges, B. Maier, Y. Zhang, M. Peliti, D. Bensimon, and V. Croquette, *Phys. Rev. Lett.* 89, 248102 (2002).
142. S. A. Rice and M. Nagasawa, "Polyelectrolyte Solutions." Academic Press, New York, 1961.

143. P. Pierce, Ph. D. thesis. Department of Chemistry, Yale University, 1958.
144. S. L. Brenner and V. A. Parsegia, *Biophys. J.* 14, 327 (1974).
145. D. Stigter, *Biophys. J.* 16, 1435 (1977).
146. I. S. Gradsshteyn and I. M. Ryzhik, "Table of Integral, Series, and Products." Academic Press, New York, 1980.
147. M. Doi and S. F. Edwards, "The Theory of Polymer Dynamics." Clarendon Press, Oxford, UK, 1986.
148. N. Becker, E. Oroudjev, S. Mutz, J. P. Cleveland, P. K. Hansma, C. Y. Hayashi, D. E. Makarov, and H. G. Hansma, *Nat. Mater.* 2, 278 (2003).
149. G. A. Jeffrey, "An Introduction to Hydrogen Bonding." Oxford University Press, New York, 1997.
150. M. Zuker and P. Stiegler, *Nucleic Acids Res.* 9, 133 (1981).
151. J. S. McCaskill, *Biopolymers* 29, 1105 (1990).
152. A. E. Walter, D. H. Turner, J. Kim, M. H. Lyttle, P. Muller, D. H. Mathews, and M. Zuker, *Proc. Natl. Acad. Sci. U.S.A.* 91, 9218 (1994).
153. I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster, *Monatshefte fur Chemie* 125, 167 (1994).
154. D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner, *J. Mol. Biol.* 288, 911 (1999).
155. I. L. Hofacker, *Nucleic Acids Res.* 31, 3429 (2003).
156. U. Gerland, R. Bundschuh, and T. Hwa, *Biophys. J.* 81, 1324 (2001).
157. U. Gerland, R. Bundschuh, and T. Hwa, "Mechanical Pulling Through a Nanopore Can Reveal the Secondary Structure of Single RNA Molecules." (2003), e-print: cond-mat/0306126, http://de.arxiv.org/abs/cond-mat/0306126.
158. J. F. Marko and E. D. Siggia, *Macromolecules*, 28, 8759 (1995).
159. H. Zhou, Y. Zhang, and Z.-C. Ou-Yang, *Phys. Rev. E* 62, 1045 (2000).
160. J. Wilhelm and E. Frey, *Phys. Rev. Lett.* 77, 2581 (1996).
161. K. Kroy and E. Frey, *Phys. Rev. Lett.* 77, 306 (1996).
162. J. Samuel and S. Sinha, *Phys. Rev. E* 66, 050801 (2002).
163. J. Kierfeld and R. Lipowsky, *Europhys. J.* 62, 285 (2003).
164. M. C. Williams, J. R. Wenner, I. Rouzina, and V. A. Bloomfield, *Biophys. J.* 80, 874 (2001).
165. I. Rouzina and V. A. Bloomfield, *Biophys. J.* 80, 882 (2001).
166. I. Rouzina and V. A. Bloomfield, *Biophys. J.* 80, 894 (2001).
167. A. Lebrun and R. Lavery, *Nucleic Acids Res.* 24, 2260 (1996).
168. P. Cizeau and J. L. Viovy, *Biopolymers* 42, 383 (1997).
169. A. Ahsan, J. Rudnick, and R. Bruinsma, *Biophys. J.* 74, 132 (1998).
170. J. F. Marko, *Phys. Rev. E* 57, 2134 (1998).
171. S. Panyukov and Y. Rabin, *Europhys. Lett.* 57, 512 (2002).
172. P. Y. Lai and Z. C. Zhou, *Physica A* 321, 170 (2003).
173. C. Storm and P. C. Nelson, *Phys. Rev. E* 67, 051896 (2003).
174. R. L. Ornstein, R. Rein, D. L. Breen, and R. D. MacElroy, *Biopolymers*, 17, 2341 (1978).
175. T. B. Liverpool, R. Golestanian, and K. Kremer, *Phys. Rev. Lett.* 80, 405 (1998).
176. R. Golestanian and T. B. Liverpool, *Phys. Rev. E* 62, 5488 (2000).
177. Z. Zhou and P. Y. Lai, *Chem. Phys. Lett.* 346, 449 (2001).
178. B. Mergell, M. R. Ejtehadi, and R. Everaers, *Phys. Rev. E* 66, 011903 (2002).
179. B. Mergell, M. R. Ejtehadi, and R. Everaers, *Phys. Rev. E* 68 021911 (2003).
180. P. Y. Lai and Z. C. Zhou, *J. Chem. Phys.* 118, 11189 (2003).
181. F. B. Fuller, *Proc. Natl. Acad. Sci. U.S.A.* 68, 815 (1971).
182. H. G. Patterton and T. von Holt, *J. Mol. Biol.* 229, 623 (1993).
183. R. M. Harland, H. Weintraub, and S. L. Mcknight, *Nature* (*London*) 302, 38 (1983).
184. H. Weintraub, *Cell* 32, 1191 (1983).
185. E. K. Hoffman, S. P. Trusko, M. Murphy, and D. L. George, *Proc. Natl. Acad. Sci. U.S.A.* 87, 2705 (1990).
186. A. V. Vologodskii, "Topology and Physics of Circular DNA." CRC, Boca Raton, 1992.
187. G. Caglugareanu, *Rev. Math. Pure Appl.* 4, 5 (1959).
188. J. H. White, *American J. Math.* 9, 693 (1969).
189. F. H. C. Crick, *Proc. Natl. Acad. Sci. U.S.A.* 73, 2639 (1976).
190. J. H. White and J. M. Bauer, *Proc. Natl. Acad. Sci. U.S.A.* 85, 772 (1988).
191. F. B. Fuller, *Proc. Natl. Acad. Sci. U.S.A.* 75, 3557 (1978).
192. J. F. Marko and E. D. Siggia, *Phys. Rev. E* 52, 2912 (1995).
193. J. F. Marko, *Phys. Rev. E* 55, 1758 (1997).
194. J. F. Marko, *Europhys. Lett.* 38, 183 (1997).
195. B. Fain, R. Rudnick, and S. Östlund, *Phys. Rev. E* 55, 7364 (1997).
196. C. Bouchiat and M. Mézard, *Phys. Rev. Lett.* 80, 1556 (1998).
197. C. Bouchiat and M. Mézard, *Eur. Phys. J. E* 2, 377 (2000).
198. R. D. Kamien, T. C. Lubensky, P. Nelson, and C. S. O'hern, *Europhys. Lett.* 38, 237 (1997).
199. J. D. Moroz and P. Nelson, *Proc. Natl. Acad. Sci. U.S.A.* 94, 14418 (1997).
200. J. D. Moroz and P. Nelson, *Macromolecules* 31, 6333 (1998).
201. H. Zhou and Z.-C. Ou-Yang, *J. Chem. Phys.* 110, 1247 (1999).
202. D. Garrivier and B. Fourcade, *Europhys. Lett.* 49, 390 (2000).
203. S. Panyukov and Y. Rabin, *Phys. Rev. Lett.* 85, 2404 (2000).

204. S. Panyukov and Y. Rabin, *Phys. Rev. E* 62, 7135 (2000).
205. A. V. Vologodskii and J. F. Marko, *Biophys. J.* 73, 123 (1997).
206. Y. Zhang, H. Zhou, and Z.-C. Ou-Yang, *Biophys. J.* 78, 1979 (2000).
207. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *J. Chem. Phys.* 21, 1087 (1953).
208. A. V. Vologodskii, A. V. Lukashin, M. D. Frank-Kamenetskii, and V. V. Anshelevich, *Sov. Phys. JETP* 39, 1059 (1974).
209. B. A. Harris and S. C. Harvey, *J. Comput. Chem.* 20, 813 (1999).
210. W. R. Bauer and C. J. Benham, *J. Mol. Biol.* 234, 1184 (1993).
211. W. R. Bauer, H. Ohtsubo, E. Ohtsubo, and C. J. Benham, *J. Mol. Biol.* 253, 438 (1995).
212. M. Li and Z.-C. Ou-Yang, unpublished.
213. R. P. Feynman and A. R. Hibbs, "Quantum Mechanics and Path Integrals." McGraw Hill, New York, 1965.
214. F. W. Wiegel, "Introduction to Path-Integral Methods in Physics and Polymer Sciences." World Scientific, Singapore, 1986.
215. H. Kleinert, "Path Integrals in Quantum Mechanics, Statistics, and Polymer Physics." World Scientific, Singapore, 1990.

CHAPTER 10

Self-Consistent P3M Simulation of Ion Channels

Campbell Millar, Asen Asenov

Device Modelling Group, Department of Electronics and Electrical Engineering, University of Glasgow, Glasgow, Scotland

CONTENTS

ISBN: 1-58883-048-9

Handbook of Theoretical and Computational Nanotechnology
Edited by Michael Rieth and Wolfram Schommers
Volume 6: Pages (489–543)

1. THE BEGINNING

The biological importance of the class of proteins known as ion channels cannot be overstated. Ion channels are the route by which a cellular organism maintains the balance of electrolytes required for cell function and survival. To this date, no organism has been discovered that does not rely on channel proteins for its continued existence.

Ion channels are macromolecular pores embedded in cell membranes, which evolved over 3 billion years ago [1] as a method of transferring electrolytes through the newly evolved cell membrane (Fig. 1). Nature developed a staggering array of mechanisms to allow the new organisms to survive. As the lipid membrane is designed to permit the separation of a cell's internal environment from external conditions, the disposal of ionic wastes and the replacement of denuded resources promoted the development of a cross-membrane transport mechanism with which homeostasis could be maintained.

In the simplest sense, ion channels are proteins, spanning the membrane and containing small holes (~3–10 Å) that various substances can traverse. An example is shown in Fig. 2. However, in reality such a simple mechanism would make life untenable in anything other than the most benign of conditions and would certainly give up the energy efficiency that is a direct result of the evolution of the membrane. The simple gambit of providing a transmembrane conduit too small for the majority of solutes to enter would still allow the smaller ions, which are the basis of cell function, to diffuse freely. Therefore, the ionic channel evolved a host of properties. These amazing molecules can select a large ion over a smaller one and switch on and off in response to mechanical strain, voltages, chemical triggers, and pH changes.

The understanding of these molecules is vital to the medical and pharmaceutical industries because approximately 60% of current manufactured drugs act either directly or indirectly on channel proteins [2], and several diseases, designated channelopathies [3], are the direct result of changes in the properties of specific ion channels. In addition, more than 60% of genetic material codes are for channel proteins, making the study of these molecules important for the understanding of genetics.

As well as the obvious biological importance of these proteins, ion channels are becoming of interest to the engineering and nanotechnology communities. The ion conduction characteristics and specifically the current-voltage (IV) response of channels mimic the behavior of several types of semiconductor devices, most notably those of diodes and transistors. It has been recently shown [4–6] that ion channels, at least in large numbers, can be interfaced directly to semiconductors. This development leads toward the direct integration of biological components into electronic circuits and the ability to measure biological processes *in vivo*

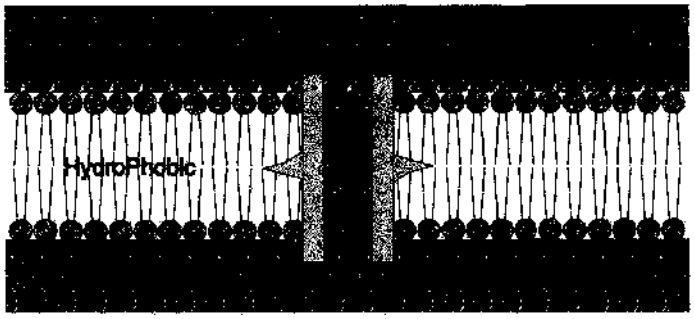


Figure 1. The lipid membrane. Phospholipids form a bilayer in which proteins can embed and bridge the ion-inaccessible membrane.

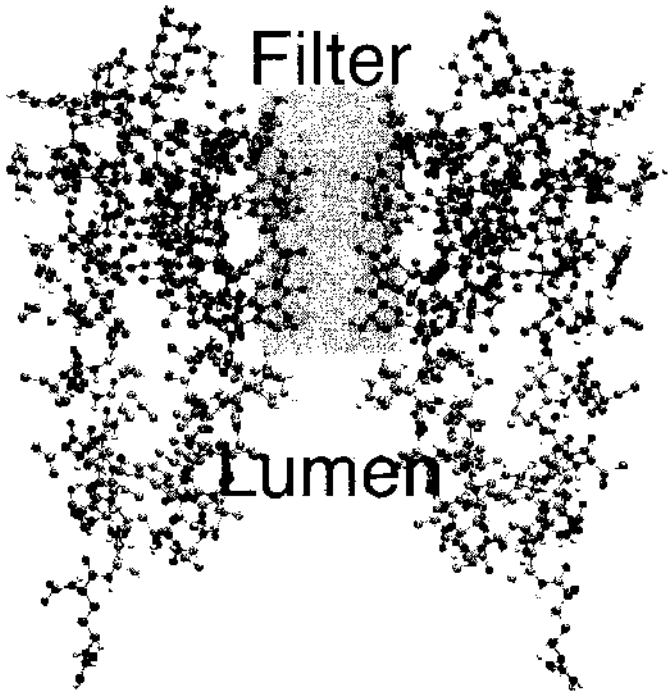


Figure 2. The structure of a typical channel protein. The pore of Kcsa is extremely small (~4 Å) and is lined with charged hydrophilic side chains. This molecule is split into two parts, a very narrow selectivity filter in the pore and a wider lumen region thought to be responsible for this channel's gating behavior.

through the connection of highly integrated or embedded sensor platforms coupled directly to cells.

Figure 3 shows a scale representation of a Kcsa ion channel affixed as the second gate in a 4-nm MOSFET structure [7]. As nanotechnology shrinks, the scales on which semiconductor devices are fabricated reach the dimensions of typical biological macromolecules such as DNA and ion channel proteins (typically below 10 nm). This creates the possibility of a direct nano–bio interface on a molecular level that would allow semiconductor components to be coupled to biological molecules in such a way that they could be used to directly measure signal propagation in neurons, to bridge the gaps left by trauma to nerve cells of many types, as sensors or in biocomputing.

Despite the immense interest by the scientific community in channel proteins, relatively little is known about the mechanisms by which they perform the various biological functions developed over the course of their evolution. For example, the Kcsa ion channel from *Streptomyces lividans* [8, 9] is capable of preferentially allowing potassium ions to traverse its pores while strongly rejecting other anions such as sodium, despite their smaller size [9–11]. Various mechanisms have been suggested to explain this behavior [1, 10, 12, 13], but the most plausible explanation is that there is a relation between the structure of the protein and the function of the selectivity filter of the channel. Unfortunately, the structures of only a handful of channel proteins [14] are known to some degree of accuracy at this time. To complicate matters, channel proteins are known to change configuration during events such as gating (when a channel opens and closes).

The majority of this demonstration will use Kcsa as a test for the simulation technology, which is a voltage-gated potassium channel from the bacillus *Streptomyces lividans* [10, 15]. Kcsa is four dual helical subunits arranged as a tetrameric structure where the transmembrane segments support the internal selectivity filter (see Fig. 2). The general structure of the protein consists of two regions: an extremely narrow pore/filter region on the intracellular side of the membrane and a larger lumen region formed by the α helical transmembrane segments. The lumen region was putatively considered to be the gate region of the channel [16, 17], an insight recently proven when an X-ray structure for the open state of the Kcsa protein was recently discovered [18].

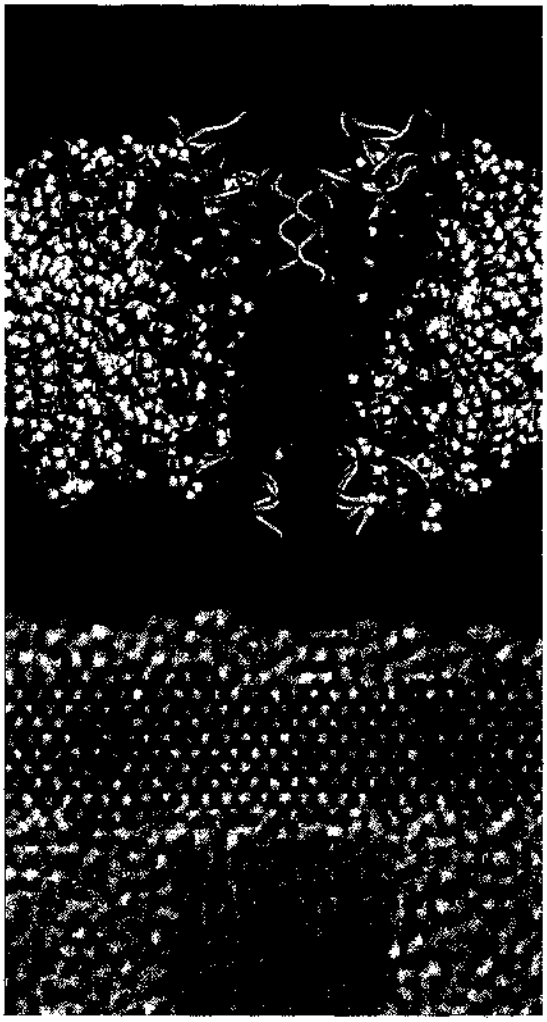


Figure 3. A scale representation of a Kcsa ion channel protein as the second gate in a 4-nm double gate MOSFET. The transistor source, gate, and drain are labeled.

Experimental data for ion conduction in Kcsa are limited when compared with other proteins such as gramicidin A; only two papers show the current-voltage relationship for the channel [9, 11]. However, the body of structural information available for Kcsa and the fact that it displays many of the important features of channels in general such as selectivity, gating, and current rectification make this molecule the ideal choice for modeling purposes.

2. SIMULATION TECHNIQUES

The field of ion channel simulation is still relatively young. The first attempts at simulating the conduction of ions through channel proteins were made in the late 1980s and were generally based upon the Poisson-Nernst-Planck theory (PNP), also known as drift diffusion (DD). This continuum approach has been successfully used for many years to stipulate the function of ion channel proteins from their structure [19–22]. Because of the increase in the computational power of modern computer systems, it is now possible to simulate the trajectories of individual ions and atoms in channel simulations, which allows a much more detailed study of ion conduction, albeit at a substantially increased computational overhead.

Currently, four main simulation approaches are used for the simulation of ion conduction in channel proteins: PNP, Brownian dynamics, particle Monte Carlo, and molecular dynamics. The use of equivalent circuit models of ion conduction paths [1, 23] has been popular in the past. However, because these models lack any predictive power and are phenomenologically fitted to experimental data, they have been superseded by the more advanced numerical

Table 1. A tabular comparison of the four main groups of ion channel simulation methods from the perspective of simulating the conducting behavior of channel proteins.

Particle based			
Continuum based			
Equivalent circuit	PNP	Langevin, particle-MC	MD
Extremely fast	Fast	Slower than PNP	Extremely demanding
Based on Ohm's law	Mesoscopic	Scales with population	Scales with population
Neglects nonlinearities	Steady state	Mesoscopic	Microscopic
Phenomenological	Solvent?	Steady state	Transient
	Monolithic implementation	Solvent?	Solvent!
		Modular implementation	Modular implementation
			Useful for IV?

Source: Reprinted with permission from [23], M. Saraniti (MCM 2003, Berlin).

simulation techniques. Table 1 shows a comparison [23] of the current simulation techniques used on ion channel problems and relates their relative strengths and weaknesses.

2.1. Poisson-Nernst-Planck

In this approach, the flow of ions is represented as a continuous charge density. The corresponding Nernst-Planck (or current continuity) equations are self-consistently coupled to the Poisson equation to give the flux through the simulation domain at equilibrium. The efficiency of PNP methods and the ease with which protein structures can be represented make it an attractive alternative to more complex simulation techniques. In addition, the availability of commercial technology computer-aided design (TCAD) packages that can efficiently solving the PNP equations means that channel simulation domains can be prototyped extremely quickly [24, 25].

There are some problems that have recently become apparent with the use of continuum methods for simulating what are, essentially, discrete charge transport problems:

- The PNP theory assumes that the whole of the simulation domain is in steady state, which may not necessarily be the case in extremely small pores, such as those found in Kcsa and gA [14], because the pore region can contain vacuum, vapor, or liquid water.
- The continuous nature of the PNP solution renders it incapable of modeling the direct ion–ion interactions that occur on such small scales and means that effects from Coulomb blockade and charge single filing, which may affect ion supply to channel pores and mediate permeation, cannot be investigated.
- This approach does not offer the possibility of investigating the effect of solvent on ion conduction. Because all pores contain water molecules, it may be necessary to model their effect on the ions traversing the pore [26].

2.2. Brownian Dynamics and Particle Monte Carlo

Brownian dynamics (BD) simulations, based on the Langevin equation, allow the modeling of ions in an atomistic fashion, making the resolution of inter-ion interactions possible. This technique has been used to model the transport of multiple ions through realistic protein structures and has proved extremely successful [12, 27–29] in its ability to resolve the interaction of ions with the protein. Brownian simulations require relatively few input parameters, usually only the diffusion coefficients of the ions and the structure and dielectric constant of the protein. The efficiency of BD methods and their atomistic nature make them an attractive alternative to continuum methods.

There are, however, issues with the most commonly reported BD simulations in that they are incapable of correctly modeling the electrochemical potentials that drive ions across the membrane, are applied by the electrodes in patch clamp experiments [1, 30] or arise from variations in the average ion concentration. The use of direct Coulomb interaction in these simulations provides excellent resolution of the inter-ion forces on short ranges but

typically does not include the long-range interactions that come from the potentials applied at contacts.

Once again, the issue of resolving the impact of solvent molecules is a problem. Because BD algorithms treat the ion interaction with the surrounding water molecules through a stochastic kick of a distribution determined by the ion's diffusion coefficient, the representation of individual water molecules in the pore presents a difficulty.

Recently, self-consistent Brownian dynamics simulations have been developed, which use the solution of the Poisson equation to deal with long range electrostatics [25, 31–33]. This removes the difficulty in treating electrochemical or applied potentials and allows the replication of the bulk properties of ionic solutions from an atomistic scale [25, 32].

Both self-consistent [34] and non-self-consistent particle Monte-Carlo [35] methods have been used to simulate the transport of ions through channels. These methods treat the interaction with the solvent through scattering rates, which probabilistically determine the free flight time of the ion before scattering by a collision with a water molecule. The ability to accurately simulate the motion of particles away from equilibrium makes these techniques an attractive alternative to BD. However, they require very small (fs) time steps, which introduces difficulties when simulating channel IV characteristics requiring the accumulation of statistics over hundreds of nanoseconds.

2.3. Molecular Dynamics

Molecular dynamics (MD) techniques have been extensively employed in the investigation of the structure and function relationship of ion channels. These methods treat all of the individual atoms in the simulation domain as finite sized particles moving in a Newtonian fashion. Electrostatic interactions between the particles are generally treated via pairwise Coulomb and Van Der Waals interactions, and methods such as Ewald summation are used [36] to treat long–range interactions. The fact that both water and proteins are included on the atomic scale means that MD simulations can model both the trajectories of ions in channels and the interactions with water molecules and the protein itself.

The fact that MD techniques follow the behavior of each atom in the simulation domain allows them to be used to model the behavior of the channel protein itself, a use for which there is much demand [17, 26, 29]. The fully atomistic approach gives MD the ability to make significant *ab initio* predictions of the functional properties, such as selectivity, of channel proteins [10, 36] and the conformational changes that govern effects such as gating [16]. However, this predictive power comes at a price.

The accurate treatment of the trajectories and interaction of many particles mean that MD simulations are *extremely* computationally expensive, with a practical limit on the number of atoms that can reasonably be simulated of approximately 50,000. In addition, the atomic resolution of the technique may require that MD use a subfemtosecond time step. These two factors mean that MD can reasonably simulate up to approximately 10 nanoseconds of real time [14]. Ten nanoseconds is a sufficiently long time to study the behavior of ions in small regions of the channel, such as the selectivity filter, and to recover parameters such as diffusion coefficients for the constrained ions, but is insufficient to reproduce the IV characteristics.

The treatment of long-range electrostatics, such as those from applied or electrochemical potentials, is problematic in MD simulations. This is partly due to the periodic boundary conditions that tend to be used in MD simulations to expand the effective size of the simulation domain [36]. The introduction of complex boundaries and electrodes will require large numbers of image charges to be taken into account, thus increasing the complexity of the simulator and the computational cost.

These factors combined mean that the results of MD simulations cannot be directly compared with experimental measurements, because experimental data typically have a minimum temporal resolution of around 10 μs. To accurately simulate the IV characteristics for a channel, multiple ions must cross the pore, so that a time-averaged current can be measured. MD simulations simply cannot be run for a sufficiently long time on current computing platforms to allow multiple ions to cross the membrane.

2.4. Summary

The previous discussion ultimately calls into question the viability of both molecular dynamics and continuum methods as tools for the investigation of the conducting behavior of channel proteins. This focuses the attention on the intermediate level of BD and MC simulation techniques. Both methods require parameterization to correctly model the behavior of ions in pores. Molecular dynamics studies can provide information about diffusion coefficients for ions in constrained pores, describe anisotropy in those coefficients, model the behavior of proteins, and describe solvation effects relevant to ionic mobilities. Therefore, MD simulations are going to be required for the correct parameterization of the coarser-grained BD and MC simulation methods, despite their inability to model IV characteristics.

Continuum methods are currently extremely useful for the prototyping of simulation domains before their simulation with more computationally expensive particle mesh techniques. Although current PNP and DD simulations are incapable of taking into account the atomistic effects that are an important feature of ion transport, it is possible that they may in turn be reparameterized with concentration-dependent mobility models, based, for example, on Brownian or Monte Carlo simulations. This could eventually allow continuum methods to accurately replicate channel IV curves.

3. BROWNIAN DYNAMICS

3.1. The Langevin Equation

The movements of small particles in a medium can be described by the theory of Brownian motion [37, 38]. Brownian motion, in the physical sense, is a statistical manifestation of the transfer of thermal kinetic energy between particles, and the medium that they occupy, through interparticle collisions [39]. The original formulation of the Langevin equation assumes that particles interact only with their surrounding medium and not with each other. In the simulation of an ionic solution, this is clearly not true because all of the relevant species are charged and hence interact through long-range Coulombic interactions. It is therefore necessary to modify the Langevin equation in order to introduce interparticle electrostatic interactions. In many cases, an externally applied electric field also acts on the ions and has to be take into account to correctly simulate the behavior of ions.

A modified Langevin equation proposed in Ref. [40] combines the original description of thermal motion with the effect of Coulombic interactions between charged ions and longer-range electric fields from the environment or externally applied potentials.

$$m\frac{d\mathbf{v}(t)}{dt} = q\mathbf{E}(\mathbf{r}, t) - \frac{m\mathbf{v}(t)}{\tau} + \mathbf{F}(t) \tag{1}$$

where m is the mass of the particle, $\mathbf{v}$ is the particle's velocity, $\mathbf{E}$ is the electric field, q is the charge on the particle, τ is the momentum relaxation, or free flight, time, and $\mathbf{F}(t)$ is a rapidly fluctuating random force. This extension of the Langevin equation has the advantage of describing the interactions of a large number of ions and to define a particle drift in terms of the experimentally measured mobility of the particle. The right hand side of this equation can be viewed as three distinct parts:

- The first term $q\mathbf{E}(\mathbf{r}, t)$ is the force applied to the ion by the electric field. This field has two components, the first related to all other ions in the simulation domain and the second related to the externally applied potentials in the dielectric environment of the simulation. In the P3M solution of the Poisson equation, the two components are superimposed in the solution of the electrostatic potential. Short-range force corrections or very high resolution in the Poisson solution is necessary to accurately resolve the interparticle interactions.
- The second term $[m\mathbf{v}(t)]/\tau$ is a velocity-dependent frictional force that models the dissipative effects of physical interactions between the ion and the surrounding water molecules. The momentum relaxation time (or free flight time) τ can be obtained from

the mobility μ and is defined as the time for which the average velocity takes to elax to $1/e$ of the equilibrium velocity, after a step switching of the electric field.

- The final term, $\mathbf{F}(t)$, is a rapidly fluctuating force representing the stochastic kick the ion receives due to collisions with the water molecules in the solution. This function is usually represented as delta-correlated white noise with a mean value of zero, where the magnitude of the fluctuating force is determined from the thermal energy c the solvent via the equipartition theorem. The delicate balance between the frictiona and rapidly fluctuating force causes the particle velocity to stabilize as a Boltzmann veocity distribution of the correct shape at thermal equilibrium.

Equation (1) can be solved in a constant electric field to give the time dependene of the velocity. The ensemble average of Eq. (1) eliminates the fluctuating force $\mathbf{F}(t)$, which averages to zero. The resulting equation is trivial to solve for a constant field E applied at $t = 0$ to give

$$\langle \mathbf{v}(t) \rangle = \frac{q\tau}{m}\mathbf{E} + \left(\langle \mathbf{v}_0 \rangle - \frac{q\tau}{m}\mathbf{E} \right) e^{-\frac{t}{\tau}} \tag{2}$$

where $\langle \mathbf{v}_0 \rangle$ is the average initial particle velocity. This time dependence shows that the average velocity will relax to the correct equilibrium value of $(q\tau/m)\mathbf{E}$ after an initial tranient. The mobility μ is simply a constant of proportionality between the average particle velocity and the electric field in equilibrium [41] and is defined as

$$\mu \equiv q\tau/m \tag{3}$$

with units of $(m^2V^{-1}s^{-1})$, where q is the charge on the particle, and m is the mass. Thus, Eq. (1) reproduces the microscopic formulation of Ohm's law. However, because the value of τ is almost impossible to define for any real ionic solution, the mobility is usudly a phenomenological parameter that is determined experimentally. Alternatively, the mobility can be calculated from the particle's diffusion coefficient via the Einstein relation

$$D = |k_b T/q|\mu \tag{4}$$

where k_b is the Boltzmann constant and T is the temperature of the simulation. A discrete time approximation of Eq. (1) is employed for the integration of the equations of motion of the individual ions in our P3M simulations. The implementation of a simple Euler integration over a short time interval Δt gives

$$\mathbf{v}(t + \Delta t) = \left[\frac{q\mathbf{E}(t)}{m} - \frac{\mathbf{v}(t)}{\tau} \right] \Delta t + \Delta \mathbf{V}(t) \tag{5}$$

where

$$\Delta \mathbf{V}(t) = \frac{1}{m} \int_t^{t+\Delta t} \mathbf{F}(t')dt' \tag{6}$$

Because the random force is considered to be rapidly varying on any time scale, it requires different treatment from the other terms. Simple Euler integration is not applicable because it does not approach linear limits when $\Delta t \to 0$. The result $\Delta \mathbf{V}(t)$ from the integation function is delta-correlated Wiener process and, like $\Delta \mathbf{F}(t)$, averages to zero over a long enough period of time.

The statistical properties of integral $\Delta \mathbf{V}(t)$ are obtained from the central limit theorem, which states that the integral of any sufficiently large number of uncorrelated random events will have a Gaussian distribution and can be viewed as a random walk in velocity space The variance of the Gaussian distribution of $\Delta \mathbf{V}(t)$ is related to the particle diffusion coefficient in velocity space via the equivalent of the diffusion length.

$$\langle (\Delta \mathbf{V}(t))^2 \rangle = 2D_v \Delta t = 2\frac{k_b T \Delta t}{m\tau} \tag{7}$$

The random walk in velocity space and the associated diffusion coefficient must be related to τ on very short time scales according to the fluctuation dissipation theorem because there is correlation between the stochastic term and the frictional term [39, 42]. It can be shown [39, 40, 43] that the diffusion coefficient in velocity space is related to the diffusion coefficient in real space D_x, which can be experimentally determined, through the relationship.

$$D_x = D_v \tau^2 \tag{8}$$

By introducing a discrete time $t_n = n\Delta t$, the integral form of the Langevin Eq. (5) can be rewritten as a finite difference equation in terms of $\mathbf{v}_n = \mathbf{v}(n\Delta t)$ to give

$$\mathbf{v}_{n+1} = \gamma \mathbf{v}_n + \frac{q}{m}\mathbf{E}_n \Delta t + \Delta \mathbf{V} \tag{9}$$

where $\mathbf{E}_n = \mathbf{E}(n\Delta t)$ is the electric field at the particle's position at time $n\Delta t$; γ can be interpreted as a discrete time equivalent of the frictional coefficient related to the current particle velocity and obtained via simple Euler integration.

$$\gamma = \left(1 - \frac{\Delta t}{\tau}\right) \tag{10}$$

For a constant field, the discrete time approximation of the Langevin equation can be solved to give the velocity of the particle at a given time step n to give

$$\langle \mathbf{v}_n \rangle = \mu \mathbf{E} + (\mathbf{v}_0 - \mu \mathbf{E})\gamma^n \tag{11}$$

It is obvious from Eq. (11) that the velocity will relax to its correct value only if the time step Δt is less than the momentum relaxation time τ. The similarity with Eq. (2) is clear, and this discrete time equation nicely approximates the time dependence of the transient portion of the particle acceleration in that for $\Delta t/\tau < 1$, $\gamma^n \to \exp[-(\Delta t)/\tau]$ as $n \to \infty$, therefore $\langle \mathbf{v}_n \rangle$ will tend toward the correct value $\mu \mathbf{E}$ as time progresses.

It is possible to improve the accuracy of the discrete time approximation. By using the average velocity $[(\mathbf{v}_n + \mathbf{v}_{n+1})/2]$ of the particle over Δt in the frictional term in Eq. (5) to give

$$\beta \mathbf{v}_n = \gamma \mathbf{v}_n + \frac{q}{m}\mathbf{E}_n \Delta t + \Delta \mathbf{V}_n \tag{12}$$

where

$$\beta = 1 + \frac{\Delta t}{2\tau} \tag{13}$$

this leads to a replacement for the frictional coefficient γ with γ'

$$\gamma' = \frac{1 - (\Delta t/2\tau)}{1 + (\Delta t/2\tau)} \tag{14}$$

which gives a better approximation to the exponential term in Eq. (2) with half of the error due to discretization.

It is clear that on time scales where $t \gg \tau$ the velocity will have relaxed to $\mu \mathbf{E}$, that is, the equilibrium velocity for the particle in the local electric field. If we are not concerned with the detailed dynamics of the ions, the discrete Langevin equation can simplified, removing dependence on τ, leading to a classical random walk in position

$$\Delta \mathbf{x} = \mu \mathbf{E} \Delta t + \Delta \mathbf{X} \tag{15}$$

where $\Delta \mathbf{X}$ is a Gaussian distributed Wiener process with a variance given by Eq. (16), and D_x is the diffusion coefficient of the particle in real space and has the units $m^2 s^{-1}$.

$$\langle (\Delta \mathbf{X}(t))^2 \rangle = \sqrt{2D_x \Delta t} = 2\frac{k_b T \tau \Delta t}{m} \tag{16}$$

This approximation is now identical to the drift diffusion equation but for discrete particles, where the variance of the Wiener process ($\Delta \mathbf{X}$) is the diffusion length, or mean squared displacement, of the particle in the time Δt.

When we introduce a discrete time $t = n\Delta t$, this random walk in turn leads to a discrete form of the Langevin equation that can be easily simulated.

$$x_{t-1} = x_t + \mu \mathbf{E}(t)\Delta t + \Delta \mathbf{X} \tag{17}$$

Equation (17) is valid on any time scale, provided that $\Delta \mathbf{X}$ is smaller than the spatial resolution of the electric field. This restriction is related to the coupling between the motion of the ion and the solution of the self-consistent electric field, which is discussed in the next section. Although in this formulation all information on the details of the particle's kinetic energy and therefore energy distribution has been lost, the liberty in choosing a larger Δt proves extremely useful in the simulation of long time scale properties (such as IV characteristics). The diffusion coefficient (D_x) is widely reported in literature and tables of physical constants.

In addition, Δt should not approach the momentum relaxation time. Because τ can be viewed as the time taken for the particle velocity to get to $1 - (1/e)$ of the equilibrium velocity, this means that after one τ the particle should have a velocity of only $0.632 \times \mu \mathbf{E}$. It is clear that a choice of time step closer to 5τ is required to achieve a velocity of >99% of $\mu \mathbf{E}$.

3.2. Ion Dynamics

When a comparison is made with the classical version of the Langevin equation as used in Einstein's seminal work [37], there is a marked difference between the methods of calculating τ. It is interesting to point out that Einstein's original formalism uses Stokes's Law (Eq. 18) to calculate the relaxation time of a particle,

$$\tau = m/6\pi a\eta \tag{18}$$

where a is the particle's radius, and η is the viscosity of the suspending medium.

The original work by Einstein was developed to explain the motion observed by Brown [38] and justified by comparison with experiments on the diffusion of sugars in solution, based on the assumption of a spherical particle's motion through a continuous enveloping medium. Einstein also proposed his methodology as a means of calculating the size of small particles in suspension. Table 2 shows the radius of a potassium ion as calculated from the Stokes's Law formalism for a value of $\eta = 0.891 \times 10^{-3}$ Ns/m^2. Current thinking indicates

Table 2. Momentum relaxation times for potassium ions calculated from Eq. (3) and the effective radius of the potassium ion from Stokes's Law/Langevin equation, using the experimentally determined mobility of potassium of 7.62×10^{-8} m^{2}V^{-1}s^{-1}. The mass of the ion water complex is calculated by simply adding the mass of the water molecules (18 amu) to the mass of a bare potassium ion (39 amu); once the relaxation time has been calculated, this can then be used in Eq. (18), along with the simple combined mass to calculate the effective radius of the complex. Although this is not a rigorous method of determining the radius of a small particle, it neatly illustrates some of the problems encountered in calculating physical parameters of ions in solution.

Number of water mols	τ (s)	Ionic radius (Å)
0	30.8×10^{-15}	0.65
2	50.93×10^{-15}	1.25
4	87.8×10^{-15}	1.85
6	0.116×10^{-12}	2.45
8	0.144×10^{-12}	3.05
10	0.172×10^{-12}	3.65
12	0.200×10^{-12}	4.25
14	0.229×10^{-12}	4.85
16	0.257×10^{-12}	5.45

the size of the hydration shell of a potassium ion to be somewhere between 6 and 12 water molecules per ion [1, 39, 44–47], which is reasonable from a thermodynamic point of view. However, the fact that the hydration shell has two distinct layers, an inner sphere of highly oriented water and a more diffuse outer layer, one of which is susceptible to external friction and local field effects, makes the concept of a hydrated radius difficult to quantify. It is clear that an ion with a hydrated radius of ~8 times its ionic radius, when 16 water molecules are associated with it (considering the radius of a water molecule to be ~1.2 Å), is incorrect.

Table 2 shows the values of τ for K^+ ions with varying numbers of bound water molecules as calculated from Eq. (3), which represents the Drude model originally used in metals [48] and relates the relaxation time to the mobility and mass of the ion.

With all this in mind, the question of independently estimating the mass and relaxation time of diffusing ions remains difficult. This creates problems when both these quantities are needed in the discrete time approximation of the Langevin equation (Fig. 4), Eqs. (5) or (9). In order to reproduce Ohm's Law, the ratio of τ/m must be fixed by Eq. (4). However, the integral form of the Langevin equation [Eq. (15)] does not require these parameters to be explicitly defined, thus removing these problems.

3.3. Practical Implementation

Despite the simplicity and efficiency of the Brownian methods presented in the previous section, the implementation of these algorithms for the simulation of ions in solution is complicated. Great care must be taken to ensure the efficacy of the various portions of the algorithms because they have proven rather sensitive to statistical errors, especially those introduced through poor random-number generation. Therefore, before the added complication of the self-consistent solution, including long-range electric fields or short-range interparticle interactions, are added to simulations, it is important to check that the numerical implementation of the methodologies reproduces properly the expected spatial evolution of a diffusing ensemble of particles, the drift in an electric field, and the corresponding energy distribution.

Based on Eq. (5), it is relatively simple to implement the full Langevin equation computationally, provided some restrictions are satisfied. As mentioned before, the time step Δt must be less than the momentum relaxation time/free flight time τ, and the particle must not be allowed, in a single time step, to move more than the spatial resolution of the electric field, as determined by the discretization of the Poisson equation. Also, Δt should not be smaller

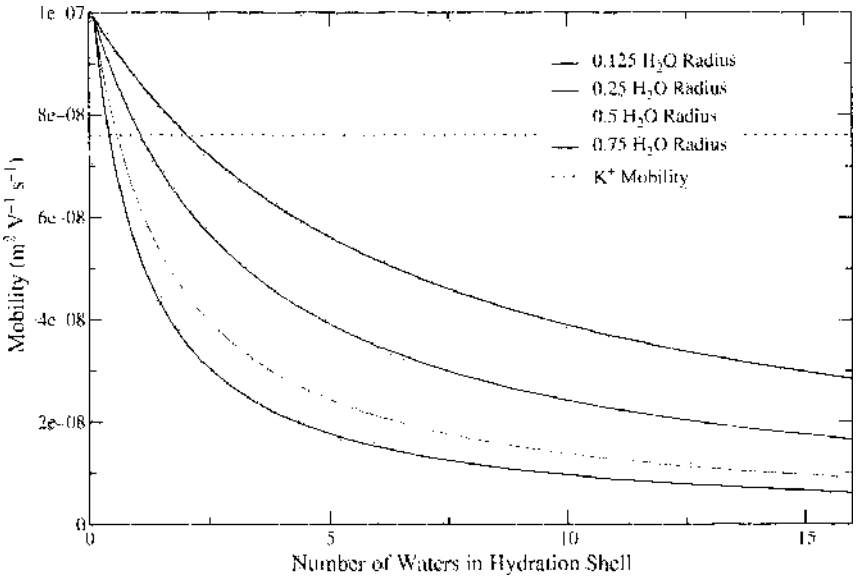


Figure 4. The mobility of K^+ ions calculated from Stokes's law friction in the Langevin equation. The dotted line shows the experimentally determined mobility of K^+. The curves are plotted with various proportions of the H_2O molecule's radius added to the crystal radius of the potassium ion. This shows how poor an approximation the continuum frictional model is when applied to very small ions.

than the typical collision time between an ion and the solvent molecules, which would render the usage of Brownian dynamics questionable.

Figure 5 shows a plot of the velocity distribution obtained from the 3D implementation of Eq. (5) under a constant applied field in the x direction. The particles were allowed to relax to steady-state conditions before the velocity distribution was extracted. The distribution of the particles is in accordance with the Maxwell-Boltzmann distribution and is shifted in the direction of the applied field by exactly the energy that the particles acquire from the field.

The implementation of Eq. (17) must also be tested. Because it is a discrete analogue of the drift diffusion equation, the code was tested against a 1D diffusion problem. It is possible to derive an expression for the time dependence of a concentration profile from an initial delta function

$$C(x, 0) = S\delta(0) \tag{19}$$

which gives

$$C(x, t) = \frac{S}{\sqrt{\pi Dt}} \exp\left(\frac{-x^2}{4Dt}\right) \tag{20}$$

where S is the number of particles in the simulation.

Figure 6 shows results of the diffusion simulation and proves that the diffusive behavior of the particles exactly reproduces the theoretical spatial distribution given by Eq. (20) over time.

As an additional check, in Fig. 7, we also compare the evolution of the standard deviation of the concentration profile calculated both analytically and numerically, which are in excellent agreement.

In the practical implementation of Brownian algorithms, efficiency is one of the major benefits compared with methods such as Monte Carlo. However, in Eqs. (9) and (17), it is easy to see that there are very few calculations required to perform a basic simulation when one removes the random component of the motion. The choice of random number generator in a Brownian framework therefore determines, almost exclusively, the computational effort involved in tracing the motion of the particles.

There are several available pseudorandom number generator algorithms [49] in common use, and their behavior has been comprehensively studied and characterized [50]. However, the variation in the computational effort required and the quality of the random numbers

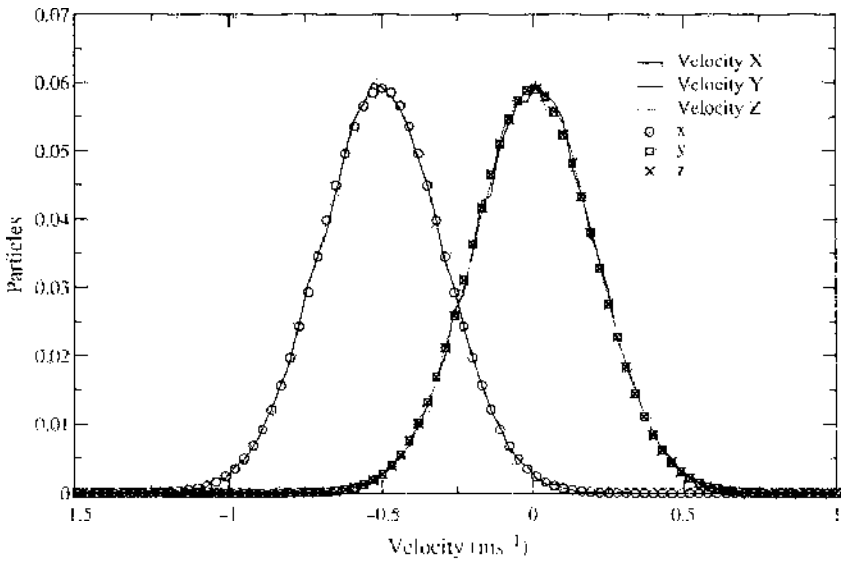


Figure 5. The distribution of Brownian particles in 3D under the influence of a 0.5 Vm^{-1} electric field applied in the X-direction. The figure shows the drift from a constant electric field compared with Maxwell-Boltzmann distributions shifted to the expected velocities (x, y, z, symbol legends). In this example, the particle's mobility was an arbitrary value of 1 $m^2V^{-1}s^{-1}$ for clarity. Note that the particle distribution is shifted in the X direction but not in Y or Z and that the distribution of velocities around the average does not change with drift.

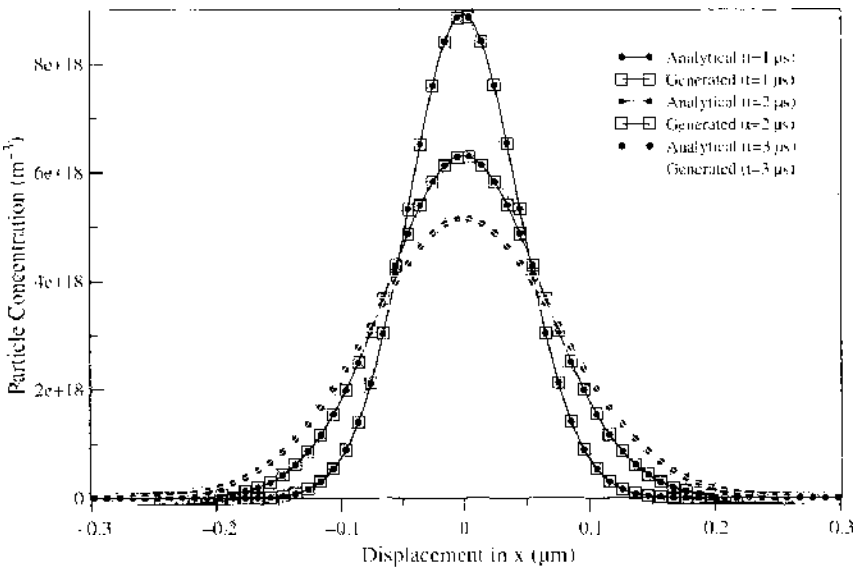


Figure 6. Particle concentration profile with time. From an initial delta distributed concentration at 0 using a particle diffusion coefficient of $D = 1 \times 10^{-5}$ m^{2}s^{-1}, this test was performed using 10,000 particles in 3D. The concentration profile is shown only in the X direction. Note that the Brownian code almost exactly reproduces the expected analytical result.

produced by the various techniques mean that they are not always applicable for Brownian dynamics simulations.

The ability to produce pseudorandom number sequences with good zero means is essential to the success of any stochastic simulation. One of the most commonly used random number generators is the ran2 generator from the "Numerical Recipes" [49] series, which is based around the linear congruential methods developed many years ago. Although this algorithm seems attractive to use in a Brownian simulation, it can be shown that the essence of the linear congruential generation method can introduce correlations when the numbers used are mapped into multiple dimensions [50]. Although this is unlikely to be a problem in the case of the full Langevin methodology, it can introduce unusual effects in the random walk formalism that can be extremely difficult to quantify. Therefore, it is sensible to give up some of the computational efficiency of the simpler random number generators in favor of more robust algorithms that carry a higher computational load.[1]

The simulations detailed here use the ran2 and ran4 uniform random number generators from Ref. [49]. The choice of generator is dependent on the available computational resources.

The period of a random number generator can cause some confusion because it is usually assumed to be the time at which the pseudorandom sequence "wraps" or begins to generate the same sequence once again. This is not necessarily the case, as often the period is meant as the minimum number of values that must be generated from the sequence before correlation can be seen between them. Therefore, it is important to choose a random number generator with a long enough period; although the random sequence may continue to be different, the behavior of the system in which the numbers are used may become correlated on long time scales. A good rule of thumb is to not attempt to use more than 5% of the period of the generator in a single simulation.

The second generator used in this work is the ran4 generator, again from Ref. [49]. It is a more complex generator that is based around pseudo-DES encryption and hashing of large integers. Although this generator produces *extremely* good random deviates, when compared with the expected uniform distribution, it is substantially slower than simpler generation

[1] The standard random number generators supplied with languages, such as the C/C++ and Fortran 90 rand routines, should be avoided since the language standards simply state that a random number generator should be included with no minimum specification for the algorithm quality. This introduces a platform dependence in the simulator that may cause incorrect simulation results due to poor random generation.

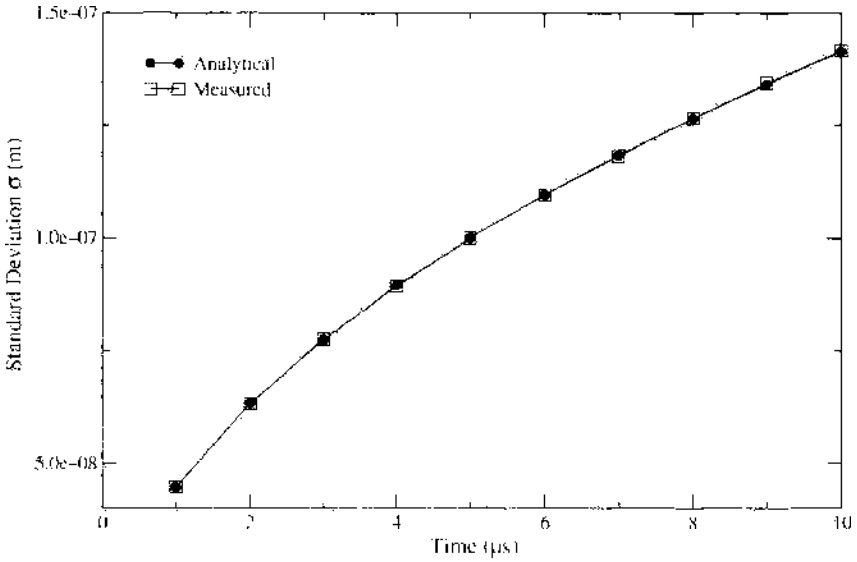


Figure 7. The standard deviation σ of particle positions in Brownian diffusion with time, compared with the expected values calculated from Eq. (20).

algorithms (about half the speed of the ran2 generator). The choice of algorithm is a trade-off between quality and speed. The simple test simulations can be carried out on relatively low-powered workstations by using ran2. However, in the latter demonstrations, the vast majority of the computational effort is expended elsewhere in the code, and since these simulations require a powerful computer server, the use of the ran4 generator is acceptable.

Once uniformly distributed random numbers have been generated, it is necessary to transform the distribution of the deviates to a Gaussian distribution of the variance appropriate for generation of $\Delta\mathbf{V}$ and $\Delta\mathbf{X}$. Gaussian distributions are generated by using the Box-Muller transformation to change distribution of the uniform random deviates to a Gaussian. A multiplication factor is added to the original implementation from Ref. [49] to change the standard deviation of the distribution from 1 to the required σ. Although this method generates nicely distributed numbers, it can be somewhat inefficient because it is based on a rejection technique.

The electric field causes the drift of the particles in the Langevin equation. However, no details have, so far, been given of how the field can be calculated in a real system. In the examples in this section, only a constant electric field has been considered, and the interactions between the charged particles has been ignored. In the case of Brownian simulations of ionic channels, the correct treatment of both the long- and short-range electric fields has to be included in a self-consistent manner. It is likely, for example, that electrostatic interactions between the pore walls and ions play a major role in the important phenomena of selectivity and rectification. The correct treatment of the electric field requires the solution of the Poisson equation.

In that the pores of some ion channels may be as small as 3–4 Å [17, 18, 51], the combination of factors may present a problem. To be able to recover transport characteristics such as IV curves, the slow transport time of ions across membranes (≈nanoseconds) and the very short values for τ shown in Table 2 make clear that the Poisson equation will have to be solved extremely efficiently if self-consistent Brownian simulations are to be used to study the transport of ions through channel proteins.

4. THE SOLUTION OF THE POISSON EQUATION

In the particle-particle-particle-mesh (P3M) approach, electrostatic interactions are treated via the 3D solution of the Poisson equation in a domain that contains the particles, electrical contacts, and regions of varying dielectrics representing the lipid, protein, and water. However, there are considerable difficulties associated with solving the Poisson equation in

complex simulation domains when the dielectric constant of materials rapidly changes (such as at the water–lipid interface) and where the driving field from an applied voltage has to be taken into account through appropriate boundary conditions. Furthermore, it becomes necessary to solve the Poisson equation frequently, in such a way as to remain self-consistent with the motion of the charged particles present in the simulation. Because the Poisson equation is impossible to solve analytically, in anything other than very simple cases a numerical approach has to be adopted.

Mesh-based Poisson solvers have been heavily used for many years for the simulation of systems under quasi-electrostatic conditions. The corresponding discretization and numerical solution of both the linear and nonlinear Poisson equations have long been of great importance and interest to scientists in many fields. The distinct advantages of the mesh-based solution of the Poisson equation are that the long-range electric field and local variations due to changes in dielectric are easily taken care of via boundary conditions and fine-grained discretization. However, the inherent band limits imposed by the discrete sampling of a continuous problem mean that the resolution of the rapidly varying short-range Coulomb forces may become poor unless extremely fine meshes (on the order of Å) are used.

In most cases where particle simulations are being performed, the solution of the Poisson equation constitutes the bulk of the computational effort associated with the simulation. In 3D simulations this can easily account for more than 98% of the total execution time. Therefore, it is highly desirable that an efficient numerical method is used.

In the following section, the design and implementation of a highly efficient multigrid Poisson solver is discussed. Of the currently available solution methods, multigrid has proven to be one of the most robust and efficient algorithms.

4.1. The Poisson Equation

The purpose of solving the Poisson equation is to ensure the correct coupling of individual particle motion to the long-range forces from applied potentials and the distribution of the remaining particles. Therefore, for a self-consistent simulation, the charge from the distribution of particles contained in the simulation domain must be included in the final solution of the Poisson equation on each time step. Then the positions and velocities of all particles can be updated, based on the forces calculated from the resulting potential distribution.

At first sight, it may appear simpler to solve the Poisson equation only once for the potential distribution arising from applied voltages and summing the interparticle Coulomb forces for the contribution of the remaining particles.

$$\hat{\mathbf{F}}_{(1,2)} = \frac{q_1 q_2 \hat{\mathbf{r}}_{(1,2)}}{4\pi\epsilon_0\epsilon_r r^3} \tag{71}$$

However, this approach has difficulties associated with the calculation of the Coulomb forces through a series of materials with different dielectrics. The calculation of the interparticle forces requires an operation count of at least $N(N-1)\alpha$, where α is the number of operations required to resolve a single Coulomb interaction, and N is the number of particles in the simulation. In addition, a serious drawback of direct force evaluation is that they may require the summation of infinite numbers of image charges at solution domain boundaries, massively increasing the computation time. This problem is dealt with naturally in Poisson solution via simple Neumann boundaries. Although, in principle, the direct Coulomb method can be used, provided the number of particles is small [52]. However, the unavoidable overhead from evaluating the Coulomb forces in ion channel simulations is increased because of the inclusion of different materials. The fact that ϵ_r becomes spatially dependent drastically increases the computational effort required to evaluate the force between two particles, because a geometric description of the changing values in the dielectric constant must be obtained from a structural representation for every particle pair. However, it may still be desirable to use Coulomb interactions in small simulations because they are not prone to the problems with short-range and self-forces that can occur in mesh-based solutions.

Particle-mesh (PM) simulation schemes [40, 52–54] have a distinct advantage over direct force evaluation methods in that they substantially reduce the computational overhead involved in the self-consistent simulation of particle systems. The PM concept is based on the estimation of the local electric field from the global numerical solution of the discretized Poisson equation. In the ideal case, the computational overhead should drop from $N(N-1)$ to $N_g \log(N_g) + N\beta$ where N_g is the number of grid points in the discretization and β is the overhead associated with the manipulation of the particles.

The self-consistent PM simulation consists of three distinct steps:

1. Assign the charges to the mesh by using their position and a weighting function.
2. Solve the Poisson equation using the method of choice.
3. Calculate the $\mathbf{E}_{(x,y,z)}$ for each particle and move it accordingly.

Most current charged particle transport simulation methods depend on the numerical solution of the linear Poisson equation.

$$\nabla \cdot (\varepsilon_0 \varepsilon_r \nabla \Psi) = -\rho \tag{22}$$

where ψ is the electrostatic potential, ρ is the charge density, ε_r is the relative dielectric permittivity of the material, and ε_0 is the permittivity of free space. Because Eq. (22) cannot be solved analytically for any but the simplest of test cases, it must be solved numerically. Such solution will enable the coupling of particle motion/charge density to the electric field in a fully self-consistent fashion, with the motion and changing distribution of carriers in a discrete time particle simulation. Equation (23) shows explicitly the partial differentials in the Poisson equation in a form that becomes the starting point for the discretization and numerical solution of Eq. (22).

$$\frac{\partial}{\partial x}\left(\varepsilon_{(x,y,z)}\frac{\partial \Psi_{(x,y,z)}}{\partial x}\right) + \frac{\partial}{\partial y}\left(\varepsilon_{(x,y,z)}\frac{\partial \Psi_{(x,y,z)}}{\partial y}\right) + \frac{\partial}{\partial z}\left(\varepsilon_{(x,y,z)}\frac{\partial \Psi_{(x,y,z)}}{\partial z}\right) = -\rho_{(x,y,z)} \tag{23}$$

Note that for simplicity $\varepsilon_0\varepsilon_r$ has been replaced with the single value $\varepsilon_{(x,y,z)}$ to denote the value of the dielectric constant at the point (x, y, z). This notation is used from here on. To solve Eq. (23) computationally, a discretization has to be performed. Both finite element and finite difference techniques can be used for the discretization of the Poisson equation for the purposes of particle mesh simulation. Finite difference discretization on a Cartesian grid is the method of choice because this greatly simplifies the numerical solution of the problem.

By introducing a discrete grid in 3D space represented by a mesh spacing h, Eq. (23) can be discretized by using simple box integration, which is a modification of the approach based upon the integral form of the Poisson equation found in Gauss's law.

$$\oint \varepsilon_{(x,y,z)} E \cdot ds = \rho_{(x,y,z)} \tag{24}$$

Figure 8 illustrates the implementation of the box integration scheme, the nomenclature of the node indexes, and the assignment of the dielectric constant. The dielectric is assumed to be constant over the distance from a mesh point to its neighbor, where the dielectric of a cell volume is determined by a reference point at the bottom left of a cell, and the discrete Laplacian coefficients are calculated by integrating the flux between points, through the plane of the dielectric boundary, using Gauss's law.

The equation arising from the discretization of the Poisson equation for a point x, y, z is shown next.

$$\begin{aligned} -\rho_{(x,y,z)}\, hx\, hy\, hz = {} & \frac{\left(\varepsilon_{(x-1,y,z)}\Psi_{(x-1,y,z)} - 2\varepsilon_{(x,y,z)}\Psi_{(x,y,z)} + \varepsilon_{(x+1,y,z)}\Psi_{(x+1,y,z)}\right)}{hx} hx\, hy \\ & + \frac{\left(\varepsilon_{(x,y-1,z)}\Psi_{(x,y-1,z)} - 2\varepsilon_{(x,y,z)}\Psi_{(x,y,z)} + \varepsilon_{(x,y+1,z)}\Psi_{(x,y+1,z)}\right)}{hy} hx\, hz \\ & + \frac{\left(\varepsilon_{(x,y,z-1)}\Psi_{(x,y,z-1)} - 2\varepsilon_{(x,y,z)}\Psi_{(x,y,z)} + \varepsilon_{(x,y,z+1)}\Psi_{(x,y,z+1)}\right)}{hz} hx\, hy \end{aligned} \tag{25}$$

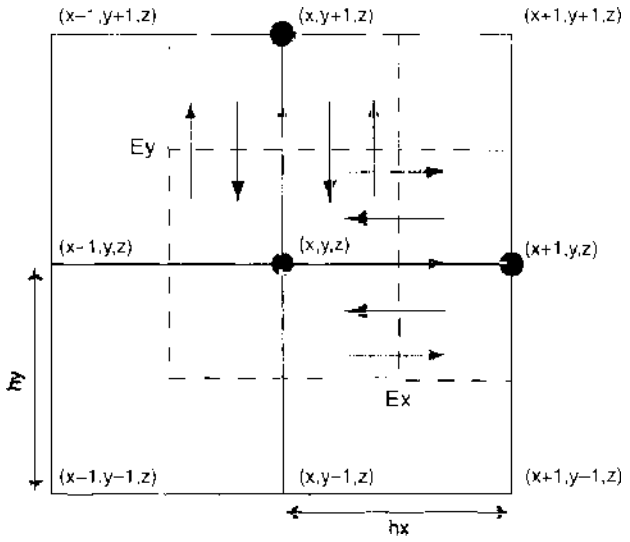


Figure 8. The Gaussian box integration discretization used in the calculation of the Laplacian. The dielectric for a cell containing the point x, y, z is assumed to be constant over the range $\pm h/2$ when calculating the flux to and from that point. For simplicity, a slice in the x, y plane is shown.

where $\rho_{(x,y,z)}$ is the charge density at the volume surrounding the node x, y, z used in the Gauss integration.

The equations for each node of the grid can be assembled as a system of simultaneous algebraic equations that can be combined and written in matrix form as

$$\mathbf{Ax} = \mathbf{b} \tag{26}$$

where $\mathbf{x}$ is the unknown solution vector of size $l \times m \times n$ representing the potential Ψ, $\mathbf{b}$ is a vector of size $l \times m \times n$ containing the values of the charge density $\rho_{(x,y,z)}$, and $\mathbf{A}$ becomes a matrix of size $(l \times m \times n)^2$ containing the discretization coefficients calculated from Eq. (25).

The matrix $\mathbf{A}$ is a sparse, seven-diagonal matrix, which is convenient, considering the size of the full matrix. This is particularly a problem in Poisson solution methods using matrix inversion because the inverse of $\mathbf{A}$ is not sparse. For example, the storage of the inversion of the discretization matrix of a $50 \times 50 \times 50$ Poisson solution requires ~116 Gb of memory.[2] Such large memory requirements are the primary reason that matrix inversion methods tend not to be used in 3D simulations; the efficient implementation of such a method requires a simulation platform that is capable of holding the entirety of the inverted matrix in main memory and cache memory. This enormous consumption of system resources means that, although matrix inversions are extremely fast (once the initial inversion has been calculated) and highly parallelizable, they are a relatively rarely used method in simulations where the resolution of the Poisson equation requires a discretization of a large number of points in each spatial direction.

It is an unfortunate consequence of using a discrete mesh to represent the electrostatic potential solution that some inaccuracy is introduced on account of band limiting in the solution from the discrete sampling. The Nyquist theorem limits the wavelength component of any of the fluctuations in the solution on a mesh of spacing h to $2h$. Because this sets the high-frequency limit of the components of the potential, it can be a particular problem when the resolution of short-range forces between two interacting particles becomes important, on account of the large number of high frequencies in the $1/r$ Coulomb interaction at short range. In effect, the sampling of the mesh has the property of smearing the charge from a point particle to a cloud of radius $h/2$. This can be unacceptable in some cases, especially where there are only a small number of particles in the simulation domain or the limit of one particle per mesh cell is approached.

[2] Based on an 8-byte double value to give $(50 \times 50 \times 50)^2 \times 8 = 116.4153$ Gb.

In the previous section, some of the limitations imposed by the use of the Brownian dynamics algorithm were mentioned, most notably that a particle must not be allowed to move more than the spatial resolution of the electric field in a given time step. This restriction has a direct bearing on the construction of the Poisson mesh topology. A fine grid is necessary to resolve the short-range forces while keeping the size of the mesh and the time step in the Brownian algorithm to an optimum. As we will see later, the solution of the Poisson equation entails the vast majority of the computational time expended in a Brownian simulation. Therefore, care must be taken not to overrefine the mesh and dramatically increase the computational effort required to solve the potential to a given error. Overrefinement of the mesh also necessitates the reduction of the Brownian time step because of the restrictions placed on the particle engine by the electric fields, resulting in more iterations required to simulate biologically relevant times.

There is one more factor that must be taken into account in the construction of the Poisson mesh. The Debye length determines the screening length for an environment with a given amount of free charge.

$$l_d = \sqrt{\frac{\epsilon_0 \epsilon_r k_b T}{(zq)^2 n}} \tag{27}$$

where k_b is the Boltzmann constant, T is the temperature, z is the valence of the ion, q is the electron charge, and n is the number density of ions. Effectively, the Debye length is a measure of the ability of the mobile charge to counteract the electric field that occurs from electrostatic potential differences [1]. Because of the unknown finite size of ions in their hydrated state, this form of the Debye length is unlikely to give a physically realistic value in that it is derived for a single ionic species under ideal conditions. It is therefore used only as a guide when selecting mesh spacings. This can become a restricting factor in the case of ion channel simulations because of the extremely high concentration of charged particles present in an ionic solution. For example, in a 10 nm^3 box of 1 M KCl, there are 602 particles of each type. This translates to a charge density of 6.02×10^{26} Cm^{-3} for each ionic species, giving a Debye length of 4.36 Å.[3] Therefore, the maximum mesh spacing should be close to or lower than this value for the short-range electrostatic forces to be evaluated properly.

4.2. Charge Assignment

The first step in the self-consistent solution of the Poisson equation is to ensure the correct inclusion of the finite charges associated with the individual ions in the simulation domain in the right-hand side of the Poisson equation. This requires mapping a single point charge (SPC) of an ion onto the discrete mesh, used for the numerical solution via some interpolation function. Currently, there are three charge assignment schemes commonly in use: the nearest grid point (NGP), cloud in cell (CIC), and triangular shaped cloud (TSC) schemes [53, 55].

The simplest method of charge assignment is the NGP scheme. In this approach, the particle's charge is simply assigned to the closest mesh point in space. Here, the only computational overhead is associated with the determination of the relative position of the particle and the nodes of the mesh. In the case of a uniform mesh, this is a trivial exercise requiring very little computation, but in the case of a large, nonuniform topology, the problem can become somewhat unwieldy despite its algorithmic simplicity.

The NGP scheme has the advantage that it best preserves the pointlike nature of the particles but has the serious disadvantage that it inherently introduces a position-dependent error in calculating the electric field. This can have serious repercussions in the resolution of short-range particle interactions in situations where high concentrations of carriers and small mesh spacings are used in the simulations. In addition, NGP introduces a large erroneous

[3] Calculated for $T = 300$ K and $n = 6.02 \times 10^{26}$.

self-force, which in conjunction with the discontinuous first derivate of the potential can cause extremely unpredictable particle behavior. Hence, NGP is rarely used in modern simulations, where the small reduction in computer time is overshadowed by the problems associated with the method.

The CIC assignment scheme is possibly the most commonly used method of charge assignment in P3M simulations. It uses a linear interpolation scheme to weight the amount of charge assigned to the grid points in the mesh cell surrounding the particle's position. Equation (28) shows the interpolation function for a rectilinear mesh

$$\rho_{(x,y,z)} = \frac{q_p}{v_{(x,y,z)}}\left(1 - \frac{|x - p_x|}{hx}\right)\left(1 - \frac{|y - p_y|}{hy}\right)\left(1 - \frac{|z - p_z|}{hz}\right) \tag{28}$$

where $p_{(x,y,z)}$ is the particle's position in space; hx, hy, and hz are the mesh spacing in each of the cardinal directions; q_p is the charge on the particle; and $v_{(x,y,z)}$ is the volume associated with the node x, y, z. This method can be used on a nonrectangular mesh because there are no restrictions on whether the mesh spacing has to be uniform in each of the directions. The relatively low computational overhead of this charge assignment scheme and the accuracy of the positional representation of the particle make it an attractive option. However, the pointlike nature of the particle has now been lost as far as the electric field at short range is concerned, and the particle is now represented as a cloud of charge of dimensions (hx, hy, hz). On irregular grids, the characteristic "size" of the particle will vary with the mesh spacing.

The third commonly used charge assignment method, the triangular-shaped cloud (TSC) [53], uses a quadratic weighting function to assign charge to the mesh. This requires the assignment of charge to the 27 nodes nearest the particle. This method also gives a continuous derivative of the potential and can be extended to larger numbers of points by using higher-order polynomial weighting functions. However, the additional smearing of the charge to something akin to a Gaussian shape over $3h$ means that the resolution of short-range forces is substantially decreased. Also, the use of larger numbers of abscissas in the assignment introduces problems close to material boundaries where the dielectric changes rapidly, as is often the case in biological simulations.

The effect of the smearing of charge in the assignment process has been largely ignored in self-consistent simulations. The fact that the particle charge is smeared over the volume of the mesh cell means that, provided the particle does not occupy the exact center of the mesh cell, a potential gradient will exist across a single cell from a single charge. The corresponding electric field then exerts a force on the particle that is due to its own charge. These "self-forces" can be extremely large. Hitherto, the effects of these forces have been largely ignored in conventional semiconductor device and plasma simulations, due to the large number of particles present in the simulation domain and each mesh cell.

The self-force does not significantly affect the results of simulations until the limit of one particle per mesh cell is approached because the smoothing from the Poisson equation tends to mask the effect when multiple particles per cell are present. Figures 9 and 10 show the position of a single potassium ion accelerated for 2 ns by the self-force in the absence of Brownian motion and an inital velocity of 0 ms^{-1}. The initial position of the particle is directly on a mesh node, in which case the CIC charge assignment places all of the charge from the particle to that node. This creates a potential gradient diagonally across the cell and moves the particle. Over the course of the self-consistent integration of the particle trajectory, the particle moves toward the center of the cell, where the charge assignment equally distributes the charge across all eight of the nodes in the mesh cell. This gives an equal potential at each point, and therefore the corresponding field and self-force on the particle become zero.

As yet, there is no concrete solution for this problem, which has proved remarkably resistant to analytical treatment. There are several factors controlling the magnitude of the self-force: the mesh spacing in three dimensions, the dielectric of the materials, and the proximity to Dirichlet boundaries, as well as variations introduced by the particle position and charge assignment scheme. Laux [55] did propose that the problem could be overcome by always

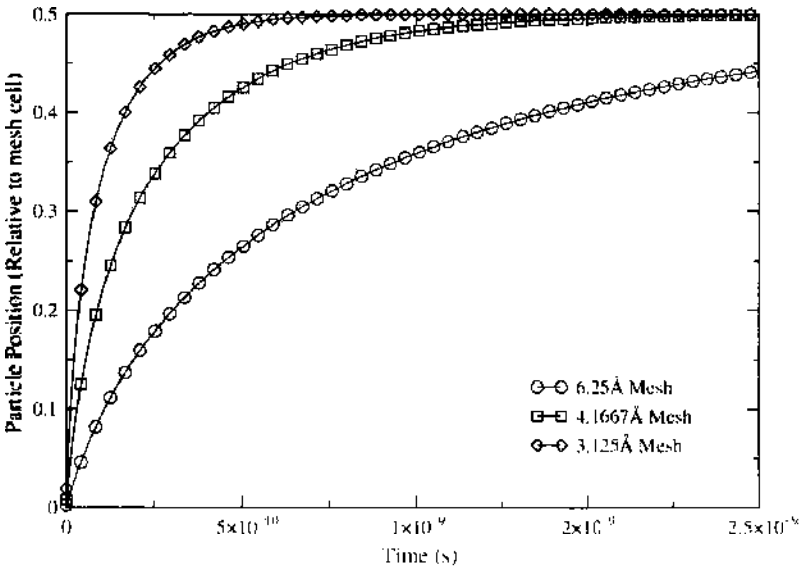


Figure 9. The trajectory due to self-force of a single potassium ion in a $20 \times 10 \times 10$ nm simulation domain with initial position at a mesh node in the middle of the simulation domain. There are no random scattering events to perturb the trajectory of the particle. The particle position is shown with time for three different uniform mesh spacings. Note how the final position is at 0.5 h_x, that is, the point where the field from the charge assignment error cancels out.

calculating the electric field for a particle at the center of the mesh cell that it occupies. This is practicable but introduces errors in the short-range force that are acceptable when there are a large number of particles present and the electric field is relatively smooth. However, in the simulation of ionic channels, this is a rare case. Usually, the number of particles in the simulation domains are of the same order as the number of grid points. Figure 11 shows an example of the potential calculated for an electroneutral solution of KCl. Although the average potential over the entire domain is zero, the local potential variations due to individual ions can create large electric fields. The extremely small mesh spacings required to resolve the atomic detail of pore compounds ensures that self-forces can be large, and therefore the associated uncertainty in the integration of the equations of motion of particles can lead to erroneous particle behavior.

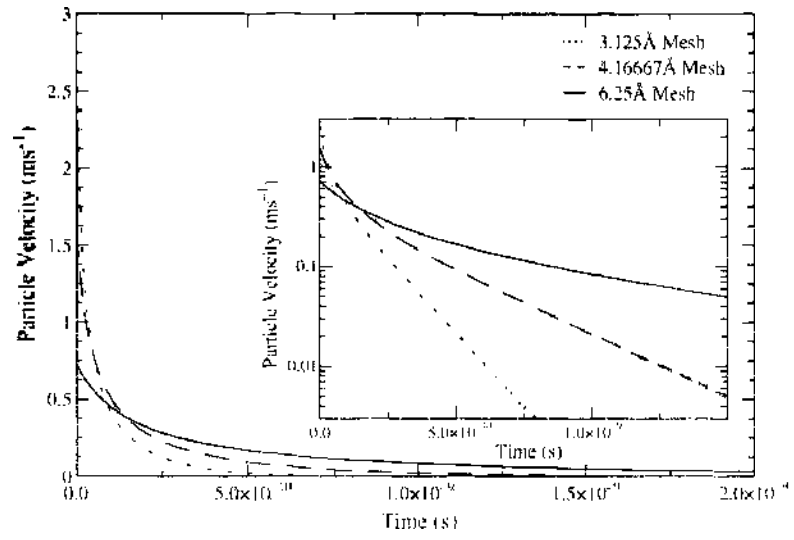


Figure 10. Shows the particle velocity corresponding to Fig. 9 with both linear and log scales. Note the exponential decrease in the velocity.

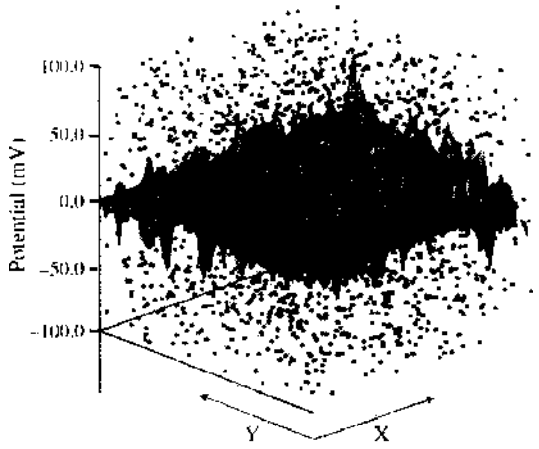


Figure 11. The potential solution for a simple electroneutral ionic simulation. Despite the fact that the potential averages to 0 local variations due to the presence of individual ions can create potential variations of up to 100 mV.

4.3. The Numerical Solution of the Poisson Equation

Iterative or biconjugate gradient methods (BiCGSTAB) [49] are typically used to solve the Poisson equation in 3D mesh-based simulations. However, the need for efficiency in the Poisson solver of an ion channel simulation has prompted the use of the the multigrid technique, which is known to be one of the fastest solution methods available [56].

The multigrid method has its roots in the standard relaxation solution techniques that have been used for a long time in the solution of partial differential equations. The fact that relaxation techniques have been known for a long time to have error smoothing properties [57, 58] led Fedorenko [59, 60] and later Bakhvalov [61] to investigate two-level approaches using relaxation methods for error reduction rather than direct solution, making them the first multiple-grid methods. Although the properties of multiple-grid methods were well known at the time, the actual computational efficiency of these solution methods was first noted by Brandt [62–64]. In addition, his development of local Fourier analysis enabled original adaptive techniques to be modified to solve the general cases of elliptic differential equations, leading directly to the explosion in multigrid solution techniques in the 1980s.

The essence of the multigrid technique lies in a slight paradigm shift in the approach to the problem. Instead of treating the solution as a gestalt and attempting to solve it directly, multigrid deals with the reduction of the error in an initial guess via the reduction of error components in Fourier space [56, 65]. The error is calculated via the residual equation, and frequency smoothing operators are then used to correct the initial guess at the solution based on the calculated defect in that guess.

By using the notation defined in Eq. (26), where **A** is, in our case, the matrix, arising from the discretization of the Poisson equation, containing the Laplacian for the problem, **b** is a vector containing the known portion of the problem, and **u** is a vector containing the solution of the equation, we define the general problem

$$\mathbf{Au} = \mathbf{b} \tag{29}$$

Obviously, the exact solution of the system is unknown initially, but we can define an approximation to the solution **v**, which may have been just computed from a previous iteration or can be an arbitrary guess at the final solution. We can then calculate a measure of the error in the initial guess (**e**), known as the algebraic error.

$$\mathbf{e} = \mathbf{u} - \mathbf{v} \tag{30}$$

However, in a real situation, the algebraic error is just as inaccessible as the final solution to the problem. Therefore, we must use a computable measure of the quality of the approximation as a solution to the problem, known as the residual (**r**). The following Eq. (31) shows

that the residual can be used as a measure of the defect in the initial guess, in the same way as the error.

$$\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{v} \tag{31}$$

Although $|\mathbf{r}| = 0$ only when $|\mathbf{e}| = 0$, however, it may not be true that when $\mathbf{r}$ is small locally that $\mathbf{e}$ is small. If the residual is then rewritten as

$$\mathbf{A}\mathbf{v} = \mathbf{b} - \mathbf{r} \tag{32}$$

then we can substitute for $\mathbf{b}$ with Eq. (29) and for $\mathbf{v}$ with Eq. (30) to give

$$\mathbf{A}\mathbf{e} = \mathbf{r} \tag{33}$$

Equation (33) is called the residual equation. This is extremely important in the functioning of the multigrid method because it allows us to directly measure the error in a particular solution and therefore correct an approximation of the final solution based on the calculated defect in that guess. The residual equation provides some insight as to how the multigrid algorithm is going to be constructed, but before it can be laid out in full, there are some further concepts that must be introduced.

As previously mentioned, multigrid does not rely on the direct solution of a problem; rather, it uses the smoothing properties of standard relaxation techniques to remove error components in frequency space. On a particular grid (Ω_h with mesh spacing h), a high frequency is defined as being at or above half the Nyquist frequency ($f_n/2$) of the particular mesh, and a low frequency is below $f_n/2$.

Standard relaxation techniques such as Jacobi or Gauss-Seidel iteration techniques [49, 65] can be used to completely solve systems of linear equations. However, they are usually inefficient because low-frequency components in the solution slow their convergence drastically. These solution techniques are, in effect, low-pass filters, the characteristics of which are of great importance for the efficiency of the multigrid solution method.

Figures 12–14 show the effect of three common smoothing operators used in multigrid methods: red-black Gauss-Seidel (RBGS), lexicographical Gauss-Seidel (GS-LEX), and weighted Jacobi (w-JAC) for problems of different frequencies. The efficiency of smoothing

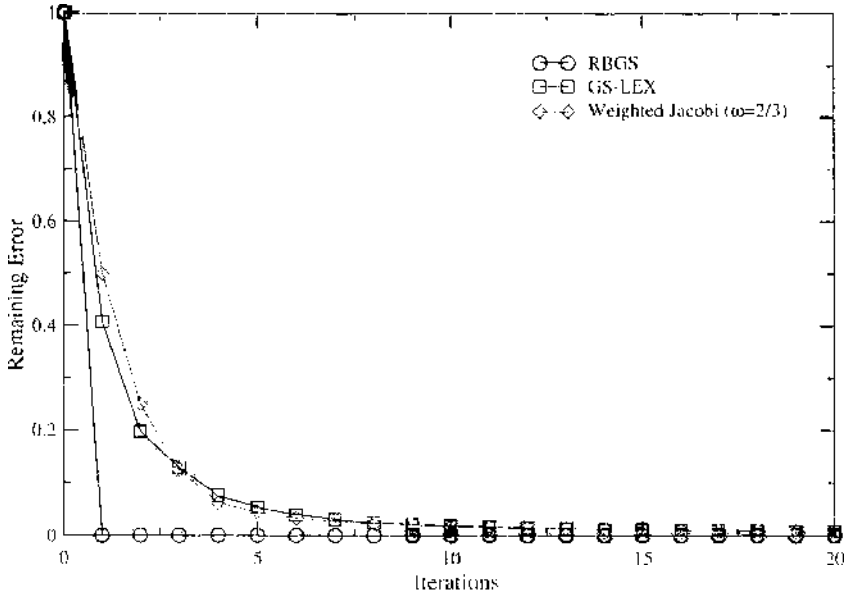


Figure 12. The smoothing properties of three relaxation methods: Red-black Gauss Seidel Relaxation (RBGS), lexicographical Gauss Seidel (GS-LEX), and weighted Jacobi relaxation (w-JAC) with an overrelaxation parameter of 2/3. The plot shows the reduction in error for problems with components at f_n. The high frequencies for a grid level are strongly damped by all of the smoothing methods, but especially by RBGS, thus reducing aliased components when the smoothed problem is transferred from level Ω_h to level Ω_{2h}.

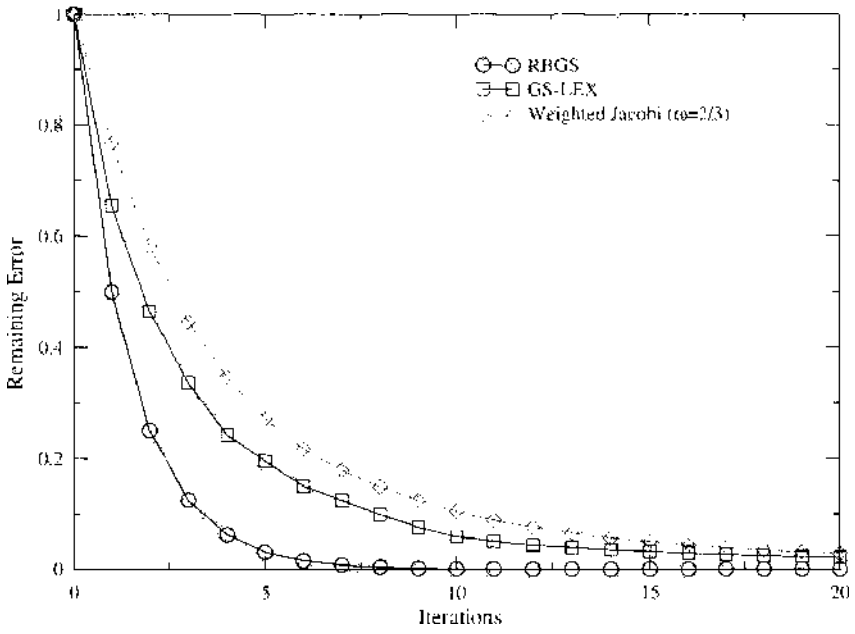


Figure 13. Relaxation smoothing for $f = f_n/2$. Note RBGS reduces the error by a factor of 0.5 per iteration.

methods can be characterized in terms of their reduction per iteration of the magnitude of an error at a given frequency. For these test cases, the problem is simplified by using the discretization of the Laplace equation, rather than the full Poisson equation, Dirichlet boundary conditions of zero are applied at x_{max} and x_{min}, and the initial guess is a sine wave of the desired frequency. Therefore, the reduction in the error components at that frequency can be defined as the maximum error in the computed value $|v_{max}|$, allowing a direct measure of the efficiency of smoothing.

Figures 12–14 show that the efficiency of the three methods changes dramatically with the frequency of interest. It is for this reason that iterative solvers experience deterioration in their solution convergence rate as the number of iterations increases. Because the strongly

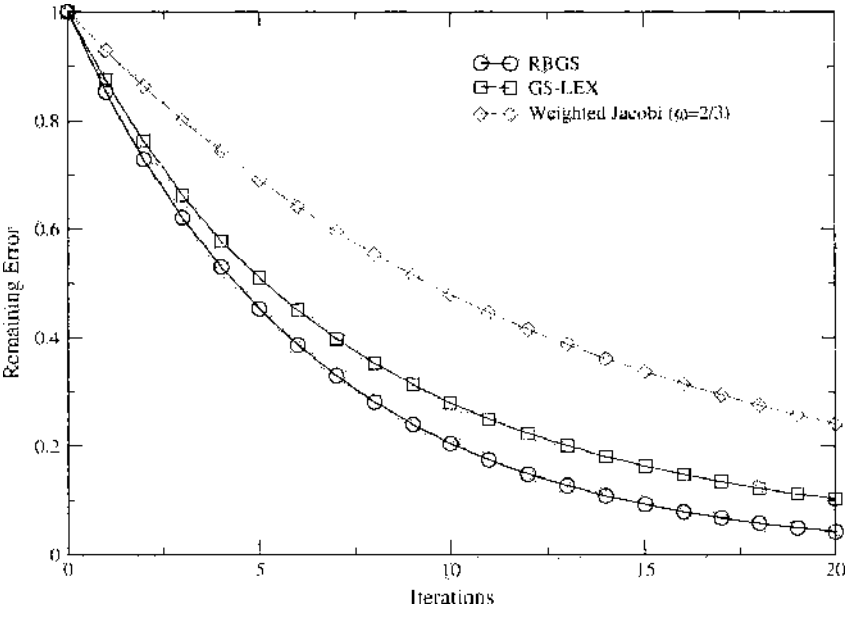


Figure 14. Error smoothing for $f_r = f_n/4$. All methods suffer from reduced efficiency, but RBGS still shows a marked improvement over other methods.

damped high frequencies are eliminated very efficiently in the initial few iterations, only problematic low-frequency components remain, slowing the convergence of the solver.

The multigrid method relies on the fact that high-frequency components of the error are damped extremely quickly. The difficulty lies in how to treat the components with frequencies below $f_n/2$. The solution is to resample the problem onto a coarser mesh, where the frequencies between $f_n/2$ and $f_n/4$ have moved from the low- to the high-frequency region, therefore allowing their efficient reduction. However, the reduction of high frequencies must be treated with care because any remaining power at these frequencies will cause aliasing when the problem is resampled.

The frequency that is of greatest importance for the convergence on a particular mesh is that at the bottom of the high-frequency bracket (for example, $f_n/2$), which is the point corresponding to the worst convergence of all of the smoothers for the subset of the problem that is to be solved on that particular level. The reduction of the problem onto a coarser grid level (Ω_{2h}) means that the convergence rate of the smoother at $f_n/2$ is the primary controlling factor in the efficiency with which the entire problem can be solved. Figure 13 shows the convergence rates for the three smoothers at $f_n/2$. RBGS is by far the most efficient in reducing the error at this point, offering a reduction by a factor of 0.5 on each iteration, compared with 0.66 for GS-LEX and 0.77 for w-JAC. Because $f_n/2$ is the frequency most likely to leave spurious frequency components that may be introduced as aliased frequencies to Ω_{2h}, the overall convergence of the multigrid algorithm is primarily determined by the smoothing operator performance on a given level.

Figure 15 shows the flow of control for a complete generic multigrid algorithm with an arbitrary number of grid levels. The algorithm uses the smoothing properties of the relaxation methods to reduce high-frequency error components on a particular level before resampling the problem via a restriction operator to a smaller grid, where lower-frequency error components can be removed. This process is continued on successively smaller grids

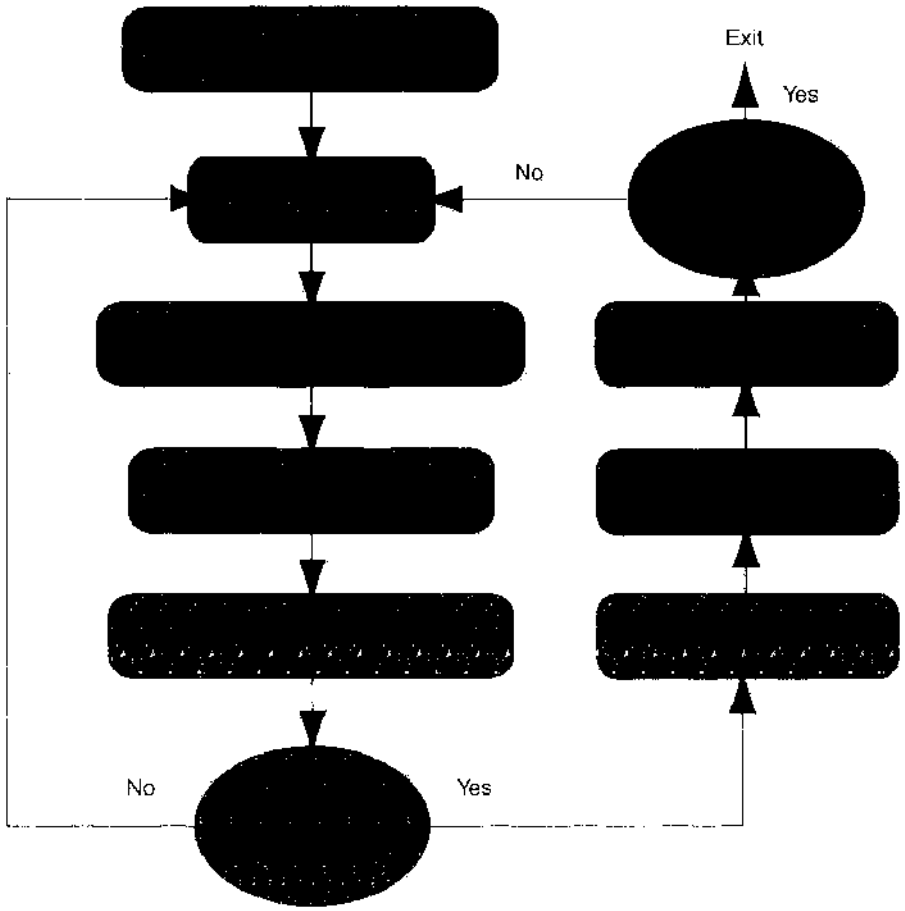


Figure 15. The flow of control for a V-cycle multigrid algorithm.

until either a defined minimum size is reached or the problem cannot be restricted further. On the lowest grid level Ω_{n2h} (where n is the number of levels in the multigrid implementation and can vary depending on the initial grid size and grid coarsening scheme), the problematic low-frequency components can be eliminated almost immediately by conventional iteration methods, and the error correction calculated from the residual equation [Eq. (33)] can then be transferred to the higher levels and used to correct the initial guess at the solution.

The choice of intergrid transfer operators can have an impact on the performance and convergence of the solution although their effect on the overall performance of the algorithm tends to be secondary to the efficiency of the smoothing operator. However, the specific optimum methodologies are highly problem-dependent, and there are several different choices.

The restriction transfer operator is used to resample the problem from grid Ω_h to grid Ω_{2h} and tends to be determined by the coarsening and smoothing methods used in the rest of the algorithm. The simplest restriction operator is direct injection, where the value x from Ω_h is directly injected into Ω_{2h} on all points common to each grid. This is the most computationally inexpensive of the methods but may not necessarily be the best choice for a given problem.

The full-weighting and half-weighting methods are more commonly used, despite the increased computational overhead associated with them. Equations (34) and (35) show the operators for full- and half-weighting restriction (R_{2h}^{h}, that is restriction from Ω_h to Ω_{2h}) in stencil form, for a uniform three-dimensional mesh.

$$R_{2h}^{h} = \frac{1}{64}\left[\begin{array}{ccc|ccc|ccc} 1 & 2 & 1 & 2 & 4 & 2 & 1 & 2 & 1 \\ 2 & 4 & 2 & 4 & 8 & 4 & 2 & 4 & 2 \\ 1 & 2 & 1 & 2 & 4 & 2 & 1 & 2 & 1 \end{array}\right] \tag{34}$$

$$R_{2h}^{h} = \frac{1}{8}\left[\begin{array}{ccc|ccc|ccc} \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & 1 & \cdot & 1 & 2 & 1 & \cdot & 1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot \end{array}\right] \tag{35}$$

These operators are constructed from the tensor product of three one-dimensional full-weighting operators [56]. It should be noted that the change in volume of the mesh cell between Ω_h and Ω_{2h} must be taken into account in order that the solution and corrections remain of consistent magnitude between the various levels.

For the majority of this work, the half-weighting operator has been used because it entails a lower computational overhead than full weighting, and in concert with the RBGS smoothing operator it can be simplified to a single point injection with half weighting. This is due to the fact that the residual at every second point is zero after a single RBGS iteration. However, care must be taken to ensure that the ordering of the RBGS smoothing and the coarsening scheme matches, or the restricted residual will be zero over the entire grid.

The counterpart to the restriction operator (R_{2h}^{h}) is the prolongation operator (P_{h}^{2h}) used to transfer the error correction from Ω_{2h} to Ω_h. Again, the simplest method is direct injection of the improved error estimate, but this can introduce high frequencies in the correction of the error estimate. Therefore, the most common method is to use a trilinear interpolation scheme, a direct analogue of the full-weighting restriction method. Equation (36) shows the form of P_{h}^{2h}, again in stencil notation. To indicate that this is now a distributive operator rather than constrictive, the direction of the containing brackets is reversed.

$$P_{h}^{2h} = \frac{1}{8}\left]\begin{array}{ccc|ccc|ccc} 1 & 2 & 1 & 2 & 4 & 2 & 1 & 2 & 1 \\ 2 & 4 & 2 & 4 & 8 & 4 & 2 & 4 & 2 \\ 1 & 2 & 1 & 2 & 4 & 2 & 1 & 2 & 1 \end{array}\right[\tag{36}$$

Two specific implementations of the multigrid algorithm can be seen in Fig. 16. The V-cycle is the simplest extension of the basic two-grid method, in which the residual is progressively restricted down through the grid levels and the error is calculated from Ω_h until Ω_{n2h} is reached. The error estimate is then solved completely on this level, transferred as a correction to the estimate on the levels above, until finally the initial solution itself is corrected.

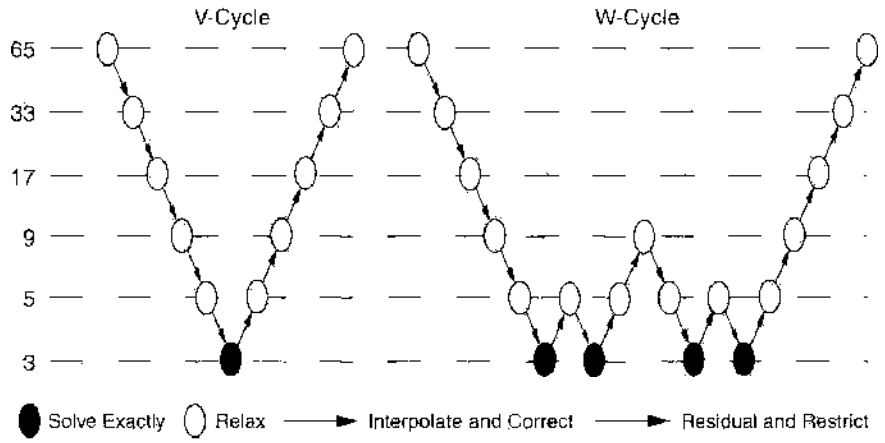


Figure 16. The multigrid V and W-cycle corresponding to $\gamma = 1$ and $\gamma = 2$ in the multigrid algorithm.

The principle of the W-cycle is similar to the V-cycle in that the initial solution is treated identically. However, when the correction is calculated on level Ω^n, it is transferred upward only for a number of levels smaller than n; an improved error estimate is then calculated on the higher level and treated in the same fashion as before. Figure 16 shows a second-order W-cycle ($\gamma = 2$), where a cycle with $\gamma = 1$ is equivalent to a V-cycle, which can be extended up to $\gamma = (n - 2)$ by progressively increasing, then decreasing, the level that the W-cycle recurses to on each subcycle.

Figure 17 shows the improvement in solution time as a function of γ for a simple pore test structure, shown in Fig. 18, which uses the dielectrics specified in Table 3 and has Dirichlet boundaries of 200 mV and 0 mV at x_{min} and x_{max}. There is relatively little improvement in efficiency of the solver for this simple structure, but, like most tuning parameters in multigrid methods, the optimum value for γ is highly problem-dependent and can be determined only empirically. For the majority of this work, $\gamma = 2$ is used because it generally proves to be more efficient for working with channel-like structures. The three-dimensional solution of the problem is shown in Fig. 19.

The efficiency of the solution method for the Poisson equation is of great importance in the simulation of ionic solutions. From Section 3, we have seen that the time step in the simulations is limited by various factors, including the resolution of the Poisson mesh. The

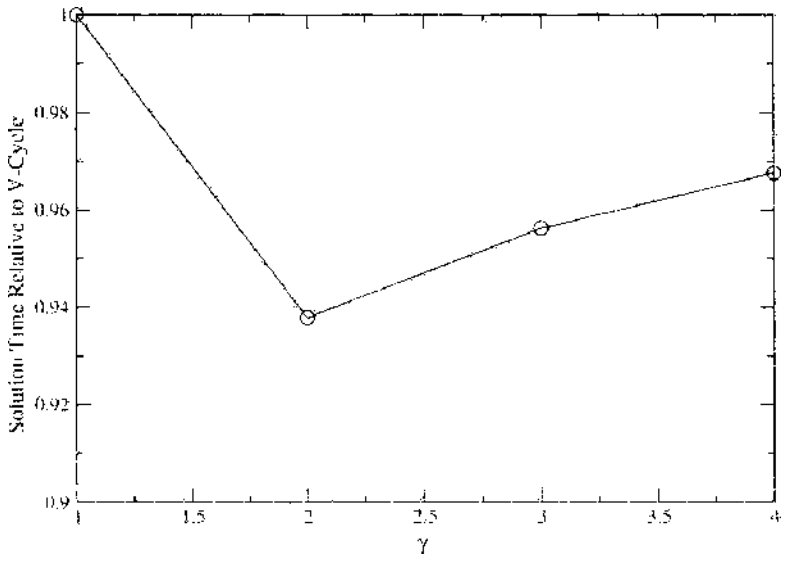


Figure 17. W-cycle performance for varying values of γ. In this case, the optimum is achieved at $\gamma = 2$, but there is relatively little improvement in the solution time. However, this optimum value is highly problem dependent and can be determined only by experimentation.

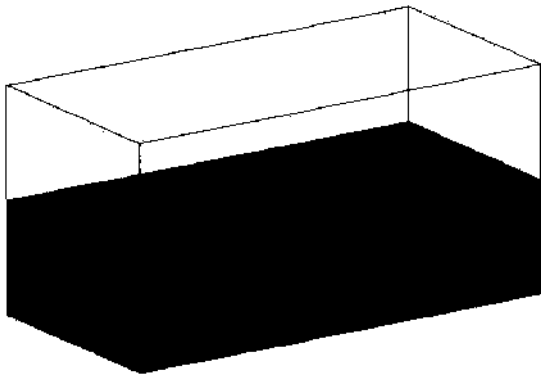


Figure 18. The poisson solver test structure. A 15 × 10 × 10 nm region of water with a 3 nm lipid membrane region and a 4 Å pore. Material color keys are shown in Table 3.

very short relaxation time τ for ions, in combination with the long time scales necessary to study biological processes, means that the Poisson equation must be solved many times in a simulation. Figure 20 shows a comparison between the commonly used BiCGSTAB solution method and multigrid implementations on two platforms for a simple channel structure solution. It is clearly obvious that the multigrid method is far more efficient than BiCGSTAB by a factor of ~10. The test problem was also solved with an SOR method, but this was ~100 times slower to converge than multigrid. In addition, the inherent simplicity of multigrid means that it tends to be a much more numerically stable solution method. The large changes in dielectric constant in biological simulations (2–80) mean that BiCGSTAB in particular may require the use of a preconditioner to guarantee convergence for a given geometry. Multigrid, however, has displayed no such convergence problems, provided the simulation geometry is properly represented on each of the grid levels.

5. SELF-CONSISTENT SIMULATIONS

The complexity involved in the self-consistent simulation of ion channels stems from the fact that the problem contains some of the more extreme situations that can arise in P3M simulations.

To recover a current-voltage (IV) characteristic for a given channel structure, the simulator must be capable of resolving features from Å to tens of nm, including the cell membrane and the ionic solutions on both sides, as well as maintaining the correct behavior of very small numbers of ions at high concentrations and retaining the ability to do so on time scales from fractions of a ps to μs for the total simulation time. In any situation where a simulation contains a very small number of particles or particularly small simulation domains and mesh spacings, the simplest of these simulation tasks can become problematic.

This section discusses the specific optimizations of the multigrid algorithm required to reduce the compute time of each iteration to a level where real time in the range of microseconds can be simulated by using picosecond time steps. The very low mobility of ions ($\sim 10^{-8}$ $m^2V^{-1}s^{-1}$) [1] means that, although a channel may only be several nm long, the conduction of ions through the pore is a slow process. In addition, the arrival of ions at the mouth of the pore is a relatively rare event, with typically fewer than a hundred ions traversing the membrane per microsecond [21, 24, 66]. Because the solution of the Poisson

Table 3. The values of the dielectric constant and color key for the materials used during the course of this project.

Material	Dielectric (ϵ_r)	Color
Water	80.0	Blue
Lipid	2.0	Red
Protein	20.0	Green

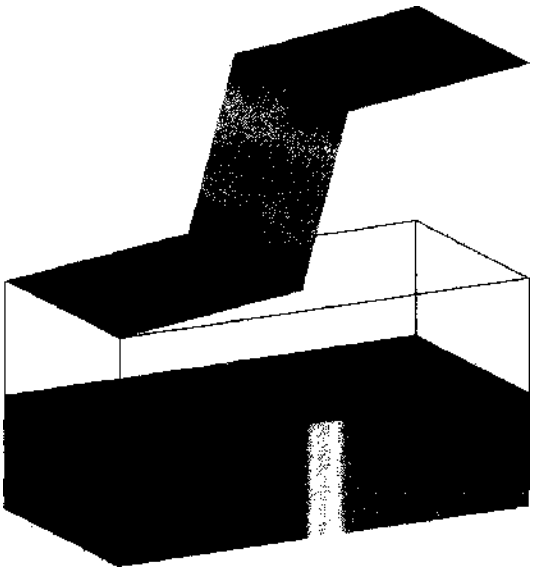


Figure 19. The potential solution for the structure shown in Fig. 18 with dielectric constants from Table 3.

equation accounts for the vast majority of the computational time, the efficiency of the solver will determine to what extent simulations longer than microseconds, capable of resolving the multiple ion traversal events necessary for the reliable estimation of the current, are practicable.

The simulation of an ion channel must be performed under conditions similar, if not identical, to the conditions found in experimental measurements; otherwise, the accuracy of the simulator cannot be assessed. The simulation domain (shown in Fig. 21) and particle behavior are of sufficient complexity to render simple qualitative analysis of particle conduction useless as a guide to the quality of a model. Therefore, the ability to replicate experimental conditions is essential.

The modeling of particle interface conditions in particle simulations has proved problematic in many fields [53, 67]. The necessity of reproducing the experimental conditions found in patch clamp measurements of channel conduction requires an accurate method for including the effect of the ionic solution surrounding the simulation domain. We present

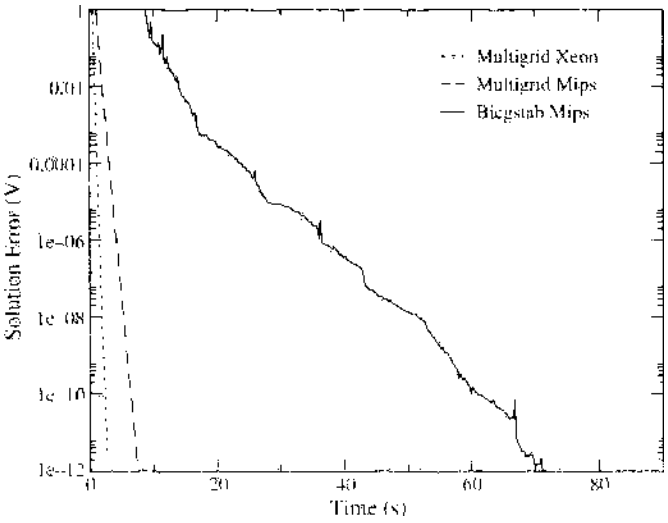


Figure 20. A comparison of the solution time of the BiCGSTAB and multigrid Poisson solvers for a simple channel structure on a $65 \times 65 \times 65$ mesh. Results are shown for an SGI origin 300 Mips r14K 500 MHz and a Compaq Evo6000 Intel Xeon 2.2 GHz. Multigrid is obviously substantially more efficient than BiCGSTAB. Results were also generated using a successive overrelaxation method, but these have been removed since the solution time was ~700 s.

Figure 21. A schematic representation of the test domain used with the Brownian code. Red lines indicate Neumann boundary conditions with particle reflection, Black lines are Dirichlet boundaries. The particle contacts are shown simply as regions where the ion exit boundary is aligned with the external Dirichlet boundary. Materials are shown as blocks, which directly corresponds to their representation in the simulation.

a novel method for the treatment of particle interface that ensures that the bulk solution concentration of ions is maintained at the boundary with the simulation domain, despite the low numbers of particles present in channel simulations. This method drastically reduces the number of particles that have to be simulated when compared with the normal "reservoir" methods [67].

The current measurements of ion channel proteins have restricted bandwidth, with a temporal resolution of less than 1 μs. The comparison with simulation results, where the total simulated time is around the same, is difficult unless time-averaged IV properties are used. However, the recovery of reliable IV characteristics from both experiment and simulation can be problematic because of the very low signal-to-noise ratio encountered in measuring ionic currents. The simple charge-counting method of current determination can be used only if the simulation time is long enough to ensure statistically reliable averaging. Unfortunately, this easily implemented method requires a very large number of iterations to smooth the thermal noise sufficiently at a given applied potential. It is therefore necessary to implement a more advanced current measurement technique based on the Ramo-Shockley theorem [68], which takes into account the simultaneous contribution of all particles in the simulation domain to the current.

5.1. Boundary Conditions in the Poisson Equation

The ability to apply different potentials at the boundaries of the simulation domain is necessary to model the effect of external electrodes. In the multigrid solver, the matrix **A** obtained from the discretization of the Poisson equation requires modification to correctly represent boundary conditions for the electrostatic potential. Because **A** is sparse in the case of the 3D Poisson equation, there are only seven diagonals that need be stored for each level of the multigrid algorithm. To represent the boundary conditions these coefficients must be manipulated on all grid levels. The storage of the seven matrix coefficients in seven three-dimensional arrays is the simplest method of relating these coefficients to the spatial position of the Poisson solution nodes. These will be referred to as **Am** for the main diagonal of the **A** matrix, **Al** for the coefficient applied to the point to the left, and so on through **Ar**, **Au**, **Ad**. **Af**, and **Ab** for right, up, down, front, and back.

Two types of boundary conditions are used in this work: Dirichlet (or fixed) and Neumann (or floating) boundaries. The Dirichlet boundary condition, used to represent a fixed potential that is the equivalent of an equipotential electrode surface, is the simpler to implement. However, great care must be taken when using a multigrid solver to ensure that the Dirichlet boundaries are correctly represented on all grid levels of the multigrid algorithm. This is vital to the convergence of the solution to the correct value because the nodes with fixed potential

are the reference by which the relaxation "propagates" the solution across a mesh to the final solution. If an electrode is mistakenly removed from the simulation domain through the restriction of the problem to a smaller mesh, then the error estimate on that level will contain no contribution from the applied potential at this point. Although this may not affect the quality of the final solution, because the electrode will, at least, be represented on one of the higher levels, it will dramatically slow the convergence of the solver. Therefore, care must be taken when using electrodes with complex geometries or electrodes that cover a small number of nodes.

The Dirichlet boundary as a constant potential boundary condition is easily dealt with in the Poisson solver by setting the main diagonal coefficient (**Am**) belonging to the contact node to 1, all other coefficients to 0, and the potential and corresponding node in the charge density matrix to the desired value. This causes the relaxation operator to leave the potential at the contact unmodified during the smoothing and solution operations. A similar effect can be achieved by using property flags to specify when a contact node is to remain untouched. The specification of Dirichlet nodes, while necessary for solver convergence, should be done with care. The fixing of the potential leads to an erroneous electric field at the boundary, and Dirichlet boundaries should be positioned as far away as possible from the simulation region of interest.

The Neumann boundary condition specifies the derivative of the potential at the boundary, reflecting the fact that the potential is allowed to vary at the boundary nodes. Effectively, the Neumann condition with zero derivative applied in these simulations assumes that there is a mirror image of the simulation domain on the other side of the boundary. Therefore, when a particle approaches the boundary, the electric field at the boundary is zero because of the contribution from its mirror particle. This allows predictable numerical behavior in the Poisson solution while allowing the potential to vary according to the self-consistent solution in particle-based simulations. The Neumann boundary condition tends to be combined with reflective particle boundaries, where the particle interaction with the boundary is treated as a Newtonian reflection from the plane of the boundary. The use of Neumann boundary conditions—and the associated image forces near the boundary—introduces errors in the short-range force but maintains the value of long-range fields. Therefore, they introduce relatively small errors in the electrostatics of a simulation, and placement of the boundaries, particularly on parallel surfaces, is not as critical as in the case of Dirichlet points.

The application of Neumann conditions is performed as part of the matrix coefficient calculation in the discretization of the Poisson equation. Because the discretization described in Section 4 assumes that the flux and dielectric are constant over an eight-node cell, the coefficients can be calculated by summing the contribution to the flux through a point from its neighboring nodes. Total flux away from a node is calculated by using Eq. (37), and summing the contribution to the surrounding nodes by iterating from element 1 to node $N_a - 1$ in the array, where N_a is the number of nodes in a direction.

$$Am_{(x,y,z)} = Am_{(x,y,z)} - \varepsilon_0 \varepsilon_{r(x,y,z)} \left(\frac{hy\,hz}{4hx} + \frac{hx\,hz}{4hy} + \frac{hx\,hy}{4hz} \right) \qquad x,y,z = (1 \ldots N-1) \tag{37}$$

Similarly the left, right, and other coefficients can be calculated by using Eq. (38) (given here for the left-hand coefficient as an example).

$$Al_{(x,y,z)} = Al_{(x,y,z)} + \varepsilon_0 \varepsilon_{r(x,y,z)} \left(\frac{hy\,hz}{hx} \right) \tag{38}$$

The effect of this summation process is to ensure that all of the external points in the array are automatically set up as Neumann boundaries as part of the discretization process, because the contribution from a nonexistent external point is never added to the flux.

In the discretization, the electric field is assumed to be constant over the width of a mesh cell. Therefore, the field should be calculated on a cell-by-cell basis by using Eq. (39):

$$\mathbf{E}_{(x,y,z)} = -\begin{pmatrix} \dfrac{\Psi_{(x+1,y,z)} - \Psi_{(x,y,z)}}{hx} \\ \dfrac{\Psi_{(x,y+1,z)} - \Psi_{(x,y,z)}}{hy} \\ \dfrac{\Psi_{(x,y,z+1)} - \Psi_{(x,y,z)}}{hz} \end{pmatrix} \tag{39}$$

rather than the more common Eq. (40).

$$\mathbf{E}_{(x,y,z)} = -\begin{pmatrix} \dfrac{\Psi_{(x+1,y,z)} - \Psi_{(x-1,y,z)}}{2hx} \\ \dfrac{\Psi_{(x,y+1,z)} - \Psi_{(x,y-1,z)}}{2hy} \\ \dfrac{\Psi_{(x,y,z+1)} - \Psi_{(x,y,z-1)}}{2hz} \end{pmatrix} \tag{40}$$

If Eq. (39) is used, the field at the particle position in the cell is then available via a trilinear interpolation, which should be consistent with the charge assignment scheme used. The advantage of Eq. (39) is that because the field is calculated internal to a cell, rather than at a node, the field never has to be calculated over a change in material dielectric or boundary condition, greatly reducing the computational cost of the calculation.

5.2. Parallelization and Optimization of the Multigrid Solver

The solution of the Poisson equation, which has to be recalculated after every time step, constitutes the bulk of the computational effort in these simulations (usually more than 98%). Therefore, the optimization of the solver is critical to reduce the simulation time. In this section, we discuss details of the C and C++ implementation of the multigrid algorithm and its parallelization.

The fundamental components of the multigrid algorithm are matrix operations, performed on memory arrays, which ideally occupy single contiguous blocks in main system memory during the execution of the code. The method by which these blocks are accessed seems to be an unlikely source of inefficiency because many programmers tend to rely on predefined access operators within the language in use. For example, in C and C++ and Fortran a multidimensional array can easily be accessed randomly. However, the order in which variables are accessed can be a major contributory factor affecting the efficiency of code. The native multidimensional arrays available in most languages provide a simple method of addressing elements in a matrix, and the use of three coordinate related indexes to describe a storage address easily associates matrix elements with their position in the simulation. The three-dimensional arrays used in multigrid are stored in memory in row major order. To access a particular matrix element, the pseudocode shown in Eq. (41) is used.

$$X = B + z * sizex * sizey * S + y * sizex * S + x \tag{41}$$

where B is the base address of the memory block; x, y, and z are the matrix element coordinates; *sizex*, *sizey*, and *sizez* are the number of elements in each direction of the Cartesian grid; and S is the byte size of the type of variable stored in the array. This gives a total operation count for an array access of six multiplications, three additions, and several fetches to memory.

Accessing array elements in the wrong order, a surprisingly common error, can lead to serious performance penalties due to prefetch cache misses, even in the presence of extensive compiler optimizations, such as (z, y, x rather than x, y, z). Table 4 shows a comparison of the time taken to perform 100 red-black Gauss Seidel relaxation iterations on a $33 \times 33 \times 33$

Table 4. Simulation times for 100 RBGS iterations with both compiler optimized and unoptimized code.

	Linear(s)	Out of sequence(s)	% Slowdown
Optimized	1.93	6.88	356.5
Unoptimized	6.58	11.36	172.6

array, with both compiler optimized and unoptimized code and for both linear and out-of-sequence array access. The relative performance drop is large and can primarily be attributed to the fact that the page size prefetched to cache memory is smaller than the memory offset per element in the z direction. Interestingly, compiler optimization improves the situation, probably because of the improved compiler heuristics involved in branch and execution prediction.

The operations overhead associated with matrix element access can be substantially reduced because array access in the multigrid algorithm is performed in a piecewise linear fashion. Therefore, the base address offset calculation [shown in Eq. (41)] can be simplified from several multiplications, additions, and fetches to a single addition and multiplication.

The use of iterators in code reduces the computational overhead associated with multidimensional arrays at the cost of reduced code readability and ease of use. However, the gain in efficiency makes this an attractive option. An iterator is an object with which the elements of associated arrays can be accessed directly; for example, in the multigrid relaxation algorithm, the elements surrounding the current point are all needed to calculate the current value of the potential. Because these elements occupy positions that are separated by fixed indices, the offsets to points up, down, and so on, in memory can be precalculated, thus minimizing the overhead associated with their evaluation. As the algorithm walks the array, it is simply a matter of incrementing each offset by the amount required. This reduces the operation count for a typical array element access to a single multiplication and an addition. Because iterators are in continuous use, they tend to be register (on RISC processors) or stack variables (on CISC processors) rather than variables contained in larger data structures, which avoids the overhead of memory fetches. In addition, the use of simple static integer offsets greatly increases the likelihood of correct access prediction, allowing for more efficient cache management.

This scheme can, in fact, be simplified further. As midlevel languages, the C and C++ make provision for direct access to memory addresses through constructs known as memory pointers [69, 70]. While the pointer underpins most of the operators and operations in C and C++, it can be used as a data type itself. By rewriting the multigrid operators using pointers to access the memory locations of the required points and simply initializing these pointers with the correct static offsets for surrounding nodes, it is possible to reduce the computational overhead to a single addition by the use of increment operators. Because these operators can be completely evaluated at compile time, the additional multiplication of the size of data variable present in the previous method is removed. In addition, the fact that the contents of a pointer variable is the "real" address of the data to be manipulated means that cache prefetch is simplified as much as possible.

Because the use of low-level data types such as pointers means a substantial increase in the complexity of the hand-written code, the direct access iterators are written using C rather than C++. The higher-level data structures available in C++ mean that overall code design and maintenance are simplified. But in a situation where code performance is the paramount issue, the use of such structures is irrelevant because code is being written at a lower level than the language is designed for.

Table 5 shows the overall effect that array access operator optimization has on the efficiency of the multigrid algorithm. The normal array access method is taken as the reference, and the execution time for 20 multigrid cycles, on a $65 \times 65 \times 65$ node grid, is measured for implementations using each of the access methods described. The ~20% increase from the use of iterators in C++ is attractive, but the 41% code efficiency increase due to the use of direct access iterators is excellent. Because multigrid is such a large proportion of the simulation computation time, this can be directly equated to a 41% decrease in the

Table 5. Execution times for 20 multigrid cycles on a single 500 MHz MIPS R14K processor (SGI Origin 300). The percentage of speedup due to the two optimized array access methods are shown relative to the normal access method.

Access method	Time(s)	% Speedup
Array	72.74	0
C++ Iterators	58.29	19.86
C Direct	42.65	41.36

overall execution time of the code. It may appear unusual that the performance increases from operator simplification are not quite as large as the operations count suggests that they should be, but without specific details of the compile time optimizations, which are commercially sensitive and therefore not available, it is difficult to assess the contribution that this makes to the performance of the code.

5.3. Particle Maintenance in the Contacts

Particle contact implementation is an issue in any particle-based simulation. Ion channel simulation requires two types of boundary conditions to be considered from a particle perspective: reflective surfaces and regions where the particle concentration must remain constant [71].

Reflective particle boundaries are simple to implement. They are treated as reflective planes where the particle undergoes specular reflection in which energy is conserved. They correspond directly to the application of Neumann boundary conditions in the solution of the Poisson equation. The implementation of fixed concentration boundaries is substantially more complex, because they must mimic the interface of the simulation domain with the bulk electrolyte solution.

Gonzalez and Pardo [67] showed that the maintenance of so-called Ohmic contact regions can seriously affect the quality of results from particle-based Monte Carlo semiconductor device simulations. The problem of correctly modeling the interface between semiconducting materials and the metal contacts used to apply potentials across these devices is essentially the same as treating the boundary between a channel simulation domain and the surrounding ionic solution. The contact region must behave as a particle source and sink in such a manner that its behavior is consistent with the ionic distribution and electric field inside the simulation domain, as well as maintain the bulk solution concentration at the boundary.

There is no standard model for the simulation of particle transmission boundaries. Three methods based on the injection of particles at the simulation boundary plane are in common use and use an injection velocity based on either Maxwellian, displaced Maxwellian, or velocity-weighted Maxwellian distributions [67]. By far the most effective of these schemes uses the velocity-weighted Maxwellian distribution to assign a velocity to carriers arriving at the boundary with randomly distributed arrival times through the simulation. In this method, the Maxwellian distribution is offset by the average velocity of the carriers currently occupying the simulation domain. With the small number of particles contained in a realistic channel simulation domain, the average particle velocity is extremely noisy. This can result in oversupply or undersupply of particles to the simulation away from the steady-state average. Although after perturbation the simulation does eventually return to a steady state, the use of this method can seriously increase the time that the simulation takes to relax to equilibrium from initial conditions, thus increasing the overall simulation time.

The fourth method, known as the reservoir method, uses a small region adjacent to the simulation domain as a particle source and sink. Particles are injected into the far side of the reservoir by using a Maxwellian distribution, thus keeping this unphysical boundary as far as possible from the region of interest and allowing the particles' momentum and energy to equilibrate correctly to the conditions at the interface neighboring the simulation domain.

The Poisson equation is usually not solved in the reservoir, but instead the field is fixed at all points inside the region to the value of the average electric field at the reservoir–simulation interface. In some cases, it may become necessary to solve the Poisson equation

inside the reservoir, especially when the simulation domain is short. In this case, particles in the reservoir are held at thermal equilibrium, but the local field is solved semi-self-consistently inside the reservoir [72], thus ensuring the correct spatial distribution of carriers while leaving their velocities unaffected. The typically small number of particles in the reservoirs of a Brownian channel simulation necessitates that the Poisson equation is solved in the reservoirs in order to achieve the correct spatial distribution of ions, thus increasing the processing overhead. We present a mixed solution approach, which combines both Maxwellian injection and the use of reservoirs, capable of maintaining particle concentrations in simulation domains to within a few percent of the expected values, even when the simulation domain is extremely small or contains very few particles.

In a typical channel simulation, the electric field in solution-filled regions is small because the majority of the potential drops across the lipid membrane and the pore, both of which have high resistance. Therefore, it is possible to place Dirichlet boundaries relatively close to the active simulation region. Experimentation with the simulator has shown that, provided the boundary is more than 10 nm from the channel, the very short Debye length in concentrated ionic solutions ensures negligible effect on channel conductivity from the boundary.

Because of the small size of the simulation domain and the low number of ions, local field variations can dominate over long-range fields. Therefore, in order that the ions in a reservoir should have the correct spatial distribution when they enter the simulation, the Poisson equation must be solved in any reservoir region. However, the methods proposed in both Ref. [67] and Ref. [72] rely on diffusion to maintain the particle supply to the simulation. Although this is acceptable in semiconductor simulations, because of the extremely high carrier density in metal contacts, in ionic channel simulations, where the concentration is relatively uniform throughout the entire simulation domain, these approaches tend to result in particle depletion.

Because the solution of the Poisson equation constitutes the majority of the computational time in a Brownian simulation, minimizing the number of nodes required by the particle contact is desirable. In our method, the reservoir region is extended into the simulation domain only as far as the first mesh cell, and the Poisson equation is solved fully self-consistently in this volume. However, because of the small number of particles in this volume and the need for charge neutrality to be maintained over the reservoir as a whole, the region is then further subdivided in slices parallel to the plane of the contact to give finer control over the spatial distribution of the ions.

Figure 22 shows a schematic representation of the described particle contact. Particles are permitted to leave the simulation domain only through the plane of the Dirichlet

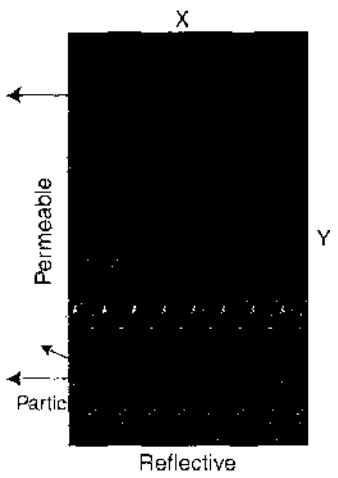


Figure 22. A schematic representation of the particle contact maintenance method. Particles are permitted to leave the simulation domain only through permeable boundaries and are reflected from all others. A small region at the edge of the simulation domain is subdivided in steps no smaller than σ for the particles. The concentration of particles in these subdivisions is then maintained artificially.

boundary and are reflected specularly from any Neumann boundaries. The optimum level of subdivision of the boundary mesh cell can be determined empirically, but provided each of the subdivisions is close to the average displacement of the particles in a time step σ (see Section 3), the method remains stable by ensuring that the majority of particles will not diffuse more than one subdivision in a time step. Thus, the required particle concentration can be calculated for the volume of each slice in the plane of the electrode and is found to remain sufficiently constant and consistent with the particle drift in the remainder of the simulation domain.

Particle injection is determined by comparing the actual particle concentration in a subdivision with the desired equilibrium value. If an integer number of particles is required in a subdivision, then they are created with a uniformly distributed random position inside the volume. Fractional particle requirements are summed until the remainder becomes greater than 1, at which point a particle is created in the current working subdivision. By removing particles from the simulation only when they pass the Dirichlet boundary and checking subdivision concentrations from the inner cell boundary to the outer, we can ensure that the majority of particles requiring injection are in the subdivision closest to the boundary because the flux in and out of the other subdivision volumes remains virtually constant over a long period of time. When a particle requires injection from the external boundary, it is created at a random position on the plane of the boundary with a velocity vector determined from its diffusion coefficient (see Section 3). The velocity is then scaled by a uniform random deviate between zero and one, to simulate an arrival time at some point during the current time step in the simulation. This scaled velocity is then used to evaluate the final position of the particle.

This computationally inexpensive method of treating contacts maintains particle concentrations extremely well, even in very small volumes or simulation domains where there are very few particles. Figure 23 shows a plot of the number of particles in a simulation domain over the course of a 0.2 μs simulation. The initial distribution of particles was specifically chosen to be far from the equilibrium particle concentration as a test of the contact region's ability to correct depleted simulation domains. The test structure was a simple lipid membrane with a 4-Å pore and an applied voltage of 200 mV in order to test the ability of the described contact procedure to cope with particle flux associated with drift in an electric field. The particle contacts are set to model a 1 M bulk solution of KCl, which should give a final particle count in the simulation volume of 2143[4] ions.

As can be seen in Figs. 23 and 24, despite the initial undersupply of particles, the count steadily increases until an average value of 2147 is reached. The fluctuation in concentration over the period of the simulation has a final standard deviation of 19.94 particles. The described method, as well as being robust enough to reliably repopulate a simulation, can maintain particle concentrations to within less than 1% of the expected value with fluctuations of approximately $\pm 1\%$.

The overall computational efficiency of a simulation can be greatly increased by using a sensible initial distribution of particles. The results shown in Fig. 23 indicate that approximately two-thirds of the simulation time has been effectively wasted because no IV data can be gathered reliably until the particle concentration stabilizes.

5.4. Current Measurement

The measurement of current in ionic channel simulations can prove problematic. Because of the low mobility of carriers, the drift current corresponding to typical applied voltages in experimental conditions has a substantially smaller magnitude than the thermal noise associated with Brownian motion. In measuring the current in the simulations, a lot of averaging is required to ensure that the mean fluctuation from the thermal motion of the particles tends toward zero. Because the error in data recovered from a Brownian simulation will vary as $\sqrt{N}$, where N is the number of particles, the key is to maximize the number

[4] Based on the solvent accessible volume of 1.78×10^{-24} m^3, a particle concentration of 6.02×10^{26} m^{-3}, and two species of particle, to give 2143 particles total.

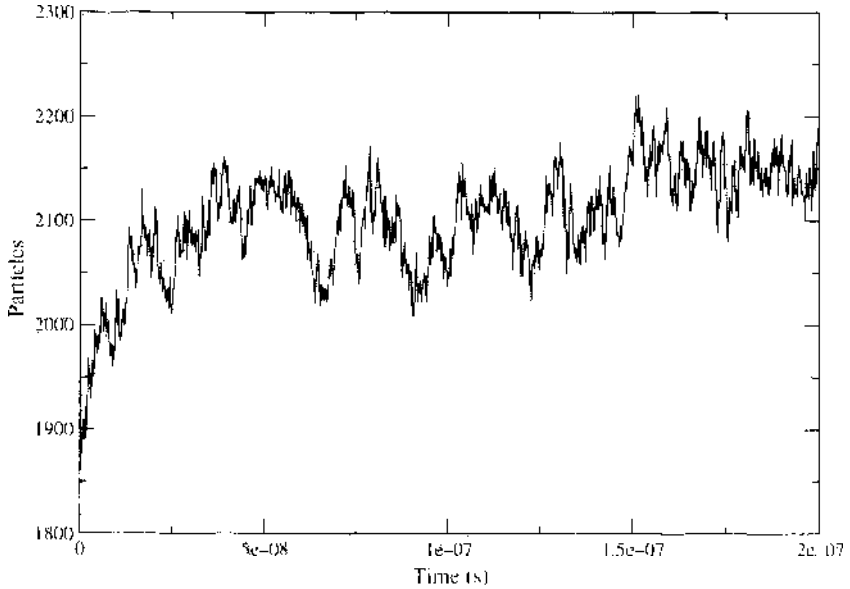


Figure 23. The particle count for a simple 4 Å pore simulation domain. Initially the simulation is given 1800 particles, uniformly distributed through the solvent-accessible regions of the domain. After simulating, the particle numbers reach ~2140, where the expected particle count is 2143.

of particles used to calculate the current. Therefore, there are two potential solutions to this problem: simulate for a very long period of time or simulate using a large number of particles. However, neither of these methods is practical in this case, where the simulation time is limited and the number of particles cannot be increased.

Table 6 gives some values for the type of signal-to-noise ratio (SNR) that can be expected in an ionic simulation current. The table shows a comparison of the displacement at equilibrium due to a linear electric field and the average thermal displacement for K^+ ions at 25 °C. The SNRs shown are the "best case" scenario for a single ion, and although a ratio of ~3.5% is very poor, the best SNR that can be expected from actual simulation results is substantially lower. At equilibrium, the average velocity of carrier ions is expected to be $\mu\mathbf{E}$, but in a complex simulation domain the field is not uniform, nor is the particle flux.

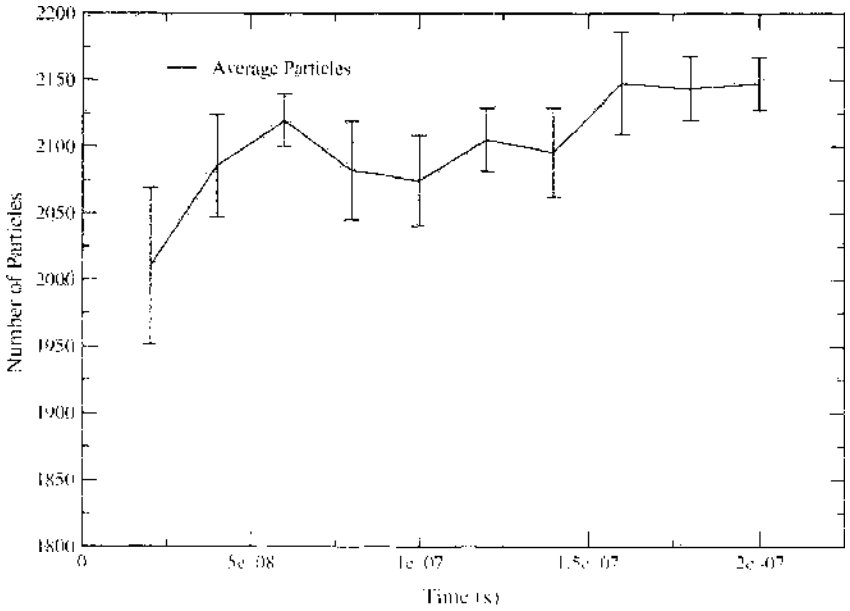


Figure 24. The average particle count for the data shown in Fig. 23. Error bars indicate the standard deviation σ of the fluctuations in particle concentration.

Table 6. Best case particle current SNR calculated for a linear field for a typical potassium ion with a mobility of 7.62×10^{-8} $m^2 V^{-1} s^{-1}$, diffusion coefficient of 1.96×10^{-9} $m^2 s^{-1}$, and time step of $\Delta t = 2ps$. There is a linear increase in the SNR based on the applied field, but at fields comparable to those found in biological problems the SNR is still extremely low.

E_{app} (Vm^{-1})	$\mu E \Delta t$ (m)	$\lvert \Delta X_{flux} \rvert$ (m)	SNR
5×10^6	7.62×10^{-13}	8.854×10^{-11}	0.0086
10×10^6	1.524×10^{-12}	8.854×10^{-11}	0.0174
15×10^6	2.286×10^{-12}	8.854×10^{-11}	0.0258
20×10^6	3.084×10^{-12}	8.854×10^{-11}	0.0348

There are three commonly used methods of current measurement in P3M simulations whose implementations require an increasing degree of code complexity. The first is the method of charge counting, where the current can be defined as

$$I = \partial q / \partial t \tag{42}$$

where q is the charge passing through a defined plane of the simulation in a given time. This method has only the advantage of simplicity, because the implementation of this equation in code is relatively simple. However, the method suffers from the fact that only a small subset of the particles in the simulation are used in the calculation of the current, and therefore the method requires long time averages to minimize noise in the estimation of the current. In addition, because the current is measured in terms of the particle flux through a plane, this method excludes the effect of the displacement current, which, due to the high capacitance of the lipid membrane, can be important in channel simulation.

The second method calculates the flux density through the simulation domain based upon the density and velocity of a carrier species using the expression

$$\mathbf{J} = -znq\mathbf{v} \tag{43}$$

where $\mathbf{J}$ is the flux density in Am^2, z is the valency of the particle, and $\mathbf{v}$ is the average particle velocity. Using this method, the noise in the current is reduced because of the averaging over all particles present in the simulation domain. In addition, it allows the separation of the current due to multiple species of ions, which can be particularly attractive in looking at the selective properties of pores. However, there are drawbacks in the fact that this method will only give a partial measure of the displacement current, since only the portion of the particle drift induced by displacement current can be measured. There are also particular complications involved in calculating the flux computationally because the number density must be calculated only over the volume accessible to carriers. The final result of the computation is a series of spatially distributed vectors, and the vector field requires integration to yield the total current.

The final current measurement technique is the Ramo-Shockley method. This technique is based upon the Ramo-Shockley theorem [73, 74], which was devised in the 1930s to calculate the currents produced in nearby conductors from electron motion in vacuum tubes, and it has been used in many fields within semiconductor research for many years. Later, De Visschere [75] showed that the theory of Ramo is generally valid in a homogeneous dielectric for any number of electrodes at constant potentials. Finally, Kim et al. provided a proof, based on Green's functions [68], that the Ramo-Shockley theorem was applicable in nonuniform dielectric environments and with temporally varying potentials. This proof showed that the Ramo-Shockley theorem can be used to measure the instantaneous current at electrodes in particle simulations of real systems, including displacement currents associated with capacitative coupling between the charge in the simulation domain and the electrodes.

The Ramo-Shockley theorem states that the current due to an electrode potential is separable from the other current contributions in the system by the comparison of the instantaneous particle velocities with a precalculated test function. Therefore, for an electrode j, the instantaneous current $i_j(t)$ can be obtained from the expression

$$i_j(t) = \frac{-\sum_i^N q_i \mathbf{v}_i \cdot \nabla f_j(\mathbf{r}_i)}{V_j} \tag{44}$$

where q_i is the charge on the individual particle i, N is the total number of particles, v_i is the velocity of the particle, $f_j(\mathbf{r}_i)$ is the value of a test function at the particle position, and V_j is the potential of the electrode in the test function. The function $f_j(\mathbf{r}_i)$ is, in the simplest case, the solution of the Poisson equation when the electrode j is set at the potential V_j, all other electrodes in the simulation are set to zero, and all mobile and fixed charges are removed from the simulation domain. It has been shown that more optimized test functions can be used instead of Ramo's original Poisson solution [76], but there is still some doubt remaining about the efficacy of such procedures [77]. The test function is purely a geometry factor that is invariant over time and allows the determination of the particle current uniquely due to the contact j. The operation $\mathbf{v} \cdot \nabla f$ gives the projection of the velocity vector onto the geometric electric field calculated from the test function for the potential V_j, usually chosen to be 1 V for simplicity. The projection gives the proportion of the particle's velocity that is due to the potential applied at solely j because the potential on all other electrodes is set at 0 when f is calculated. Therefore, the sum over all particles gives the current flowing in the simulation domain due to the contact.

The great advantage of the Ramo-Shockley method compared with the charge counting and the flux density methods is that the current calculated for an individual electrode includes the displacement current, in both equilibrium and nonequilibrium conditions. In addition, the use of all particles in the simulation domain means that the fluctuations due to thermal noise are reduced by a factor of $\sqrt{N}$ when compared with the charge-counting method.

In terms of practical implementation, the Ramo-Shockley method has both advantages and disadvantages associated with it. Compared with the algorithmic complexity of the flux density method, Ramo-Shockley is simple, but the need to store a test function for each of the electrodes in the simulation means that there may be a significant memory overhead. The test function storage array should ideally have the resolution of the finest grid in the multigrid solver to avoid errors in the measured current. In addition, there is a one-off computational overhead of calculating the test function for each electrode, but this is small when compared with the overhead associated with evaluating the vector field for the flux density and the integration required to determine the current.

Figure 25 shows 1 ns of a current trace from a typical ionic channel simulation, consisting of ~2000 particles at an applied potential of 200 mV. It is clear that the Ramo-Shockley method dramatically reduces the noise in the measured current. In fact, the Ramo-Shockley method reduces the noise by more than two orders of magnitude at 200 mV. Table 5 shows the SNR that could be reasonably expected for Brownian particles. Figure 26 shows a plot of the actual SNR in the current recovered from a single electrode in a channel simulation, for both Ramo-Shockley and the charge counting method. It can be clearly seen that the Ramo-Shockley method results in a relatively smooth increase in SNR with electrode potential. In contrast, the charge counting method shows almost no improvement in the same circumstances, underperforming the Ramo-Shockley method by more than two orders of magnitude at high applied voltages. This is due to the small number of particles that actually pass through the counting plane at any one time. Clearly Ramo-Shockley delivers a much higher SNR because it utilizes all of the information available from the simulation.

Despite the seeming simplicity of the Ramo-Shockley method, it is possibly the most difficult current measurement technique to implement in a full-scale simulation. Because Ramo-Shockley uses velocities from particles in all regions of a simulation, it is extremely sensitive to errors in either the physical modeling or implementation in code of the particle behavior and electrostatics. It can, in fact, be useful as a debugging tool when it is necessary

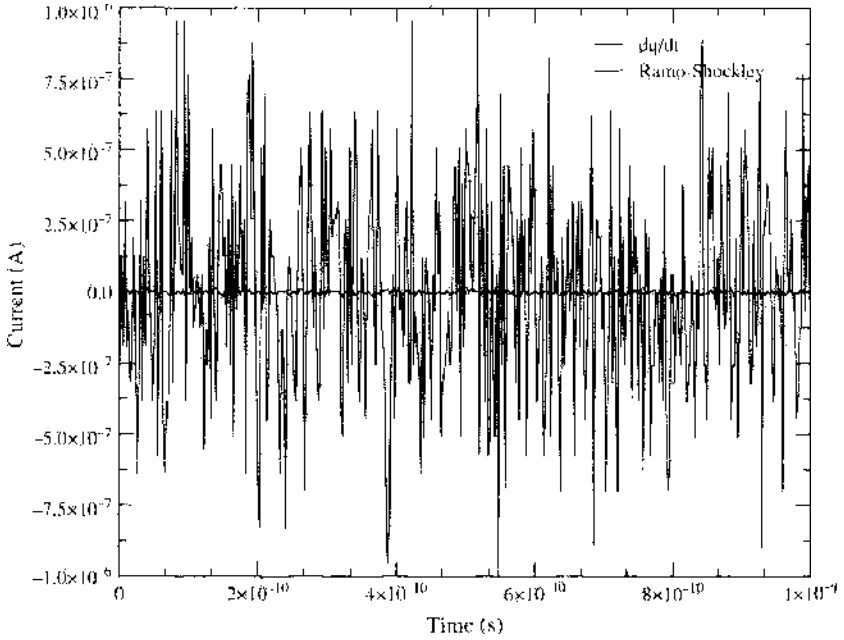


Figure 25. The current measured in 1 ns of a simple channel simulation. The plot shows the current measured by the dq/dt charge counting method and Ramo-Shockley. Note that Ramo-Shockley reduces the noise by more than one order of magnitude.

to test all of the components of a particle simulation simultaneously, in that any bugs from the interaction of previously tested components become immediately obvious. This is why Ramo-Shockley is not as widely implemented as it should be, despite the fact that it has been proven to be an extremely useful and efficient technique.

5.5. Test Results

Before application to channels, the developed simulation methodology and associated software must be tested to verify that it is correctly reproducing the behavior of ions in solution. Therefore, in this section we now apply the self-consistent Brownian technique to the simulation of the bulk behavior of solutions of two species of ions over biologically relevant times.

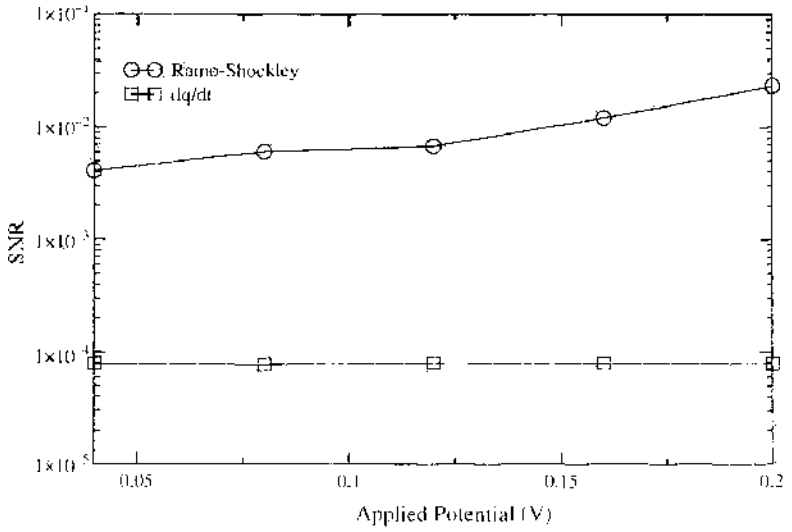


Figure 26. The actual SNR ratio for dq/dt and Ramo-Shockley, calculated using the data shown in Fig. 25. This clearly shows the level of noise reduction possible with Ramo-Shockley.

In any simulation of uniform concentration, Ohm's law must be reproduced by usin the Brownian simulation technique [25, 33, 78]. The correct drift of carriers and recovey of currents has to be checked in simple situations where there are reliable analytical soluions to the problems. The self-consistent modeling of a resistor provides an ideal test cas for the particle behavior, current measurement, and particle mesh coupling.

We have simulated a 50 nm^3 resistor (shown in Fig. 27), measured the current throug the simulation for a particle concentration of 1 M KCl, and then compared the measured curent with the expected values calculated using Ohm's law [41]. Figure 28 shows the IV urve simulated in the resistor structure where the conductivity extracted from the simulatin is within 1% of expected analytical value.

Following the confirmation of Ohm's law, a full test of the simulator's capabilites is required to confirm that not only the drift but also the diffusion of ions is correctly hanlled. To do this, the simulation of the *ab initio* formation of a Nernst potential [1, 79] wasperformed. The Nernst potential is similar to the built-in potential found in semiconductor diodes [41] and is measured using an experimental arrangement shown in Fig. 29. Two aths of ionic solutions of different concentrations are separated by a membrane permeate to only one of the ionic species present. When ions diffuse down the concentration graient across the membrane, a potential builds because the baths are no longer charge neutral. At equilibrium, the drift of particles with the electric field balances the diffusion acros the membrane, giving rise to a steady, measurable voltage difference.

The value of the equilibrium potential for a species of ion can be calculated by usin the Nernst equation

$$V_i = \frac{k_b T}{z_i q} \ln(C_1/C_2) \tag{45}$$

where the potential due to an ionic species i can be calculated based on the concentrtion difference between the two sides of the bath.

The simulation was conducted in a 100 nm^3 box with reflective particle boundaryconditions on all extremities. Dirichlet conditions at 0 V were applied on the right-hanc side of the simulation domain, and all other boundaries were allowed to float. The semipermeable membrane was modeled by defining a plane through the middle of the simulation rom which chlorine ions are reflected, and the simulation was initialized with 1 M KCl soltion in the left-hand bath and 0.1 M KCl in the right. The solutions are then allowed to elax to equilibrium for a temperature of 300 K, which typically takes 1–2 μs of simulation ime.

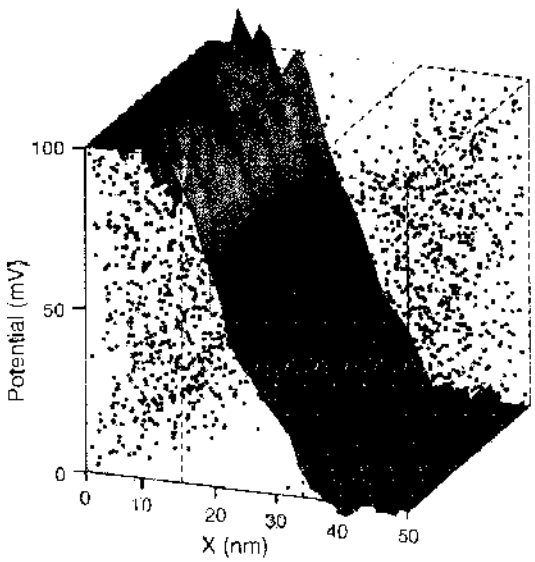


Figure 27. The structure of a 50 nm^3 resistor filled with a 1 M KCl solution. Dirichlet boundaries and prticle contacts are applied at $X = 0$ and $X = 50$ nm; some particles have been removed from the central egon for clarity.

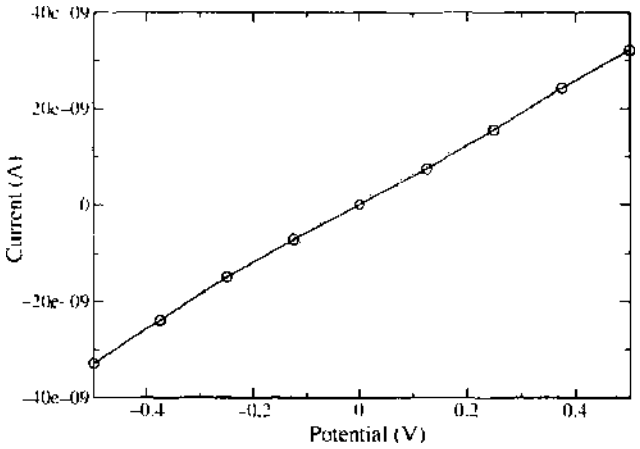


Figure 28. The current-voltage relationship measured in the 50 nm^3 resistor shown in Fig. 27. The conductivity of the solution agrees with the analytical value calculated from Ohm's law to within 1%.

The average potential drop can then be measured across the simulation and compared with the theoretical Nernst potential.

Figure 30 shows the result of the simulation. The equilibrium value of the potential is, on average, 60.2 mV, which compares favorably with the theoretically calculated value of 60.15 mV for potassium. The marginally higher equilibrium potential given by the simulation can be attributed to short-range interactions between ions that are not present in the theoretical model developed by Nernst [79]. We believe that this is the first time that such an equilibrium potential has been simulated by using P3M techniques [25, 32]. This result conclusively shows that the self-consistent Brownian simulation code is capable of resolving drift and diffusion-related ionic behavior on biologically relevant time scales.

The final test required to verify the validity of the simulation software is to check that there are no periodic artifacts being introduced in the behavior of particles, from the particle contact regions, plasma oscillations, random number generators, or elsewhere [53]. If the Brownian particles are behaving correctly, there should be no correlation in the particle motion over long times. This can be checked by calculating the autocorrelation function [80] of the particle current.

Because the particles are coupled by the electric field, artifacts will manifest as periodicity in the simulation current. Figure 31 shows the autocorrelation function of the particle current in a simple resistor simulation. As can be clearly seen, the only correlation in the particle motion occurs at a phase of 0 samples, proving that the noise in the simulation is indeed random. Had plasma oscillations or large errors due to particle injection been present, they would be instantly evident in such a plot.

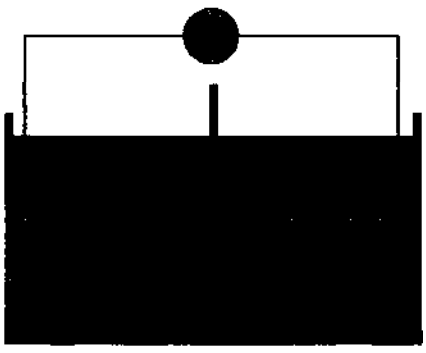


Figure 29. The Nernst potential is measured by placing two electrolytic solutions of concentration C1 and C2 on either side of a membrane that is permeable to only one species of ion in the solution. The equilibrium potential that develops can then be measured as a voltage drop across the membrane.

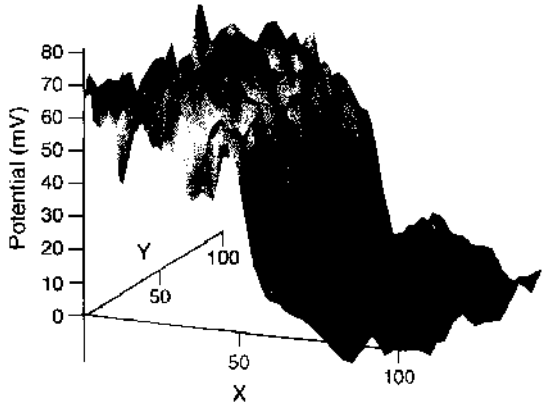


Figure 30. The equilibrium Nernst potential generated by the *ab initio* simulation of 1 M–0.1 M KCl.

We have discussed the practicalities of implementing and optimizing the self-consistent Brownian simulator and have presented a novel method for the maintenance of particle contact regions that performs extremely well, even with very low numbers of particles present in the simulations. It is now time to apply these simulation techniques to the specific problem of ion conduction in subnanometer scale pores and ion channel proteins.

6. SIMULATION RESULTS

In this section we present the results of Brownian simulations of model ion channel structures and compare them with drift diffusion (DD or PNP) simulation to further verify the simulation methodology and highlight some of the differences between continuum and particle-based simulations.

Having shown in the previous section that the self-consistent Brownian methodology is capable of accurately reproducing the behavior of ions in solution, we extend these simulations to include the lipid membrane and the structure of the channel protein itself.

Using a simplified channel structure for both the gramicidin A and Kcsa proteins, we first simulate the conduction of ions through these narrow (2–4 Å) pores by using the commercially available semiconductor drift diffusion simulator Taurus [81]. The channel structures are generated from molecular surfaces calculated by using the open source MSMS [82] software. These structures are prototyped and tested in Taurus, applying region-specific mobility modifications deduced from MD simulations of the behavior of ions in constrained pores.

The simple structure capable of replicating the current-voltage (IV) characteristics of channels, using appropriate mobility scaling, is then incorporated in the Brownian simulation

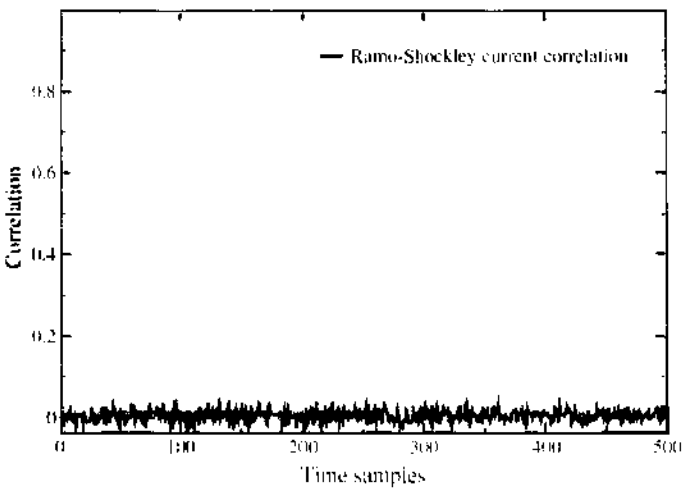


Figure 31. The particle current autocorrelation function for a simple resistor simulation.

code. The initial test of the Brownian simulator aims to determine to what extent the methodology described in Section 3.1 agrees with the continuum DD solutions obtained from Taurus. For this purpose, statistical enhancement based on particle subdivision is used. In the limit of an infinite number of particles, the Brownian dynamics should be considered to be an exact solution of the drift diffusion equation, and the results should accurately reproduce those from the Taurus simulations.

Finally, the Brownian simulation code with single charged particles representing each ion is applied to the model channel structures, and the results of these simulations are discussed in order to directly study the effects associated directly with the discreteness of the ion charge.

6.1. Ion Channel Structure

The structures of only a handful of all of the channel proteins known have been determined to within a reasonable degree of accuracy. The problems associated with accurately measuring the structure and conformation of proteins are well known and will not be discussed here. Fortunately, the coordinates of the atoms in several channel proteins are currently known [8, 18, 83, 84]. Although there is always some degree of uncertainty as to what extent these configurations remain unaltered when the protein is solvated, the available structural information is considered reliable by the majority of the ion channel simulation community.

For the purposes of demonstration, we focus on two of the best-described protein structures: the antibiotic gramicidin [51, 83, 85] and the Kcsa bacterial ion channel from *Streptomyces lividans* [8, 18, 86].

Figure 32 shows the structure of the gramicidin A (gA) protein, taken from the 1MAG structure in the protein database [84]. The gA protein has an α helical structure and is composed of two 15-residue subunits. It is a good starting point for the testing of ionic channel simulations because of its simple structure and the large body of experimental data available for the pore's conduction characteristics [21, 83, 87].

Figure 33 shows the molecular surface—which determines the particle-inaccessible boundary in the simulation domain—of gA from two perspectives. These surfaces were generated using MSMS [82] with a sample point density of 3 per Å and a probe radius of 1 Å, chosen as an intermediate value between the approximate radius of Na and K ions and

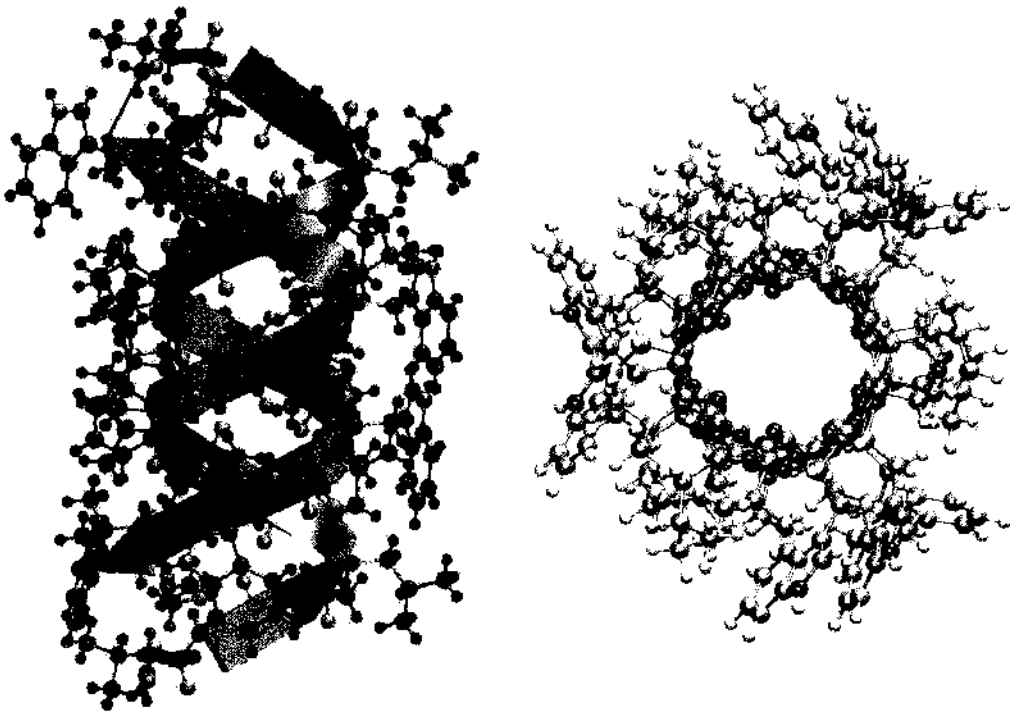


Figure 32. The structure of gramicidin A from 1MAG (Protein Database). The α helical backbone is shown in the left-hand plot. On the right is a view down the pore axis, showing the regularity of the structure.

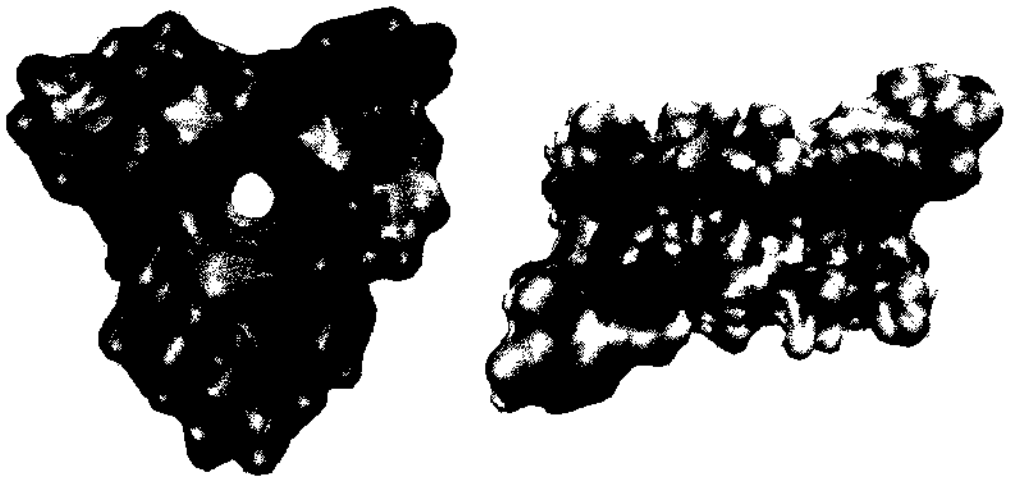


Figure 33. Molecular surfaces for two projections of the gramicidin molecule (1MAG). The pore with a diameter of 4 Å is shown on the left, and the surface is colored using the type of atom underlying that portion of the plot. On the right a molecular surface of the cross section of the gramicidin channel is shown; in this case the surface is colored by atom partial charges to show the partially charged O and N atoms inside the pore. Surfaces were generated using MSMS with a probe radius of 1 Å and a sample density of 3.0 per Å.

water. Unless otherwise stated, all molecular visualization was performed using the Visual Molecular Dynamics (VMD) software available from the University of Illinois [88]. The generation of the molecular surface allows the measurement of the approximate radius of the gA pore in this configuration and gives a value of ~2 Å. These approximate measurements are used to create a simple simulation structure that can be used to test further the self-consistent Brownian simulation technique previously described.

The Kcsa potassium channel from *Streptomyces lividans* was chosen as the primary test structure, because, at the inception of this research, it was, structurally, the best-described naturally occurring ion channel protein [15]. Moreover, a putative open structure for the Kcsa molecule has recently been discovered [18, 86], which appears to confirm the predictions of the open-pore conformation obtained from molecular dynamics simulations [16, 17]. However, there is relatively little experimental information available on the conduction characteristics of the Kcsa protein [9, 11] on account of the difficulty of working with Kcsa under controlled experimental conditions.

As Fig. 34 shows, the structure of the Kcsa channel protein can be divided into two regions of importance for the ionic conductance: the very narrow pore/selectivity filter and the much wider, water-filled lumen region. The pore itself is extremely narrow, with a radius of between ~2 and 3 Å, whereas the lumen is substantially wider with a roughly square cross section of dimensions ~38 Å × 26 Å. The MD simulations have shown that the partially charged oxygen atoms along the pore interior create a particle ladder occupied by alternating water and potassium ions, which helps maintain the structure of the pore and increase the conductivity when compared with an uncharged structure.

Figure 35 shows molecular surfaces generated with a MSMS probe radius of 1.2 Å (the approximate radius of a potassium ion [1, 43] reduced slightly to take into account the apparent overestimation of the Pauling radius of oxygen [89]). As is shown in Fig. 35, the pore is extremely narrow, just wide enough to permit ions and water to enter. The potassium "ladder" can also clearly be seen in Fig. 35.

6.2. Simulation Using Technology Computer-Aided Design Tools: The Simple Channel Model

Drift diffusion simulators can be used in commercial technology computer-aided design tools (TCAD) for the simulation of ion conduction through pores [24, 25]. TCAD tools allow the user to perform simulations while minimizing development and testing times. In this

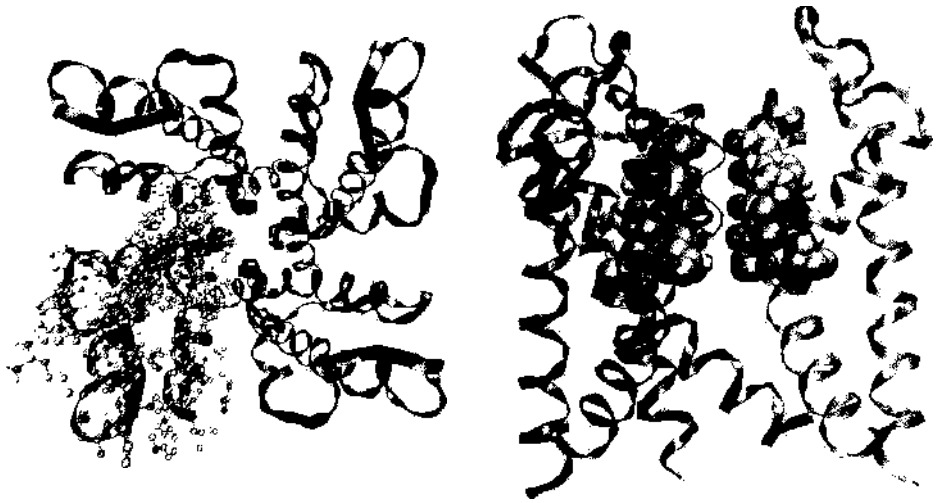


Figure 34. The open conformation of the tetrameric channel protein, Kcsa from *Streptomyces lividans*. On the right, a detail of the internal pore structure is shown. This region has a radius of approximately 2 Å and is lined with partially charged carbonyl oxygen atoms.

section, we discuss the modification of the commercial simulation package Taurus [81] for the purpose of modeling the behavior of ions in solution and pores.

Taurus is a flexible semiconductor device simulator capable of solving the coupled Poisson and current continuity equations' DD approximation, in complex 3D structures with variable material parameters. Taurus allows the user access to almost all of the physical parameters used in the simulation, such as material dielectrics and mobilities. It is therefore relatively simple to adjust the parameters for semiconductor materials to match those in biological simulations. In addition, Taurus allows the user to change the values for physical parameters in the different regions of the simulation domain, considering that the mobility of an ion traversing a channel protein is unlikely to remain constant [14, 21]. This feature is essential.

The default material parameters in Taurus were modified to mimic the materials relevant to channel simulation: water, lipid, and protein. Table 7 shows the material parameters used in the simulation in comparison with the default values as defined by Taurus for the selected

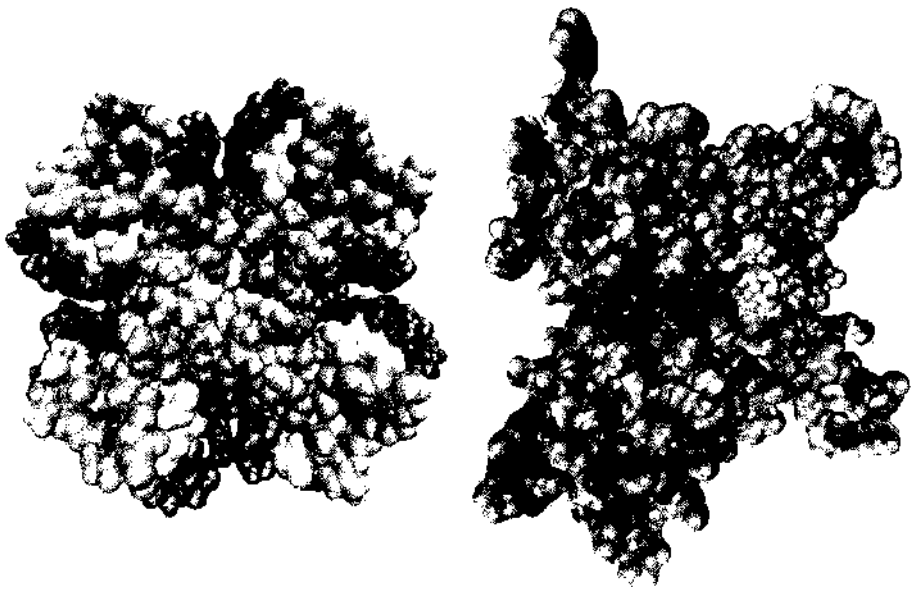


Figure 35. The molecular surface of the Kcsa channel in a putatively open state. Again, a detail of the pore structure is shown, with the carbonyl oxygens shown in red. Note that the selectivity filter potassium-binding sites reported by Sansom and Capener can be clearly seen. Surfaces were generated using MSMS with a sample density of 3.0 per Å and a probe radius of 0.9 Å.

Table 7. Redefinition of the Taurus material dielectrics to mimic the components of an ion channel simulation.

Taurus material	Taurus	Defined	Biological material
Silicon	11.9	80.0	Water
Silicon dioxide	3.9	2.0	Lipid
Silicon nitride	7.5	20.0	Protein

materials. The original mobility of electrons in silicon is ~0.14 m^2/Vs and the mobility of holes is ~0.07 m^2/Vs [41]. These values are several orders of magnitude too high to be used to model the motion of ions where the mobility is on the order of 10^{-8} m^2/Vs [1]. Therefore, the modeling of ions requires the adjustment of the mobility parameters.

Hess et al., [24] have previously shown good agreement between TCAD-based DD simulations and measured ion channel characteristics. However, this requires the rescaling of all parameters involved in the simulation with the incumbent potential for errors. With Taurus, it is possible to simply adjust those parameters requiring modification.

Figure 36 shows the simplified channel structure used in Taurus to simulate conduction through the gramicidin A protein. On account of its simple secondary structure, consisting of two alpha helices, gA has been modeled as a simple hole of appropriate length and cross section, in dielectric materials replicating the protein and lipid membrane [90].

The aim of the gA simulation is to build confidence that calibrated drift diffusion simulations using TCAD tools can replicate the gross conduction properties of ions in small pores and not to make any structure/function predictions based on the more detailed study of ion behavior in real pore structures [24]. In these simulations, the mobility of the sodium ions is fixed throughout the entire simulation domain to a value of 6×10^{-10} $m^2V^{-1}s^{-1}$, as suggested in Ref. [24]. Interestingly, this mobility is somewhat lower than the generally accepted bulk mobility for sodium of 5.19×10^{-8} $m^2V^{-1}s^{-1}$ and has been used as a fitting parameter that matches simulations to experimental measurements. The ionic solution concentration is modeled by doping the silicon region representing water with a donor concentration equivalent to 1 M NaCl.

Figure 37 shows the results of the simulation of the simplified gramicidin structure obtained from Taurus. As can be seen, the Taurus IV characteristics are remarkably close to the values of the ISE-TCAD simulations of Hess et al. [24] and agree well with the experimental measurements. The Hess group has more accurately included the atomic structure of the protein by assigning the partial charges associated with the side chains lining the pore of gramicidin to the nodes in their simulation mesh. However, the Taurus simulation

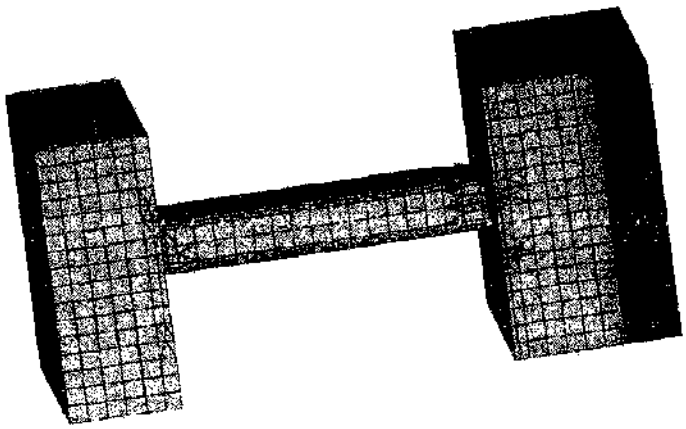


Figure 36. A simplified gramicidin A structure mesh created by Taurus. The protein and lipid materials have been removed to show the pore. The pore length is the same as in gramicidin (~3 nm), and the cross-sectional area is also approximately replicated.

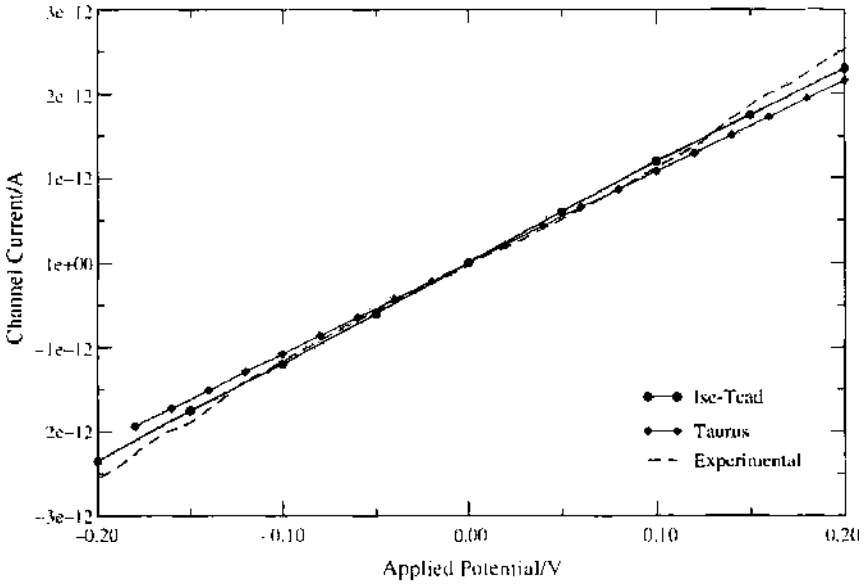


Figure 37. The simulated current through the simple structural model of gramicidin A. The data are compared with experiment and the results of the simulations by Hess et al. [24].

in our simplified structure is able to closely replicate the conductivity of the pore without the inclusion of these charges.

Having shown that TCAD tools can successfully simulate conduction through simple structures of the scale of ion channels, we move now to the simulation of more complex ion channels.

The idealized simulation domain used to represent the protein structure of the Kesa channel is shown in Fig. 38. This simple simulation domain forms the basis for the rest of the simulations presented in this chapter. The discrete mesh can be generated, in Taurus, by defining blocks of materials and specifying initial mesh point positions where the resultant mesh can be refined automatically. Alternatively, the Taurus fabrication process models can be used to etch and deposit the relevant materials and the automatic mesh generation can be allowed to build the simulation domain. Each of these methods will, eventually, generate a usable simulation structure. The mesh initialization through block specification is substantially quicker, but process modeling tends to generate less rigidly artificial shapes (smooth corners and the like).

In the continuum simulation of channel structures, it is necessary to take into account the change in the behavior of ions in constrained pores. Tieleman et al. published a review [14] that, based on data from numerous MD simulations of different pores from many published sources [91], produced an approximate guide to the change in particle diffusion

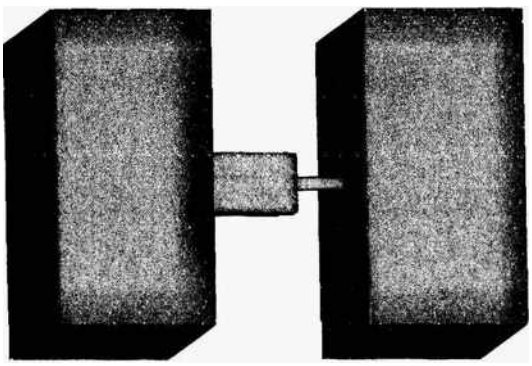


Figure 38. The structure of the Kesa protein representation in Taurus. The dimensions have been chosen to be consistent with the open structure of the channel protein. Note that the protein and lipid have been removed to show the structure of the pore and lumen regions.

coefficient with the radius of the constraining pore, as shown in Fig. 39. The simulations implemented, for example, by Hess et al. [24] and others always appear to require mobility and diffusion coefficient scaling in order to replicate experimental data. In the past, this has been done in the calibration process by fitting the mobility in order to achieve a good agreement with the experimental results. However, there is now a sufficient amount of data from MD simulations indicating a drop in mobility caused by the change in external forces on the particle, associated with their permeation through constrictions in the pore.

In our simulations of Kcsa, where the minimum pore radius is ~2 Å, a mobility scaling of 0.2 in the selectivity filter was used as an *a priori* assumption based on the data presented in Fig. 39. Also, a scaling factor of 0.6 has been adopted for the lumen region of the channel (based on a radius of ~6). Interestingly, at the time of the gA simulations, these data were not available, but preliminary Brownian simulations have shown that a mobility scaling of between 0.2 and 0.1 is required for the basic model to replicate the actual conductivity of the gA pore. This indicates that the MD simulation predictions of diffusion coefficient and mobilities are giving values with a reasonable degree of accuracy.

The results of the continuum simulations of the simplified Kcsa structure, for different ionic solution concentrations, are shown in Fig. 40. It is clear that the simulations that implement the mobility scaling, previously described, accurately replicate the conduction of potassium through the pore constriction. In these simulations, the conduction of Cl ions is completely ignored because, assuming that the selectivity of Kcsa for K over Cl is perfect, the equilibrium Cl current is zero.

Figure 41 shows a more detailed view of the positive voltage part of the characteristics for 1 M and 0.5 M KCl concentrations. Although there is very good agreement with the published experimental curves at low applied potentials, at high voltages, simulated IV characteristics deviate from the experiment, with the differences being more pronounced at solute concentrations higher than 1 M.

The deviation of simulation from experiment can be related to the discrete nature of the ions. Unfortunately, the single filing of discretely charged particles, with finite size, through a narrow pore cannot be represented by the continuum DD methodology.

The marked drop in channel conductivity shown in Fig. 42 at 1.5 M is somewhat more puzzling and also points to some discrepancy in the experimental measurements. The concentration-conductance relationship for Kcsa has been measured up to 1.6 M K concentration [9] and shows no indication of this sudden drop in conductivity above 1.0 M. Although this conductivity drop may be due to saturation of the pore and increased single filing, questions remain about the accuracy of the measured data, in that LeMasurier, Heigenbotham, and Miller [9] reported that it was extremely difficult to measure the IV curve of Kcsa above 1 M. At the same time, a comparison of the data gathered by Meuser et al. [11]

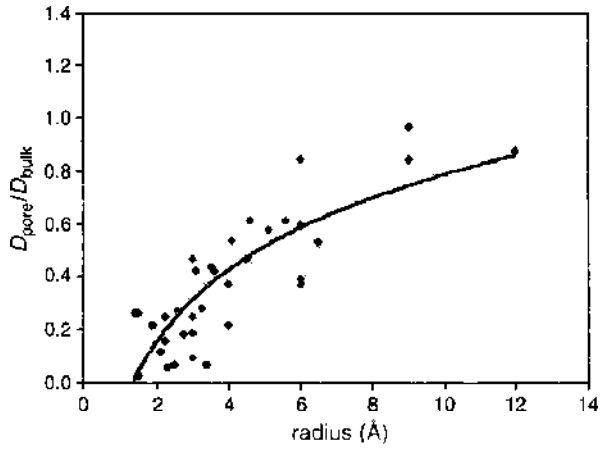


Figure 39. Ion diffusion scaling in constrained pores, replicated from Tieleman et al. [14]. The ratio of bulk diffusion coefficient to pore diffusion coefficient is shown relative to the radius of the constraining pore. The solid line has no meaning other than as a guide for the eye.

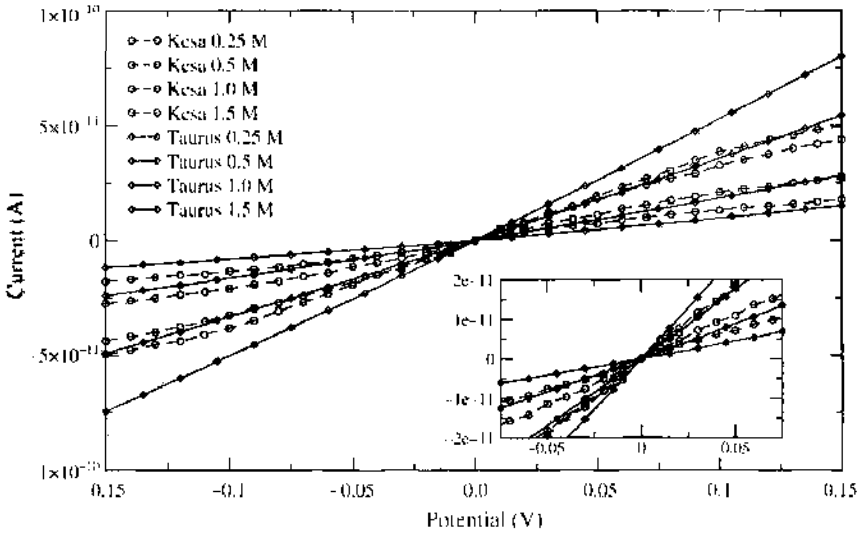


Figure 40. Taurus simulation of the simplified Kcsa structure with KCl solution. Currents shown are forward-biased simulation and experimental curves from Meuser et al. [11]. Inset shows a detail of the current at low applied potential. At very high potentials and concentrations, the continuum model appears to break down.

and by LeMasurier, Heigenbotham, and Miller [9] indicates that the forward-biased currents measured by both groups are very similar in the ~0.5–1 M concentration range.

The current rectification displayed in the results of LeMasurier, Heigenbotham, and Miller [9] is not available in the Meuser data in that they present the IV characteristics based on the conductivity of the fully open-channel state. Given the symmetry of the curves, under the reverse-biased conditions, it is likely that the channel is simply embedded in the membrane in reverse. Because the simulations use a static model of the pore, the intermediate conduction states shown in the patch clamp data for Kcsa cannot be replicated, making the Meuser data a better reference for the model, in that the data from LeMasurier, Heigenbotham, and Miller will show a slightly lower overall conductivity because of the inclusion of the intermediate conduction states.

We have shown that it is possible to replicate some of the conduction properties of the Kcsa and gramicidin A ion channels by using continuum DD simulations, simple structural models, and commercially available technology computer-aided design tools. Despite the limitations inherent to continuum methods in simulating discrete systems, they are extremely useful in the simulation of ion channels [20, 21, 24, 66, 92, 93]. The computational efficiency and the ready availability of commercial simulation software mean that

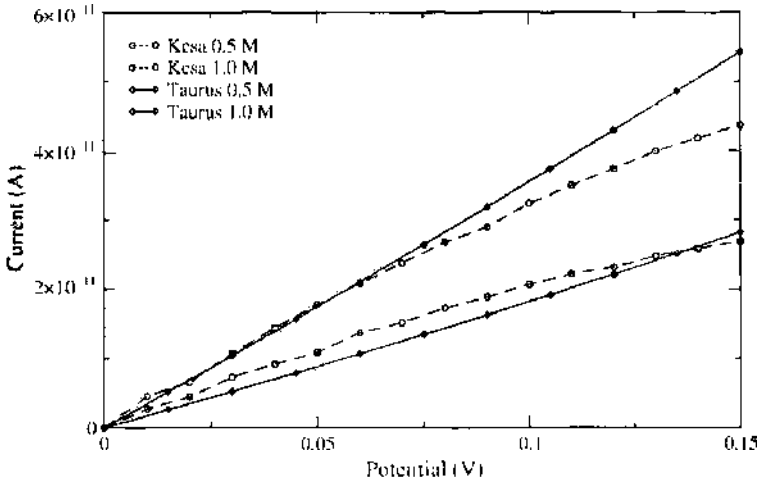


Figure 41. Kcsa simulation currents for 1 M and 0.5 M KCl solutions are shown. The currents are extremely close to experimental values but do not replicate the rolloff in conductivity at high potentials.

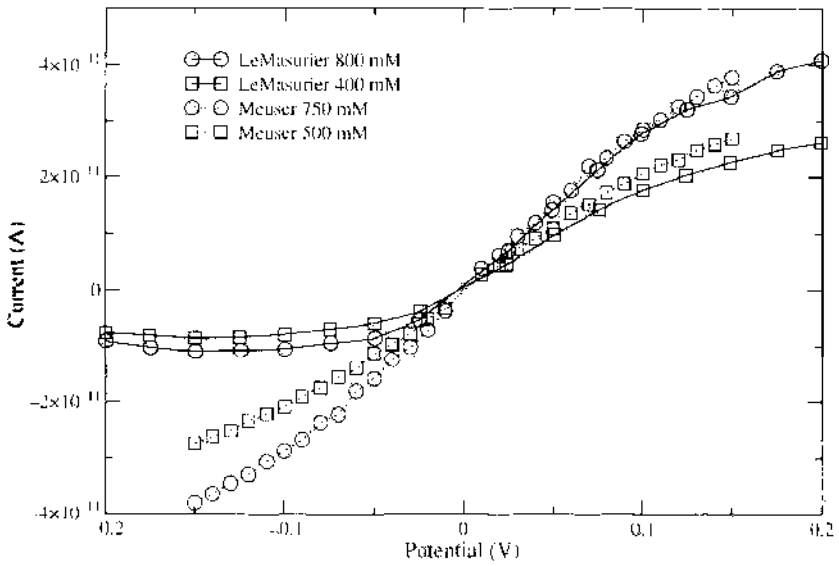


Figure 42. A comparison of the Kesa IV curves by LeMasurier and Meuser for potassium concentrations where the behavior of the channel is reliable. Note that the LeMasurier data show the rectification of the K current.

drift diffusion/Poisson-Nernst-Plank simulations of ion flux remain an attractive alternative to particle simulation. However, to remedy some of the artifacts associated with continuum simulation and to more realistically treat the processes involved in ion permeation and selectivity, it is necessary to consider ions as discrete objects and charges in particle-based simulations. Therefore, we compare the results presented in this section to the self-consistent Brownian simulation techniques utilizing the simulation methodology and tools described in the previous sections.

6.3. Brownian Simulation

Previous Brownian dynamics simulations of ion transport through channels have lacked self-consistency (including the treatment of long-range electrostatic effects that comes from the solution of the Poisson equation [12, 27–29]) in realistic simulation domains. Here we present simulations of model Kesa structures that include a fully self-consistent solution of the Poisson equation.

Our Brownian simulation method is a direct discrete analog of the continuum drift diffusion equations. By using the concept of the superparticle [94], it is possible to linearly scale the charge on a Brownian particle, provided that the number of particles in the simulation domain is modified by the inverse of the scaling factor and the total charge in the system remains constant. In a similar fashion to the ensemble Monte Carlo technique used in semiconductor device simulations, which is a method of solving the Boltzmann transport equation [94], the Brownian simulation approach can be considered as a method for solving the drift diffusion equation in the limit of an infinite number of "microparticles" with equivalent fractional charges.

By using a charge scaling factor of less than one, the number of particles in the Brownian simulation can be increased by the reciprocal of the scaling. This method of using "microparticles" reduces the noise in the simulation, making it easier to recover small currents, and as the scaling factor tends toward $1/\infty$, the self-consistent Brownian methodology becomes the solution of the continuum drift diffusion equation. This statistical enhancement of the solution methodology allows a direct comparison between the results generated in Taurus and the Brownian code, with a corresponding reduction in the resolution of discrete particle interactions.

Figure 43 compares the Brownian simulator, Taurus, and experimental data for 1 M KCl solution. In both simulations, the mobility scaling of 0.2 in the pore region and 0.6 in the lumen is used. A scaling factor of 0.1 is applied to the Brownian particles, increasing the total

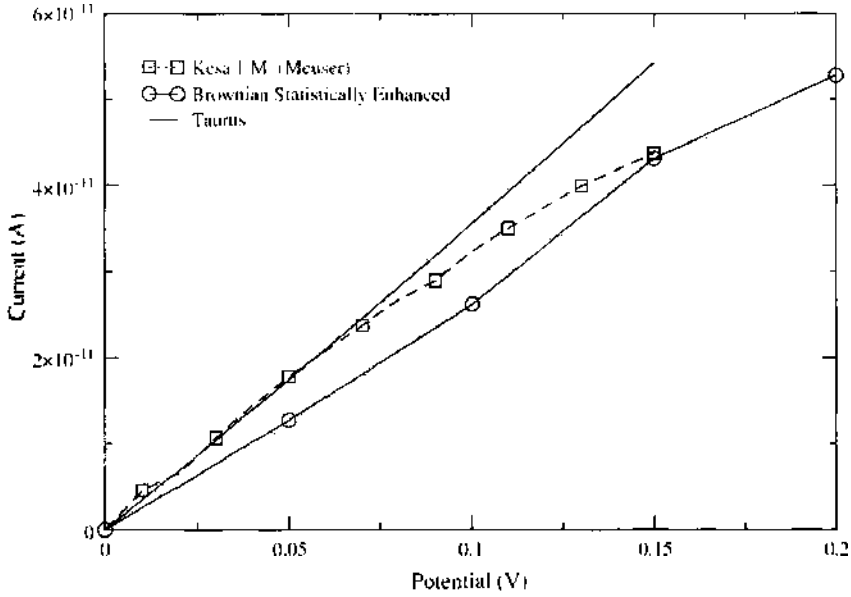


Figure 43. The statistically enhanced "microparticle" simulation of the simplified Kesa structure compared with both experiment and Taurus. A scaling factor of 0.1 is applied to the particle charge, increasing the number of particles by 10. In both cases, the mobility scaling is detailed in Section 6.2. Note that the Brownian simulation still differs from the continuum simulation method in that it still retains some ability to resolve discrete particle interactions.

number of particles from ~3000 to ~30,000, decreasing the noise in the simulation by an average factor of 3.16, and moving the solution closer to the continuum method. In addition, the increased number of particles lowers the overall simulation time required to reliably recover the current for a given applied potential at a minimally increased computational overhead. Simulations were run with a time step of 2 ps for 1 μs, requiring a total real simulation time of approximately 72 hours for 5 IV points.

As can be seen in Fig. 43, the enhanced Brownian simulation agrees well with the drift diffusion simulation from Taurus. Note, however, the reduction in channel conductivity that occurs above 0.15 V. This is due to charge/volume effects, sometimes known as Coulomb blockade, known to be of importance in ion conduction [8, 14].

Figure 44 shows the results of the Brownian simulation of the Kesa structure for 1 M KCl, using ions with exact charge and size. The overall drop in channel conductivity and the sublinear dependence at high applied potentials is primarily due to single filing in the pore. Figure 45 clearly shows this single filing of particles in a small pore. The simulations were run with a time step of 2 ps for 1 μs, taking approximately 68 hours for 5 IV points on a 500 MHz Mips R14K silicon graphics system.

The overall drop in channel conductivity shown in Fig. 44 can be attributed to two factors:

1. Due to the atomistic nature of the Brownian simulation, the electrostatic repulsion caused by the presence of an ion in the pore ensures not only the single filing of charges through the pore but also a correlation in the arrival of an ion in the high field region close to the pore mouth, which becomes less likely when the pore is occupied, thus reducing the conductivity of the channel by reducing the supply of particles at the pore mouth. This Coulomb blockade is also directly responsible for the nonlinear current voltage dependence shown in Fig. 44 because the supply of particles to the mouth of the pore now becomes a field- and occupancy-dependent problem.
2. The choice of mobility for the pore and lumen regions has been based on the data presented in Fig. 39. The data used to generate this dependence have been acquired from molecular dynamics simulations of real protein structures with realistic concentrations of electrolytes. The accurate representation of the protein structure, along with the requisite partially charged residues, means that binding sites inside the pore are generally occupied by the ion specific to the particular channel protein. These are

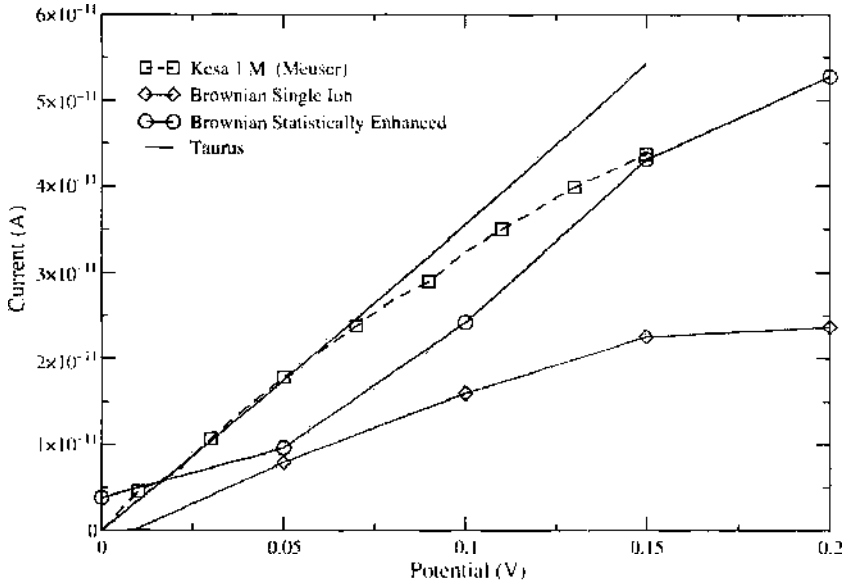


Figure 44. Single particle Brownian dynamics simulation of 1 M KCl in the simple Kesa structure. Note tht the Brownian simulation shows an excellent shape for the IV curve but that the magnitude of the current is reduced by approximately a factor of two when compared with both the Taurus and statistically enhanced Brownian simulaions.

therefore inherently multi-ion simulations containing, to a great extent, the effet of the single filing of ions in these pores. This leads to the conclusion that the overall drop in the measured current in the Brownian simulations is probably due to the "double counting" of this effect.

If the mobility scaling parameters that gave such good agreement in the continuum smulations are artificially low because the double counting of single filing, then the linear scling of the Brownian simulation currents will remove the effect of this underestimation o the actual particle mobility. Figure 46 shows the result of simply scaling the current fron the 1 M KCl Brownian simulation. The agreement with experiment is excellent. By multipying the current from the simulation by a factor of 1.8 and 1.61, the data match almost exactl the data from Meuser et al. [11] and LeMasurier, Heigenbotham, and Miller [9], respectivly.

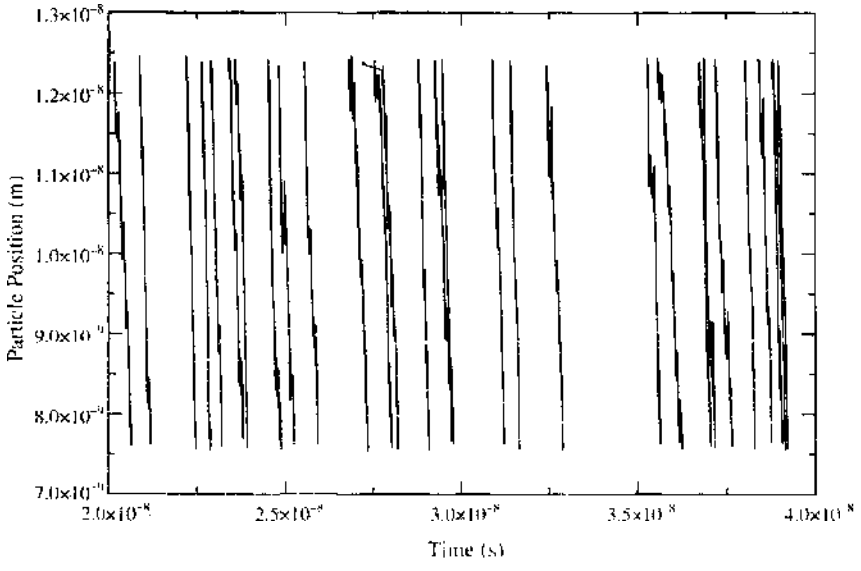


Figure 45. The single filing of charges in a 4 Å diameter pore. In this case, the pore length is increased to 32 nm to better demonstrate the effect.

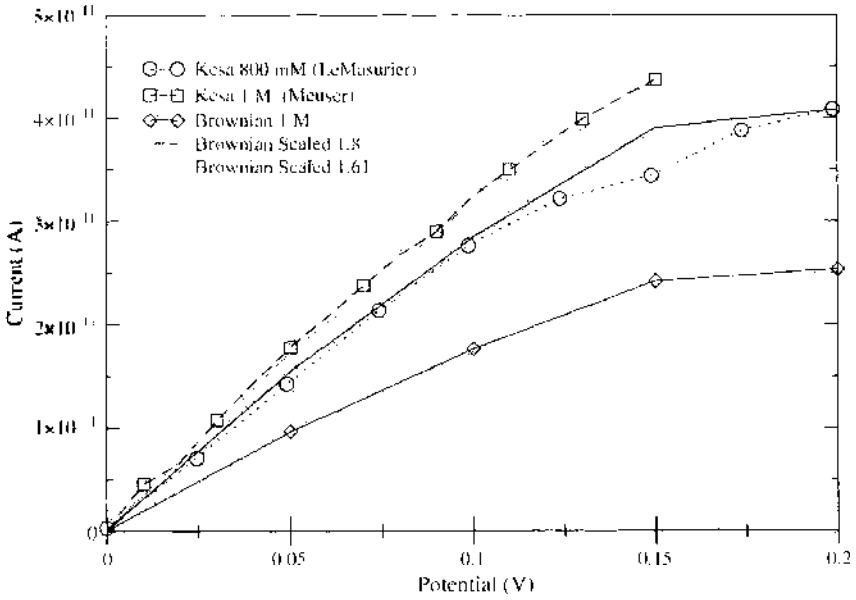


Figure 46. By scaling the current measured in the Brownian simulation of the Kesa structure, remarkable agreement is achieved with experimental values from both sources. The data were calibrated by a factor of 1.8 to match the data from Meuser et al. [11] and by 1.61 for the 800 mM data from LeMasurier et al. [9]. Note that the shape of the curve is an extremely good match in both cases. This would indicate that the Brownian simulation is accurately modeling the majority characteristics of ion conduction in this molecule.

Because this calibration of the current is performed linearly, this result indicates that the Brownian simulation code is accounting for the multi-ion interactions that are occurring in the biological systems. Some work has been specifically dedicated to the behavior of ions in constrained situations such as those found in ion channel pores [91, 95]. However, these are invariably multi-ion simulations. In order that mesoscopic simulation techniques, such as Brownian methods, can be correctly parameterized to accurately predict channel conduction properties, the fundamental properties of ions in small pores must be properly understood.

7. WHAT NEXT?

Obviously, the most pressing modification that is required is the inclusion of real protein structures. Once a generic method of incorporating a structure from either protein database [85] structural definitions or from other sources such as Gromacs topologies [96–98] is in common usage, the application of Brownian dynamics to channel problems should become substantially more common. This realistic structural information is essential to extend the predictive ability of the Brownian approach to include selectivity and proper current rectification. In addition, the inclusion of protein structures will make available the partial charges on all of the protein atoms and will therefore allow direct comparison with existing MD simulations and ultimately the simulation of realistic proteins.

Before the Brownian methodology can be viewed as complete, some aspects require further clarification. For example, there is currently no method of directly dealing with the solvation/desolvation of ions as they enter very narrow pores. The energy required to desolvate an ion in water is known [43], but without detailed knowledge of the actual geometry of the hydration shell, it will be difficult to predict the exact behavior of an ion under specific circumstances. There is also the question of whether the Wien effect [99, 100] will play a role in ion conduction, in that the electric field close to the mouth of the pore can reach magnitudes well beyond those required for an ion to display an anomalous increase in mobility.

At some point, it will be necessary to attempt MD simulations of ion and water behavior in the pores of channels so that physical parameters can be determined to better parameterize Brownian simulations. Additionally, an investigation of the formation of the electrical double layer known to exist at the lipid–solution interface [100] will need to be instigated. Because

ions are intimately involved in the structure of the double layer, it is possible that this may in fact be the particle source for the majority of ions supplied to the pore, rather than the bulk solution, as has been thought up till now. However, the double layer has been remarkably resistant to theoretical analysis, and more data, on its configuration and the behavior of particles in this region, are required before an estimate of its effect, if any, on ion conduction can be made.

In the meantime, there remains a vast amount of original science left to do in the field of biosimulation, for which this chapter has hopefully whetted your appetite.

REFERENCES

1. B. Hille, "Ion Channels of Excitable Membranes." Sinauer, Sunderland, MA, 2001.
2. M. S. P. Sansom, private communication.
3. F. M. Ashcroft, "Ion Channels and Disease." Academic Press, San Diego, CA, 2000.
4. P. Fromherz, in "Technical Digest of the International Electronics Devices Meeting." IEEE, San Francisco, 2001.
5. P. Fromherz, *Chem. Phys.* 3, 276 (2002).
6. P. Fromherz, *Phys. Rev. E* 16, AOS, New York (2003).
7. A. Asenov, A. R. Brown, J. H. Davies, S. Kaya, and G. Slavcheva, *IEEE Trans. Electron Devices* 50, 1837 (2003).
8. M. S. P. Sansom, I. H. Shrivastava, J. N. Bright, J. Tate, C. E. Capener, and P. C. Biggin, *Biochimica et Biophysica Acta* 1565, 294 (2002).
9. M. LeMasurier, L. Heigenbotham, and C. Miller, *J. Gen. Physiol.* 118, 303 (2001).
10. D. A. Doyle, J. M. Cabral, R. A. Pfuetzner, A. Kuo, J. M. Gulbis, S. L. Cohen, B. T. Chait, and R. Mackinnon, *Science* 280, 69 (1998).
11. D. Meuser, H. Splitt, R. Wagner, and H. Schrempf, *FEBS Letters* 462, 447 (1999).
12. T. W. Allen, M. Hoyles, S. Kuyucak, and S. H. Chung, *Chem. Phys. Lett.* 313, 358 (1999).
13. D. Gillespie, W. Nonner, D. Henderson, and R. S. Eisenberg, *Phys. Chem. Chem. Phys.* 4, 4763 (2002).
14. D. P. Tieleman, P. Biggin, G. R. Smith, and M. S. P. Sansom, *Quart. Reviews of Biophysics* 34, 473 (2001).
15. H. Schrempf, O. Schmidt, R. Kommarlen, S. Hinnah, D. Muller, M. Setzlar, T. Steinkamp, and R. Wagner, *EMBO J.* 14, 5170 (1995).
16. I. H. Shrivastava and M. S. P. Sansom, *European Biophys. J. with Biophys. Lett.* 31, 207 (2002).
17. P. C. Biggin and M. S. P. Sansom, *Biophysical Journal* 83, 1867 (2002).
18. Y. Jiang, A. Lee, J. Y. Chen, M. Cadene, B. T. Chait, and R. Mackinnon, *Nature* 417, 523 (2002).
19. K. Cooper, E. Jakobsson, and P. Wolynes, *Prog. Biophys. Molec. Biol.* 46, 51 (1985).
20. D. Chen, J. Lear, and B. Eisenberg, *Biophysical Journal* 72, 97 (1997).
21. U. Hollerbach, D. P. Chen, and R. S. Eisenberg, *J. Scientific Computing* 16, 373 (2001).
22. T. A. van der Straaten, J. M. Tang, U. Ravaioli, R. S. Eisenberg, and N. Aluru, *J. Comp. Elec.* 2, 231 (2002).
23. M. Saraniti, MCM 2003, Sept. 15–17, Berlin.
24. K. Hess, U. Ravaioli, N. R. Aluru, M. Gupta, and R. S. Eisenberg, *VLSI Design* 13, 179 (2001).
25. C. Millar, A. Asenov, and S. Roy, *J. Comp. Electronics* 1, 405 (2002).
26. C. E. Capener and M. S. P. Sansom, *J. Phys. Chem. B* 106, 4543 (2002).
27. S. Chung, M. Hoyles, T. Allen, and S. Kuyucak, *Biophys. J.* 75, 793 (1998).
28. T. W. Allen and S. H. Chung, *Biochimica et Biophysica Acta* 1515, 83 (2001).
29. A. Burykin, C. N. Shultz, J. Villa, and A. Warshel, *PPROTEINS: Structure, Function and Genetics* 47, 313 (2002).
30. J. Tang, F. Quandt, and R. Eisenberg, *NeuroMethods* 26, 123 (1995).
31. M. Saraniti, S. J. Wigger-Aboud, Z. Schuss, and R. S. Eisenberg, *Comp. NanoSci. & NanoTech.* San Juan (2002).
32. C. Millar, A. Asenov, and S. Roy, in "Nanotech 2003: Technical Proceedings" (M. Laudon and B. Romanowicz, Eds.). Computational Publications, Cambridge, 2003.
33. S. Wigger, M. Saraniti, and B. Eisenberg, in "Nanotech 2003: Technical Proceedings" (M. Laudon and B. Romanowicz, Eds.). Computational Publications, Cambridge, 2003.
34. T. A. van der Straaten, G. Kathwala, Z. Kuang, D. Bada, D. P. Chen, U. Ravaioli, R. S. Eisenberg, and D. Henderson, in "Nanotech 2003: Technical Proceedings" (M. Laudon and B. Romanowicz, Eds.). Computational Publications, Cambridge, 2003.
35. D. Boda, D. Henderson, and D. D. Busath, *J. Mol. Phys.* 14, 2361 (2002).
36. A. R. Leach, "Molecular Modelling: Principles and Application." Prentice Hall, New York, 2001.
37. A. Einstein, "Investigations on the Theory of the Brownian Movement." Dover, New York, 1956.
38. R. Brown, in "Private Publication." Taylor, 1828, unpublished.
39. D. A. MacQuarrie, "Statistical Mechanics." University Science Books, Sausalito, 2000.
40. C. R. Arokianathan, Ph.D. thesis, Glasgow University, 1998.
41. S. M. Sze, "Semiconductor Devices and Technology." Wiley, 1985.
42. W. T. Coffey, Y. P. Kalmykov, and J. T. Waldron, "The Langevin Equation." World Scientific, Singapore, 1996.
43. C. Kittel, "Elementary Statistical Physics" Wiley, New York, 1958.
44. R. M. Lynden-Bell, *J. Chem. Phys.* 105, 9266 (1998).
45. S. Koneshan, J. C. Rasaiah, R. M. Lynden-Bell, and J. C. Rasaiah, *J. Phys. Chem.* 102, 4193 (1998).

46. R. M. Lynden-Bell, J. C. Rasaiah, and J. P. Noworyta, *Pure Appl. Chem.* 73, 1721 (2001).
47. L. Delle Site, R. M. Lynden-Bell, and A. Alavi, *J. Molec. Liquids* 98–99, 79 (2002).
48. N. W. Ashcroft and N. D. Mermin, "Solid State Physics," W. B. Saunders, Philadelphia, 1976.
49. W. H. Press, S. A. Teukolsk, W. T. Vetterling, and B. P. Flannerey, "Numerical Recipes in C," Cambridge University Press, Cambridge, 1992.
50. W. Janke, in "Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms," Vol. 10, NIC, Jülich, 2002.
51. B. A. Wallace, in "Gromicidin and Related Ion-Channel-Forming Peptides," Wiley, London, 1999.
52. P. Gibbon and G. Sutmann, in "Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms," Vol. 10, NIC, Jülich, 2002.
53. R. W. Hockney and J. W. Eastwood, "Computer Simulation Using Particles," Institute of Physics Publishing, New York, 1994.
54. T. Darden, D. York, and L. Pedersen, *J. Chem. Phys.* 98, 10089 (1993).
55. S. E. Laux, technical report, IBM Research Division, http://www.ibm.com, 1995.
56. U. Trottenberg, C. Oosterlee, and A. Schüller, "Multigrid," Associated Press, London, 2001.
57. E. Steifel and M. R. Hesteres, *J. Res. Nat. Bureau. Stan.* 49, 409 (1952).
58. R. V. Southwell, "Relaxation Methods in Theoretical Physics," Clarendon, 1946.
59. R. P. Fedorenko, *USSR Comp. Math. Math. Phys.* 1, 1092 (1962).
60. R. P. Fedorenko, *USSR Comp. Math. Math. Phys.* 4, 227 (1964).
61. N. S. Bakhvalov, *USSR Comp. Math. Math. Phys.* 6, 101 (1966).
62. A. Brandt, in "Proceedings of the 3rd International Conference on Numerical Methods in Fluid Mechanics," Lecture Notes in Physics, Vol. 18, Springer, 1973.
63. A. Brandt, Multi-level Adpative Techniques(mlat).i. the multigrid method, technical report, IBM T. J. Watson Research Center, Yorktown Heights, NY, 1976.
64. A. Brandt, *Math. Comp.* 31, 333 (1977).
65. W. L. Briggs, "A Multigrid Tutorial" SIAM, Philadelphia, 1987.
66. R. S. Eisenberg, *J. Membrane Biol.* 150, 1 (1996).
67. T. Gonzalez and D. Pardo, *Solid-State Electronics* 39, 555 (1996).
68. H. Kim, H. S. Min, T. W. Tang, and Y. J. Park, *Solid-State Electronics* 34, 1251 (1991).
69. B. W. Kernigan and D. M. Ritchie, "The Ansi C Programming Language," Prentice Hall, New York, 1988.
70. H. Schildt, "The Complete C++ Reference," Mc-Graw Hill, Berkeley, 1999.
71. S. E. Laux and M. Fischetti, in "Monte-Carlo Device Simulation-Full Band and Beyond" (K. Hess Ed.), Kluwer, Boston, 1991.
72. F. Medjdoub, F. Dessane, J. L. Thobel, M. Zokoune, and D. Theron, *Solid-State Electronics* 48, 683 (2004).
73. S. Ramo, *Proc. IRE* 27, 584 (1939).
74. W. Shockley, *J. Appl. Phys.* 9, 635 (1938).
75. P. De Visschere, *Solid State Electronics* 33, 455 (1990).
76. P. D. Yoder, K. Gartner, U. Krumbein, and W. Fichtner, *IEEE Trans. Computer Aided Design* 16, 1082 (1997).
77. C. Jungemann and B. Meinerzhagen, Electron Devices Meeting, 2000. IEDM Technical Digest. International, 2000.
78. S. Aboud, M. Saraniti, and R. S. Eisenberg, *J. Comp. Elec.* (in press).
79. W. Nernst, "Theoretical Chemistry," 4th ed, Macmillan, New York, 1904.
80. F. G. Stremler, "Communications Systems," Addison Wesley, Reading, 1990.
81. Taurus Process and Device Manual, http://www.synapsis.com.
82. M. F. Sanner, A. J. Olson, and J. Spehner, in "Proc. 11th Symp. Comp. Geom," p. 406, ACM Press, New York, 1995.
83. O. S. Anderson, *Annu. Rev. Physiol.* 46 (1984).
84. The Protein Data Bank: http://www.rcsb.org/pdb.
85. B. A. Wallace, *Prog. Biophys. Molec. Biol.* 57, 59 (1992).
86. Y. Liu, P. Sompornpisut, and E. Perozo, *Nature Structural Biol.* 8, 883 (2001).
87. D. W. Urry, *Proc. Natl. Acad. Sci. USA* 68, 672 (1971).
88. W. Humphrey, A. Dalke, and K. Schulten, *J. Molec. Graphics* 14, 33 (1996).
89. http://www.webelements.com.
90. T. W. Allen, T. Boolug, and S. M. Chung, *Biophys. J.* 84, 2159 (2003).
91. G. R. Smith and M. S. P. Sansom, *Biophysical Chemistry* 79, 129 (1999).
92. G. Moy, B. Corry, S. Kuucak, and S. Chung, *Biophys. J.* 78, 2349 (2000).
93. G. Moy, B. Corry, S. Kuucak, and S. Chung, *Biophys. J.* 78, 2364 (2000).
94. C. Jacoboni and P. Lugli, "The Monte-Carlo Method for Semiconductor Device Simulation," chapter 5, Springer-Verlag, Heildelberg, 1989.
95. R. M. Lynden-Bell, *J. Chem. Phys.* 105, 9266 (1996).
96. D. van der Spoel, A. R. van Buuren, E. Apol, P. J. Meulenhoff, D. P. Tieleman, A. L. T. M. Sijbers, B. Hess, K. A. Feenstra, E. Lindahl, R. van Drunen, and H. J. C. Berendsen, "Gromacs User Manual version 3.1," Nijenborgh 4, 9747 AG Groningen, The Netherlands.
97. H. J. C. Berendsen, D. van der Spoel, and R. van Drunen, *Comp. Phys. Comm.* 91, 43 (1995).
98. E. Lindahl, B. Hess, and D. van der Spoel, *J. Mol. Mod.* 7, 306 (2001).
99. M. Wien, *Ann. Physik* 1, 400 (1929).
100. J. Hasted, "Aqeous Dielectrics," Chapman and Hall, London, 1973.

CHAPTER 11

Nanomagnetics in Biotechnology

Ching Jen Chen, Yousef Haik, Jhunu Chatterjee
FAMU-FSU College of Engineering, Tallahassee, Florida, USA

CONTENTS

1. INTRODUCTION

In recent times, the broad science of nanotechnology has attracted widespread attention not only from scientists and researchers but also from the federal government and the media. This interest stems from the fact that nanotechnology as an emerging science has the potential to change our everyday lives perhaps as much as the Internet did some years ago.

Although there exists a wide spectrum of applications of nanotechnology, this chapter concentrates on its potential and use in the area of biomedical applications. The application of nanomagnetic materials in biomedicine is primarily because these materials are comparable in size range with living cells and various biological molecules. Moreover, these materials are

ISBN: 1-58883-048-9

Handbook of Theoretical and Computational Nanotechnology
Edited by Michael Rieth and Wolfram Schommers
Volume 6: Pages (545–565)

also susceptible toward an external magnetic field. As a consequence of both these facors, nanomagnetic materials have attracted the interest of researchers from a variety of diverse disciplines.

Magnetic particles with sizes ranging from a few nanometers to several micrometer are being used in biomedical fields for specific applications such as cell separation [1, 2], diagnostics [3, 4], hyperthermia [5, 6], targeted drug delivery [7–9], contrast agents in magnetic resonance imaging [10, 11], and so forth, as described by various groups of workers. Other than magnetic nanomaterials, polymeric nanoparticles in the biomedical field [12–15] have also been well reported. This chapter reviews the existing research work on nanosized magnetic particles for biomedical applications and also highlights several potential applicaions in the near future.

Magnetic nanoparticles are used extensively in pharmacy, biology, and medicine as cariers of biological compounds. Biomolecules, attached with magnetic particles, can be transported and also separated quickly from complex mediums. As such, they can be used for both therapeutic and diagnostic applications. In therapeutic applications, magnetic particlesmay be used to control the release of a drug to a specific site [16] under the effect of an external magnetic field or used in the extraction of tumor cells from the organism and their curing *in vitro* (or *ex vivo*). In hyperthermia, magnetic latex has been used to generate suffiient heat in the carcinoma cells to inactivate the tumor cells [17].

In diagnostic applications, magnetic particles may be used in separation. The impotant characteristics that need to be considered for their use in easy and fast separation stes in immunoassays are large surface area, ease of chemical bonding (adsorption and chemical grafting), high magnetization, and so on. Magnetic nanoparticles have sizes ranging between 10 and 100 nm. This size is comparable with normal cells (10–100 μm), viruses (20–500nm) or genes (2 nm wide and 10–100 nm long), and this proximity in size enables the nanomagnetic particles to attach to the above entities. In nanoscale, the chemical and physical properties of the materials differ significantly from those of atomic or bulk materials 18]. In most investigations into the ferromagnetic particles in the size range of 10–30 nm, a superparamagnetic behavior has been found. Magnetic properties in nanomaterials arewell studied in some reviews [19, 20], which explain the basic physics with appropriate examples.

In most applications, nanosized iron oxide is used as the magnetic material for bioredical applications—the iron is coated with a suitable surfactant. Some of the surfactants used to coat magnetic materials are proteins or a polymer. Coating of the inorganic magetic materials ensures the migration of magnetic materials with the help of an external magetic field, protects the inorganic magnetic materials, and also provides necessary surface acivity to attach with a biological molecule. Biodegradable superparamagnetic nanoparticles containing starch, dextran, polyethylene imine, chistosan, and albumin are reportedly used in magnetic field–assisted radiotherapy [21]. Magnetic particles coated with derivatives of crylates and methacryaltes provide a hydrophilic surface, whereas particles with hydroplobic surfaces are formed with styrene as starting materials. Superparamagnetic iron oxides ether with or without a coating are being used as contrast agents [22–24] in magnetic resonance imaging. Gadolinium-loaded core shell nanoparticles have been reported as a new contrast agent for magnetic resonance imaging [25]. Ferrofluids that contain magnetic nanoparticles are also extremely important for some biomedical applications such as magnetic hyperhermia [26], hemosorption [27], and so forth.

Fundamental theories of magnetism and modeling in the magnetic materials are advaced by the pioneering work of Neel [28]. A better understanding of the magnetism of mall particles is extremely important for technical applications such as information storage, olor imaging, bioprocessing, and so on.

Magnetic particles are mostly characterized by electron microscopy for the morphcogical studies: magnetic properties are studied using vibrating sample magnetometers (VSMs) and super conducting quantum interference devices (SQUIDs). Recently, atomic orce microscopy (AFM) has been used to see the surface topography of the coated nanoparticles. Wide-angle powder X-ray crystallography and Mossbauer spectra are being used fo the structural determination. In this review, we focus on the properties of nanosized magetic particles, synthesis procedures for the particles, and the encapsulation processes established

in our center, along with different biomedical applications for these nanomagnetic particles and some future directions of research.

2. MAGNETISM IN NANOSIZED MATERIALS

2.1. Extrinsic and Intrinsic Properties

Nanostructured materials modulated on a length scale of 100 nm exhibit different and often superior properties to those of conventional materials with phase or grain sizes larger than this [18]. The magnetic property is one of the enhanced properties for the nanostructured materials. Enhanced remanance and giant coercivity of the nanostructured materials play a major role in a various technological applications. Bulk materials have two types of properties: extrinsic and intrinsic. Intrinsic properties are those that arise from interactions on an atomic scale and that are material specific and independent of sample size, shape, and microstructure. Extrinsic properties arise from longer-range interactions, and even for the same material vary with the size, shape, and microsturcture of the sample. In the case of bulk ferromagnetic materials, saturation magnetization per unit volume M_s, magnetocrystalline anisotropy constant per unit volume K, and Curie temperatures T_c are intrinsic properties, whereas remanent magnetization M_r and coercivity per unit volume H_c are extrinsic properties.

2.2. Effect of Particle Size

The magnetic body has a multidomain structure comprising uniformly magnetized regions (domains) separated from each other by domain walls. This type of structural arrangement minimizes the magnetostatic energy and makes the system stable, but a complete balance of magnetostatic energy, anisotropic energy, and exchange energy determine the domain structure and shape. When the dimension of the crystal is reduced, the domain size is also reduced. Because of the energy requirement for the domain wall formation and the balance with the magnetostatic energy, the subdivision of the domains to an optimum value is limited. In fact, there is a corresponding lower limit of the crystal size below which a single-domain structure cannot exist [29]. For typical magnetic materials, the dimension limit is 20–800 nm [30]. For a spherical crystal, the characteristic radius is given by [31] $R_{sd} = 9E_\sigma/\mu_0 M_s^2$, where M_s is the saturation magnetization and E_σ is the total domain energy per unit area and is equivalent to $2(K/A)^{1/2}$, where K is the anisotropy energy constant and A is a parameter representing exchange energy density. The change from a multidomain to single-domain structure is accompanied by large increase in coercive field ($Hc \approx K/3M_s$).

The magnetic properties of the nanometer-sized particles are different from their bulk counterparts. With decreasing particle size, an increasing number of atoms lies near or on the surface and in the interfacial regions. This surface and interfacial electronic structure actually affect the magnetic properties. The large number of atoms located on the surfaces or interfaces, for example, whose local environments differ greatly from those of the interior atoms, leads to a distinction between the intrinsic and extrinsic properties. In small ferromagnetic materials, one might expect intrinsic properties M_s (saturation magnetization per unit volume), K (magnetocrystalline anisotropy constant per unit volume), T_c (Curie temperatures), M_r (remanent magnetization), and H_c (coercivity per unit volume) to differ from their bulk counterparts in a size-dependent way. Another important property is that unlike the bulk ferromagnetic materials, which have multiple domains, sufficiently small ferromagnetic particles (nanosized) are single domained. Large surface-to-volume ratios in the nanomagnetic materials and the single-domain behavior could led to new and unusual magnetic properties. Although the critical particle size for the single-domain formation are dependent on some other factors, for most of the magnetic materials, critical size is in the range of 20–2000 nm [32].

Magnetic materials are distinguished by their ability to exhibit a long-range order of their atomic magnetic moments through an exchange interaction between the neighboring atoms below a characteristic ordering temperature: the so-called Curie temperature (T_c). Above this temperature, the long-range order is destroyed and material is said to be in a

paramagnetic state. Well below the Curie temperature, all of the ferromagnetically coupled atomic moments in a particle are aligned in essentially the same direction and give rise to a total magnetic moment that can be larger than that of a single atom. The equilibrium magnetic properties of a large assembly of noninteracting uniaxial single domain particles, each of volume V, is largely determined by the relative magnitude of three characteristic energies: the thermal energy ($E_T = K_B T$, where K_B is the Boltzman constant), the anisotropic energy ($E_A = CV$, where C is the total anisotropy energy per unit particle volume), and the magnetostatic energy $E_m = \mu H$, where H is the applied magnetic field and μ is the permeability of the medium). Thus, with decreasing particle size, the anisotropy energy decreases and for a grain size lower than a characteristic value, it may become so low as to be comparable to or lower than the thermal energy $K_B T$. This implies that the energy barrier for magnetization may be overcome and that the total magnetic moment of the particle can fluctuate thermally like a single spin. Thus, the entire spin system may be rotated, and spins within the single-domain particles may be magnetically coupled.

The magnetic behavior of such an assembly of ultrafine particles is called *superparamagnetism*. Superparamagnetism is exhibited by particles within a defined range of dimension. If the particles are too small, their magnetic and electrical properties change a lot compared to the bulk ones, and the superparamagnetic model cannot be applied. The lower-dimensional limit has been stated to be approximately 2 nm by Dormann et al. [33], and the upper limit is given, in principle, by characteristic single-domain size and structure, though the characteristic grain size of a magnetic material for superparamagnetic relaxation depends on the anisotropy constant and saturation magnetization values, and it is usually less than 20 nm for spherical particles having an uniaxial anisotropy [33].

2.3. Effect of Temperature and Measuring Time

At a temperature at which $E_T \gg E_A$ and in a zero-applied magnetic field, the direction of μ fluctuates in time, and there is no global magnetization. In the presence of a magnetic field, magnetization is observed whose field dependence exhibits no magnetic hysterisis and that resembles a classical paramagnet if the atomic moment is replaced by particle moment μ. With decreasing temperature, a blocking temperature T_B reached degree below that which ferromagnetic and history-dependent magnetic properties first appear. Such systems are called superparamagnetic systems (which exhibits magnetism in presence of the magnetic field only).

For fine magnetic materials, magnetic behavior depends on the value of the measuring time (τ_m) in the specific experimental technique with respect to the relaxation time (τ) associated with the overcoming of the energy $\tau = \tau_m \exp(CV/K_B T)$ where $\tau_m = 10^{-10}$–10^{-13} sec, K_B is the Boltzman constant, and T is the temperature. If $\tau_m \gg \tau$, the relaxation appears to have been so fast that a time average of magnetization orientation is observed in the experimental time window and the assembly of particles behaves as paramagnetic system (superparamagnetic state). On the contrary, if $\tau_m \ll \tau$, the relaxation becomes very slow and the moment gets blocked like a magnetically ordered system. The blocking temperature T_B, separating two states, is defined as the temperature at which $\tau = \tau_m$.

Blocking temperature (T_B) is considered to be an important parameter for studying nanomagnetic materials. The definition of blocking temperature for a superparamagnetic system is given by $T_B = Ea/K_B \ln(\tau_m f_0)$, where τ_m is the experimental measuring time, f_0 is a frequency factor on the order of 10^9 s^{-1}, and $K_B \ln(\tau_m f_0)$ can be treated as a constant. If the interactions between particles are considered, then $T_B = (E_A + E_{int})/K_B \ln(\tau_m f_0)$, where E_{int} is the interaction energy. Obviously, higher interaction may cause higher T_B. This temperature T_B is usually obtained by analyzing zero field cooled (ZFC) and field cooled susceptibility (FC) versus temperature curve, and the maxima in ZFC curve usually denotes the blocking temperature. For superparamagnetic particles with uniform size, a sharp blocking temperature is observed [34].

The current techniques used to measure the superparamagnetic relaxation are dc susceptibility ($\tau_m = 100$ s), ac susceptibility (10^{-2}–10^{-4} s for experiments at lower frequency, $\tau_m = 10^{-1}$–10^{-5} s for classical experiments) Mossabauer spectroscopy (10^7–10^9 s) [18], ferromagnetic resonance, and neutron diffraction, depending on the type of experiments. A typical

magnetization curve (measurement from dc susceptibility) at room temperature shows both coercivity and retentivity as a zero for superparamagnetic materials [35], and the hysterisis plot at a lower temperature and a hysterisis plot at room temperature fall under the same universal plot [20].

3. FABRICATION OF NANOMAGNETIC PARTICLES FOR BIOMEDICAL APPLICATIONS

3.1. Methods of Fabrication for Nanomagnetic Particles

There are two steps to fabricate nanomagnetic particles for biomedical applications: synthesis of inorganic nanomagnetic particles, and encapsulation of the nanomaterials by a protein or polymer or with a suitable surfactant.

For the first step, the nanoparticles are mostly formed by wet chemical methods such as chemical coprecipitation [36–38], borohydride reduction [39–41], sonochemical [42, 43], vapor trapping [44], molecular self assembly [45], and so forth. Dry methods include laser ablation, microwave plasma vapor deposition, and carbon arc process [46]. To make encapsulated magnetic particles, chemical methods are mostly used. For biomedical applications, the chemical methods are mostly reported. Magnetic nanoparticles that have high magnetic susceptibility and saturation magnetization are suitable for medical applications. Considering this magnetite is extremely suitable, as it is ferromagnetic and is much more stable than other transition metals such as Co, Ni, Fe, and so on.

For *in vivo* applications, it is important that stable organic coating surround the magnetic particles. Mostly Fe_3O_4 and γFe_2O_3 are being used as magnetic materials to make magnetic-encapsulated particles for biomedical applications. There are some reports about metallic iron being used as a magnetic material [47] encapsulated with dextran. Recent work has also established the use of polysiloxane dispersion of cobalt particles coated with silica [48] for opthalmic applications. Physical methods usually form dry powders, and wet methods give rise to particles suspended in water or in any solvent, but they can be transformed into dry powder also, depending on the requirement for the further encapsulation procedure.

Several researchers have achieved coprecipitation reaction using microemulsion. Synthesis of stable colloidal Fe_3O_4 particles [49, 50] in microemulsions have been reported. This colloidal suspension has particles in the nanometer range, and the colloid showed superparamagnetic behavior. Ferric chloride and ferrous chloride in an alkali medium was used as the aqueous medium, and the micellar system was aerosol OT and (AOT/water/iso octane) water in oil (w/o) microemulsion. The whole reaction occurred at room temperature.

Another microemulsion system with $FeCl_3$, $6H_2O$, double distilled water, cyclohexane, $FeSO_4$, $7H_2O$, and Brij-97 also formed iron oxide nanoparticles at a higher temperature. This process had significant advances in crystallization of the formed particles and thus eliminated further steps in synthesis [51]. To form mondispersed particles in colloidal medium, the magnetite particles are sometimes precipitated in a crosslinking polymer matrix or a network of gel [52]. Maghemite particles (γFe_2O_3) are also formed by the coprecipitation of acidic mixtures of ferrous and ferric chloride by ammonium hydroxide, and the effects of citrate ion on these particles have been studied. With citrate ions, particles can be made as small as 2 nm [53].

In our lab, we have synthesized γFe_2O_3 by coupling chemical coprecipitation [54] and ultrasonication. A transmission electron micrograph for these particles is shown in Fig. 1. The particles were further coated with a cationic surfactant. The particles were very well dispersed and are in the size range of 6–10 nm [54].

3.2. Methods of Encapsulation for Nanoparticles

Nanoparticles have wide application in the pharmaceutical, agricultural, textile, and cosmetics industries as delivery vehicles. In these fields of application, a drug, protein, magnetic material hormone, peptide, fertilizer, pesticide, herbicide, dye, fragrance, or other agent is encapsulated in a polymer matrix and delivered to a site either instantaneously or in a

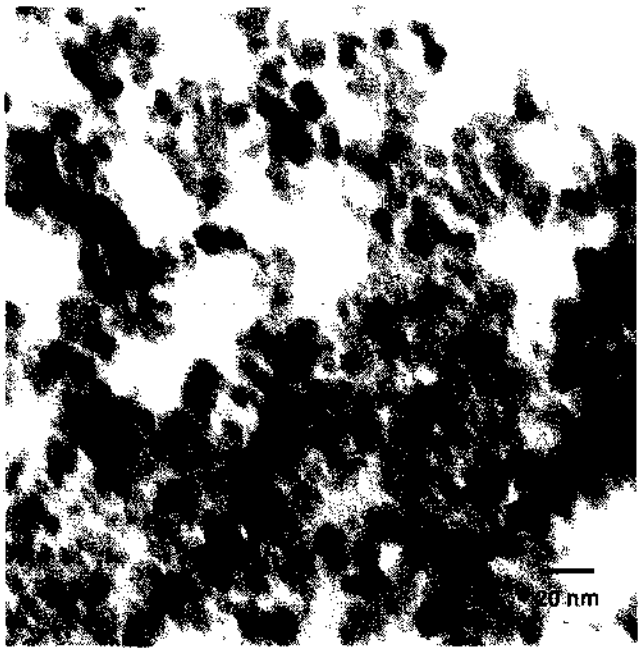


Figure 1. Transmission electron micrograph of iron oxide.

controlled manner in response to some external stimulus (i.e., pH, temperature, magnetic field, ultrasound, radiation, pressure, concentration gradients, etc.).

Many microencapsulation techniques exist that can produce a variety of particle types and sizes under various conditions. Methods typically involve solidifying emulsified liquid polymer droplets by changing temperature, evaporating solvent, or adding chemical cross-linking agents. The physical and chemical properties of the encapsulant and the material to be encapsulated can sometimes dictate the suitable methods of encapsulation, making only certain methodologies useful in certain circumstances. Brief description of several existing methods for making micron/submicron particles are given below.

3.2.1. Solvent Evaporation

The polymer is dissolved in a volatile organic solvent, such as methylene chloride. A substance to be incorporated is added to the solution, and the mixture is suspended in an aqueous solution that contains a surface active agent such as poly(vinyl alcohol). The resulting emulsion is stirred until most of the organic solvent evaporates, leaving solid microspheres. Both nano- and micron-sized particles can be generated by this process This method is useful for relatively stable polymers like polyesters and polystyrene [55].

3.2.2. Phase Inversion

Particles can be formed from polymers using a phase inversion method. Here, polymer is dissolved in a good solvent; fine particles of a substance to be incorporated, such as a drug, are mixed or dissolved in the polymer solution; and the mixture is poured into a strong nonsolvent for the polymer, to spontaneously produce polymer particles, after which either the polymer is coated on the particles or the particles are dispersed in the polymer. In general, droplets of the polymer solution are formed by forcing it through a spinneret or syringe needle (the size of these droplets determines the size of the final microcapsule) and contacting the droplets with a nonsolvent for the polymer that is highly miscible with the polymer solvent, thereby causing rapid precipitation of the outer layer of the droplet. The microcapsules must be left in contact with the nonsolvent until a substantial amount of the solvent has been replaced with nonsolvent. This process requires the formation of a droplet with dimensions established before contacting the nonsolvent.

The method can be used to produce microparticles in a wide range of sizes, including, for example, about 100 nm to about 10 μ. Exemplary polymers that can be used include

polyvinylphenol and polylactic acid. Substances that can be incorporated include, for example, imaging agents such as fluorescent dyes, or biologically active molecules such as proteins and so forth. Phase-inversion phenomenon have been applied to produce macro- and microporous polymer membranes and hollow fibers used in gas separation, ultrafiltration, ion exchange, and reverse osmosis. Structural integrity and morphological properties of these membranes are functions of polymer molecular weight, polymer concentration, solution viscosity, temperature, and solubility parameters (of polymer, solvent and nonsolvent).

3.2.3. Solvent–Nonsolvent Temperature-Induced Crystallization

In our lab we have coupled ultrasonication and a nonsolvent temperature-induced crystallization process to synthesize magnetic nanoparticles encapsulated by polymers. By this process, crystalline polymers that are insoluble at room temperature can be obtained in very high yields [56]. In this process, a polymer and iron oxide (magnetic material) are dissolved in a solvent with high boiling point and then added to a nonsolvent that is also held at high temperature. Then the mixture is brought to 0°C suddenly, and an emulsion is formed, separating two liquids. The polymer and iron oxide form a brown layer at the junction of the two liquids. The top and bottom layers of the solvent–nonsolvent mixture are pipetted out, and the particles are washed with acetone. Particles can be dispersed in water by ultrasonication. This method is very useful to make nanomagnetic encapsulated particles with a crystalline polymer.

Polymer iron oxide nanoparticle conjugate can also be prepared by the precipitation of iron oxide in presence of a polysaccharide [24]. Starch magnetic nanoparticles are produced [21] by modification of the above procedure. The starch nanobeads are smaller than corresponding dextran particles and have a higher percentage of iron oxide in them. The final size of a composite nanoparticle is very much dependent on the molecular weight of the polymer. Starch has much lower molecular weight (25–30 kDa) than dextran (40 kDa). These nanoparticles were further modified for their complexation with radionucleotides [21].

Dextran-covered superparamagnetic nanoparticles are also used as a contrast agent in magnetic resonance imaging [24, 57]. Some commercial examples are AMI-25 (feridex, endorem) [58, 59] AMI 227 (USPIO, BMS18059 and MION) [60–62].

3.3. Encapsulation with Protein

We have reported the synthesis of chemically cross-linked albumin magnetic particles for blood cell separations [63]. This method has some differences in the reported methods [64, 65] for synthesizing albumin microparticles. Briefly, 1 mL aqueous solution of 250 mg human serum albumin and 250 mg maghemite was added dropwise to a mixture of 40 mL n-hexane, 10 mL light mineral oil, and 0.5 mL of sorbitan sesquioleate to form a water-in-oil inverse emulsion system, which was sonicated for 10 minutes at 50% amplitude. Ten milliliters of 25% glutaraldehyde saturated with toluene was added to this mixture and mechanically stirred with a teflon paddle stirrer at 200 rpm for 2 h. The whole mixture was then centrifuged at 3000 rpm for 15 minutes. The supernatant was decanted, and the nanoparticles were washed repeatedly washed with petroleum ether and acetone and were finally dispersed in distilled water. These particles were further modified with avidin and biotinylated lectin to attach to the red blood cells. Transmission electron micrographs for the particles are shown in Fig. 2.

3.4. Encapsulation with Synthetic Polymer

A very dilute (0.05% w/w) solution (10 ml) of polyethylene wax and iron oxide mixture was made using Decalin at 150°C in a 25 mL screw-cap scintillation vial, by ultrasonication. The same amount of nonsolvent tetraglyme at the same temperature was added to it and sonicated at 50% amplitude for 30 s. After dissolution and mixing at 150°C, the mixture was immediately cooled at 0°C temperature. An emulsion was formed within few minutes. The solution was then cooled at room temperature. Within 45 minutes to 1 hour, polyethylene particles, along with maghemite, were found to be suspended in the emulsion. The mixture

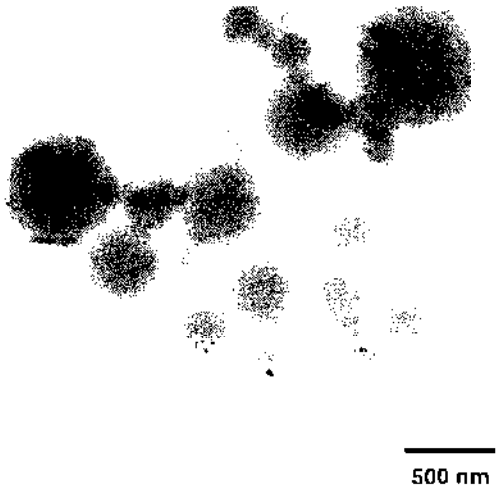


Figure 2. Transmission electron micrograph of human serum albumin magnetic particles.

was then kept at −10°C in the refrigerator for about half an hour. A very thin reddish brown layer was observed in the junction of two liquids. The top and the bottom layr of liquids were extracted with the help of a micropipette and a syringe. The particles were hen centrifuged in a microcentrifuge (because of its small quantity) to isolate them fron rest of the solvent mixture. The remaining solvent was removed and particles were washedwith acetone.

Coating the polyethylene composite particles with poly dl-lactic acid has been descibed by the authors elsewhere [66]. In brief, 0.2 mg of polyethylene nanocomposite particle and poly dl-lactic acid (1:1 by weight) were dispersed in 1 mL methylene chloride and hated at 50°C to dissolve the polylactic acid thoroughly. After cooling, a slight excess of metlanol was added to the mixture of polymers. A precipitate of coated composite particles was obtained almost immediately. Transmission electron micrographs for coated particle are shown in Fig 3.

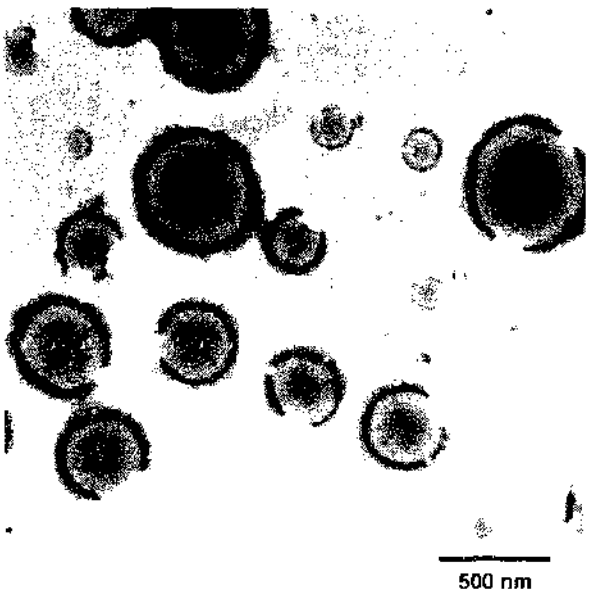


Figure 3. Transmission electron micrograph of polylactic acid-coated polyethylene sample.

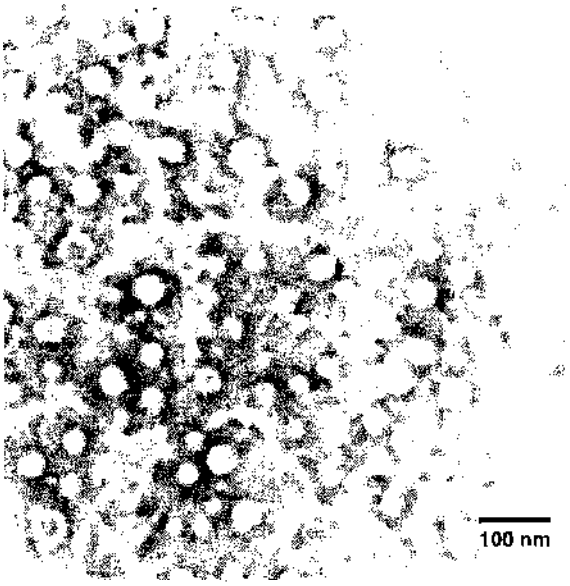


Figure 4. Transmission electron micrograph of hydroxyl propyl cellulose magnetic particles.

3.5. Encapsulation with Biodegradable Polymer

In our lab, we reported [67] a two-step synthesis of nanomagnetic gel. In the first step, hydroxypropyl cellulose particles are formed with surfactant-modified maghemite by the emulsion method, as done by Hu et al. [68], to synthesize cross-linked HPC nanoparticles. In the second step, these particles are cross-linked by a commercial crosslinking agent Zirmel M to give a network structure. It is pertinent to note that both steps are important and influence the final gel structure and that a nanomagnetic material is being formed at each step. By this mechanism of network formation, it was possible to introduce a homogeneous distribution of maghemite into the polymer matrix. One gram of hydroxy propyl cellulose (HPC) (Mol. Wt. = 370,000) was mixed with 300 mg (30% by weight of the polymer) cetyl trimethyl ammonium bromide (CTAB)-modified maghemite powder in 200 mL sodium hydroxide solution (pH = 12) and sonicated at an amplitude of 50%. The mixture was mixed at a speed of 200 rpm at 65° for 1 h. Divinyl sulphone (4% of the weight of the polymer) was added to the mix. Thus, the nanomagnetic HPC particles were formed. These particles were further crosslinked to form the nanomagnetic gels. Figure 4 shows the transmitting electron micrograph for the HPC magnetic nanoparticles.

4. CHARACTERIZATION OF NANOMAGNETIC PARTICLES

Nanomagnetic composite particles are characterized mainly in terms of their size and shape, surface charge density, zeta potential, structure, and magnetic properties. Size and shape can be determined by several techniques such as Dynamic Light Scattering [28, 69], scanning electron [70] and transmission electron microscopy [20, 30, 50, 71], AFM [56], and a particle surface charge by a zeta potential analyzer. Structural properties can be obtained by X-ray crystallography [71, 55] and magnetic properties using Mossabauer spectroscopy [20], SQUID [35, 50, 55, 71], and VSM [48, 53] surface charge by zeta potential analyzer. Structural properties can be obtained by X-ray crystallography [54, 71] and magnetic properties by using Mossbauer spectroscopy [20, 35]. Details of the characterization of magnetic composite particles produced in our laboratory are given below.

4.1. Structural and Morphological Characterization

4.1.1. X-Ray Crystallography

X-ray crystallography for pure iron oxide was performed; an X-ray powder diffractogram is shown in Fig. 5.

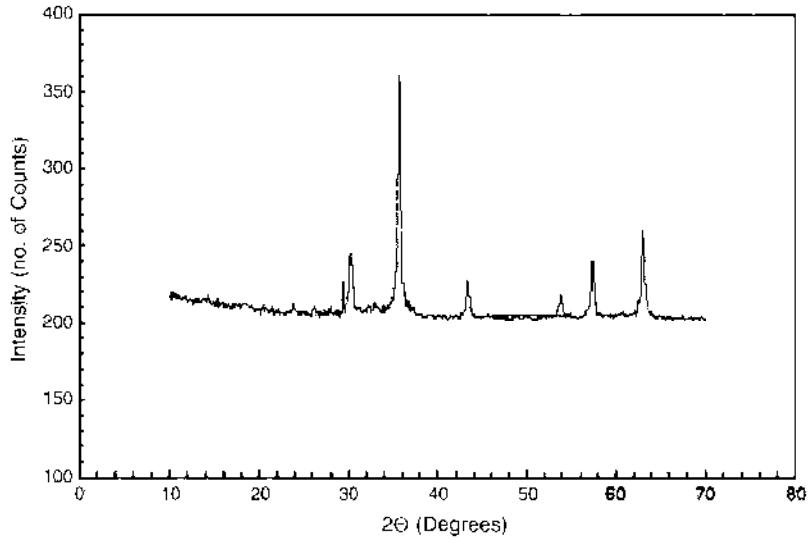


Figure 5. Wide-angle X-ray diffractogram for pure iron oxide powder.

4.1.2. Transmission Electron Microscopy

Transmission electron microscopy was performed in a JEOL 2010 microscope operated at 200 KV. The bright-field imaging technique was used to image the samples by selecting the transmitted diffraction spot to form the image. An Advanced Microscopy Techniques (AMT) charge-coupled device camera and software carry the image acquisition. One drop of composite nanoparticles in acetone was placed on a carbon-coated copper grid, dried, and observed under the microscope. The transmission electron micrographs for the cationic surfactant-coated γFe_2O_3 synthesized in our lab are shown in Fig. 6. These particles have a more uniform distribution.

4.1.3. AFM

For AFM, a D 3000 Nanoscope from Digital Instruments was used in tapping mode. One drop of the sample was placed on cleaved mica surface and observed under the microscope. Micrographs were taken in both height and amplitude mode. Mica surface was used instead

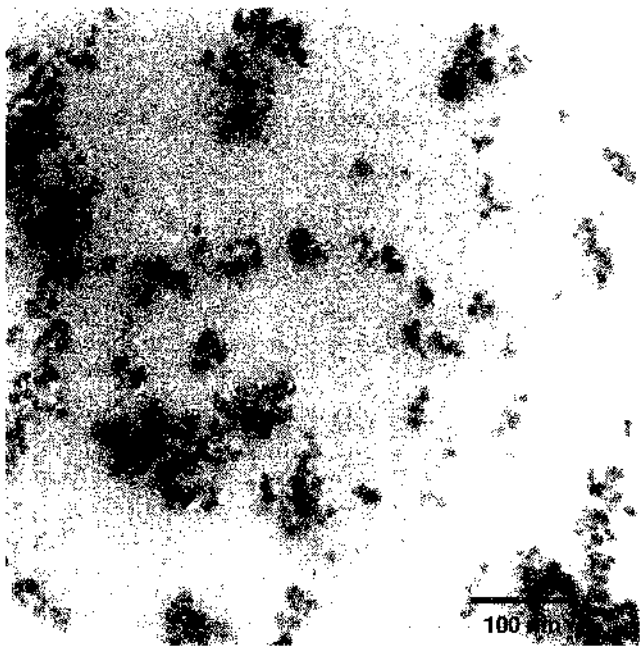


Figure 6. CTAB-coated iron oxide particles.

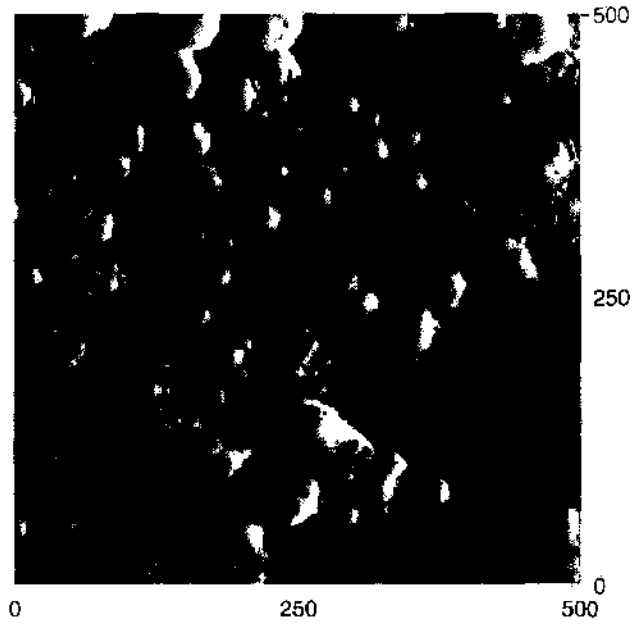


Figure 7. Atomic force micrograph of hydroxyl propyl cellulose gel magnetic particles.

of a glass slide to avoid artifacts caused by the surface of a glass slide. The micrographs for several particles are given in Figs. 7 and 8.

4.2. Magnetic Property Measurement

A Quantum Design MPMS5 DC SQUID magnetometer was used to study the magnetic properties of the maghemite particles and the polyethylene magnetic particles. Weighed amounts of samples were packed in gel capsules and placed tightly in the glass tube, ensuring no movement in either direction, and the magnetic properties were measured. We have not obtained any sharp blocking temperature for the composite particles because of the wide size distribution in the ZFC and FC plots. There was no remanent magnetization or

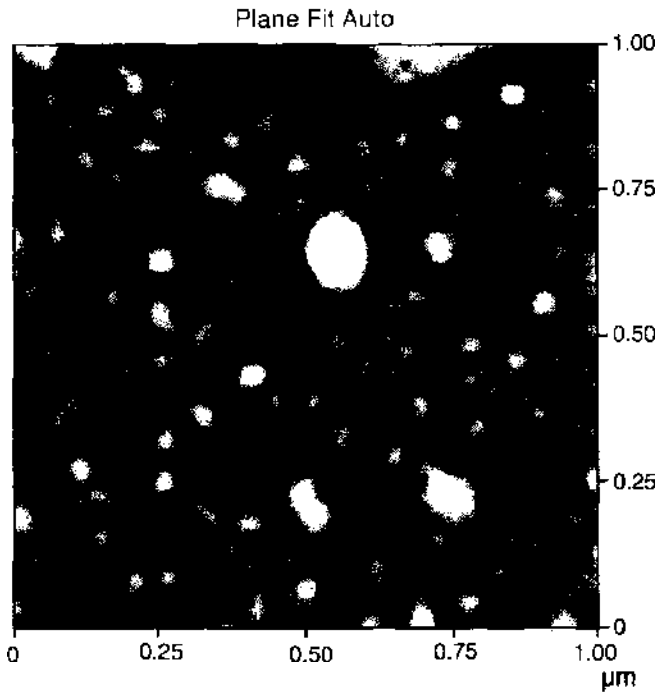


Figure 8. Atomic force micrograph of polyethylene magnetic particles.

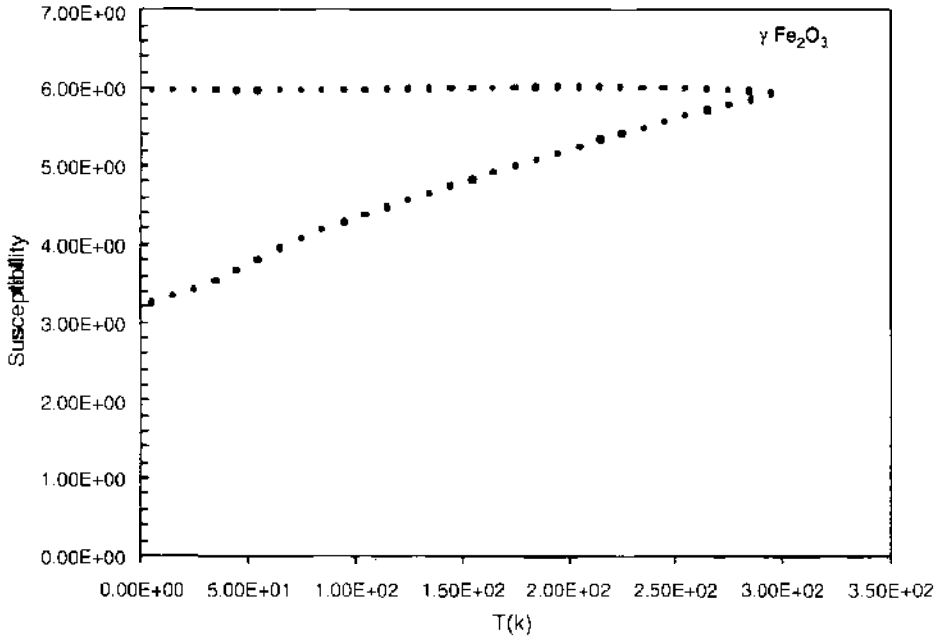


Figure 9. Temperature versus susceptibility plot for iron oxide nanoparticles (maghemite).

coercivity, as observed in the hysterisis plot obtained at 300 K. The composite particles showed superparamagnetic behavior as discussed earlier. Magnetization data for some of the composite particles are shown in Figs. 9 and 10.

5. APPLICATIONS: PRESENT AND FUTURE

Although magnetic particles have been widely used in biomedical field for wide variety of applications, it is only in the last few years, nanosized magnetic particles are being used for applications in biotechnology sector. Some of the areas of applications are described below along with some future scope.

5.1. Cell Separation and Selection for Therapeutic Use

The theoretical basis of the separation phenomenon using magnetic particles is described briefly below. When a magnetic particle is in a magnetic field, magnetic and electroviscous forces act on the particles, and at equilibrium, $F_{VIS} + F_M = 0$.

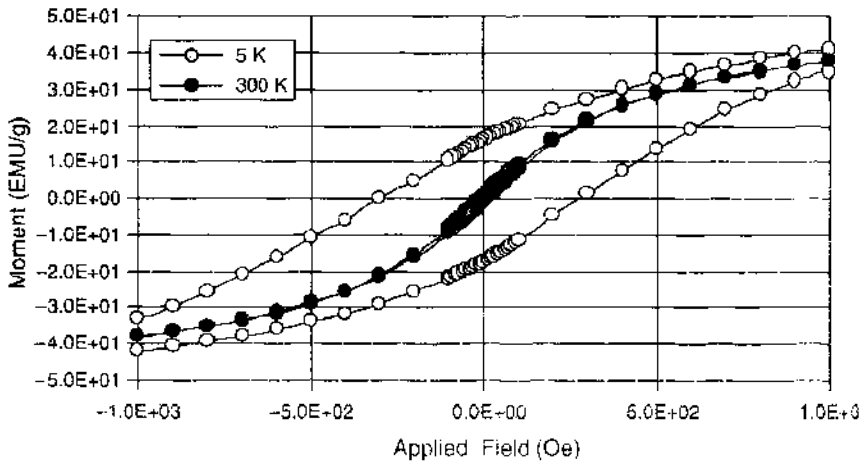


Figure 10. Magnetization versus applied field for iron oxide nanoparticles (maghemite).

F_{VIS} is basically the resistance to the displacement of a particle in a liquid medium, and it is a function of the viscosity of the medium. According to Stoke's law, $F_{VIS} = -6\pi\eta RV$, where η is viscosity of the medium, R is the hydrodynamic radius, and V is the separation velocity. The magnetic force F_M can be expressed as a function of magnetic field strength (H), gradient of the magnetic field (δH), and magnetic susceptibility ($\mu = \mu_0 M_s \cdot 4\pi R^3/3$) as follows

$$F_M = \mu H \delta H$$

From the above-stated relation between two forces (magnetic and magnetoviscous), the separation velocity $V = (2\mu_0 \cdot M_s \cdot R^2/9\eta) \cdot H \cdot \delta H$.

This shows that the separation velocity increases with the increase in the particle size. Therefore, the magnetic separation will go faster when using magnetic particles that have a larger diameter, but at the same time, nanosized magnetic particles will have negligible interference with the separated cells.

The magnetic nanoparticles can be used to separate cellular subtypes with high purity from a mixed population. Cellular therapy finds its greatest application in the field of stem cell graft engineering and in adoptive immunotherapy of cancer [72]. The applications include the purging of malignant cells from autologous stem cell products, depletion of T cells, and selection of specific lymphocyte subsets with potential antileukemic activity. With the industrial explosion in the development of monoclonal antibodies and the improved characterization of lineage- and tumor-specific antigens, it is now possible to separate many specific cell types from mixed populations in blood, bone marrow, and other tissue fluids. Autologous bone marrow or peripheral blood stem cells may be used to support the contamination of the stem cell product with residual malignant cells. Eradication of these cells should reduce the chance of relapse following transplantation. Purging strategies have been investigated in leukemia, lymphoma, neuroblastoma, and breast cancer [73, 74]. As an alternative, allogeneic stem transplantation may be used. This will most likely require T-cell depletion. T cells are the major cell type responsible for the development of graft versus host disease (GvHD) following allogeneic stem cell transplantation (29–42% of recipients develop GvHD). Engineering the graft to deplete T cells can completely abrogate GvHD even in mismatched bone marrow transplants [75, 76].

A number of methods for T-cell depletion have been investigated. Mostly these have included monoclonal antibody-based negative selection techniques, separation based on lectin-mediated agglutination, and physical separation by size and density. These methods frequently result in the loss of more than half of the hematopoietic progenitors [73]. Magnetic nanoparticles provide a promising means of both purging stem cell products in the autologous transplantation and depleting T cells in the allogeneic stem cell transplantation. Ferrite-based magnetic particles produced by chemical precipitation methods or by chemical vapor deposition have been used for cell separation methods.

There are two currently available systems for immunomagnetic methods of cell selection, but both systems have proven to be of limited value in recovering cells. The Isolex 300 i cell selection technology (Baxter) has been reported to have achieved a mean cell recovery of between 41%–69% [77–79]. This device uses magnetic microspheres on the order of 3–4 μm. The other available immunomagnetic cell selection technology is the CliniMACS device. This device uses monoclonal antibodies conjugated to 50–100-nm iron-dextran beads. This device also has the disadvantage of losing up to 50% of the stem cells during the T-cell depletion.

Nanoscale particles for immunomagnetic applications will be superior to microscale particles for several reasons. The nanoscale particles are less likely to affect the cell functions, are able to attach to more receptors at the cell surface, and are considered to be biodegradable. However, the reduced size of the particle diameter means that the magnetic moment of the particles is reduced by three orders of magnitude. Ferrite-based magnetic particles have been used because of their stability, nontoxicity, and ease of production. By controlling the wet chemical methods, the size distribution of the magnetic core can be maintained uniformly. Except for iron oxides, we have also developed silica-coated NdFeB magnetic nanoparticles. The base of the magnetic core can be expanded to include other materials such as FePt

and NiPd. The magnetic moment for these magnetic particles is considerably larger than those of the maghemite particles per unit volume of the particle. The bulk of these magnetic materials exhibits ferromagnetic properties, but when the magnetic particles are produced on a scale of less than 30 nm, they can be considered to be superparamagnetic. Controlling the size in this range will minimize particle agglomeration because of their intrinsic magnetic moments.

Encapsulation and surface modification of the particles provides a high level of specificity. The silica-coated magnetic particles will be further modified using an avidin-biotin-antibody complex. Both positive and negative cell selections can be performed. For example, monoclonal antibodies against B lymphocyte–specific antigens such as CD10, CD19, and CD20 can be used as complement for the *ex vivo* purging of lymphoma cells from autologous bone marrow before transplantation [73]. Monoclonal antibodies tagged to the surface of magnetic nanoparticles can be used to target the CD34 antigen to isolate the hematopoietic progenitor cells [78]. Reports on identifying the suitable tags for cell selection continue to appear in the literature. Future research can be directed to enhance the stability of the surfactant and increase the bonding efficiency. The stability of the surfactant can be enhanced by controlling the experimental parameters during the wet chemical synthesis procedure.

5.2. Diagnosis

Hand-held, rapid, and robust biodetection devices, such as miniaturized analysis systems, have made significant advances in the last decade [80, 81]. The concept has been developed from being a chemical sensor alternative to being a lab-on-a-chip. The use of magnetic nanoparticles in a micro-total analysis system (M-TAS) will overcome serious disadvantages encountered with current microanalysis systems. For example, as the size of a sample is successively decreased, an immediate concern relates to how representative the sample is of the specimen from which it was derived. This is a problem for inhomogeneous biological specimens that contain a diversity of constituents. This problem can be overcome by developing a site-specific agent such as magnetic nanoparticles that isolate the rare component from the flow.

Another issue is the detectability of components. Conventional microanalysis systems cannot detect small quantities on the order of picograms per liter. We are able to detect 1 pg/L of myoglobin by using magnetic nanoparticles [82]. In addition, the magnetic nanoparticles will facilitate the reactions in the microsystems. The use of nanoparticles will also eliminate the need for dilution of the reactants, which will increase detectability and maintain the reaction speed at a level that corresponds to the concentration of analytes. Furthermore, with the use of an external magnet, the nanoparticles can be easily localized. This allows continuous pumping of the reaction components through the spheres, thereby enhancing the yield of the reaction.

We have developed a magnetic immunoassay for the detection of cardiac markers [82]. Traditional detection methods have proven to be slow, bulky, and in most cases require a significant amount of sample. Magnetic particles can be prepared such that they will be only linked to a particular biomolecule in a heterogeneous solution. Then, with the use of a magnetic separator, these particles can be isolated from the solution and further processed to quantify the concentration of the biomolecules [83]. The use of magnetic nanoparticles to extract a biomarker from a solution provides a tool that is highly sensitive and that allows for miniaturization of the detection device. An important direction in the current development of analytical techniques is the miniaturization of analyzers. Generic names for these new micrometer devices include μ-TAS, lab-on-a-chip, or biochip [80].

Several factors are responsible for the continuous interest in microminiaturization of analyzers. First, a range of analytical methods has emerged for which miniaturization has obvious benefits. These include high-throughput, massively parallel testing for drug discovery [81], small hand-held portable analyzers for point of care testing, and lightweight analyzers for use in space exploration missions [84, 88]. Second, miniaturization reduces the cost of the analytical processes because the amount of reagent used per assay is much lower than in conventional analysis. Finally, an important advantage of miniaturization is the ability to

integrate all steps of a complex multistep analytical process onto a single device [89, 90]. Promising applications of magnetic nanoparticles include those in the following sections.

5.2.1. Detection of Cardiac Markers for Acute Myocardial Infarction

According to the American College of Cardiology/American Heart Association guidelines for management of patients with acute myocardial infarction (AMI), treatment should be started within less than 30 min of patient arrival. For patients with AMI, treatment with thrombolytic therapy is most beneficial if treatment is administered as early as possible.

5.2.2. Detection of Infectious Diseases for Blood Bank Applications

Current practices at blood banks require that donated blood to be sent to a regional, central blood testing lab to check for infectious diseases before the blood can be admitted to the bank. If the blood is found to be contaminated, discarding the blood safely is costly. Testing the donor using a small sample of blood before drawing a whole bag of the donor's blood will help save the time and effort of the blood bank administration and will avoid expenses associated with the rejection of blood.

5.2.3. Hyperthermia Treatment of Tumors

Hyperthermia is a rapidly developing technique in cancer therapy [91]. This treatment takes advantage of the higher sensitivity of tumor tissue to heat. Magnetic hyperthermia can minimize side effects by heating only the desired part of the organism, including tumors located deep inside the patient's body. Magnetic hyperthermia involves the introduction of magnetic particles into the desired part of the organism and heating them remotely with an alternating magnetic field. Unfortunately, with today's magnetic particles, it is impossible to control the uneven heating at the tumor site. This may cause local overheating and necrosis. Production of magnetic material with a Curie temperature of 42–43°C will enable this technology to be safely used in the treatment of tumors in conjunction with radiation therapy.

5.2.4. Hemosorption

The removal of pathogenic agents from the body by a blood purification procedure is one of the oldest therapeutic approaches in medicine [92]. Hemosorption not only allows us to eliminate toxins but also to remove viruses and bacteria. The standard technique of hemosorption involves the withdrawal of significant amount of patient's blood from the organism to the sorption column, which can be packed with different types of absorbents. The selective removal of substances of different relative molecular mass requires the use of functionalized surfaces of adsorbent particles. Carbon-based adsorbents with micronsize particles are being used in the hemosorption columns. The reduction of the adsorbent size of the particles increases the column efficiency and the sorption capacity. Magnetic nanoparticles with attached carbon or surface functionalized magnetic nanoparticles could be very useful, acting as absorbent materials. These particles can be moved or kept confined with the help of an external magnetic field.

5.2.5. Magnetic Tissue Engineering

Magnetic nanoparticles can be used to create specific spatial ordering of multiple cell types in an extracellular matrix. Magnetically tagged cells can be directed to a specific location on the cellular pattern by applying an external magnetic field.

5.2.6. Development of Technologies to Produce Macroscale, Mesoscale, and Microscale Diagnostic Device Components

Macroscale technologies to produce diagnostic devices are predominant because of the availability of manufacturing processes. Fundamental research issues related to the fabrication of the microanalysis system that uses the magnetic nanoparticles for detection include the following.

5.2.6.1. Development of Microfluidic Devices to Manipulate the Magnetic fluid Solution Scaling of devices increases the capillary forces and makes sample localization difficult; controlling the surface properties of both the nanoparticles and the microchannels becomes essential.

5.2.6.2. Development of Laminar Mixing Components to Mix the Nanospheres with the Biosample One of the proposed solutions for creating the proper mixing is to use the magnetic characteristics of the nanoparticles and to facilitate mixing with the help of an external magnetic arrangement. Alternating magnetic fields with high-density gradients located adjacent to a microfluidic channel, with spatial distribution on both sides of the channel, will be used to mix the sample and the magnetic nanoparticles. Producing a high-density gradient magnetic field for the microapplications will be an active area of research.

5.2.7. Development of Micromagnetic Separators

Production of magnetic separators with a high-density magnetic gradient is also important, in addition to producing high-magnetic-moment particles with a stable surfactant, it is equally important to produce magnetic separators that are able to capture the magnetic nanoparticles [93]. Research issues that need to be addressed include producing an efficient laminar mixing of the magnetic nanoparticles with cell solution in a confined conduit isolating the tagged cells without blocking the conduit, and producing a high-density magnetic gradient without the use of meshed wires to reduce the shear stresses on the cells. Producing magnetic nanoparticles with superior magnetic properties and uniform distribution will enable the design of a micromagnetic separator that will be able to capture all the magnetic particles. A magnetic force that is capable of attracting nanoparticles require the development of high-density gradient magnetic fields.

5.2.8. Targeted Drug Delivery

An ideal drug delivery system should work on the target organ or on specific cell type of an organ, but there are some anatomical barriers that limit the biodistribution of drugs. Because of this barrier, certain problems arise. Biodistribution of pharmaceuticals throughout the body, lack of site specificity, need of a large amount of pharmaceuticals to reach a target site, nonspecific toxicity, and other adverse effects are among some of them. Various methods are being tried such as surface modification of the drug delivery system to each a specific cell type. Magnetic guidance of a drug to a specific site is one of the most recent areas of research being tried in targeted drug delivery in cancer treatment. Freeman et al. [94] proposed for the first time that fine iron particles could be transported through the reticulo endothelial system and be concentrated at a particular point in the body with the help of an external magnetic field. This method of drug delivery can replace that of magnetic particle chemotherapy or radiation therapy.

Background and recent developments: Biomedical applications of a magnetic drug carrier have been reported for different types of cancer. Turner et al. [95] used magnetically guided ferromagnetic silicone to accomplish selective vascular occlusion and necrosis of hypernephromas in man. Use of ethylecellulose containing mitomycin (MMC) in VX-2 tumors in rabbit has been reported by Kato et al. [96], and use of magnetically guided albumin microparticles containing a drug has been reported as a site-specific drug delivery medium with the help of an external magnetic field by Widder et al. [97] In all of the above cases, magnetic fluid containing nanosized magnetic particles was used. though after encapsulationwith protein or polymer the final size of the capsule increased to micron level. In an experiment performed by Morimoto et al. [98] using I_{125} labeled magnetic albumin microsphere in the lung cancer cells of mouse, it has been shown that the accumulation of the magnetic microsphere was about fourfold compared with the control (when no magnetic field is applied) when two magnets (3000G) were applied. Other parameters such as the method of infusion and the size of the microcapsule also affect the accumulation of the magnetic particle in the site, but certainly magnetic guidance has more influence in this site specific delivery.

In the cancer treatment, experiment done by Morimoto et al. [99], the survival time of rats treated with free adriamycin (ADM) was lower than that of the rats that were treatedwith

albumin microspheres containing ADM under the effect of external magnetic field. Their experiment proved that magnetic microcapsules containing magnetic fluid exhibits sutained drug release, and the albumin matrix would be biodegraded by hydrolytic enzymes (e.g., protease). The microparticles could be directed by magnetic means into the target site (lung, kidney, etc.). The entrapped drug distribution can be further improved by the development of magnets suitable for biological applications. Site-specific drug delivery by magnetic means would eliminate the possibility of adverse effects of drug distribution throughout the body. The other advantage of the magnetic drug carrier over the other chemotherapeutic agents without magnetic materials is that the biodistribution of these therapeutic agents depends on physicochemical and biological interactions between the carriers and the human bodies, limiting their clinical applications, whereas targeting with the magnetic nanoparticles is dependent on the magnetic properties between the carrier and the magnet and not on other *in vivo* parameters.

Clinical trials for humans were first tried by Lubbe et al. [100], who used an antibiotic epirubicin coupled with a ferrofluid. In this process, magnetic particles with chemically bound drugs were infused intravenously, along with a chemotherapeutic dose, it has been found that the ferrofluid was successfully directed to tumors with the application of a magnetic field. Another area of recent interest is the use of nanomagnetic particles in gene therapy. By inserting plasmid DNA into the target cells, it is possible to treat the genetic disorders. By this process, therapeutic agents are formed in the form of protein or peptides.

However, the problem associated with this method is the inefficient delivery of the DNA to the target cell. To overcome that difficulty, the concept of magnetically driven drug delivery has been explored for the gene vectors, to enable them to reach the target cell. The enhanced delivery of therapeutic genes to a specific target site by associating gene vectors with superparamagnetic iron oxide nanoparticles drawn to the target tissue with the help of an external magnet is called magnetofection.

Schrer et al. [101] recently published their results on the efficiency of magnetofection, and according to them, the gene transfer with a small amount of DNA to the target tissue takes less than 5 min to get in magnetofection compared with 4–5 h when done in a cell culture plate placing the magnet beneath the plate, and also that this process can reduce the dose amount and the side effects. In their process, superparamagnetic iron oxide coated with polyelectrolytes is coupled with gene vectors by salt-induced colloid aggregation, and for *in vivo* application. they used iron-Nd_Boron magnets in direct contact with the target tissue [101].

The work reported by Hughes et al. [102] using commercially available magnetic particles is also in accordance with the results of Schrer et al. [101] in terms of the peak transfection level. To make these processes of magneto transfection more effective, some issues need to be explored, such as reaching the application of magnetic fields into the deeper tissues and using stronger magnets.

Either in targeted drug delivery or in gene delivery, the factors to be considered are the particle size in the ferrofluid, drug particle interaction, application of the magnetic field, strength of the magnetic field, route of infusion of the drug loaded magnetic particles, and so on.

5.2.9. Magnetic Resonance Imaging Contrast Agents

The imaging of internal organs and structures is very important for diagnostic procedures. Imaging procedures such as X-ray, magnetic resonance imaging and ultrasound all use contrast agents to improve the quality of the image. Magnetic resonance imaging is the diagnostic imaging technique of choice for the central nervous system and is widely used for the imaging of ligaments and tendons. Magnetic resonance imaging provides high-quality spatial resolution and does not use radiation. In magnetic resonance imaging, most of the signal intensity comes from the most abundant proton in tissues, the water protons. Magnetic resonance imaging contrast (between image and background and between images of adjacent tissues) usually relies on differences of proton density, the transverse relaxation time (T2), and the longitudinal relaxation time (T1). Every tissue in the human body has its own T1 and T2 values, and these are measures of different kinds of magnetic resonance relaxation

that occur after an atom has been stimulated by a radio signal in the presence of a strong magnetic field. In magnetic resonance imaging, the emitted radio signal from a particular tissue depends on a combination of T1 and T2 values. It is the variation of these relaxation times of neighboring tissues that make each tissue distinguishable in an MRI image; otherwise, the image would be all in one tone without distinguishing between them. Relaxation characteristics of normal and pathologic tissues are not always different enough to produce obvious differences in signal intensity.

The need for the development of contrast material lies in the fact that the soft tissues provide substantial contrast whereas others do not. There are couple of advantages to using a contrast agent [103]. First, pathology, which is sometimes occult on unenhanced images, becomes obvious in the presence of contrast. Second, enhancement significantly increases magnetic resonance imaging sensitivity. Third, in addition to improving delineation between normal and abnormal tissues, the pattern of contrast enhancement can improve diagnostic specificity by facilitating characterization of the lesion(s) in question. Fourth, contrast can yield physiologic and functional information in addition to lesion delineation. The contrast agent basically increases the relaxation rate (decreases the relaxation time) of the water protons in the vicinity of the tissue when it is sufficiently proximal to the component of body tissue under study, and thereby the nuclear magnetic resonance signal from the tissue of interest is enhanced relative to the background and to the surrounding tissue. Contrast agents play a significant role in improving the quality of diagnostic images by increasing the contrast between different internal structures or types of tissues in various disease states. As a consequence, contrast agents, which may be administered intravenously or orally, are widely used when available.

Most of the contrast agents that are being used are gadolinium-based chelates and are paramagnetic in nature [104–106]. These are called T1 contrast agents and have short blood time residence and are generally nonspecific. However, superparamagnetic nanoparticulate materials have more tissue specificity and possibility of targeting. These materials represent a new class of contrast agent and are called T2 contrast agents. The increase in relaxation rate by magnetic particles takes place via a complex mechanism. In the presence of a static magnetic field, superparamagnetic particles (~10nm) possess very large magnetic moments and the dipolar interaction between the magnetic cores and the surrounding solvent protons results in an increase of the transverse and longitudinal relaxation rates. Most of the superparamagnetic and ferromagnetic contrast agents are based on maghemite and magnetite, or a mixture of iron oxides commercially known as feridex, resovit, combidex, and so forth.

Currently available imaging techniques can be of limited usefulness in visualizing certain soft-tissue structures, such as the liver and the lymphatic system, which are among the principal sites where metastases of many common cancers, including colon, prostate, and breast cancer, are discovered. Magnetic resonance imaging exams of the liver produced with contrast agents (Feridex I.V) provide more diagnostic information and permit the identification of smaller abnormalities than images produced by magnetic resonance imaging studies without contrast agents. Our product was the first organ-specific magnetic resonance imaging contrast agent designed specifically for the liver. Magnetic resonance imaging exams of lymph nodes using a contrast agent can provide increased confidence in the evaluation of lymph nodes as part of the staging of metastatic disease. An magnetic resonance imaging contrast agent that localizes to and causes contrast enhancement of the lymph nodes, such as Combidex, is reported to allow for more accurate disease diagnosis.

Magnetic dextran nanoparticles have been widely used for last 10 years [107] as a magnetic resonance contrast agent. These are nontoxic, biodegradable and rapidly turn over to body iron. Superparamagnetic dextran-coated particles reported by Tanimoto et al. [108] are very small and produce much larger local field distortions as compared with paramagnetic contrast agents, and relatively lower concentrations are sufficient to produce the required contrast.

Magnetodendrimers have been recently reported by Bulte et al. [109]. These are a new class of cellular magnetic resonance contrast agents and they attach with a variety of mamalian cells. Recently, ultrasmall superparamagnetic iron oxide particles have been tested

and compared to albumin-$(GdDTPA)_{30}$ for tumor microvessel characterization [110]. Ultrasmall superparamagnetic iron oxide particles show positive enhancement of T1-weighted imaging (imaging in which tissue T1 is used to control the brightness of the image pixels).

6. CONCLUSION

Magnetic nanoparticles have been used in many kinds of different biomedical applications. Then most common use is for separation of various biological products, often in association with immunological interaction. Most of the characteristics required by the magnetic nanoparticles are similar in a lot of applications, such as size small enough for binding, large enough for magnetic attraction, no remanence, and so forth. The advantages of using a magnetic particle medium are, on the one hand, the speed of detection and production of the analysis result, and on the other hand, the simplicity of separation, which makes the system easy to automate.

REFERENCES

1. K. Sugibayahsi, Y. Morimoto, T. Nadai, and Y. Kato, *Chem. Pharm. Bull.* 25, 3433 (1977).
2. J. Ugelstad, A. Berge, T. Ellingsen, O. Auno, L. Kilaas, T. N. Nilsen, and R.Schmid, *Macromol. Symp.* 17, 177 (1988).
3. E. F. Rossomando and J. Hadjimichael, US Pat. 5 158 871, University of Connecticut.
4. M. Mary, in "Clinical Application of Magnetic Carrier" (U. Hafeli, W. Schutt, M. Zborowski, and J. Teller, Eds.), p. 303, Plenum Press, New York, 1997.
5. B. M. Lilly, I. A. Brezovich, and W. J. Atkinson, *Radiology* 154, 243 (1985).
6. D. C. F. Chan, D. B. Kirpotin, and P. A. Bunn, *J. Magn. Mag. Mater.* 122, 374 (1993).
7. A. E. Senyei, K. J. Widder, and G. Czerlinski, *J. Appl. Phys.* 49, 3578 (1978).
8. J. P. Iannotti, T. C. Baradet, M. Tobin, A. Alavi, and M. Staum, *J. Orthop. Res.* 9, 432 (1991).
9. P. K. Gupta, C. T. Hung, and D. G. Perrier, *J. Pharm. Sci.* 76, 141 (1987).
10. M. Ohgushi, K. Nagayama, and A. Wada, *J. Magn. Reson.* 29, 501 (1978).
11. C. W. Jung and P. Jacobs, *Magn. Reson. Imaging* 13, 661 (1995).
12. C. A. Mirkin, R. L. Letsinger, R. C. Mucic, and J. J. Storhoff, *Nature* 382, 607 (1996).
13. P. Couvreur, L. Grislain, B. Kante, M. Ronald, and P. Spaiser, *J. Pharm. Sci.* 71, 790 (1982).
14. L. Peppas-Brannon, *Int. J. Pharm.* 116, 1 (1995).
15. V. Schafer, H. Briesen, R. Von Andressen, and A. M. Steffan, *Pharm. Res.* 9, 541 (1992).
16. R. Langer, *Science* 249, 1527 (1990).
17. A. Jordan, P. Wust, R. Scholz, B. Tesche, H. Fachling, T. Mitrovics, T. Vogl, J. Navarro Cervos, and R. Felix, *Int. J. Hyperthermia* 12, 705 (1996).
18. L. J. Dorman, D. Fiorani, and E. Trone, *Adv. Chem. Phys.* 48, 283 (1997).
19. L. D. Leslie-Pelecky and R. D. Rieke, *Chem. Mater.* 8, 1770 (1996).
20. J. K. Vassiliou, V. Mehrotra, M. W. Russell, E. P. Giannelis, R. D. McMichael, R. D. Shull, and R. F. Ziolo, *J. Appl. Phys.* 73, 5109 (1993).
21. C. Gruttner, J. Teller, W. Schutt, F. Westphal, C. Schumichen, and B. R. Paulke, "Scientific and Clinical Application of Magnetic Carrier" (U. Hafeli, W. Schutt, M. Zborowski, Eds). Plenum Press, 1996.
22. J. P. Iannotti, T. C. Baradet, M. Tobin, A. Alavi, and M. J. Staum, *J. Orthop. Res.* 9, 432 (1991).
23. K. Turetschek, P. L. Timothy, P. L. Roberts, E. Floyd, A. Preda, V. Novikov, D. M. Shames, W. O. Carter, R. C. Brasch, *J. Mag. Reson. Imaging*, 13, 882 (2001).
24. R. S. Molday and D. Mackenzie, *J. Immunol. Methods* 52, 253 (1982).
25. C. H. Reynolds, N. Anan, K. Beshah, J. H. Huber, S. H. Shaber, R. E. Lenkinski, and J. A. Wortman, *J. Am. Chem. Soc.* 122, 8940 (2000).
26. A. Jordan, P. Wust, R. Scholz, H. Fachling, T. Mitrovics, J. Krause, and R. Felix, in "Scientific and Clinical Application of Magnetic Carrier" (U. Hafeli, W. Schutt, M. Zborowski, and J. Teller, Eds.), p. 569, Plenum Press, New York, 1996.
27. M. V. Kutushov, A. A. Kuznetsov, V. I. Filippov, and O. A. Kuznetsov, in "Scientific and Clinical Application of Magnetic Carrier" (U. Hafeli, W. Schutt, M. Zborowski, and J. Teller, Eds.), Plenum, 1996.
28. L. Neel, *Ann. Geophys.* 5, 99 (1949).
29. C. Kittel, *Phys. Rev.* 70, 965 (1946).
30. J. L. Dormann and D. Fiorani, eds. "Magnetic Properties of Fine Particles," Amsterdam, 1992.
31. H. Zijlstra, in "Ferromagnetic Materials," (E. P. Wohlfarth, Ed.), Vol. 3, Amsterdam, 1982.
32. S. Morup, "Nanomagnetism," (A. Hernando, Ed.), p. 93, Kluwer Academic, Dordrecht, 1993.
33. J. L. Dorman, D. Fiorani, and E. Trone in "Adv. in Chem. Phys." (I. Prigogine and S. A. Rice Eds.), p. 288, Wiley, New York, 1997.
34. J. Dai, J. Wang, C. Sangregorio, J. Fang, E. Carpenter, and J. Tang, *J. Appl. Phys.* 87, 7397 (2000).
35. R. F. Ziolo, E. P. Giannelis, B. A. Weinstein, M. P. O'Horo, B. N. Ganguly, V. Mehotra, M. W. Russell, and D. R. Huffman, *Science* 257, 219 (1992).

36. D. Vollath, D. V. Szabo, R. D. Taylor, and J. O. Wills, *J. Mater. Res.* 12, 2175 (1997).
37. E. K. Shalafalla and G. W. Reimers, *IEEE Trans. Magn.* 16 (1980).
38. F. A. Tourinho, R. Franck, and R. J. Massart, *J. Mater. Sci.* 25, 3249 (1990).
39. A. Bee, R. Massert, and S. Neveu, *J. Magn. Mag. Mater.* 149, 6 (1995).
40. H. Schlesinger, *J. Am. Chem. Soc.* 75, 215 (1953).
41. L. Yiping, *J. Magn. Mater.* 79, 321 (1989).
42. S. Wells et al., *J. Phys. Condens. Mater.* 1, 8199 (1989).
43. K. S. Suslick, T. Hyeon, and M. Fang, *Chem. Mater.* 8, 2172 (1996).
44. Q. Wang, H. Xia, and C. Zhang, *J. Appl. Polym. Sci.* 80, 1478 (2001).
45. C. F. Kernizan, K. J. Klabunde, C. M. Sorensen, and G. C. Hadjipanayis, *Chem. Mater.* 2, 70 (1990).
46. J. P. Chen, K. M. Lee, C. M. Sorensen, K. J. Klabunde, and G. C. Hadjipanayis, *Phys. Review B* 75, 5876 (1994).
47. M. E. McHenry, S. A. Majetich, M. DeGraef, J. O. Artman, and S. W. Staley, *Phys. Rev. B* 49, 11358 (1994).
48. S. H. Kilcoyne and A. Gorisek, *J. Magn. Magn. Mater.* 177, 1457 (1998).
49. J. P. Stevenson, M. Rutnakornpiutk, M. Vadala, A. R. Esker, J. S. Riffle, S. W. Charles, S. Wells, and J. P. Dailey, *J. Magn. Magn. Mater.* 225, 47 (2001).
50. K. M. Lee, C. M. Sorensen, K. J. Klabunde, and G. C. Hadjipanayis, *IEEE Trans. Magn.* 28, 3180 (1992).
51. J. A. Lopez-Perez, M. A. Quintela-Lopez, J. Mira, and J. Rivas, *IEEE Trans. Magn.* 33, 4359.
52. J. Mira, J. A. Lopez-Perez, M. A. Lopez-Quintela, and J. Rivas, *Mater. Sci. Forum* 235, 297 (1997).
53. J. Lee, T. Isobe, and M. Senna, *J. Coll. Int. Sci.* 177, 490 (1996).
54. A. Bee, R. Massert, and S. Neveu, *J. Magn. Magn. Mater.* 149, 6 (1995).
55. J. Chatterjee, Y. Haik, and C.-J. Chen, *J. Magn. Magn. Mater.* 257, 113 (2003).
56. E. Mathiowitz, *J. Scan. Micros.* 4, 329 (1990).
57. J. Chatterjee, Y. Haik, and C.-J. Chen, *J. Magn. Magn. Mater.* 246, 382 (2002).
58. L. Josephson, J. Lewis, P. Jacobs, P. F. Hahn, and D. D. Stark, *Magn. Reson. Imaging* 6, 647 (1988).
59. D. D. Stark, R. Weissleder, and G. Elizondo, *Radiology* 168, 297 (1988).
60. C. W. Jung and P. Jacobs, *Magn. Reson. Imaging* 13, 661 (1995).
61. R. Weissleder, G. Elizondo, J. Wittenberg, C. A. Rabito, H. H. Benegele, and L. Josephson, *Radiology* 175, 489 (1990).
62. S. J. McLachlan, M. R. Morris, and M. A. Lucas, *Radiology* 4, 301 (1994).
63. J. Chatterjee, Y. Haik, and C.-J. Chen, *J. Magn. Magn. Mater.* 225, 21 (2001).
64. P. K. Gupta and C. T. Hung, *J. Microencapsulation* 6, 427 (1989).
65. P. K. Gupta, C. T. Hung, F. C. Lam, and D. G. Perrier, *Int. J. Pharmaceutics* 43, 167 (1988).
66. J. Chatterjee, Y. Haik, and C.-J. Chen, *J. Disp. Sc. Tech.* 23, 563 (2002).
67. J. Chatterjee, Y. Haik, and C.-J. Chen, *Coll. Polym. Sci.* 281, 892 (2003).
68. Z. Hu, X. Lu, J. Gao, and C. Wang, *Adv. Mater.* 12, 1173 (2000).
69. P. A. Dresco, S. Z. Vladimir, R. J. Gambini, and C. Benjamin, *Langmuir* 15, 1945 (1999).
70. Y. S. Kang, S. Risbud, J. F. Rabolt, and P. Stroeve, *Chem. Mater.* 8, 2209 (1996).
71. T. Hyeon, S. S. Lee, J. Park, Y. Chung, and H. B. Na, *J. Am. Chem. Soc.* 123, 12798 (2001).
72. J. G. Gribben, *New Engl. J. Med.* 325, 1525 (1991).
73. S. S. Farag, Therapeutic application of immunomagnetic cells: A review, in "Proceedings of the Fourth International Conference on the Scientific and Clinical Applications of Magnetic Carriers," Tallahassee, Florida, 2002. p. 48.
74. Y. Reisner and M. Martelli, *Curr. Opin. Immunol.* 12, 536 (2000).
75. F. Aversa, *New Engl. J. Med.* 339, 1186 (1998).
76. T. A. Lane, *Blood* 85, 275 (1995).
77. C. J. Civin, *Clin. Oncol.* 14, 2224 (1996).
78. M. Korbling, *Clin. Hematology* 12, 41 (1999).
79. A. Manz, N. Graber, and H. M. Widmer, *Sensors Actuators* B1, 244 (1990).
80. J. P. Devilin, "High Throughput Screening," Dekker, 1997.
81. Y. Haik, M. Cordovaz, C.-J. Chen, and J. Chatterjee, *Euro. Cells Mater.* 3, 41 (2002).
82. M. Cordovez, "Point of Care Diagnosis of Acute Myocardial Infarction using Magnetic Immunoassays," Master's thesis, Florida State University, 2001.
83. Y. Haik, C.-J. Chen, and V. Pai, in "Biomagnetic Fluid Dynamics" (W. Shyy, Ed.), Cambridge University Press, Cambridge, UK, 1999.
84. H. Becker and W. Dietz, *Proc. SPIE-Microfluidic Devices and Systems* 3515, 177 (1998).
85. R. J. Jackman, J. L. Wilbur, and G. M. Whitesides, *Science* 269, 664 (1995).
86. R. M. McCormick, R. J. Nelson, M. G. Alonso-Amigo, D. J. Benvegnu, and H. H. Hooper, *Anal. Chem.* 69, 2626 (1997).
87. D. J. Harrison, A. Manz, Z. Fan, H. Ludi, and H. M. Widmer, *Anal. Chem.* 64, 1926 (1992).
88. M. U. Kopp, A. J. de Mello, and A. Manz, *Science* 280, 1046 (1998).
89. P. Wilding, M. A. Schoffner, and L. J. Krick, *Clin. Chem.* 40, 1815 (1994).
90. M. A. Burns, B. N. Johnson, S. N. Brahmasandra, K. Handique, J. R. Webster, M. Krishnan, T. S. Sammarco, P. M. Man, D. Jones, D. Heldsinger, C. H. Mastrangelo, and D. T. Burke, *Science* 282, 484 (1998).
91. K. Hofer, Micro-electro-mechanical systems (MEMS), in "Proceedings of the Fourth International Conference on the Scientific and Clinical Applications of Magnetic Carriers" (A. P. Lee, Ed.), pp. 78–80, Tallahassee, FL, 2002.

92. D. Falkenhagen. *Art. Organs* 20, 420 (1996).
93. M. I. Kilani, P. C. Galambos, Y. Haik, and C. J. Chen, "Proc. ASME Int. Mech. Eng. Cong. Expos." Nov. 11–16, New York, 2001.
94. M. W. Freeman, A. Aarot, and H. H. L. Watson, *J. Appl. Phys.* 31, 404 (1960).
95. R. D. Turner, R. W. Rand, J. R. Bentson, and J. A. mosso, *J. Urol.* 113, 455 (1975).
96. T. Kato, *Gan To Kagakuryho* 8, 698 (1980).
97. K. J. Wider, A. J. Senyei, D. G. Scarpelli, *Proc. Soc. Exp. Biol. Med.* 158, 141 (1978).
98. Y. Morimoto and S. Fujimoto, *CRC Crit. Rev. Therapeutic Drug Carrier Sys.* 2, 19 (1985).
99. Y. Morimoto, M. Okumura, K. Sugibayashi, and Y. Kato, *J. Pharmacobiodynamics* 4, 624 (1981b).
100. A. S. Lubbe, C. Alexiou, and C. Bergemann, *J. Surg. Res.* 95, 200 (1997).
101. F. Schrer, M. Anton, U. Schillinger, J. Hanke, A. Kruger, B. Gansbacher, and C. Plank, *Gene Ther.* 9, 102 (2002).
102. C. Hughes, J. Galea-Laurie, F. Farzaneh, and D. Darling, *Mol. Ther.* 3, 623 (2001).
103. M. Ohgushi, K. Nagayama, and A. Wada, *J. Magn. Reson.* 29, 599 (1978).
104. H. J. Weinmann, R. C. Brasch, W. R. Press, and G. E. Wesby, *Am. J. Roentgenol* 142, 619 (1984).
105. C. Tilcock, E. Unger, P. Collis, and P. MacDougall, *Radiology* 171, 77 (1989).
106. P. Caravan, J. J. Ellison, T. J. McMurry, and R. B. Lauffer, *Chem. Rev.* 99, 2293 (1999).
107. R. Lawaczeck, H. Bauer, T. Frenzel, M. Hasegawa, Y. Ito, K. Kito, N. Miwa, H. Tsutsui, H. Vigler, and H. J. Weinman, *Acta Radiol.* 38, 564 (1997).
108. A. Tanimoto, in "Microspheres, Microcapsules and Liposomes" (Reza Arshady, Ed.), Vol. 3, p. 525, Citus Books, London, 2001.
109. E. Strable, J. W. M. Bulte, B. Moskowitz, et al., *Chem. Mater.* 13, 2201 (2001).
110. K. Turetschek, S. Huber, E. Floyed, et al., *Radiology* 218, 562 (2001).

CHAPTER 12

Computational and Theoretical Approaches to Unraveling the Permeation Dynamics in Biological Nanotubes

Shin-Ho Chung
Department of Theoretical Physics, Australian National University, Canberra, Australia

D. Peter Tieleman
Department of Biological Sciences, University of Calgary, Canada

CONTENTS

1. INTRODUCTION

All living cells are surrounded by a thin membrane consisting of an aggregation of two layers of phospholipid molecules, called the *lipid bilayer*, each molecule being composed of a polar head and a hydrocarbon tail. This thin membrane effectively partitions the external

ISBN: 1-58883-048-9

Handbook of Theoretical and Computational Nanotechnology
Edited by Michael Rieth and Wolfram Schommers
Volume 6: Pages (567–618)

medium from the internal medium. The dielectric constant of the hydrophobic interior of the membrane is about 2, whereas that of the electrolyte solutions on either side of the membrane is 80. Thus, no charged particles, such as Na^+, K^+, and Cl^- ions, can jump across the membrane. The amount of energy needed to transport one monovalent ion, in either direction across the membrane, known as the *Born energy*, is enormously high. For a living cell to function, however, the proper ionic gradient has to be maintained, and ions at times must move across the membrane to maintain the potential difference across the membrane and to generate synaptic or action potentials. To do so, nature has devised specialized, large, transmembrane protein molecules and inserted them densely across the membrane. These transmembrane proteins, known as *ion channels*, form water-filled passages through which ions can freely move in and out.

Ionic channels of living membranes play a crucial role in the existence of all living organisms. All electrical activities in the nervous system, including communication between cells and the influence of hormones and drugs on cell function, are regulated by opening and closing of membrane ion channels. Because these channels are the elementary building blocks of brain function, understanding their mechanisms at a molecular level is a fundamental problem in neurobiology. Moreover, elucidation of how single channels work will ultimately help us find the causes of, and potentially cures for, a number of neurological and muscular disorders.

The measurement of ionic currents flowing through single channels in cell membrane has been made possible by the giga-seal patch-clamp technique [1, 2]. A tight seal between the rim of the electrode tip and the cell membrane drastically reduces the leakage current and extraneous background noise, enabling the resolution of the discrete changes in conductance that occur when single channels open or close. The technique has so far proven to be a powerful tool for characterizing biologically important currents. It appears that all different main types of membrane ion channels have been discovered and characterized in the past 20 years. These fall into three broad classes. There are ligand-gated channels: these channels open when one or possibly two chemical molecules, known as *neurotransmitters*, secreted from nerve terminals, bind to the channel. Some channels, the so-called voltage-gated channels, open in response to step-changes in the electric field. Dipoles or charged residues in the channel protein interact with the sudden change in the membrane potential, and this interaction energy causes the channel to open briefly. The importance of the third class of ion channels, the second-messenger–mediated channels, is now beginning to be realized. A channel is activated either chemically or electrically, and this activation in turn triggers a cascade of biochemical events inside the cell. One of the products released in turn activates a cluster of these second-messenger–mediated ion channels.

All of these different types of channels obey a few general principles. A channel in the closed state acquires a small amount of energy, either binding energy or electrostatic energy, and goes into the open state. When a channel is in the open state, or a higher energy state, ions can flow through the central pore under the influence of chemical or electrical driving force. Some drugs and toxins can interfere with this process, making a channel less likely to open or more likely to stay in the open conformation once it opens. Others may physically occlude the ionic pathways, thus impeding the flow of ions across the open channel. A channel, once open, stays open for various intervals, usually a few milliseconds. How long it will stay open and when it will open again is determined by two random variables. In other words, it obeys the first-order Markovian process. A channel exists, mostly, in one of two states: fully closed or fully open. When a channel is fully open, we see an ionic current of about 1 pA (10^{-12} A).

Despite the wealth of information accumulated over the past two decades, some important questions about how biological channels work remain unanswered. The first among these questions is the detailed dynamical processes underlying the permeation of ions across an open channel. An ion in electrolyte solution does not move freely but drags along a shell of water molecules that are semipermanently bound to it. To give an example, the electrostatic binding energy of a water molecule in the first hydration shell of a Na^+ ion is about 10^{-19} J or 24 kT in room temperature units. This is a huge energy compared to the average kinetic energy of ions. To move across a narrow conduit, such as the selectivity filter of the potassium channel, the ion–water geometry needs to be rearranged, with partially dehydrated

ions interacting electrostatically with the charged residues on the protein wall. Second, all biological ion channels are selectively permeable to a specific ion. Channels generally discriminate anions from cations; some channels select sodium ions but reject potassium ions or *vice versa*. This selectivity mechanism needs to be understood in terms of the interactions of the permeating ions with the surrounding water and protein molecules. Third, what determines the upper-limit of channel conductance? To be functionally effective, a channel must process a large number of ions but, at the same time, it has to be highly selective to specific ionic species. It is a challenge to unravel the design that most satisfactorily reconciles these conflicting requirements. Fourth, what kind of structural changes take place when a channel makes transitions from the closed conformation to the open conformation? Whether the gating of biological ion channels is steric or controlled electrostatically, such as by rotating the orientation of a ring of dipoles guarding the channel gate, remains to be elucidated. Finally, the tertiary structures of all known ionic channels need to be determined. The solutions proposed for these unsolved questions must be consistent with the channel structure deduced by protein chemists.

In the past few years, the field of ion channels has entered into a rapid phase of development. Recently, the molecular structures of the *Streptomyces lividans* KcsA potassium channel, as well as several other potassium channels, two mechanosensitive channels, and a chloride channel, have been determined from crystallographic analyses [3–7]. It is expected that crystal structures of other ion channels will follow these discoveries. Also, as new analytical methods have been developed and the available computational power increased, theoretical models of ion permeation have become increasingly sophisticated. Now it has become possible to relate the atomic structure of an ion channel to its function, through the fundamental laws of physics operating in electrolyte solutions. Many aspects of macroscopic observable properties of ion channels are being addressed by molecular and stochastic dynamics simulations. Intuitive and hand-waving explanations of the permeation and selectivity of ions are beginning to be replaced by quantitative statements based on rigorous physical laws.

Here we give an overview of recent advances in biophysics of ion channels, placing a special emphasis on theoretical approaches that are currently under development. Computational methods of solving complex biological problems, such as permeation, selectivity, and gating mechanisms in ion channels, will increasingly play prominent roles as the speed of computers increases. Our aim in this chapter is to provide some understanding of various methods that have been proposed for treating time-dependent, nonequilibrium processes that underlie the flow of currents across biological ion channels. This review chapter is primarily devoted to the three computational approaches—molecular dynamics, stochastic dynamics and continuum theories—to unravel the inner workings of biomolecules. We give intuitive explanations of the physics underlying each of the methods, referring to more comprehensive publications for mathematical details. We discuss the merits and shortcomings of each computational approach. Detailed accounts of recent experimental findings on ion channels are not given here; the reader is referred to the latest edition of Hille [8], which provides an excellent source of information in this regard.

The chapter is organized as follows. We first describe the principles underlying the continuum theories, stochastic dynamics, and molecular dynamics, stressing the strengths and weaknesses of each approach. We then discuss briefly how these computational tools have been applied in studying selectivity and permeation of ions in biological channels. We do not attempt to give an exhaustive review of the literature, which is given elsewhere [9–15]. Instead, we discuss a few selected examples to illustrate how the theories are applied in modeling of ion channels. We conclude the paper by highlighting future directions for research and identifying the problems in different approaches that need to be resolved.

2. MODELS OF ION CHANNELS

2.1. Conceptual Frameworks for Studying Ion Channels

Henri Poincaré once said: "The goal of science is order. Science is constructed from facts just as a house is constructed from stones, but an accumulation of facts is no more a science than a pile of stones is a house." This statement aptly applies to the field of ion channel

research. Thus, the ultimate aim of theoretical biophysicists has been to provide a comprehensive physical description of biological ion channels. Such a theoretical model, once successfully formulated, will link channel structure to channel function through the fundamental processes operating in electrolyte solutions. It will also concisely summarize the data, by interlacing all those seemingly unrelated and disparate observations into a connected whole. The theory will elucidate the detailed mechanisms of ion permeation—where the binding sites are in the channel, how fast an ion moves across the channel, and where the rate-limiting steps are in conduction. Finally, it will make predictions that can be confirmed or refuted experimentally.

The tools of physics that are employed in this endeavor, from fundamental to phenomenological, are *ab initio* and classical molecular dynamics, stochastic dynamics, and continuum theories. In molecular dynamics simulations, trajectories of all the atoms in a system are followed using Newton's equation of motion. In *ab initio* molecular dynamics, the interaction between the atoms are determined from first-principles electronic structure calculations. As there are no free parameters in this approach, it represents the ultimate approach to modeling of biomolecular systems. But because of the extremely demanding nature of computations, its applications are limited to very small systems at present. In classical molecular dynamics, simulations are carried out using empirically determined pairwise interaction potentials between the atoms. Although it is possible to model an entire ion channel in this way, it is not feasible to simulate the system long enough to see permeation of ions across a model channel and to determine its conductance, which is the most important channel property. For that purpose, one has to go further down to stochastic dynamics, where water molecules that form the bulk of the system in ion channels are integrated out and only the ions themselves are explicitly simulated. The continuum electro-diffusion theory of Poisson–Nernst–Planck equations makes one further simplification known as the mean-field approximation. Here, ions are treated not as discrete entities but as continuous charge densities that represent the space-time average of the microscopic motion of ions. In the Poisson–Nernst–Planck theory, the flux of an ionic species is described by the Nernst–Planck equation that combines Ohm's law with Fick's law of diffusion, and the potential at each position is determined from the solution of Poisson's equation using the total charge density (ions plus fixed charges). The Poisson–Nernst–Planck theory thus incorporates the channel structure, and its solution yields the potential, concentration, and flux of ions in the system in a self-consistent manner.

There is one other approach that has been fruitfully employed to model biological ion channels, namely, the reaction rate theory [8]. In this approach, an ion channel is represented by a series of ion binding sites separated by barriers, and ions are assumed to hop from one binding site to another, with the probability of each hop determined by the height of the energy barrier. Many useful insights have been gleaned in the past about the mechanisms of ion permeation using this approach. The merits and demerits of this theory have been debated extensively in the literature [12, 16–19], to which the interested reader is referred. We will not discuss the rate theories further in this chapter because the model parameters have no direct physical relation to the channel structure, whereas our focus is on the structure–function relationships.

2.2. Continuum Theories

2.2.1. Ohm's and Fick's Laws and Nernst Equation

The flow of ions across a channel is controlled by the potential and concentration differences on the two sides of the cell membrane. For a pore whose radius is relatively large, the current across it will increase linearly with an applied potential. Also, the current here should increase linearly with ionic concentrations in the baths surrounding the pore. In the absence of an applied potential across the pore, ions should flow from the region of a high concentration to a low concentration. The quantitative relationships between the current, applied potential, and ionic concentrations are given by Ohm's and Fick's law.

Ions in electrolyte solutions undergo random Brownian motion with an average collision time τ. When an electric field $\mathbf{E}$ is applied, an ion of mass m_i and charge e accelerates, on

the average, for a time τ, gaining a drift velocity

$$v_d = \frac{e\mathbf{E}}{m_i}\tau \tag{1}$$

With n_i ions per unit volume (where n_i is known as the number density), this leads to a current density of

$$\mathbf{J} = n_i e v_d = n_i \frac{e^2}{m_i}\tau\mathbf{E} \tag{2}$$

The relationship between the current density and the potential gradient (or the electric field) is expressed by Ohm's law

$$\mathbf{J} = -g\nabla\varphi = g\mathbf{E} \tag{3}$$

where g, the conductivity of the electrolyte solution, is $n_i e^2\tau/m_i$. To make contact with the more familiar expression $I = GV$, where I is the channel current, V is the membrane potential, and G is the channel conductance, we consider a cylindrical channel with radius r and length L. Assuming a uniform current density and electric field across the channel, we have $J = I/\pi r^2$ and $E = V/L$. Substituting these values in Eq. (3) yields Ohm's law with the conductance given by

$$G = \pi r^2 g/L \tag{4}$$

The physical basis of the constant of proportionality featuring in Eq. (3) can easily be understood in terms of the microscopic motion of ions in water. Noting that the mean collision time τ is related to the mean free path l and average velocity of the ions $\bar{v}$ by $\tau = l/\bar{v}$ and making use of the equipartition theorem, we can derive a simple microscopic expression for the conductivity in electrolytes:

$$g = \frac{n_i}{n_w}\frac{e^2}{\pi(r_i^2 + r_w^2)}\left(\frac{m_w}{3kTm_i(m_i + m_w)}\right)^{1/2} \tag{5}$$

Here r_w and r_i denote the radius of water molecules and ions, respectively, and n is the number density of ions or water molecules, which is related to the concentration c (in moles/liter) by $n = 1000N_A c$ where N_A is Avogadro's number. As ions in electrolyte solutions are surrounded by a shell of water molecules, the effective mass and radius of an ion must include the hydration waters.

Ohm's law, simple as it may be, can provide us with useful insights about the permeation mechanisms across a transmembrane pore. As an example, we consider the gramicidin A channel, a cylindrical pore whose radius r and length L are approximately 2 Å and 25 Å, respectively. The experimentally determined conductivity of 150 mM K^+ ions is $g = 8.4 \times 10^{-3}$ S/cm in the conventional units of Siemens (S) for conductance. Substituting these values in Eq. (4), we obtain $G = 42$ pS. For an applied potential of 200 mV, the current across the pore is then expected to be 8.4 pA. This is about 3 times larger than the current measured experimentally in gramicidin A (O. S. Andersen, personal communication). This example illustrates that ion permeation across channels is not just a passive process as envisioned in Ohm's law—ions moving from one side of the membrane to the other under a uniform driving field, confined by the channel walls but not interacting with them. In fact, ions do interact with the fixed and induced surface charges on the channel walls that creates energy wells and barriers along the permeation path, the net effect of which is to attenuate the current from that of a purely passive pore. Thus, a correct calculation of the ion-channel interactions is of utmost importance in order to obtain reliable results from a permeation model.

Fick's law provides a similar relationship between the flux of ions and the concentration gradient across a channel

$$\mathbf{J}_n = -D\nabla n \tag{6}$$

where D is the diffusion coefficient, which is specific to each ion and depends on temperature and has units $m^2\ s^{-1}$. Here the subscript n denotes number flux, which is related to the current density by $\mathbf{J} = e\,\mathbf{J}_n$. The minus sign is because the movement of ions is from a region of high concentration to a region of low concentration. As shown by Einstein in 1905, the underlying physics is the same as in Ohm's law, namely, the Brownian motion of ions. The probability distribution in the location of a given particle whose initial position was known would relax as a Gaussian. That is, the shape of the probability distribution will become broader as a function of time, with its center located in the original position. The area beneath the curve, normalized to unity, represents the probability of finding the particle somewhere in the system. The mean square displacement from the initial position grows with time as $\langle \mathbf{r}^2 \rangle = 6Dt + c$ due to the fluctuations, where c, a constant, is zero for a random or Markovian evolution. In the case of ion channels, when one side of the membrane has a higher concentration than the other ($n_1 > n_2$), ions will flow to the other side with a flux, $J_n = D(n_1 - n_2)/L$, where L is the channel length.

There are impermeant anions A^- inside of a cell, but K^+ and Cl^- ions can move across the membrane via ionic channels. Whenever a charged particle is constrained in its movements, a Donnan equilibrium arises. For electroneutrality to exist in the extracellular and intracellular media, the concentrations of K^+ ions inside and outside must obey a Donnan relationship of the form:

$$[K]_i^2 = [K]_o^2 + [A][K] \tag{7}$$

where the subscripts i and o denote inside and outside, respectively. Thus, in the living cell, the concentration of K^+ ions is higher inside than outside and the other way around for Cl^- ions. Also, the asymmetrical distribution of ions across the cell wall gives rise to the membrane potential, which is typically about 70 mV negative inside with respect to outside. Thus, there are both potential and concentration gradients driving the ions across an ion channel. This situation is described by the Nernst–Planck equation, also known as the Nernst transport equation, that combines Ohm's and Fick's laws

$$\mathbf{J}_n = -D\nabla n - \frac{g}{e}\nabla\phi = -D\left(\nabla n + \frac{ne}{kT}\nabla\phi\right) \tag{8}$$

In writing this expression, we make use of the Einstein relation, $g = ne^2D/kT$, which relates the conductivity to the diffusion coefficient. Although the Nernst–Planck equation is primarily used to describe current flow, in the special case of a vanishing current, it make an important statement about the electrochemical equilibrium in cells. Using $\mathbf{J}_n = 0$ in Eq (8) and integrating once, we obtain the celebrated Nernst equation

$$\phi_1 - \phi_2 = -\frac{kT}{e}\ln(n_1 - n_2) = -59\log\frac{c_1}{c_2}(\text{mV}) \tag{9}$$

that gives the potential difference required to maintain the equilibrium when the conentrations are different on the two faces of the membrane. The numerical factor in Eq (9) is obtained using $T = 298$ K. In practice, the Nernst equation is often used to estimate the membrane potential generated by asymmetric solutions in cells. It predicts about a 59 mV change in potential for a 10-fold increase in $[K]_o$.

2.2.2. Poisson–Boltzmann Theory

In recent years, two continuum theories of electrolytes have been fruitfully applied for studying membrane ion channels (for reviews, see Refs. [9, 12, 13, 20]). The so-called Poisson–Boltzmann theory and Poisson–Nernst–Planck theory were originally developed early in the 20th century for bulk electrolytes, and their validity has been firmly established since then [21]. The Poisson–Boltzmann theory has become an important tool for studying proteins and membranes, leading to many insights on the key role played by electrostatic interactions [22]. In ion channels, this theory was initially used to include the effects of ionic atmosphere on the potential energy profile of an ion in schematic channels [23–25]. More recently,

the Poisson–Boltzmann calculation of potential energy profiles has been extended to realistic channel structures in numerous articles [26–35].

The Poisson–Boltzmann theory provides a classical electrostatic description of a system in which fixed external charges, represented by a density ρ_{ex}, are surrounded by mobile ions in a dielectric medium. At equilibrium, the distribution of mobile ions in the system is given by a continuous charge density ρ_{el}, given for monovalent ions, by the Boltzmann factor

$$\rho_{el} = \sum_{\nu} e n_{0\nu} \exp\left[\frac{-e\phi}{kT}\right] \tag{10}$$

where $n_{0\nu}$ is the reference number density of ions of species ν. The average electric potential $\phi(\mathbf{r})$ featuring in Eq. (10) is obtained from the solution of Poisson's equation

$$\varepsilon_0 \nabla \cdot [\varepsilon \nabla \phi] = -\rho_{el} - \rho_{ex} \tag{11}$$

By combining Eq. (10) and Eq. (11), we obtain the following Poisson–Boltzmann equation:

$$\varepsilon_0 \nabla \cdot [\varepsilon \nabla \phi] = 2en_0 \sinh\left[\frac{-e\phi}{kT}\right] \tag{12}$$

This equation cannot be solved analytically, except for a few special cases. To simplify the problem, we make the assumption that the interaction between the ions is small compared with the thermal energy of the ions. That is, we assume $\sinh[e\phi/kT] \ll 1$ and $\sinh[e\phi/kT] \approx [e\phi/kT]$. Then, Eq. (12) for a bulk electrolyte with no fixed charges (i.e., $\rho_{ex} = 0$) becomes

$$\nabla^2 \phi = \kappa^2 \phi \tag{13}$$

where $1/\kappa$ is the Debye screening length or Debye radius, given by

$$\frac{1}{\kappa} = \sqrt{\frac{\varepsilon_0 \varepsilon kT}{2e^2 n_0}} \tag{14}$$

Equation (14) is a linearized form of the Poisson–Boltzmann equation proposed by Debye and Hückel [36]. This equation is easy to solve with the boundary conditions that as $\mathbf{r} \to \infty$, $\phi \to 0$. At room temperature ($T = 298$ K) in water ($\varepsilon = 80$), the Debye length is related to concentration as $\kappa^{-1} = 3.07/\sqrt{c_0}$ Å. For $c_0 = 0.15, 0.30, 0.50$, and 1.00, the Debye lengths are, respectively, 7.9, 5.6, 4.3, and 3.1 Å. The Debye–Hückel theory provides many intuitive pictures and useful insights on the behavior of electrolyte solutions. The solution of Eq. (13) gives the following screened Coulomb potential around a central ion of diameter a:

$$\phi = \frac{e \exp[-\kappa(r-a)]}{4\pi\varepsilon_0 \varepsilon (1+\kappa a) r} \tag{15}$$

The radial density of the screening charge $p(r)$ is proportional to this potential

$$p(r) = 4\pi r^2 \rho_{el} = -4\pi r^2 \varepsilon_0 \varepsilon \kappa^2 \phi = \frac{-e\kappa^2}{1+\kappa a} r \exp[-\kappa(r-a)] \tag{16}$$

In words, when a cation is located at a fixed point, the density of anions around it increases, peaking at one Debye length or $r = \kappa^{-1}$, and then decays exponentially to the background number density. By integrating Eq. (16), we can deduce how big a sphere needs to be to shield the effect of a counter charge located at the origin. The volume integral of this shielding charge is given by

$$q(r) = -\left[1 - \frac{1+\kappa r}{1+\kappa a} \exp[-\kappa(r-a)]\right] \tag{17}$$

revealing that $-q(r)/e$ increases monotonically with r. A sphere of $r = 1/\kappa$ will contain an excess counter charge about 0.25 e, whereas a sphere with $r = 3/\kappa$ will contain an excess charge of 0.80 e. It is important to note that, for a $c_0 = 150$ mM electrolyte under bulk conditions, a length scale of ≈30 Å is required for nearly complete screening of an ionic charge.

2.2.3. Poisson–Nernst–Planck Theory

The Poisson–Boltzmann theory discussed in the previous section is limited to equilibrium situations. To describe nonequilibrium processes, such as ion transport, another continuum theory that is widely known as the Poisson–Nernst–Planck theory is used. This theory, which combines the drift-diffusion equation given in Eq. (8) and Poisson's equation given in Eq. (11), provides a premium description of ion transport problems in many branches of physics and chemistry [21, 37–40]. In recent years, the theory has also been fruitfully applied in ion channels, especially by Eisenberg and his colleagues [41–47]. Here we give a brief outline of the theory and refer to the recent articles by Eisenberg [12, 13] and others [9, 15, 20, 48] for further details and references.

When ions move across narrow channels, they interact with the charge groups on the protein walls, and the electric potential at any given position changes due to their motion. To account for these effects, the potential φ on the right-hand side of Eq. (8) has to be calculated from Poisson's equation. In the Poisson–Nernst–Planck theory, these two equations are solved simultaneously, yielding the potential, concentration, and flux of ions in the system. Here, the flux of ions is described by a continuous quantity corresponding to macroscopic averages of microscopic motion of individual ions.

Due to their nonlinear nature, the two coupled equations are notoriously difficult to solve analytically, except for a few special cases, such as the classical Goldman–Hodgkin–Katz equation [8]. For recent discussions of the analytical treatment of the Poisson–Nernst–Planck equations, the reader is referred to Syganow and von Kitzing [49–51].

When $\mathbf{J}_\nu = 0$ in the drift-diffusion equation,

$$\mathbf{J}_\nu = -D_\nu\left(\nabla n_\nu + \frac{en_\nu}{kT}\nabla\phi\right) \tag{18}$$

then, the Poisson–Nernst–Planck equations trivially reduce to the Poisson–Boltzmann equation with the density given by the Boltzmann factor,

$$n_\nu = n_{0\nu}\exp(-\psi_\nu), \qquad \psi_\nu = \frac{e\phi}{kT} \tag{19}$$

where $n_{0\nu}$ denotes a reference density and ψ_ν is the potential energy expressed in a dimensionless form. Using Eq. (19) for n_ν as an integrating factor in Eq. (18), it can be rewritten as

$$\mathbf{J}_\nu = -D_\nu\exp(-\psi_\nu)\nabla[n_\nu\exp(\psi_\nu)] \tag{20}$$

Let the number densities, $\eta_\nu = n_\nu\exp(\psi_\nu)$, at the boundary points $z = 0$ and L be $\eta_{\nu 0}$ and $\eta_{\nu L}$. Under steady-state conditions and assuming a uniform flux $\mathbf{J}_\nu$ in the z direction, Eq. (20) reduces to 1-D and integrating it from 0 to L gives

$$J_\nu = -D_\nu\frac{\eta_{\nu L} - \eta_{\nu 0}}{\int_0^L \exp[\psi_\nu(z)]\,dz} \tag{21}$$

If one assumes that the potential drops linearly from $z = 0$ to L, Eq. (21) integrates easily and gives the Goldman–Hodgkin–Huxley equation [8]. A similar expression for the density can be obtained by integrating Eq. (20) from 0 to z and using Eq. (21) to eliminate J_ν/D_ν

$$n_\nu(z) = \exp[-\psi_\nu(z)]\left[\eta_{\nu 0} + (\eta_{\nu L} - \eta_{\nu 0})\cdot\frac{\int_0^z \exp[\psi_\nu(z)]\,dz}{\int_0^L \exp[\psi_\nu(z)]\,dz}\right] \tag{22}$$

Finally, substituting Eq. (22) in Poisson's equation, we obtain a Poisson–Nernst–Planck expression for the potential

$$\varepsilon_0\frac{d}{dz}\left[\varepsilon(z)\frac{d}{dz}\phi(z)\right] = -\sum_\nu e\exp[-\psi_\nu(z)]\left[\eta_{\nu 0} + (\eta_{\nu L} - \eta_{\nu 0})\frac{\int_0^z \exp[\psi_\nu(z)]\,dz}{\int_0^L \exp[\psi_\nu(z)]\,dz}\right] - \rho_{ex} \tag{23}$$

There are no known analytical solutions of Eq. (23). With the advent of high-speed computers, the Poisson–Nernst–Planck equations can readily be solved numerically, enabling us to compute the current across ion channels [12, 44]. The Poisson–Nernst–Planck theory is perhaps the simplest form of a nonequilibrium theory that takes into account the shape of the channel, the magnitude and location of charge residues in the channel protein, and the applied electric field and asymmetrical ionic concentrations in the two sides of the channel.

To calculate the current across a channel, it is placed in a simulation system, and a reservoir with a fixed number of ions is attached at each end of the channel. The simulation system is then divided into small rectangular grids, and the Poisson–Nernst–Planck equations are solved at grid points using a finite difference algorithm [44, 48]. The grid size has to be optimized for efficient running of the program. A smaller grid size improves accuracy of the results but also takes a much longer computational time. The required inputs for the algorithm are (i) the channel shape, (ii) the dimensions of the reservoirs and ionic concentrations in each reservoir, (iii) the dielectric constant of the protein and the solution, (iv) the locations and strengths of charges on the channel wall, (v) the membrane potential, and (vi) the diffusion coefficients of cations and anions. Once these parameters are specified, the solutions of the Poisson–Nernst–Planck equations give the concentration and potential throughout the system as well as the ionic currents through the channel.

2.2.4. Modification of the Continuum Theories

The expression for the Debye shielding given in Eq. (16) provides an intuitive guide to the range of validity of the Poisson–Nernst–Planck theory and where and how it fails. An ion near the protein induces surfaces charges of the same polarity on the protein–water interface. When a cation in an electrolyte solution is placed near a slab of protein, water molecules near the ion align themselves such that the oxygen atoms, with their partial negative charges, are positioned nearest to the ion. Because polar or carbonyl groups on the protein wall cannot rotate as freely as water molecules, there will be excesses of hydrogen atoms at the water-protein interface. Viewed from the ion, these excess hydrogen atoms at the boundary appear as surface charges, exerting a repulsive force on it. Macroscopically, we say that a charge q located at a distance d from a slab of protein surface induces surface charges on the dielectric boundary. For an idealized infinite plane, the magnitude of repulsive force this ion experiences is the same as when we place another charge q', at the other side, at a distance d from the surface, and remove the boundary. The magnitude of this image charge q' is related to the relative permittivities of the protein ($\varepsilon_p = 2$) and water ($\varepsilon_w = 80$), given by

$$q' = \frac{\varepsilon_w - \varepsilon_p}{\varepsilon_w + \varepsilon_p} q \tag{24}$$

As the ion comes nearer to the boundary, the repulsive image force it experiences grows as d^2. A similar repulsive force acts on an ion that is about to enter an ion channel. Only, the force is more than an order of magnitude larger in this case because the dielectric boundary is now wrapped around the ion, enhancing the effect. Thus, an ion entering a pore formed by membrane proteins encounters an energy barrier due to the induced surface charges, the height of which increases rapidly with decreasing radius of the pore. This energy barrier, sometimes called self-energy or reaction field, plays an important role in determining permeation properties of ions across a narrow pore. For example, saturation of conductance cannot be explained if one ignores the self-energy barrier of ions. One of the shortcomings of the Poissson–Nernst–Planck theory is in its failure to calculate correctly the magnitude of the induced surface charges and the resulting self-energy barriers.

If the dimensions of the system one is dealing with are much larger than the Debye length, the Poisson–Nernst–Planck theory will give accurate results. As an example, consider ion diffusion across a cylindrical channel whose radius is 30 Å. The charge of an ion near the central axis of the pore will completely be screened by the counterions, so that there will be no induced surface charges on the water–protein boundary due to this ion. Thus, no self-energy

barriers will be encountered by the permeating ions. In this situation, the Poisson–Nernst–Planck theory will correctly predict the current across the pore under various driving forces. Now, let us consider a similar cylindrical channel with a radius of 3 Å, spanning a 30 Å-thick membrane, whose volume is $\sim 10^3$ Å^3. At physiological concentrations, an anion or a cation is expected to be in this pore only 10% of the time on average. As an ion enters the pore, it induces surface charges of the same polarity, which exert a repulsive force on it, pushing it out. Normally, there are no counterions in the pore to screen its charge and hence cancel the repulsive force. In the Poisson–Nernst–Planck theory, the simulation system is divided into small cubic cells with a volume, say, 1 Å^3, each of which contains 10^{-4} of a cationand an equal fraction of an anion. The effect of the fractional charge in each grid on the protein wall is completely screened by the equal and opposite fractional charge present in it. These fractional charges diffuse across the pore, one cubic grid to the other, under the influence of the membrane potential, unencumbered by any induced surface charges on the protein boundary. In short, the ion–protein interaction, which is the dominant force in the process of ion permeation across a narrow pore, is completely ignored in the Poisson–Nernst–Planck theory.

Detailed studies comparing Poisson–Nernst–Planck theory with Brownian dynamics[48] and lattice Monte Carlo simulations [52] have revealed that the Poisson–Nernst–Plancktheory can reliably be applied to study the permeation process across pores whose radius is about 2 Debye lengths (or 16 Å for a 150 mM solution). The magnitude of errors increases rapidly as the radius of the pore becomes smaller. Thus, application of the continuumtheories to biological ion channels, which have radii much smaller than 2 Debye lengths, is not justified. A second problem in this respect is how to treat the ion–ion interactions in the Poisson–Nernst–Planck theory, which are known to play an important role in multi-ion channels such as potassium and calcium channels. These problems need to be resolvedsatisfactorily before this theory can be used to study the structure–function relations in ion channels.

Given the computational simplicity of the continuum theories compared to simulation approaches and their widespread use, it would be desirable to find a solution to the problem of dielectric self-energy highlighted above. A possible approach is to include an explicitself-energy term in the formalism that would mimic the effect of the induced surface charges on ions entering the channel. Such a term would prevent the buildup of large countrion concentrations around a test ion and may eradicate the spurious shielding effects seen in the continuum models. Indeed, there have been several attempts to improve the accuray of the standard Poisson–Boltzmann and Poisson–Nernst–Planck theories by including a specific dielectric self-energy term [53–57]. Near a water–protein interface, a mobile monovalen ion with charge e induces polarization charges of the same sign on the boundary [see Eq. (4)]. These charges, in turn, generate a reaction potential ϕ_R that acts to repel the ion from the boundary. The potential energy due to this reaction field, called the dielectric self-enery, is given by

$$U_{rR} = \frac{1}{2} z_r e \phi_R(\mathbf{r}) \tag{25}$$

Because ϕ_R is itself proportional to e, U_{rR} depends on the square of the ionic charge and hence it is always positive regardless of the valence of ions. Thus, the dielectric self-energy contribution to the total potential energy of an ion is the same for cations and anions. This of course is incompatible with treatment of the mobile ions in the Poisson–Boltzmann theory as a continuous charge density distributed according to the Boltzmann factor

$$n_r(\mathbf{r}) = n_{0r} \exp\left(\frac{-U_r}{kT}\right), \qquad U_r = z_r e \phi(\mathbf{r}) \tag{26}$$

Thus, the Poisson–Boltzmann equation cannot lead to a suppression of ionic densitiesnear a dielectric boundary because it fails to take into account the dielectric self-energy distibution. In theory, the lack of dielectric self-energy in the Poisson–Boltzmann equation ca be

remedied by adding Eq. (25) to Eq. (26) such that

$$n_\nu(\mathbf{r}) = n_{0\nu} \exp\left[\frac{-(U_\nu + U_{\nu R})}{kT}\right] \tag{27}$$

The modified Poisson–Boltzmann equation then becomes

$$\varepsilon_0 \nabla \cdot [\varepsilon(\mathbf{r}) \nabla \phi(\mathbf{r})] = -2e n_{0\nu} \sinh\left[\frac{e\phi(\mathbf{r})}{kT}\right] \exp\left[\frac{-U_R}{kT}\right] - \rho_{ex} \tag{28}$$

As an ion approaches a dielectric boundary, U_R grows rapidly, and the exponential factor in Eq. (28) provides a natural mechanism to suppress the ion densities. For a single ion, the self-energy is well defined. We may first solve Poisson's equation for the single ion and substitute the computed reaction potential in Eq. (25). In a dilute electrolyte solution, where the effect of the reaction field created on one ion by another can be neglected, this simple recipe should be quite adequate for the self-energy correction. However, how to calculate U_R in a many-body system in a self-consistent manner is problematic. At higher concentrations, the approximation of the reaction potential with that of a single ion becomes less accurate. Incorporating the effect of the reaction field created by one ion on another ion is a difficult problem that poses a formidable theoretical challenge.

The Poisson–Nernst–Planck equations can be modified in a similar manner to the Poisson–Boltzmann equation by including a specific dielectric self-energy term in the potential energy of an ion:

$$\mathbf{J}_\nu = -D_\nu \left[\nabla n_\nu + \frac{z_\nu e n_\nu}{kT} \nabla\left(\frac{\phi + \phi_R}{2}\right)\right] \tag{29}$$

The analogy between Eq. (29) and the modified Poisson–Boltzmann equation [Eq. (28)] can be seen most clearly when the current vanishes. Then, Eq. (29) can be easily integrated yielding the modified Boltzmann factor given in Eq. (27).

2.3. Stochastic Dynamics

2.3.1. Langevin Equation

Of the several tools in statistical mechanics that treat the dynamics of nonequilibrium systems, the most widely known is perhaps the theory of Brownian motion. The behavior of an ion in an electrolyte solution undergoing a random type of motion can be described by the dynamical equation of motion in classical mechanics, known as the Langevin equation. The simplest form of stochastic dynamics, which deals with many-particle systems such as liquids and solutions, is Brownian dynamics. Brownian motion is triggered by the presence of a background noise or fluctuations, and the energy gained by a Brownian particle is dissipated in the medium.

When an impulsive force is imparted on a macroscopic particle with mass m in a fluid medium, its subsequent motion is given by Newton's equation of motion

$$m\frac{d\mathbf{v}}{dt} = -m\gamma\mathbf{v} \tag{30}$$

where $m\gamma$ is the friction coefficient, which is roughly proportional to the viscosity of the medium and the radius of the particle. The integral of Eq. (30) is simply given by $\mathbf{v} = \mathbf{v}_0 \exp(-\gamma t)$. Thus, the velocity of the particle exponentially decays to zero due to the frictional force. In other words, the energy of the particle is dissipated to the surrounding molecules in the fluid. For an ion in water, however, the situation is different because it has a similar mass as water molecules. The mean squared velocity of the ion in thermal equilibrium does not decay but remains at $3\,kT/m$ (where k and T are the Boltzmann constant and temperature in Kelvin). This background motion of the ion is brought about by random forces, caused by incessant collisions with the surrounding molecules. Thus, a random or fluctuation force $\mathbf{F}_R$ needs to be added to the right-hand side of Eq. (30) to account for the

motion of an ion performing Brownian motion

$$m\frac{d\mathbf{v}}{dt} = -m\gamma\mathbf{v} + \mathbf{F}_R \tag{31}$$

Equation (31), known as the Langevin equation. is the fundamental equation that descibes the random motion of ions in an electrolyte solution. Both the dissipative and fluctuating forces stem from the same underlying mechanisms, namely, the random bombardmeit of ions by water molecules in thermal motion, and their interrelationship is described b the fluctuation–dissipation theorem of statistical mechanics [58, 59].

The above description of the Langevin equation applies to ions freely diffusing in wter. For ions in the vicinity of a channel, there is an additional, systematic force that influnces their motion, namely, the electric force. It originates from four different sources. First, there is the electric field resulting from the membrane potential. Second, there are ixed charges in the channel protein, and the electric field emanating from them will ad to the field generated by the membrane potential. Third, charges carried by all the ios in electrolyte solution contribute to the total electric field. Finally, whenever any of thesions comes near the protein wall, it induces surface charges of the same polarity at the wter–protein interface. This last component of the force exerted on an ion plays a crucial rle in influencing the motion of an ion attempting to traverse across a narrow pore formed b the protein wall. Each of these four components of the electric force acting on an ion neels to be computed, summed, and added as a third term to the right-hand side of Eq. (31).

For the basic steps in the solution of the Langevin equation and the implementatin of the equation in the Brownian dynamics algorithm, see Refs. [60–62]. When the rndom force featuring in Eq. (31) cannot be considered uncorrelated in time, as is required i the Brownian dynamics approach, then it becomes necessary to invoke the concept of memory function, and the resulting equation is known as the generalized Lagevin equation 59]. Algorithms for generalized Langevin dynamics are given in Refs. [63, 64].

2.3.2. Brownian Dynamics Simulations in Ion Channels

To carry out Brownian dynamics simulations of ion channels, one needs to specify the boundaries of the system. This is a relatively simple problem for one-dimensional Brownian dyamics simulations [9, 65, 66], but requires addition of reservoirs to the channel system i the more realistic case of three-dimensional Brownian dynamics simulations. Here we desribe a simple stochastic boundary that has been used successfully in applications of Brownian dynamics simulations to a number of ion channels [67–72].

Figure 1 shows a schematic illustration of a Brownian dynamics simulation assembly An ion channel representing the potassium channel is placed at the center of the assembly The positions of all the atoms forming the channel are give by its x-ray structure, and the charge on each atom is assigned. Then, a large reservoir with a fixed number of K^+ (or Na^+) and Cl^- ions is attached at each end of the channel (Fig. 1A). The membrane potential is imposd by applying a uniform electric field across the channel (Fig. 1B). This is equivalent to placng a pair of large plates far away from the channel and applying a potential difference between the two plates. Because the space between the electrodes is filled with electrolyte solu ons, each reservoir is in isopotential. That is, the average potential anywhere in the reservir is identical to the applied potential at the voltage plate on that side. A potential dropthus occurs across the channel. When an ion strikes the reservoir boundary during simulatios, it is elastically scattered back into the reservoir. This operation is equivalent to letting a ion enter the reservoir whenever one leaves the simulation system. Thus, the concentratios of ions in the reservoirs are maintained at the desired values at all times. During simulatios of current measurements, the chosen concentration values in the reservoirs are maintaind by recycling ions from one side to the other whenever there is an imbalance due to a condution event. This process mimics the current flow through a closed circuit.

The number of ions that must be placed in each reservoir for a chosen concentrtion depends on the size of the reservoir. Because the computational cost is directly proportonal to the number of ions in the simulation system, it is desirable to have a small reservoir. A the

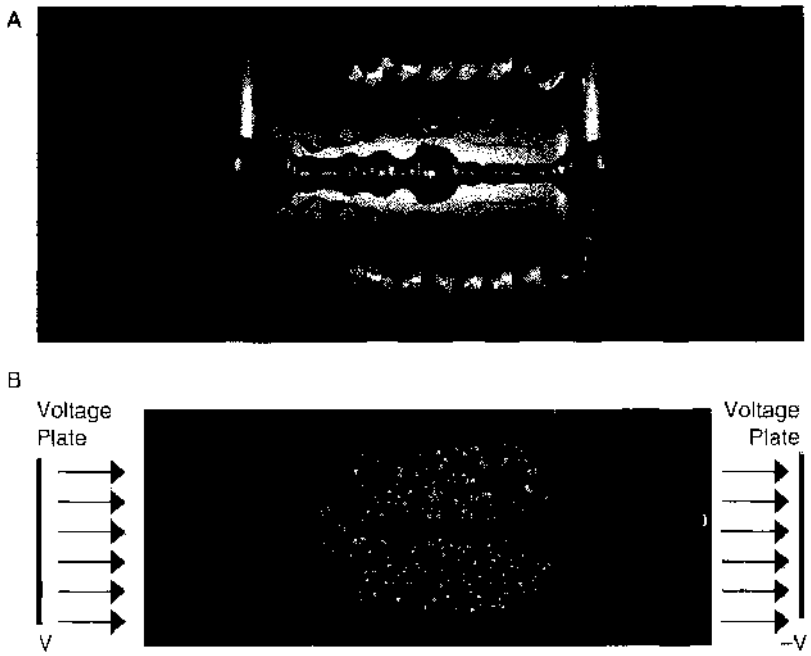


Figure 1. Simulation system using the model potassium channel. (A) Two of the four subunits of the full experimentally determined KscA potassium channel are embedded in a medium of a low dielectric medium, representing the lipid bilayer. To carry out Brownian dynamics, a large reservoir containing a fixed number of K^+ and Cl^- ions is connected to each end of the channel. (B) The potential across the channel is created by applying a constant electric field. This is equivalent to applying a potential difference across two plane plates, located at a long distance away from the channel. Reprinted with permission from [71], S. H. Chung et al., *Biophys. J.* 82, 628 (2002). © 2002, Biophysical Society.

same time, it must be large enough such that the ions in the system are in conditions similar to those in bulk electrolyte solutions. For example, the number of ions near the entrance of the pore should fluctuate according to the binomial distribution. To meet these requirements, an elaborate treatment of boundaries using a grand canonical Monte Carlo method was proposed [73]. Subsequently, Corry et al. [74] showed that, provided the dimensions of the reservoirs are about 3–4 Debye lengths, the simple stochastic boundary as described above gives the same results as the more sophisticated method proposed by Im et al. [73].

In Brownian dynamics simulations, the Langevin equation is solved repeatedly to trace the trajectory of every ion in the assembly. Snapshots of the simulation system are taken at short time intervals for millions of time-steps. At each time-step, the Langevin equation is integrated to obtain the velocity of each ion to determine to which position the ion will move in the next time-step. Once the magnitude and direction of the force acting on each ion is computed, then one can determine which position each ion would have moved to in a short time interval. The new coordinates of all ions in the assembly are deduced, and the calculation is repeated. By repeating this process for a sufficiently long period of time, usually many microseconds, one can deduce how many ions move across the channel in a fixed period of simulation time. Because of the random and frictional forces featuring in the Langevin equation, the procedure for integrating Eq. (31) to obtain the velocity and position of the particle is somewhat complicated. For the detailed methods for integrating the Langevin equation, the reader is referred to Hoyles et al. [62] and Kuyucak et al. [15].

The choice of the time-step is very important for the reliability of Brownian dynamics simulations. If the time-step is chosen too short, the number of times the Langevin equation needs to be solved for a given simulation time increases. On the other hand, the accuracy is compromised if the time-step is too long, although the computational cost is reduced. To understand how a particle undergoing Brownian motion behaves, we integrate Eq. (31) twice to obtain the expression for the mean square displacement [58]

$$\langle x^2 \rangle = \frac{2kT}{m\gamma}\left[t - \gamma^{-1}\left[1 - \exp(-\gamma t)\right]\right] \tag{32}$$

Here γ^{-1} is the relaxation time constant, which corresponds to the time required or a particle that is suddenly displaced to relax back to the original equilibrium position. Fr K^+ and Cl , γ^{-1} is about 30 fs. If we look at a potassium ion once every hundred of fs,(t is much larger than γ^{-1}), its mean square displacement becomes proportional to t

$$\langle x^2 \rangle = \frac{2kT}{m\gamma} t \tag{33}$$

In other words, the ion behaves like a diffusing particle executing a random walk. In conrast, if we examine the same ion at short time intervals, say a few fs (t is much smaller than γ^{-1}), the mean square displacement is

$$\langle x^2 \rangle = \frac{kT}{m} t^2 \tag{34}$$

That is, the ion in a short initial time interval behaves like a free particle moving witl the constant thermal velocity of $(kT/m)^{1/2}$.

In implementing the Brownian dynamics algorithm for ion channels, we need to conider the behavior of ions in these short and long time intervals. When the force experienced by an ion is changing rapidly, as it is when the channel geometry undergoes sudden changs, it is desirable to use a short time step of 1–2 fs. On the other hand, a long time-step of up to 100 fs can be used in the reservoirs with no loss of accuracy. In the algorithm of Chung et al. [69], a short time step of 2 fs is used when an ion is in the channel region where the foce is expected to change rapidly, and a long time step of 100 fs is used when an ion is elsewhere. If an ion is in the channel at the beginning of a 100-fs period, it is simulated by 50 hort time-steps instead of one long time-step; so synchronization among the ions is maintaned. One can simplify the algorithm by omitting the inertial term $m|d\mathbf{v}/dt|$ from Eq. (31 and take time-steps larger than the relaxation time constant. Because of its simplicity, this orm is sometimes used in the literature [73]. Its applicability to ion channels, however, is linited because the rapidly changing forces inside a channel demand use of short time steps.

2.3.3. Validation of the Brownian Dynamics Algorithm

The behavior of the interacting ions deduced from the Brownian dynamics simulaions accords with physical reality. Figure 2 shows the mean square displacement $\langle x^2 \rangle$, the vebcity distribution, and the velocity autocorrelation function $\langle v(0)v(s) \rangle$ obtained from one smulation lasting 500,000 time-steps. Theoretically, $\langle x^2 \rangle$ should be a linear function of tine as given in Eq. (33). The measured slopes for Na^+ and Cl^- shown in Fig. 2A (solid lines are about 7% lower than the predicted values for a bulk solution, which is due to ions scattering back from the boundary, retarding their free diffusion. The velocity distributios of Na^+ and Cl in the system are shown in Fig. 2B. From the equipartition theorem, hese distributions should be Maxwellian. The measured distributions (circles) match closely hose computed from the theoretical distribution (solid lines). Theoretically, the velocity autocorrelation function should be of the form:

$$\langle v(0)v(s) \rangle = \frac{kT}{m} \exp(-\gamma|s|) \tag{35}$$

Thus, regardless of the initial velocity, the successive velocities will be correlated ever a.ime interval on the order of γ^{-1}, the relaxation time constant of the ion. The measured funcions (circles) shown in Fig. 2C decay exponentially, as predicted from Eq. (35) (solid lines). Thus, we conclude that Brownian dynamics simulations can faithfully characterize the motin of ions in a solution confined to a small reservoir.

2.3.4. Uses of Brownian Dynamics in Ion Channels

The ability to compute current flow across ion channels confers a distinct advantae to Brownian dynamics simulations compared to other simulation techniques. Thus, two obious applications of Brownian dynamics in ion channels are in calculation of the current–vctage and conductance–concentration curves, which can directly be compared to the physiological

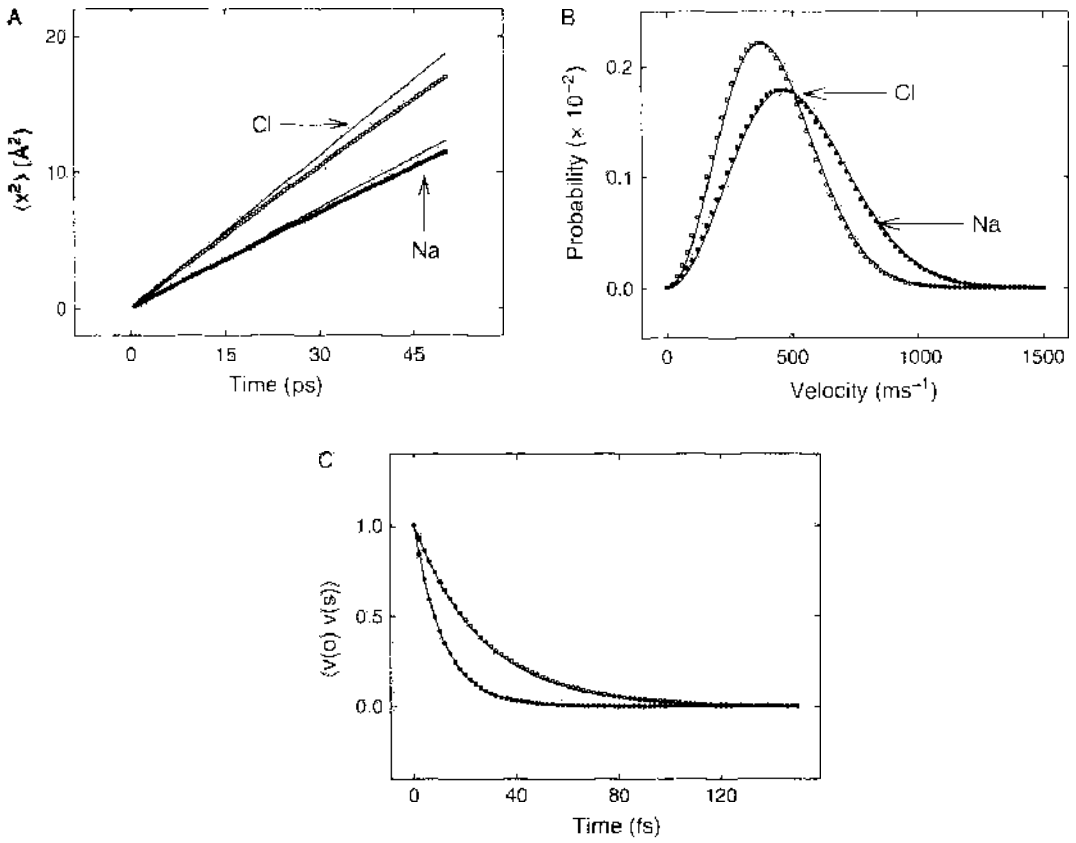


Figure 2. Validation of the Brownian dynamics algorithm. Results of the Brownian dynamics simulations are indicated by open circles for chloride ions and filled circles for sodium ions. (A) The mean-square displacements of ions measured during a simulation period is compared with the diffusion formula, $\langle x^2 \rangle = 2Dt$ (solid line). (B) The velocity distributions of ions are compared with the Maxwellian distribution (solid line). (C) Comparison of velocity autocorrelation functions of ions with Eq. (35). Reprinted with permission from [67], S. C. Li et al., *Biophys. J.* 74, 37 (1998). © 1998, Biophysical Society.

measurements to assess the reliability and predictive power of the method. By simulating a mixture of ions, one can also study selectivity ratios of ions. Because ions with the same valence are treated on an equal footing in Brownian dynamics, this can be used in a straightforward manner only in studies of valence selectivity. For monovalent ions, one has to supplement the potential profiles of ions with free-energy differences obtained from molecular dynamics simulations.

In addition to simple counting of ions crossing the channel, one can carry out a trajectory analysis of ions in the system to determine their average concentrations. This is useful in finding the binding sites and the average number of ions in the channel, both of which are experimentally observable quantities. Blocking of channels by larger molecules or divalent ions can be studied in a similar manner in Brownian dynamics. A prime example is the mole fraction effect, where increasing concentration of one type of ions leads to a reduction in the conductance of another. More refined analysis of Brownian dynamics trajectories can reveal the important transitions that take place during conduction events. This would be very important in interpreting the results of simulations in terms of the rate theories. Alternatively, one can animate the trajectories of ions obtained from the Brownian dynamics simulations and directly watch the permeation of ions in real space-time.

2.3.5. Limitations of Brownian Dynamics

Brownian dynamics makes one very important simplifying assumption. That is, water is treated not explicitly but as continuum. This is both the strength and the limitation of the method. With this simplification, it is possible to trace the trajectories of about 100 ions interacting with a dielectric boundary for hundreds of microseconds, a period long enough to deduce the conductance of an ion channel. A crucial issue is whether such a continuum

approximation in a narrow, biological nanotube can be justified or not. Molecular dynamics studies of water confined in narrow pores have shown that the boundary imposes an order on the water molecules, reducing their polarizability significantly [75–78]. From these studies, it appears that one should use very low values of ε for channel waters in solving Poisson's equation. The role of the dielectric constant in this context, however, is simply to reduce the strength of the electric field by $1/\varepsilon$. Thus, in choosing an effective ε value for channel waters, one has to consider by how much the screening of an ion's field is reduced in a channel compared to the bulk environment. As long as the first hydration shell of ions remains intact in a channel, use of continuum electrostatics with an ε closer to the bulk value may be justified. This criterion is generally satisfied in biological ion channels, including the narrow selectivity filter regions where protein atoms substitute for water completing the solvation shell. A justification of these ideas and extraction of effective ε values from molecular dynamics simulations or otherwise remain as future problems.

In Brownian dynamics, it is assumed that the water–protein interface is rigid and that the positions of atoms forming the channel do not change. Proteins forming channels are certainly not static, but whether their motion or flexibility plays an important role in ion permeation is not known. Such motions are unlikely to be relevant, as a large conformational change of proteins is a much slower process than ion permeation. Nevertheless, the effects of structural polarizability on the dynamics of ion permeation needs to be explored both experimentally and from molecular dynamics simulations. If found to be important, some of the motions of the protein, such as the bending of carbonyl groups, can readily be incorporated in Brownian dynamics modeling of ion channels.

Finally, as pointed out above, size-dependent selectivity among ions with the same valence cannot be understood within the Brownian dynamics framework, and one has to appeal to molecular dynamics or semimicroscopic Monte Carlo simulations [79] for that purpose.

2.4. Molecular Dynamics

2.4.1. General Formalism

Molecular dynamics has now become one of the important tools in modeling of biomolecular systems [80], and availability of several user-friendly packages such as AMBER [81], CHARMM [82], GROMOS [83], and GROMACS [84] have made the method accessible to any researcher. With the increasing speed of modern computers, it may eventually become possible to study the biological processes at a microscopic level and relate the function of a biological ion channel to its underlying molecular structure.

In molecular dynamics simulations, we follow the trajectories of N particles interacting via a many-body potential $V(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$ using Newton's equation of motion:

$$m_i \frac{d^2\mathbf{r}_i}{dt^2} = \mathbf{F}_i \tag{36}$$

where m_i and $\mathbf{r}_i$ denote the mass and position of the ith particle, and the force on it is given by the gradient of the potential V. Because all the atoms in the system (including water molecules) are represented explicitly in molecular dynamics, there are no frictional or random forces to deal with as in stochastic dynamics. This makes the integration of Eq. (36) rather trivial—given the positions and velocities of the particles at time t, at a later time $t + \Delta t$ they are updated to

$$\begin{aligned} \mathbf{r}_i(t + \Delta t) &= \mathbf{r}_i(t) + \mathbf{v}_i(t)\Delta t + \frac{\mathbf{F}_i(t)}{2m_i}\Delta t^2 \\ \mathbf{v}_i(t + \Delta t) &= \mathbf{v}_i(t) + \frac{\mathbf{F}_i(t)}{m_i}\Delta t \end{aligned} \tag{37}$$

although in practice more sophisticated algorithms are typically used. At every time step, the potential function is recalculated using the new positions of the particles, and this process is iterated for a number of steps until a statistically satisfactory data set is generated.

The trajectory data thus generated are stored at certain intervals, which are analyzed later to determine the structural and dynamical properties of a system. Quantities such as free energy, mean square displacement, radial distribution, and other correlation functions are calculated from an ensemble average over several simulations or parts of a long simulation.

The interactions between all atoms in the system are described by empirical potentials. An example of a frequently used potential function is:

$$V(\mathrm{r}^{\wedge}) = \sum_{\text{bonds}} \frac{k_i}{2}(l_i - l_{i,0})^2 + \sum_{\text{angles}} \frac{k_i}{2}(\theta_i - \theta_{i,0})^2 + \sum_{\text{torsions}} \frac{V_n}{2}[1 + \cos(n\omega - \gamma)] + \sum_{i=1}^{N}\sum_{i<j}\left\{4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}\right\} \tag{38}$$

This potential function contains harmonic terms for bonds and angles, a cosine expansion for torsion angles, and Lennard–Jones and Coulomb interactions for nonbonded interactions. The constants k_i are harmonic force constants, l_i is the current, $l_{i,0}$ the reference bond length, θ_i the current, $\theta_{i,0}$ the reference angle, V_n, n, γ, and ω describe dihedral angles (rotations around a central bond), ε and σ are Lennard–Jones parameters (a different pair for each possible combination of two different atom types), q_i and q_j are (partial) atomic charges, and r_{ij} is the distance between two atoms. Using this potential function, the forces (the derivative of the potential with respect to position) on all atoms in the system of interest are calculated and used to solve classical equations of motion to generate a trajectory of all atoms in time. An example of a system is shown in Fig. 3: a simulation of a model of the peptide channel alamethicin in a realistic environment consisting of lipids, water, and ions.

The primary result of the simulation is a trajectory of all atoms in time, from which specific details of the system can be analyzed. This is an exciting idea, because atoms can be followed as they move in real time on a timescale of up to ~100 ns, although longer simulations have also been reported. In principle, any properties that depend on coordinates, velocities, or forces can be calculated, given sufficient simulation time. No assumptions are required about the nature of the solvent, there is no need to choose dielectric boundaries because all atoms are explicitly present, and in principle all interactions (water–ions, water–protein, water–lipid, lipid–protein, etc.) are incorporated. This method automatically includes the dynamics of the ion channel protein itself as well as any dynamic effects of the lipids on the ion channel.

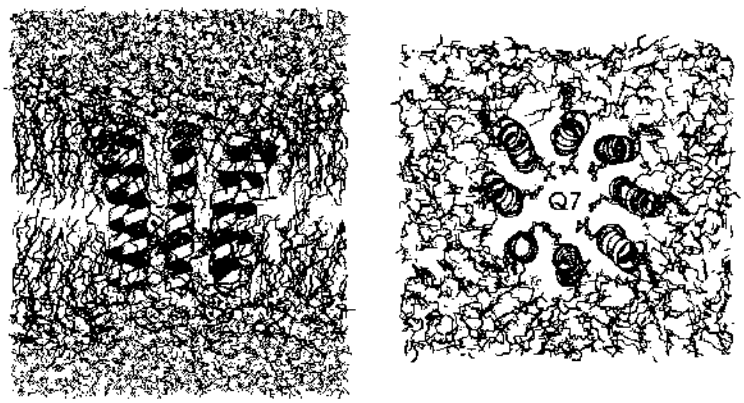


Figure 3. Snapshot of the alamethicin N8 system starting structure after a 25-ps run with position restraints on the peptide. Gln-7 and Phl-20 are drawn as ball and stick, the rest of alamethicin as ribbons, and the phosphatidylcholine headgroup is highlighted in the lipids. Reprinted with permission from [96], D. P. Tieleman et al., *Biophys. J.* 83, 2393 (2002). © 2002, Biophysical Society.

Molecular dynamics simulations have been applied by many groups, to a large selection of ion channels as well as to homology models of complex ion channels [14, 85]. Molecular dynamics simulations so far have been useful for a number of problems:

1. Molecular dynamics simulations have shown that the properties of water and ions in narrow channels change significantly compared to bulk. In particular, in many ases water molecules are strongly oriented due to local electric fields from the protein. Examples of this may be found in porin [77] and channels formed by parallel helix bundles [14]. In addition, water and ion diffusion coefficients are significantly lower than in bulk, which is relevant for coarse-grained simulations.
2. Molecular dynamics simulations have provided insight into the actual process of ion motion in potassium channels, as well as into local structural changes that may explain the experimentally observed differences between sodium and potassium in the potassium channel or different types of ions in the gramicidin channel (see below).
3. Molecular dynamics simulations have been useful to construct models of channels for which the structure is not known, when such simulations are combined with other modeling techniques and experimental data [86, 87].
4. Molecular dynamics simulations have begun to give detailed insight into the interactions of small molecules and toxins with ion channels [88].
5. Molecular dynamics simulations can be used to make models of states of ion channels that are not present in crystal structures such as the KcsA potassium channel [71 89].
6. Molecular dynamics simulations give insight into the effect of the environment (lipids) on the channel protein and *vice versa* [90, 91]. This is an important aspect because there are few other techniques available to study this directly.
7. Molecular dynamics simulations can be used to provide parameters and other information for more coarse-grained simulations that can directly be linked to macroscopic properties such as selectivity and current–voltage curves that can experimentaly be measured. This is potentially a very powerful use of molecular dynamics simulations that has been applied in a number of cases. Two recent examples can be found in interesting studies of OmpF [92] and the KcsA potassium channel [93].

2.4.2. General Properties of Water, Ions, and Protons in Ion Channels

Water Water properties in narrow spaces differ from bulk water in a number of ways depending on such details as the geometry of the space, polarity, and the local electrostatic potential. Because water motions are fast and fall well within the timescale accessible by molecular dynamics simulations, many researchers have described water properties in a range of channel proteins as well as in simplified models of pores. In particular, the diffusion coefficient of water in narrow pores, its orientation, and its behavior in single-file conditions and in pores lined by hydrophobic residues have received considerable attention.

Diffusion Coefficients In general, inside narrow pores, water diffusion coefficient are found to be lower. The exact degree of reduction depends on the details of the environment, as well as to some extent on the method used to calculate the diffusion coefficient. Diffusion coefficients measured from the mean square displacement show a radius-dependent decease of up to a factor of 10 compared to bulk, as shown in Fig. 4 [14].

Strictly speaking, these values are not diffusion coefficients because the Einstein relation used to calculate them is only valid in isotropic solutions. Instead, these values contain a significant component due to the local thermodynamic potential, which one can correct for. This becomes particularly important in permeation theories that require both local diffusion coefficients and the potential of mean force [92]. The "true" diffusion coefficients appear typically less reduced compared to bulk than values obtained from the Einstein relation. Single-file diffusion (isotropic in one dimension) follows a different scaling law, in which the mean square displacement is proportional to the square root of time rather than time itself. This can be verified in simulations of nanotubes. The behavior of water transport through carbon nanotubes and model hydrophobic channels has been characterized in great detail [94, 95].

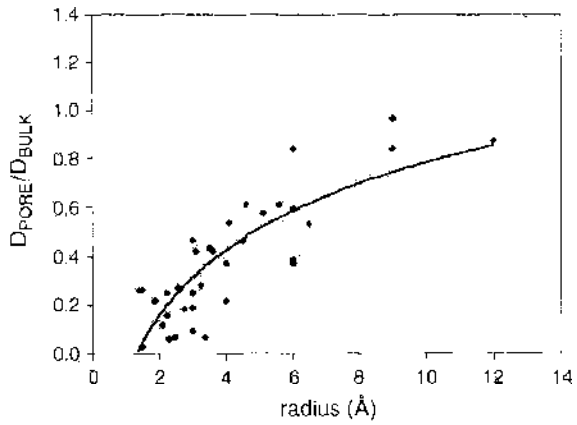


Figure 4. Water diffusion in pores. The ratio of self-diffusion coefficient of water molecules in pores to water in the bulk state (D_{pore}/D_{bulk}) is plotted as a function of pore radius. The data is a compilation summarizing results from simulations of water in smooth cylindrical cavities, simple hydrophilic cylinders, model protein pores, alamethicin channels, models of the nicotinic acetylcholine receptor, and the KcsA potassium channel. The fitted line is simply to guide the eye. Reprinted with permission from [114]. D. P. Tieleman et al., *Q. Rev. Biophys.* 34, 473 (2001). © 2001, Cambridge University Press.

Water Orientation Inside Channels Strong local electric fields are present in many channels. Examples are porin [77], in which the narrowest part of the pore is lined on one side by positively charged residues and on the opposite side by negatively charged residues, and channels formed by parallel alpha-helices such as alamethicin [96], as shown in Fig. 5. These channels show a remarkable degree of water orientation due to interactions with charged residues or the helix dipole moments.

More generally, narrow channels necessarily distort the hydrogen bonding network of bulk water. Perhaps more interestingly, hydrophobic walls or protein walls in general cause a range of new phenomena on a nanoscale that have received considerable attention recently.

Water in Hydrophobic Pores Many biological channels appear to have water-filled pores lined by hydrophobic residues. This has been observed in proteins as varied as rhodopsin, aquaporins, the mechanosensitive channel known as MscL, and the nicotinic acetylcholine receptor [94]. This seems counterintuitive at first sight, but one possible reason for the presence of these hydrophobic channels is that a hydrophobic pore allows control of gating by reversible filling and emptying. Several simulation papers have studied the behavior of

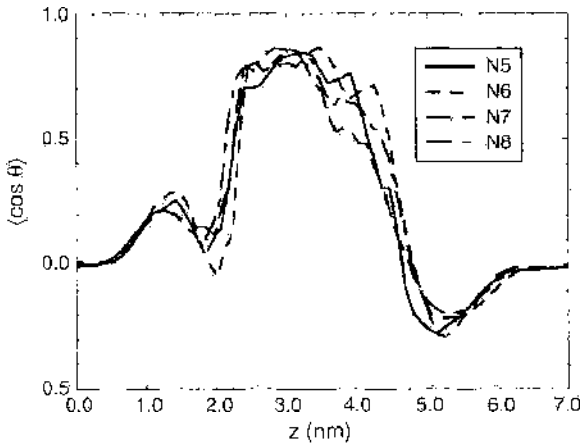


Figure 5. Orientational order of water molecules along the pore axis for N5, N6, N7, and N8 alamethicin model bundles consisting of five to eight helices. The N4 channel does not form a continuous channel, but the orientation of the water molecules is similar. The peaks at ~1.5 and 5.5 nm are due to ordering of water molecules against the lipid interface with the hydrogen atoms pointing toward the bilayer center and the oxygen atoms pointing toward the water layer. Note that details of the orientation profile depend on the details of the treatment of electrostatic interactions. Reprinted with permission from [96]. D. P. Tieleman et al., *Biophys. J.* 83, 2393 (2002). © 2002, Biophysical Society.

water in such hydrophobic channels in simple model systems. One such system is the carbon nanotube [95]. This tube alternates between completely filled and completely empty, with very rapid filling and emptying.

A second approach uses "methane" spheres to build a channel, in which the polarity of the walls can be tuned by added dipoles. Beckstein et al. [94] performed simulations of an artificial channel in a membrane made of "methane-like" atoms, basically hydrophobic spheres that were harmonically restrained to generate pores of varying lengths and geometry (see Fig. 6). They found that there is a limiting radius of the pore below which water no longer fills the channel. Only a slight increase in the radius will make the channel fill with water. Introducing dipoles in the channel walls to make the pore lining more hydrophilic makes it more favorable for water to fill the pore. Interestingly, the pore seems to fill and empty in rapid bursts [97]. This appears to be due to density fluctuations in the bulk solution [94].

A third recent study considered water dynamics in a narrow, perfectly cylindrical channel in a dielectric slab [98]. This study solved a number of technical problems with combining atomistic detail (water) with continuum electrostatics concepts (low dielectric slab as membrane). It also showed that at a certain (realistic) threshold radius, the channel fills intermittently with water. The probability of this happening increased strongly when an ion was present, which might also be related to gating mechanisms.

Ions Molecular dynamics simulations allow a detailed investigation of ion diffusion coefficients, binding sites in channels, and permeation pathways. Umbrella sampling and free-energy perturbation have been used in several ion channels to calculate potentials of mean force for ion permeation, which give the energetic barriers and wells that determine diffusion through a channel. The KcsA potassium channel and gramicidin A are the most crucial test grounds for such simulations. Several groups have drawn attention to the extreme sensitivity of simulation results on ion–water and ion–protein interaction parameters [14, 85, 99, 100]. The sensitivity arises because the key determinant of ion behavior in narrow channels is the difference in solvation free-energy between the ion in solution and in the protein, a difference between large numbers. When comparing different ions, the relative difference in this quantity becomes important, which is a difficult balance to correctly parameterize.

Despite these difficulties, several groups have reported similar results for the different interactions between sodium and potassium and the selectivity filter of the KcsA potassium channel. There now are several detailed free-energy calculations of sodium and potassium in the selectivity filter of the potassium channel [93, 101, 102]. Berneche and Roux [102]

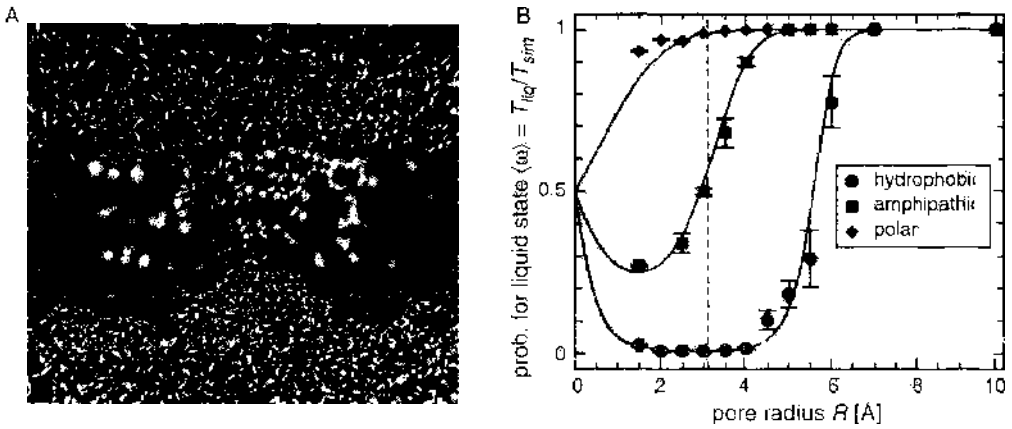


Figure 6. Hydrophobic gating in a model nanopore. (A) A simple model of a nanopore (cyan), embedded in a membrane-mimetic slab (gold), with water molecules (red/white) on either side and within the pore. B) Pore openness versus radius for a hydrophobic nanopore (black) compared with a nanopore with a polar lining (red). The simulations ranged from 52 to 164 ns in length. The polar lining was formed by two dipoles of magnitude 1.52 Å aligned parallel to the pore lining. The vertical broken line represents the pore radius in the gating region of the (closed) nicotinic acetylcholine receptor. Reprinted with permission from O. Beckstein et al., *FEBS Lett.* 555, 85 (2003). © 2003, Elsevier.

specifically parameterized the interactions of potassium with the carbonyl oxygen in the filter to reproduce interaction energies and distances of small molecule fragments and pointed out several inconsistencies in the interaction parameters between ions and water and ions and protein fragments in common force fields. The KcsA potassium channel is of particular interest because the relative occupancy of different monovalent cations in the different sites in the filter is known [103]. In gramicidin A, the interactions are yet again different because the carbonyl dipoles do not point directly at the permeating ions [104]. It is quite challenging (and might not be possible) to develop universal parameters for ions, protein backbone, and water in a classical molecular dynamics context that correctly reproduce all the relevant interactions.

Despite these difficulties, many simulations of gramicidin A have identified the same binding sites, and excellent results have been obtained for finding binding sites for potassium in the KcsA potassium channel [71, 102] and, based on these results, the conductance of the KcsA potassium channel [105]. The bacterial porin OmpF has also been a useful test system for simulations, and details of ion parameters are less likely to be crucial in this large pore in which ions are always solvated by water [92]. Simulations of alamethicin models in different salt concentrations have shown that several ions are nearly permanently present in the channel [106], suggesting ways to engineer channel versions with altered selectivity.

Continuum theories often require diffusion coefficients as input. These can be calculated from molecular dynamics simulations, and similar results are found as for water (see above; the same caveat with regards to the method used to calculate diffusion coefficients applies in this case).

Protons Several smaller channels conduct protons, including gramicidin A, the Influenza A M2 channel, and the engineered LS2 channel. Proton transport presents a special challenge to molecular dynamics simulations because protons move between different water molecules and are not easily treated by a classical potential function. There have been several approaches to extending classical simulations to include protons. The simplest method is to infer the possible behavior of a proton from the behavior of water, which is treated with standard classical potential functions. Although simple, this is not a direct way of studying the behavior of protons.

A next step up in complexity is to use a special water model that can dissociate. Several papers by Pomes and coworkers have investigated proton transport in a number of environments, including gramicidin A, using the PM6 water model [107]. PM6 is a polarizable and dissociable empirical force field consisting of O^{2-} and H^{+} moieties [108]. Voth and coworkers [109, 110] have applied a different method to model proton channels and to models of the Influenza A M2 channel [109, 110]. This model is based on the empirical valence bond (EVB) theory of Warshel [111].

A relatively new method used to simulate proton transport is the QHOP method of Lill and Helms [112], used by de Groot et al. [113] to study proton selectivity in aquaporin 1 (Aqp1), or rather to find an explanation why the water channel aquaporin does not conduct protons. This important problem has recently attracted significant attention, with two other papers presenting similar calculations using different methods, both pointing at electrostatic effects as the reason why protons are not readily transported by aquaporins, although there still are significant differences in interpretation [93, 111, 114].

Finally, one of the most advanced quantum mechanical methods used in channels might be the Car–Parinnello method, a molecular dynamics method that includes part of the system in quantum mechanical detail (using plane wave-based density functional theory, in this case). Dellago et al. [115] have compared the Car–Parinnello molecular dynamics method with the MS-EVB (multi-state empirical valence bond) model in proton transport through carbon nanotubes over short periods of 5–10 ps. Both methods gave comparable results, but the MS-EVB method is much cheaper computationally. No doubt we will see further applications of both these methods to ion channels in the near future.

2.4.3. Limitations of Molecular Dynamics

Despite the power of molecular dynamics due to its ability to provide detailed microscopic dynamics, there are several important caveats and limitations that have to be taken into

account. An obvious limitation is the combination of system size and simulation length, which is mainly determined by available computer power and software efficiency. In particular, the maximum timescale of the order of hundreds of nanoseconds is not enough to determine accurately the average number of ions passing through a channel, except for very wide channels such as porins or simplified geometrical models. This means that by timescale alone, one of our primary objectives, connecting atomic models with current–voltage curves, is still mostly out of reach of molecular dynamics simulations. Obviously, current–voltage curves by direct molecular dynamics simulation would require a number of simulations at different external voltages.

This in itself is not trivial; certain important aspects of ion channel function are hard to incorporate in molecular dynamics simulations of periodic systems with tens of thousands of atoms. One problem is that incorporating transmembrane potential differences is not straightforward, although a reasonable and promising approximation has been developed [116, 117]. Clearly, this is crucial if we want to calculate current–voltage curves. A second problem is that ionic concentrations are difficult to model, but in many cases essential for the behavior of complex ion-channel proteins. Even uniform low-salt concentrations are not straightforward to represent in a simulation, because it is difficult to sample the motions of the 27 K^+ and 27 Cl^- ions that would make up a 0.15 M KCl solution in a simulation with 10,000 water molecules. Such a simulation would also ignore the effect that the lipids have on the local salt concentration near the bilayer [118], which differs significantly from the bulk concentration. Biologically relevant concentrations of calcium or protons are even more problematic: in a typical simulation system, a physiological calcium concentration in the micromolar range would correspond to much fewer than one ion. Concentration gradients are also not routinely simulated, because the use of periodic boundary conditions typically means there is only a single water phase. Although this is not a fundamental problem, simulations become significantly more complex if two explicit water phases with two different ion concentrations would be incorporated in the model [119].

Modeling the effect of pH has similar problems with concentration and the additional problem that it is hard to model effectively proton transfer between different groups in a classical potential. Usually, pH is incorporated by calculating the pK_a of ionizable residues followed by adjusting the protonation state of ionizable amino acids according to the desired pH, although this is not a trivial calculation [120].

There are also limitations inherent in the specific choice of algorithms used. For example, it is now quite clear that electrostatic interactions must be accurately calculated, but due to their $1/r$ dependence (long-ranged compared to the system size) this entails a certain degree of approximation. Several methods have been proposed [99, 121], the most popular of which currently is Particle Mesh Ewald [122], although this may not be the final answer for membrane systems due to its artificial symmetry.

More fundamentally, the simple potential functions used might not be accurate enough for important details of ion–protein interactions across a range of ions [100]. Although any potential energy function could be used in principle, including much more complex versions than the one given in Eq. (38), parameterizing more complex functions is a daunting task, and there might not be sufficient experimental data to test the parameters. In principle, one could treat the entire system (channel, water, ions, membrane) quantum mechanically, taking the distribution of all electrons in the system into account. This is currently not possible in practice and in general not desirable even if it would be possible, because it seems unlikely electronic detail everywhere in the model is relevant for the process (ion conduction) that we would like to study. Mixed quantum mechanics and molecular mechanics simulations are increasingly used in computational studies of proteins, such as enzymes in which only the active site is treated by quantum mechanics, often by density functional or semiempirical methods [123], or rhodopsin where the chromophore and its direct environment are treated by quantum mechanics [124]. This approach has not been widely applied to ion channels [125] and, indeed, it remains unclear whether this level of detail will prove necessary.

Finally, the starting models used for simulations are rather crucial at the moment. Most simulations of ion channels have been carried out on a handful of high-resolution structures,

including gramicidin A [126], OmpF [127], the mechanosensitive channel MscL [128], and the KcsA potassium channel [129]. From an ion-channel perspective, the KcsA potassium channel is by far the most interesting of these because of its close relationship to physiologically relevant eukaryotic ion channels. The crystal structure initially had a fairly low resolution, which caused some uncertainty in the starting structures for the simulations. More recently simulations based on the 2 Å resolution structure have begun to appear in the literature, permitting a comparison with the 3.2 Å structure [129]. Simulations have also been done on homology models of various channels, in which case it becomes even more important to consider carefully the sensitivity of the results obtained to changes in the model [130].

3. COMPUTATIONAL STUDIES ON ION CHANNELS

Here we highlight some of the computational studies carried out on biological ion channels. We will first discuss in some detail two of the most thoroughly studied ion channels. These are gramicidin A, an antibiotic produced by *Bacillus brevis*, and the KcsA potassium channel, whose x-ray structure has been unveiled recently. In the absence of structural information for biological ion channels, gramicidin A has been the main focus of theoretical investigations for a long time. The recent determination of the crystal structure of the KcsA potassium channel [3] has now shifted the attention to biological ion channels, including potassium channels, the ClC Cl^- channels and the mechanosensitive channels. Because there are only a few computational studies on these channels reported in the literature, they are only mentioned briefly in this chapter.

3.1. The Gramicidin Channel

The gramicidin A pore is formed by a linear polypeptide consisting of 15 hydrophobic amino acids in an alternating L–D sequence. It was one of the first antibiotics to be discovered by Hotchkiss and Dubos [131]. The peptide in submicromolar concentrations increases the conductance of lipid bilayers by more than seven orders of magnitude. This enormous increase in conductance owes to the fact that, in membranes, the dimer of the antibiotic peptide gramicidin A forms a cation-selective pore. To form a pore, two gramicidin peptides link briefly head-to-head by hydrogen bonds between their formyl end groups [132], creating a cylindrical passage across the membrane with length 25 Å and radius 2 Å. This narrow lumen size mandates that solutes even as small as urea are not allowed to move across the pore. When an ionic passage across the membrane is formed, stepwise conductance increases in lipid bilayers can be recorded, each channel-opening event lasting between 30 ms and 60 s [133–136]. Gramicidin A has long served as a model for theoretical work on ion channels, partly because of its simple structure and partly because it was the only high-resolution structure known. There is a vast literature on theoretical studies of gramicidin A, and we only present a few papers that illustrate important points learned from gramicidin A.

Its physiological properties are characterized by linear current–voltage curves and relatively large half-saturation concentrations. These observations point to a lack of substantial barriers within the channel. In addition, NMR studies indicate well-defined binding sites near the pore entrances. This wealth of functional data has been matched with an atomic resolution structure since 1971 [132]. For these reasons, the gramicidin channel has played a prominent role in development of permeation models in ion channels. There is extensive literature on modeling of the gramicidin channel, which can be traced from several review articles [10, 11, 126, 137–142].

3.1.1. Semimicroscopic Studies

Modeling of the gramicidin channel evolved from simple electrostatic calculations with rigid dielectric boundaries [143–145] to complicated all-atom molecular dynamics simulations with the channel imbedded in a lipid bilayer and solvated with water [146–148]. Some of the outstanding issues in the gramicidin channel that have been investigated using various theories will briefly be addressed in this chapter.

It is not entirely clear if the lower resolution permeation theories such as the Poisson–Nernst–Planck equations [12, 13] and Brownian dynamics simulations [9, 68, 69], in which the forces are calculated from Poisson's equation, can reliably be applied in modeling the gramicidin channel [10, 11, 35]. First, inferences made from macroscopic electrostatics are valid only in the regions that are large compared to the diameters of water and ion molecules. In the gramicidin channel, whose radius is only 2 Å, the representation of the dielectric as a continuous medium is a poor approximation. Moreover, the water molecules form a linear chain in the pore. Second, these theories assume that the structure of the channel remains rigid, not influenced by the presence of a permeating ion, although the recent nuclear magnetic resonance experiments appear to justify this assumption [149, 150]. Despite these reservations about the applicability of the continuum theories to this channel, the Poisson–Nernst–Planck theory has recently been applied to it, giving an apparently successful description of the current–voltage relations [44–46].

Using Brownian dynamics simulations, Edwards et al. [151] deduced the potential energy profile encountered by permeating ions that correctly reproduces the experimental observations. An example of a potential profile for a K^+ ion constructed from this inverse method is given in Fig. 7A. The profile they derived has two energy wells with the depth U_w, separated by a central barrier with the height U_b. These are two parameters that needed to be determined from the Brownian dynamics simulations by fitting the calculated conductance under different applied potentials and concentrations to the available physiological data. In the

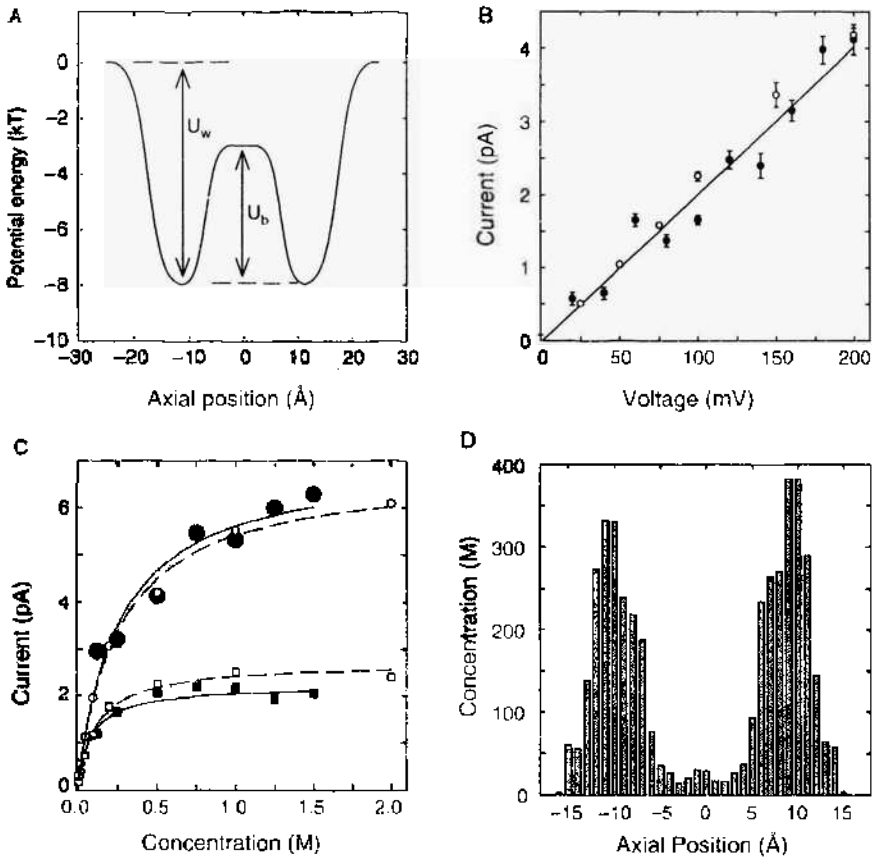


Figure 7. Properties of the gramicidin channel deduced from Brownian dynamics simulations. (A) When the shape of the potential profile experienced by a K^+ ion moving across the pore shown in the figure is incorporated into Brownian dynamics, the simulated results closely match the experimental observations. (B) The current–voltage relationship obtained with Brownian dynamics simulations by using 500 mM KCL (filled circles and solid line) is compared with the experimental data of Andersen (open circles). (C) The current–concentration curves are obtained with two different applied potentials, 100 mV (bottom curve) and 200 mV (top curve). The experimental data of Andersen are superimposed on the figure (open circles). Both simulated and experimental data are fitted with a Michaelis-Menten curve. (D) The dwell histogram for K^+ ions in the pore is obtained with an applied potential. Reprinted with permission from [151], S. Edwards et al., *Biophys. J.* 83, 1348 (2002). © 2002, Biophysical Society

profile illustrated in Fig. 7A, these two parameters, U_w and U_b, are 8 kT and 5 kT, respectively. The current–voltage–concentration profiles obtained with Brownian dynamics simulations using the profile illustrated in Fig. 7A are shown in (B) and (C). A current–voltage curve obtained from simulation of a symmetrical 500 mM KCl solution (filled circles) is compared with the experimental measurements (open circles) in Fig. 7B. The current saturates at two applied potentials: 100 mV (bottom curve of C) and 200 mV (top curve of C). The experimental data of Andersen are superimposed on the figures (open symbols). The calculated curves, fitted with Michaelis–Menten curves, follow the experimental results closely at both driving potentials. The calculated half-saturation value of 250 mM is in good agreement with the experimental value of 230 mM at 200 mV. The corresponding values at 100 mV are 90 mM (simulated) and 105 mM (experimental). Figure 7D shows the concentration profiles of K^+ ions in the gramicidin channel obtained from a symmetric 500 mM KCl solution. The ion concentration is very large at the binding sites and depressed in the middle. There are on average 0.75 ions at each site, indicating that the channel is often multiply occupied. Thus, a satisfactory description of the data can be obtained from the energy profile illustrated in (A).

3.1.2. Molecular Dynamics Studies

In a pioneering study, Åqvist and Warshel [152] calculated a free-energy profile for sodium in the gramicidin channel based on the electrostatic protein-dipoles-Langevin-dipoles (PDLD) model, a semimicroscopic model midway between an all-atom representation and a continuum model, and found reasonable well depths and barriers. They compared the results from this model with free-energy perturbation calculations of the solvation of Na^+ in bulk water and inside the channel, yielding similar results when the induced dipoles due to the low dielectric membrane environment were taken into account. An important aspect of this work is that the low dielectric environment from the membrane was taken into account explicitly. This allows an estimate of the significant effect of this term (about 10 kcal/mol), which has not always been included in later simulations.

Several molecular dynamics simulations have suggested that ion permeation is coupled to motions of the channel. The backbone of the gramicidin peptide exhibits some flexibility with the carbonyl oxygens swinging up to 10–20° [147, 153] to provide adequate solvation for a nearby ion. Roux and Karplus [11] found that there is a "peristaltic" change in conformation as a cation passes along the channel, displacing the carbonyl oxygens of the peptide backbone toward the channel axis. Deformability of the channel seems to play an important role in the dynamics and energetics of permeation of the channel by water and by ions. However, recent high-resolution nuclear magnetic resonance studies of cation transport in the gramicidin channel [149, 150] find that the gramicidin peptide remains rather rigid upon cation binding, and the ion is solvated by just two carbonyl oxygens and two water molecules. A reaction path simulation by Elber et al. [154] also showed that the motion of a permeating sodium ion is coupled to motions of water and the channel. Their calculations suggest that Na^+ does not take a straight path through the channel and questions the validity of a potential of mean force calculation for a single ion at different locations in the channel.

De Groot et al. [155] used molecular dynamics simulations to explain why a particular form of gramicidin makes an excellent water channel. Gramicidin A is found in at least two different forms, and there has been some controversy regarding which form is the main ion channel form, although the so-called head-to-head dimer is now almost universally thought to be the ion channel form [156]. The simulations of de Groot et al. [155] suggest that by removing a formyl group from the N-terminus, the double helix form rather than the head-to-head dimer becomes dominant and is a better channel for water transport.

The gramicidin channel was used to test the effect of the anesthetic halothane on its structure and dynamics [157]. Interactions of membrane-soluble anesthetics with ion channels could be a general mechanism for certain forms of anesthesia. They found some changes in the dynamics of the channel in the close presence of halothane, which might be a mechanism for general anesthetics that could modulate ion channels in neuronal membranes in a similar fashion.

The ability to obtain accurate potentials of mean force for ion permeation is crucial to link atomistic simulations to macroscopic observables like current–voltage curves. To date, all

such molecular dynamics calculations have lead to central energy barriers that are too high to allow ion permeation through the gramicidin channel at the observed rates. To give an example, the lowest translocation barrier calculated for Na^+ ions is about 20 kT [158, 159], which would completely suppress flow of ions. The height of this central barrier is reduced somewhat when several corrections are made to take into account the finite size of the simulation box and the underestimated dielectric constant of the lipid interior [160]. However, a reduction in the height of the central barrier is not sufficiently large to replicate accurately the experimental data.

Molecular dynamics simulations deduce the locations of the binding sites in the gramicidin channel consistent with the experiments. The selectivity sequences among monovalent cations, calculated from the free-energy differences, are in agreement with the experimental sequence $Cs^+ > K^+ > Na^+ > Li^+$ [161, 162]. Through the gramicidin channel, the larger ions with smaller hydration energies conduct better, just as in bulk electrolytes.

There are now several high-resolution structures of gramicidin A, solved under different experimental conditions by different methods, that differ in some structural details. Allen et al. [104] simulated two recent structures, one from solid-state nuclear magnetic resonance [163] and the other from solution nuclear magnetic resonance in micelles [164]. These structures differ in side-chain rotamers as well as in backbone structure. Extensive molecular dynamics simulations (100 ns) have shown that both these structures are relevant, but exist in a dynamic equilibrium. They were able to calculate the relative population of the two states and also showed that relaxation of the structure has a significant effect on the size of the barrier for ion permeation. A "relaxed" structure from molecular dynamics simulations has a lower barrier for cation permeation and gives better agreement with the experimental conductance rates than calculations based on the high-resolution structure directly [104]. This somewhat alters the results of Allen et al. [104] who compared the potential of mean force for cation permeation based on simulations of gramicidin A with different force fields. These simulations gave too high a barrier, which might therefore be related to properties of the structure used for these simulations [104].

With the appearance of the crystal structure of the KcsA potassium channel, the theoretical interests in the gramicidin channel have appreciably diminished. Nevertheless, because the gramicidin channel has such a simple structure and is so rich in physiological data, it can still play an important role in the development of permeation models. After all, if a model works in the more complicated case of KcsA but fails in the case of gramicidin A, it would lose much credibility.

3.2. The KcsA Potassium Channel

The unraveling of the crystal structure of the KcsA potassium channel by Doyle et al. 3] at 3.2 Å resolution was a landmark event that will have a lasting impact on ion channel studies. As the KcsA potassium channel is the first biological ion channel whose tertiary structure has been elucidated, it has been a prime target for simulation and modeling studies. These include the mechanisms underlying the permeation of ions across the channel, the basis of ion selectivity, and the conformational changes that occur in the KcsA protein when the channel opens. The KcsA potassium channel and other bacterial channels share their main features with eukaryotic potassium channels, based on evidence from toxin binding studies [165], conductance measurements [166], and from direct substitution of the KcsA filter in Shaker and inward rectifier potassium channels [167]. This last experiment is especially impressive: replacing the entire pore section of the voltage-gated Shaker channel by the pH-gated KcsA results in a voltage-gated hybrid channel. Very recently, Morais-Cabral et al. [168] succeeded in refining the crystal structure of KcsA to 2.0 Å, which allowed direct observation of the binding sites of K^+ ions in and near the selectivity filter, as well as the water molecules that are coordinated with these K^+ ions [169]. These latest results could have ramifications in much wider fields as they turned the KcsA channel into a microscope laboratory for studying the ion–water–protein interactions.

Additional potassium channel structures have confirmed that the central pore region is quite conserved structurally. MthK is a calcium-gated potassium channel that was trapped

in the open state by the presence of calcium [170]. KirBAC is an inward rectifying channel that was crystallized in a closed state [171], and KvAP is the first voltage-gated channel to have its high-resolution structure solved [172, 173]. This structure has led to very interesting controversies because of the unexpected orientation of the S1–S4 domain, which contains the voltage sensor [174].

3.2.1. The Structure of the Potassium Channel

The structure of the KcsA potassium channel determined from x-ray diffraction consists of 396 amino acid residues, or 3504 atoms excluding polar hydrogens. The overall shape of the channel resembles a truncated cone with a central pore running down the center, as shown in Fig. 8. The wider end of the cone corresponds to the extracellular mouth of the channel. The transbilayer pore is formed by a bundle of eight transmembrane helices, four M1 and four M2 helices. The selectivity filter with the potassium channel signature motif TVGYG (threonine-valine-glycine-tyrosine-glycine) is located near the extracellular mouth of the channel. This filter contains distinct ion binding sites that are well resolved in the crystal structures. Below the selectivity filter is a central water-filled cavity, which also shows a well-resolved ion-binding site in the high-resolution structure. The pore-lining M2 helices constrict the intracellular mouth to form a gate region. In this region, the pore radius falls to about 1.1 Å, that is less than the Pauling radius of a K^+ ion, 1.3 Å.

The KcsA potassium channel is known to be activated at low intracellular pH [175, 176]. Paramagnetic spin resonance studies [177, 178] have indicated that the crystal structure of the KcsA potassium channel corresponds to a closed conduction state, and the transmembrane helices forming the intracellular pore move away from the channel axis during gating. Also, in the crystal structure, the radius of the narrowest section of the pore on the intracellular side is 1.2 Å, smaller than the radius of the potassium ion (1.33 Å). To study the conductance properties of the channel, therefore, one needs to construct an open-state model by increasing the intrapore radius of the structure.

In the crystal structure, K^+ ions are observed to occupy four sites in the filter region with approximately equal probabilities [168]. These sites are in between the planes defined by the carbonyl and hydroxyl oxygens of the Y_{78}, G_{77}, V_{76}, and T_{75} residues. Thus, a K^+ ion in one of these sites is solvated by eight oxygens from the neighboring residues. A similar eightfold coordination of ions with water molecules is observed in the cavity and at the extracellular mouth of the channel [169]. These results are interpreted as two K^+ ions, with a water

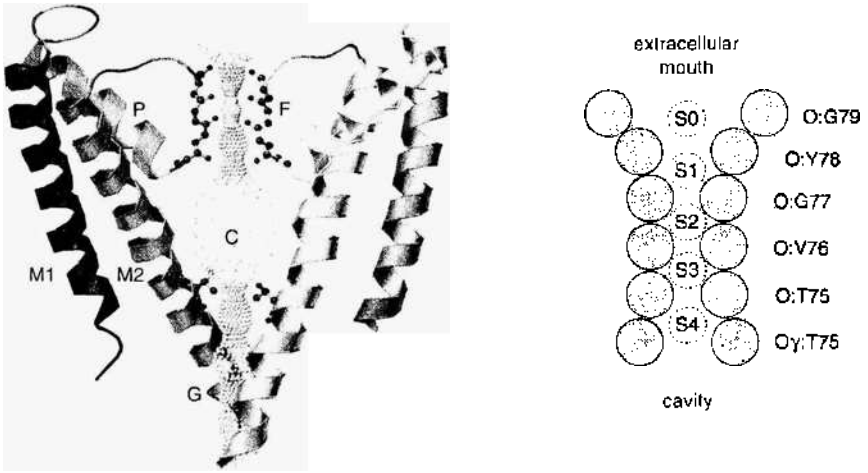


Figure 8. The bacterial (KcsA) potassium channel. (A) Two of the four subunits are shown, viewed perpendicular to the pore axis, with the selectivity filter at the top of the diagram and the intracellular gate region at the bottom. The lining of the channel is shown as a dotted surface. The M1, M2, and P helices are labeled, along with the selectivity filter (F), the central cavity (C), and the intracellular gate (G). (B) The ion binding sites (dotted circle) in the selectivity filter, together with the carbonyl oxygens (gray balls) of the residues lining the pore are indicated. Modified from [14].

molecule between them, permanently occupying either the sites 1–3 or 2–4, and oscillating between these two configurations without any significant free energy barriers. Appearance of a third K^+ ion in the filter disrupts this equilibrium, starting a conduction event. The fact that K^+ ions are eightfold coordinated with oxygen atoms at all sites from entry to exit makes the permeation process through the filter energetically very smooth, especially during the critical dehydration and rehydration steps.

3.2.2. Brownian Dynamics Studies of Ion Permeation

To determine currents flowing across the channel, several groups [69, 71, 72, 99, 105, 179] have carried out Brownian dynamics simulations on the KcsA channel using the experimentally determined channel structure. A qualitative picture of the permeation dynamics can be gleaned from the energy landscape encountered by an ion moving across the channel. Figure 9A illustrates the potential energy of a potassium ion brought into the channel from the intracellular reservoir in the absence of other ions [71]. The energy well experienced by a single ion is deep, reaching 67 kT within the selectivity filter with respect to the intracellular reservoir. The well is deep enough to accommodate two ions. Figure 9B shows the energy profile experienced by a third ion entering the channel already housing two ions. The position of the resident ions when the test ion reaches the center of the channel ($z = 0$ Å) is indicated as upward arrows. In constructing this profile, the resident ions in the channel are allowed to adjust their positions so that both axial and radial components of the force on them are minimized. Thus, the potential profile for multiple ion systems corresponds to the total electrostatic energy required to bring in the charge on the test ion from an infinite distance in infinitesimal amounts. With two ions in the selectivity filter, a third ion entering the channel sees an energy well, created by the mouth dipoles near the intracellular entrance. The ion needs to overcome a central barrier of 7–8 kT to move toward the wide cavity region. This is expected to be the rate-limiting step for conduction. Once in the cavity, the stable equilibrium established by the resident ion is disrupted, and the outermost ion is expelled from the channel. The electrostatic profiles illustrated in Fig. 9 ignore the influence of ionic atmosphere and other dynamic effects such as entropy. Because anions do not enter the channel, neglect of the counter ions is justified. From the profile illustrated in the figure, it is clear that conduction will not take place unless there are two or three resident ions in the channel. For a single ion to traverse the selectivity filter, it has to gain an energy of nearly 70 kT to exit the energy well shown in Fig. 9A.

Brownian dynamics simulations carried out by Chung et al. [71] show that there are three regions in the selectivity filter and cavity where K^+ ions dwell preferentially, as illustrated

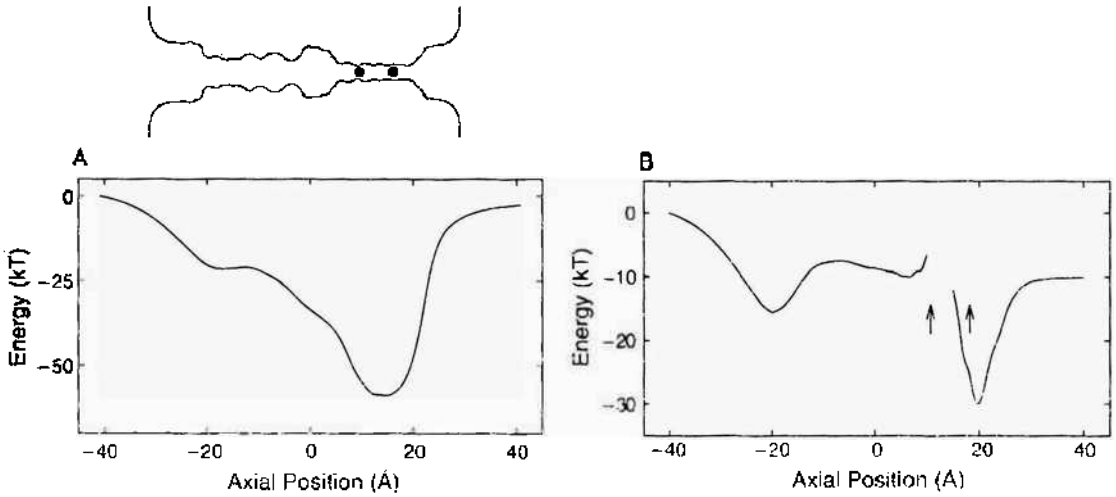


Figure 9. Electrostatic energy profile of an ion traversing the potassium channel. (A) The profile reveals potential energy encountered by a single K^+ ion as it moves across the pore. (B) The potential energy profile is constructed for a third K^+ moving from the intracellular end of the selectivity filter to the extracellular space, given that the filter is already occupied by two resident ions. The locations of the resident ions are shown as filled circles in the inset. Reprinted with permission from [71], S. H. Chung et al., *Biophys. J.* 82, 628 (2002). © 2002, Biophysical Society.

in the histogram in Fig. 10. To construct the histogram, the channel is divided into 100 thin sections, and the average number of ions over a simulation period of 0.1 μs is plotted. The locations of their maxima are indicated schematically in the inset. The location of their maxima are $z = 4.5$ Å (inside the cavity), 12.3 Å (near the T_{75} carbonyl oxygen), and 19.2 Å (near the Y_{78} carbonyl oxygen). Also, there is another prominent peak in the histogram, centered at -21 Å. Here, guarding the intracellular gate are four negatively charged amino acid residues (E_{118}). The average number of ions is 2.9 in the selectivity filter and the cavity and 0.9 near the intercellular entrance. The ion in the cavity region tends to reside mostly off-axis, as a result of the strong interaction with the oxygen atoms of the T_{74} residue at the base of the pore helix. The positions where ions dwell preferentially are in close agreement with the positions observed in Rb^+ x-ray diffraction maps [3].

There is now a general consensus on the process by which a conduction event across the potassium takes place. The most common situation in the conducting state of the channel has one ion in the left (intracellular) half and two ions in the right (extracellular) half. The ion waiting near the intracellular mouth overcomes a small energy barrier in the intracellular pore to enter the chamber region. Because this system is unstable in the presence of an applied potential, the rightmost ion is ejected from the channel. Another ion enters the intracellular mouth, leaving the system in its original configuration. Naturally, conduction of ions depends on their concentration, applied potential, and the ionization state of the charged residues at the channel entry, and many other states can be involved in the conduction process depending on the values of these variables. For example, at high concentrations and potentials, before the outermost of the three ions on the right side of the channel is ejected from the channel, a fourth ion may enter the left side of the channel. A common feature of all these conduction events is that the presence of 3 K^+ ions on the right side of the channel triggers a conduction event. This prediction of the Brownian dynamics simulations on the mechanism of ion permeation across the filter region has been confirmed by the latest experiments on the KcsA channel [168].

In Fig. 11, the current–voltage and current–concentration curves obtained from Brownian dynamics simulations are shown [71]. The radius of the intracellular gate is fixed at 4 Å. The outward conductance at 150 mV is 147 $\pm$ 7 pS. The relationship is linear when the applied potential is in the physiological range ($V < 150$ mV) but deviates from Ohm's law at a higher applied potential (Fig. 11A). In broad agreement with those determined experimentally [166, 175, 180–182], the current saturates with an increasing ionic concentration (Fig. 11B). This arises because the average time τ it takes for an ion to transit the channel has two components: the time needed for an ion to access the channel (τ_1), and the time it takes an ion to traverse the channel (τ_2). The first process, τ_1, is dependent both on the electric field E across the channel and the ionic concentration $[c]$ in the reservoir, whereas the second process, τ_2, depends solely on the electric field. Thus,

$$\tau_1 = \frac{k_1}{[c]E}, \qquad \tau_2 = \frac{k_2}{E} \tag{39}$$

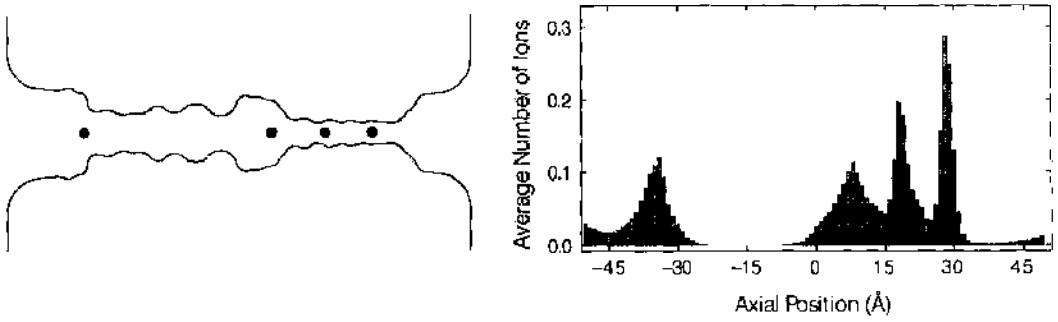


Figure 10. Average number of ions in the potassium channel. The channel is divided into 100 sections, and the average number of ions in each section is calculated over a simulation period of 1 μs. The inset shows the locations of the binding sites (filled circles) in the channel. Reprinted with permission from [71], S. H. Chung et al., *Biophys. J.* 82, 628 (2002). © 2002, Biophysical Society.

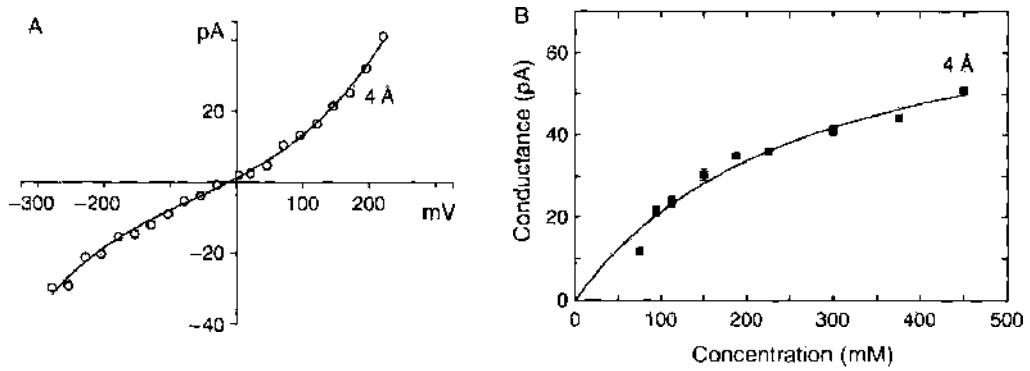


Figure 11. The current–voltage–concentration profile of the potassium channel with the intrapore radius of 4 Å. (A) The magnitude of current passing through the channel with symmetrical solution of 300 mM KCL in both reservoirs is plotted against the applied potential. (B) The outward current are obtained with symmetrical solutions of varying concentrations of KCl in the reservoirs. Reprinted with permission from [71]. S. H. Chung et al., *Biophys. J.* 82, 628 (2002). © 2002, Biophysical Society.

where k_1 and k_2 are constants. The current carried by K^+ ions across the channel is

$$I = \frac{e}{(\tau_1 + \tau_2)} = \frac{eE}{k_1/[c] + k_2} \tag{40}$$

For large concentrations, Eq. (40) approaches a maximum value that we denote by $I_{max} = eE/k_2$. Factoring out k_2 and letting $K_s = k_1/k_2$, Eq. (40) can be written in the form

$$I = \frac{I_{max}}{1 + K_s/[c]} \tag{41}$$

This is the Michaelis–Menten equation, which adequately describes the current–concentration curves obtained from ionic channels. As an example, we give the average times it takes for an ion to move through the various sections of the channel with an applied potential of 250 mV and concentration of 300 mM: the time it takes for an ion to enter the channel from the intracellular side is about 5 ns, the time for this ion to transit across the channel to reach the cavity is 6 ns, and finally, the time for the right-most ion to exit the filter is less than 1 ns. The last two processes, being independent of concentration, are the rate-limiting steps in conduction at high concentrations. It is worth emphasizing that permeation across the filter is much faster than in other parts of the channel. Thus, although the filter plays a crucial role in selecting the K^+ ions, its role in influencing their conductance properties is minimal.

There are many different types of potassium channels, which differ widely in their conductances and gating characteristics [8]. Conductance levels of various types of potassium channels range from 4 to 270 pS. Despite this diversity, they all share the common feature of being highly selective to potassium ions. The amino acid sequence of the peptide chains lining the selectivity filter of all potassium channels is known to be highly conserved [3, 165, 167, 182, 183]. This has important ramifications for modeling of other potassium channels because one can start from the KcsA structure as a template. An example of this approach is the work of Chung et al. [72]. A simplified model of the KcsA channel was developed that reproduced the calculated properties of the atomic-detail model. Then, using such simplified models, they studied how changes from this structure influences its permeation properties. A model channel is generated by rotating the solid curve shown in Fig. 12A around the symmetry axis by 180°. The shape of the channel corresponds approximately to that of the open-state KcsA channel with the full atomic details (see Inset of Fig. 12). Sets of dipoles of various strengths with fourfold symmetry, shown as filled circles and diamonds in Fig. 12A, are placed on the simplified channel model such that their net effect on an ion traversing the pore will be approximately the same as that of an ion traversing the atomic-detail model. Starting from the prototype channel, the aperture of the intracellular pore entrance is expanded from 1.5 to 5.0 Å in steps of 0.5 Å, as shown in Fig. 12B.

The potential energy profiles, obtained in the presence of two resident ions in the selectivity filter, are illustrated in Fig. 13A for three channels with radii 2, 3, and 4 Å. Two prominent

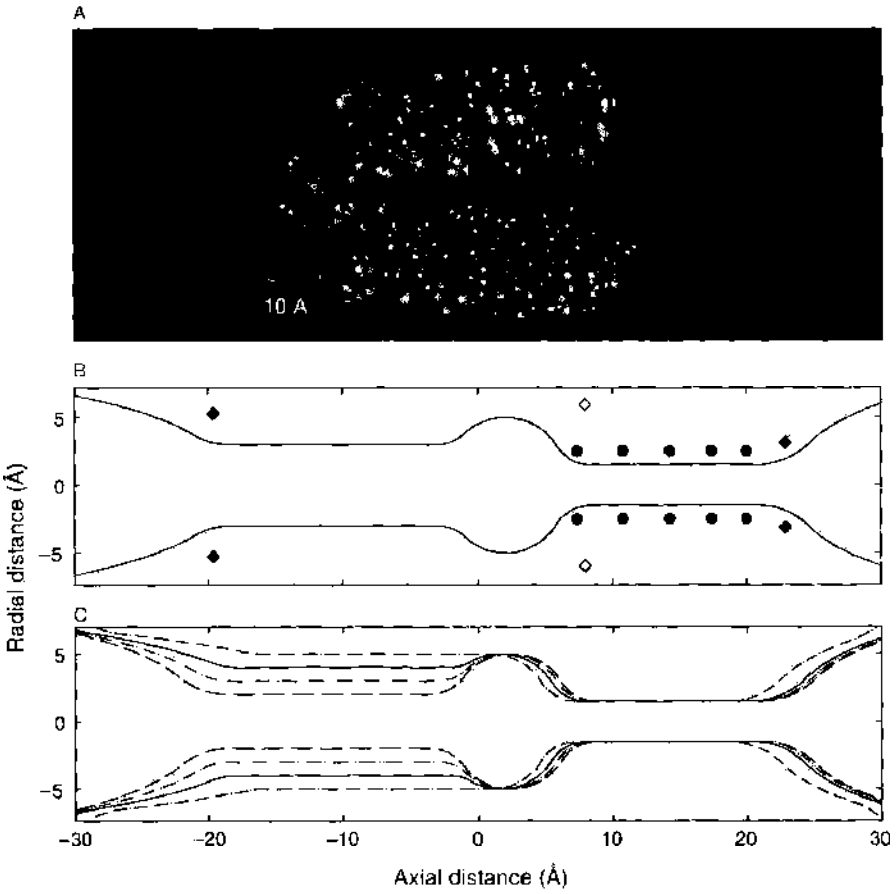


Figure 12. Models of potassium channels. (A) Two of the four subunits of the KcsA potassium channel are shown. (B) The shape of the channel is modified such that the minimal radius of the pore is 3 Å. Solid line shows the outline of a simplified model channel. The positions of dipoles on the channel wall are indicated. Filled circles are 10 of the 20 carboynl and hydroxyl oxygen atoms, open diamonds are N termini of the helix dipole, and filled diamonds are mouth dipoles. (C) Outlines of a set of model channels are superimposed. The radius of the channel facing the intracellular space (left side) is systematically changed from 1.5 to 5.0 Å. Reprinted with permission from [72], S. H. Chung et al., *Biophys. J.* 83, 263 (2002). © 2002, Biophysical Society.

features of these profiles are the energy barrier centered at $z = -10$ Å and the accompanying well at $z = -20$ Å. The height of the energy barrier ΔU an ion needs to traverse the channel from left to right decreases progressively from 7.7 to 0.8 kT as the radius of the intrapore gate is widened from 2 to 5 Å (Fig. 13B). For an outward conduction to take place, an ion sitting in the energy well needs to climb out of the barrier and proceed toward the central cavity. Thus, the enhancement in the barrier height is expected to make an ion's transit from the inner well to the cavity harder, thus retarding the conduction rate. In the inset, the log of the current across the channel is plotted against ΔU, which clearly shows the exponential decrease in conductance with the increasing barrier height. As illustrated in Fig. 13C, the current under the influence of an applied electric field of 2×10^7 V/m (about 200 mV) increases steeply with increasing intrapore radius of the channel. The outward current increases from 0.18 to 48 pA as the radius increases from 2 to 5 Å. With an applied field of 10^7 V/m, the outward and inward currents with the 5 Å channel are 21 pA (187 pS) and 33 pA (210 pS), respectively. Thus, changing the radius of the intrapore region from 2 to 5 Å increases the channel conductance by nearly two orders of magnitude, sufficient to explain the range observed in nature. This gives hope that individual potassium channels can be modeled using Brownian dynamics by taking into account available structural and physiological data.

3.2.3. Molecular Dynamics Studies of Ion Permeation

There are now numerous molecular dynamics simulations based on the initial structure reported by Doyle et al. [3], with varying degrees of approximation of the protein and its

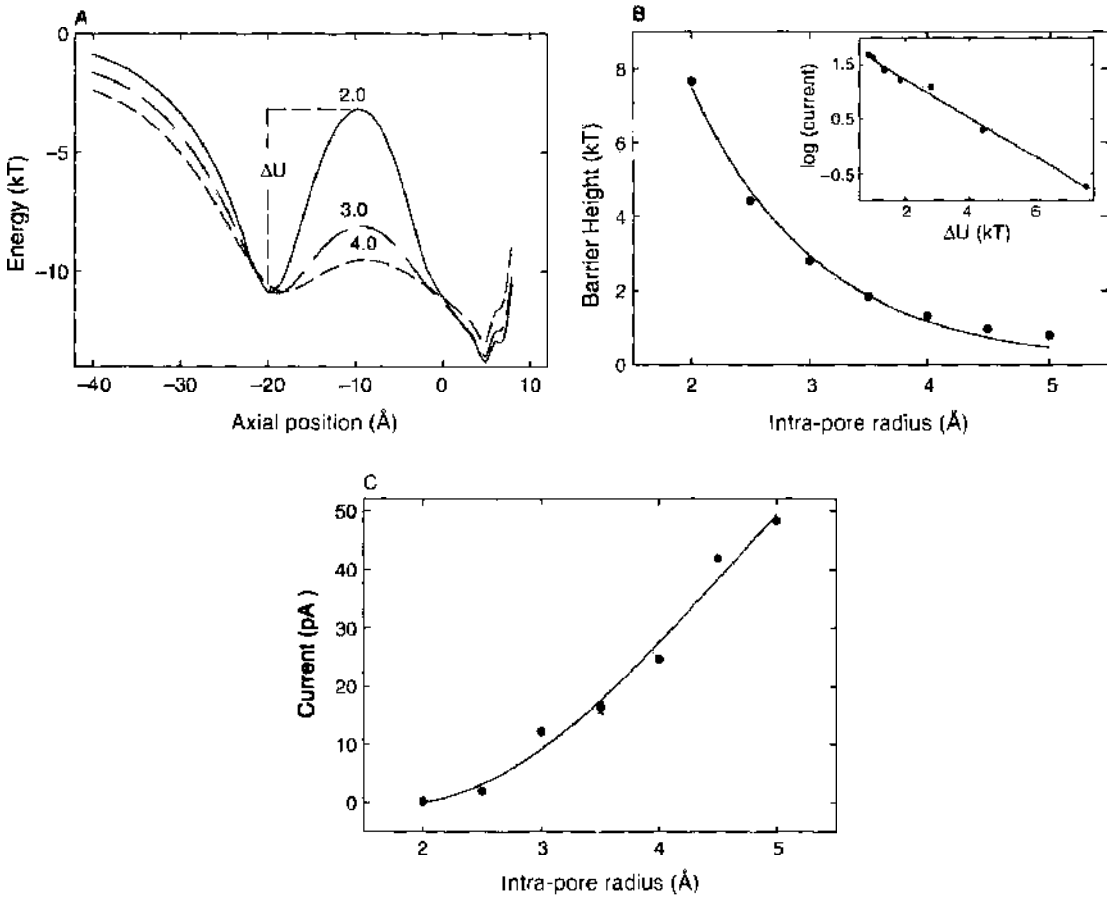


Figure 13. Rate-limiting factors in outward currents across potassium channels. (A) Potential energy profiles encountered by an ion traversing along the central axis of the channel when there are two other ions in or near the selectivity filter are shown for the channels with radii 2 Å (solid line), 3 Å (long-dashed line), and 4 Å (dashed line). Ions need to climb over the energy barrier, whose height is denoted as ΔU, to move across the channel. (B) The barrier height ΔU is plotted against the intrapore radius. In the inset, the outward current in the logarithmic scale is plotted against ΔU. (C) The dependence of outward channel currents on the intrapore radius of the channel is illustrated. The applied field to obtain the current is 2×10^7 V/m. Reprinted with permission from [72], S. H. Chung et al., *Biophys. J.* 83, 263 (2002). © 2002, Biophysical Society.

environment [85, 129, 184]. A simplified model of the KcsA potassium channel with an atomistic filter and the remainder of the protein treated as a hydrophobic continuum was used by Allen et al. [185]. The whole protein, with restraints on parts of the protein to compensate for the missing membrane environment, with water molecules within the pore at either mouth, has been simulated in a number of studies [186]. At the next step up in complexity, the unrestrained protein has been simulated embedded in a bilayer-mimetic environment made up of a "slab" of octane molecules [187] or of "hydrocarbon-like atoms" [101]. Finally, several studies have used a more realistic representation of the environment of the KcsA potassium channel, including a fully solvated phospholipid bilayer [102, 188, 189], and other more recent studies include one using the high-resolution 2 Å structure [129]. Below we summarize some of the results obtained.

3.2.4. Remodeling the Channel Structure

Potassium channels, as well as many other ligand-gated or voltage-gated channels, are controlled by sensitive mechanisms that regulate opening, closing, and in many cases numerous intermediate states. As noted in the previous section, the structure of the KcsA potassium channel [3] is in a closed conformation, so that a natural question is to ask what the open state might look like. Several simulation and modeling papers have tried to address this by various nonequilibrium mechanisms. Biggin and Sansom [190] placed a spherical virtual "balloon" near the intracellular half of the channel that only interacted with the channel, not with water or lipids. By slowly "inflating" this balloon, the channel helices are forced apart,

opening the channel, as shown in Fig. 14. Similar methods in creating open-state structures of the potassium channel based on displacing the helices at their narrowest point were used by others [71, 99, 179]. Experimental support for conformational changes came from electron spin resonance measurements on the KcsA potassium channel [177] and from accessibility studies on a voltage-gated potassium channel that show which part of the pore lining helices is accessible from the intracellular side in the open state [191]. An elegant study of the classical phenomenon of N-type inactivation in a voltage-gated potassium channel also suggests the intracellular side of the channel must be able to open quite substantially compared to the KcsA potassium channel [192]. This was confirmed by the crystal structure of MthK, a calcium-gated channel that was crystalized in the presence of calcium, in an open structure [193]. The conformational changes compared to the closed structures of KirBAC and KcsA are quite substantial. Although they are larger than most predictions by simulation, several of the simulated/modeled structures are quite reasonable.

The KvAP channel, another type of the potassium channel, is an interesting case. Voltage gated potassium channels open and close under the influence of the transmembrane potential. They consist of six transmembrane helices, two of which are homologous to other channels whose structures have been solved. It is also known that S4 carries the "voltage sensor," highly conserved arginines that are crucial for voltage-dependence of voltage-gated potassium channels. The recent crystal structures of this channel and its isolated voltage-sensing "paddle" (composed of segments S1 to S4) challenged existing models of voltage gating, but also raised a number of questions about the structure of the physiologically relevant state.

Monticelli et al. [194] investigated a possible gating mechanism based on the crystal structures in a 10-ns steered molecular dynamics simulation of KvAP in a membrane-mimetic octane layer. In this simulation, the structure of the full KvAP protein was modified by restraining the S2–S4 domain to the conformation of the isolated high-resolution paddle structure. After an initial relaxation, the paddle tips are pulled through the membrane from the intracellular to the extracellular side, corresponding to a putative change from closed to open. This simulation suggested that if an exposed S4 helix is moved through the membrane interior, it would likely create large water defects and possibly drag anions with it into the membrane. It also showed that even this extreme motion has little effect on the stability of

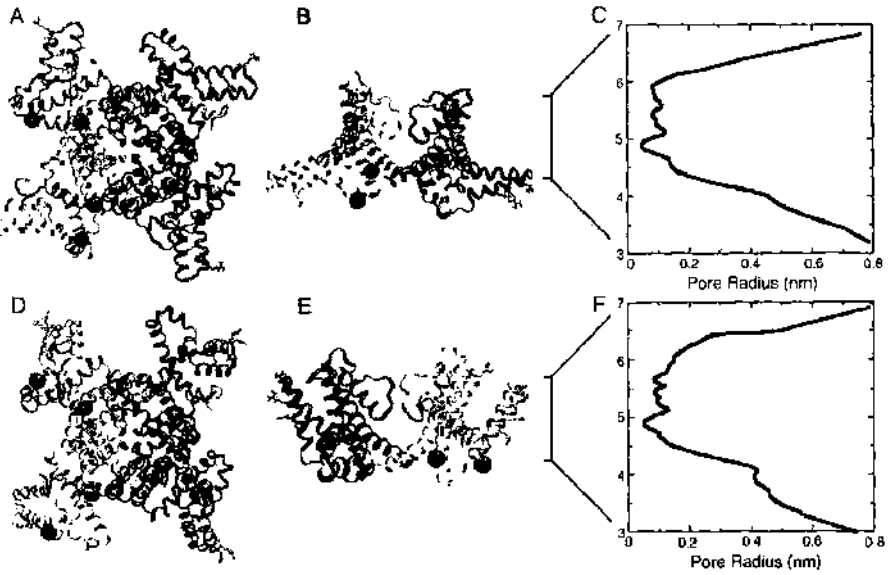


Figure 14. Structures before (A to C) and after (D to F) pulling the tip of the paddle to the extracellular side of the protein. Views are shown from the cytoplasmic side (A and D) and from the plane of the membrane (B and E). Large spheres indicate the positions of the C-alpha carbon of residues S132 and K147 located near the ends of S4 and S5, respectively. As the paddle is pulled up, the S4-S5 linker moves in a clockwise direction (as viewed from the cytosol, A and D), and S132 moves from being below K147 to almost parallel (as viewed from the plane of the membrane, B and E). Curves illustrated in (C) and (F) show the radius of the central portion of the channel. Reproduced with permission from [194], L. Monticelli et al., *FEBS Lett.* 564, 325 (2002). © 2002, Elsevier.

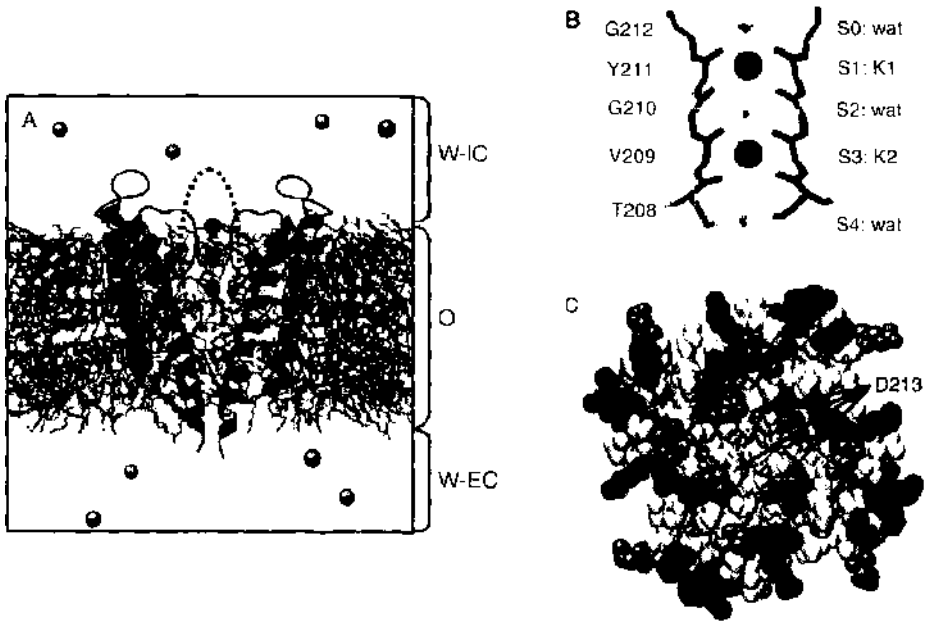


Figure 15. The inner surface of the pore for the three model states: closed (the final structure after 1-ns of simulation based on the crystal structure), expanded (the final structure after simulation Ex3 in the original paper), and relaxed (Rx1 after 11-ns of simulation in the original paper). Reprinted with permission from [190], P. C. Biggin et al., *Biophys. J.* 83, 1867 (2002). © 2002. Biophysical Society.

the pore domain, enforcing the idea that the S1–S4 domain is relatively flexible, as illustrated in Fig. 14.

As simulations reach longer timescales and become more accurate, and as more structural information about different states becomes available, the study of large-scale conformational changes by computer simulation will no doubt become increasingly interesting and might in fact turn out to be one of at most a handful of techniques that can link different structures in time and tie them to their functional roles.

3.2.5. Diffusion of Ions in the Channel

Given the presence of multiple K^+ ions within the selectivity filter of the KcsA potassium channel, a number of simulations have looked at the spontaneous motions of different configurations of K^+ ions and water molecules in the filter. Several studies also considered different types of cations in the selectivity filter: Na^+, K^+, Rb^+, Cs^+. Although these simulations are too short, by several orders of magnitude, to give direct information about the conductance of the channel, they provide information about the local interactions in the filter and other parts of the protein. In the crystal structure, the two K^+ ions within the filter occupied S1 and (S3 or S4). The recent 2.0 Å structure actually shows 7 different binding sites, with ion density (at high K^+ concentration) in sites S1–S4 as well as two more sites somewhat outside the filter on the extracellular side and 1 ion in the cavity. Simulations usually have shown K^+ ions at S2 and S4, but sampling and starting structures are an issue.

Interestingly, two independent simulations predicted there was a favorable location for a potassium ion outside the filter, which was confirmed by the recent high-resolution crystal structure [102, 195]. As discussed below, several free-energy calculations have suggested that the difference in free energy between K^+ ions at S2 and S4 and at S1 and S3 is quite low. This is consistent with the high permeation rate of potassium ions.

In multi-nanosecond simulations, concerted motions of the K^+ ions in the filter are usually seen. This is illustrated in Fig. 16, from which it can be seen that the K–W–K (i.e. ion–water–ion) triplet moves in a concerted fashion [195]. This is direct evidence for concerted single-file motion within a potassium channel selectivity filter. Clearly, this complicates attempts to simulate ion flow through potassium channels as a diffusion process. It is also significant that in most simulations, small (generally ~0.5 Å) changes in conformation of the backbone carbonyls occur. In particular, a "flipping" of the carbonyl of V76 is observed. This is likely

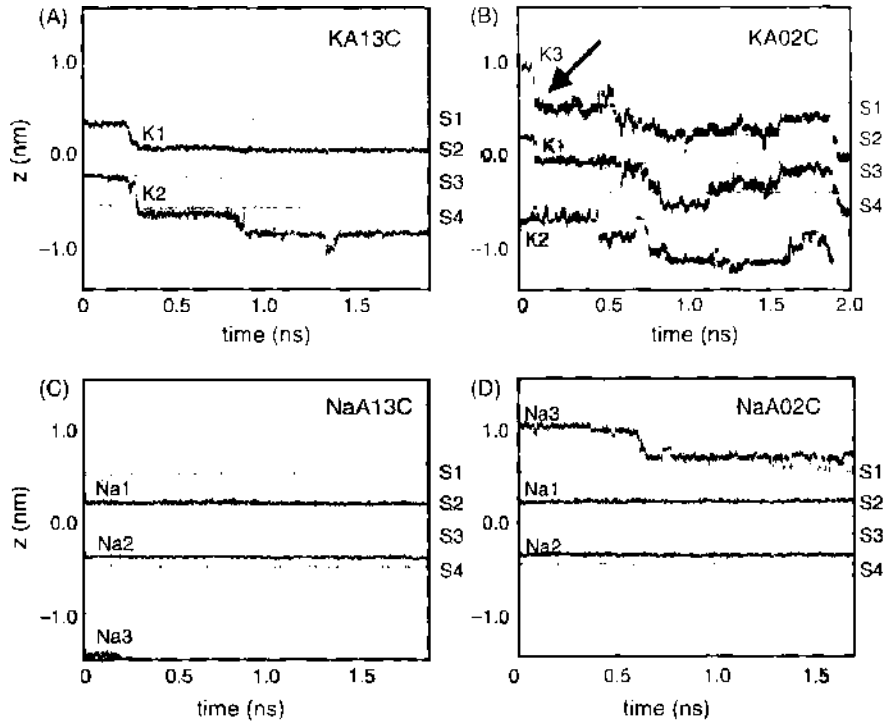


Figure 16. Trajectories (along the pore axis) of K^+ ions (thick black lines) and water molecules (gray lines) for two simulations with different starting configurations in sites S1–S4: (A) KA13C; (B) KA02C. Note that, for clarity, not all water molecules within the filter are shown. The locations on z (pore axis) of the four sites (S1 to S4) defined by the geometric center of the eight oxygen atoms are indicated by the thin black lines. At each point in time, the origin of the coordinate system is defined as the center of gravity of the 16 oxygen atoms that line the selectivity filter. The black arrow in (B) indicates the time at which a K^+ ion enters the selectivity filter from the extracellular mouth (S0) of the channel. Reprinted with permission from [195], I. H. Shrivastava et al., *Biophys. J.* 83, 633 (2002). © 2002, Biophysical Society.

to be significant, because it suggests that the conformation of the selectivity filter is not static, but can undergo dynamic changes on a timescale comparable to that of passage of the ions through the filter. Indeed, at low potassium concentrations (3 mM), ions are seen in the crystal structure mainly at S1 and S4, with some deformation of the filter consistent with observations in molecular dynamics simulations. Simulations based on the higher resolution crystal structure but using the same force field give very similar results [129, 195].

3.2.6. Energetics of Permeation

A number of groups have used atomistic simulations to explore the energetics of permeation of the KcsA potassium channel. Allen et al. [78] have calculated free-energy profiles for K^+ and Na^+ ions in a simplified model of a potassium channel, based on a channel-shaped hydrophobic pore onto which a model of the KcsA filter is grafted. Their results broadly support the "rigid filter" model of potassium channel selectivity. However, the sensitivity of the results to initial assumptions of the rigidity of the filter is a little unclear. In a subsequent paper, the same authors [196], using a complete model of the protein (but omitting the surrounding bilayer), found that the free-energy differences between K^+ and Na^+ were about half those with the simplified model. Several other groups have calculated potentials of mean force for ions in the selectivity filter. Åqvist and Luzhkov [101] showed that occupancy of sites S2 and S4 of the filter (see Fig. 8B) by two K^+ ions was more favorable (by ~2 kcal/mol) than occupancy of sites S1 and S3. Other configurations were of higher free energy. Thus, a permeation model based on switching of pairs of K^+ ions between these two configurations was proposed. Berneche and Roux [105] used umbrella sampling to calculate a two-dimensional free-energy map describing possible pathways for translocating ions and suggest a plausible mechanism involving correlated motions of at least three ions and water on a relatively flat energy landscape. Based on this map, they used Brownian

dynamics simulations to calculate current–voltage curves that gave excellent agreement with experimental measurements of the conductance of the KcsA potassium channel [105]. Combined, these two papers show the power of a hierarchical approach to link atomistic models to macroscopic observables, through potential of mean force calculations.

Another recent study [93] calculated potentials of mean force using free-energy perturbation in KcsA. As such calculations are becoming increasingly feasible on standard computers, a closer link between microscopic simulations and experimental measurements becomes possible.

3.2.7. Selectivity

Why are potassium channels so selective for potassium over sodium? The key differences between potassium and sodium appear to be only a small difference in radius and in polarizability. On the basis of the x-ray structure of the KcsA potassium channel, it has been suggested that a "rigid" selectivity filter provides stronger cation–oxygen interactions for K^+ ions than for Na^+ ions. Thus, the energetic cost of dehydrating K^+ ions is repaid by ion–protein interactions, whereas ion–protein interactions are too weak to balance the cost of dehydrating Na^+ ions. Several simulations have tried to address this question. The accuracy of ion–protein and ion–water interactions is a key concern for this type of calculation, and errors in common parameters can be quite substantial [197, 198]. For instance, an accurate balance between ion–carbonyl and ion–water interactions is difficult to achieve for different types of cations, and specific parameterization of the filter region has been used by Berneche and Roux [197] and for a different force field for simulations of the voltage gated potassium channel KvAP by Monticelli et al. [194].

The deciding factor for selectivity in channels is that of the free energy of permeation or how the free energy of the system varies as different species of ion pass through the channel. Such calculations can yield the difference between two species of ions at a particular location, in addition to the full potential of mean force for moving a particular type of ion. Allen et al. [196] calculated that the free energy (for a $K^+ \rightarrow Na^+$ transformation) is positive within the filter region, which means it is more favorable for a potassium ion to be in the filter than it is for a sodium ion. However, the exact figure arrived at was somewhat sensitive to the nature of the restraints applied to the filter during the simulation. Åqvist and Luzhkov [101, 199] have performed more detailed free-energy perturbation calculations. Their results also supported the "rigid filter" model of potassium-channel selectivity. However, it should be noted that in all three of these simulation studies, it is not clear that the filter had time to fully "relax" around the different species of cation. Longer molecular dynamics simulations of the KcsA channel with K^+ ions or with Na^+ ions in the filter suggest that the filter may be able to alter its conformation such that Na^+ ions can bind tightly within (and thus block) the filter. The geometry of interaction with Na^+ ions with the filter appears to be different from the geometry of interaction of K^+ ions [129, 195]. Examples of snapshots with Rb^+ and Na^+ in the filter are shown in Fig. 17.

Furthermore, long simulations with either K^+ or Na^+ ions at the extracellular mouth of the filter suggest a degree of selectivity in terms of which ions enter the filter [195, 200]. It is clear that very careful simulations are required to obtain the correct balance of ion–water, ion–protein, and protein deformation energies. There is experimental data for other cations, such as Rb^+ [103, 168]. In principle, these could be simulated too, but they require additional testing of parameters because they are not commonly used in biomolecular simulations.

One of the most recent and quite elaborate studies on the behavior of the selectivity filter and ions in the selectivity filter was carried out by Domene and Sansom [129], in which different cations (Na^+, K^+, Rb^+, Cs^+) were simulated in the high-resolution crystal structure of the KcsA potassium channel. This work can be compared to similar simulations on Na^+ and K^+ in the selectivity filter based on the 1998 crystal structure [195]. These older simulations missed several refinements, such as the inclusion of water molecules behind the selectivity filter. Nonetheless, in both simulations, significant flexibility of the filter was observed, as well as concerted motions of ions and water. In particular, pronounced distortions of the filter occur when no ions are present, which is also found from crystallographic studies at low salt concentration [103]. The two most readily permeant ions, K^+ and Rb^+, are similar

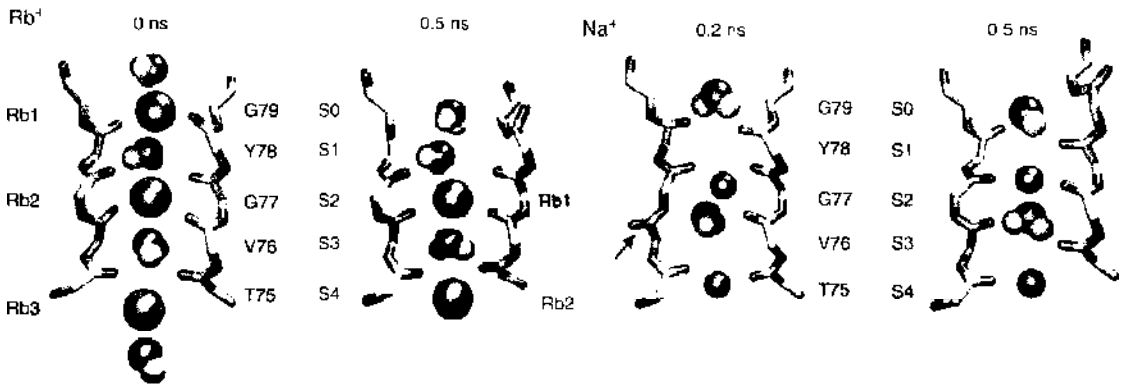


Figure 17. Different effects of cations on the filter, showing snapshots of Rb^+ (Configuration 2) and Na^+ (Configuration 1) in the filter. In both cases, the filter regions (backbone atoms only) of two subunits plus the ions and water molecules within the filter are shown. For the Na^+ simulation, the flip of the carbonyl of V76 is indicated with an arrow. Reprinted with permission from [129], C. Domene et al., *Biophys. J.* 85, 2787 (2003). © 2003, Biophysical Society.

in their interactions with the selectivity filter, whereas Na^+ ions tend to distort the filter by binding to a ring of four carbonyl oxygens (see Fig. 17). The larger Cs^+ ions result in a small degree of expansion of the filter relative to the x-ray structure and show some tendency to bind within the gate region of the channel, near the cavity. Although these results are very interesting, it should be kept in mind, however, that ion parameters are difficult to obtain with high accuracy. Further studies in this area would certainly be worthwhile.

The recent structure of the calcium-gated potassium channel MthK in an open state suggests which considerable conformational changes take place upon gating. This new structure has not yet been exploited in published simulation studies (it also has a relatively low resolution, which makes simulations more difficult), but this will only be a matter of time. Several groups have built models of what an open version of the KcsA channel might look like, using a variety of methods including purely theoretical methods [190] and extensive mutagenesis with spin labeling for electron spin resonance measurements [178].

3.3. ClC Cl^- Channels

Dutzler et al. [5, 6] unveiled the crystal structure of a transmembrane protein in bacteria, known as the ClC channel, that has subsequently been shown to be a transporter, not an ion channel [201, 202]. Nevertheless, many amino acid sequences of the prokaryotic ClC protein are conserved in their eukaryotic ClC relatives, which are selectively permeable to Cl^- ions. The ClC family of Cl^- channel is present in virtually all tissues—in muscle, heart, brain, kidney, and liver—and is widely expressed in most mammalian cells. By allowing Cl^- ions to cross the membrane, ClC channels perform diverse physiological roles, such as transepithelial transport, control of cellular excitability, cell volume regulation, and regulation of intracellular pH. Genetic alterations of some of these genes are known to be associated with myotonia congenita [203], a muscle disease characterized by stiffness upon sudden movement; Dent's disease [204], an inherited kidney disorder; Bartter's syndrome [205], a salt-wasting renal tubular disorder; and diabetes.

The prototype channel, known as ClC-0, first discovered by Miller [206], is found in *Torpedo electroplax*. Since then, nine different human ClC genes and four plant and bacterial ClC genes have been identified. Many of the salient properties of ClC-type channels have been uncovered using the techniques of molecular cloning and subsequent heterologous expression [207]. Conduction properties differ among the isoforms. For example, the current–voltage relationships measured from ClC-0 and ClC-2 are linear [206, 208], whereas those measured from other isoforms are either inwardly rectifying (ClC-1) [209] or outwardly rectifying (ClC-3, ClC-4 and ClC-5) [210–214].

The prokaryotic ClC channel is a bulky, large protein, composed of about 14,000 atoms including polar hydrogens. As shown in Fig. 18A, the protein has two independent ion-conducting pathways—it is a double-barreled dimer formed by identical subunits in contact

Figure 18. The crystal structure of the bacterial ClC channel. (A) The atomic detail of the crystal structure of the *E. coli* channel shows that the protein is a dimer, composed of two identical subunits. Arrows indicate the pore openings. (B) The front half of one of the two identical subunits is removed to reveal the ion conducting path. Several atoms are occluding the pore. (C) The residues that block the ion conducting path are removed to create an open state model. Modified from [217].

at a broad interface. From the intracellular and extracellular side of the protein, the wide entrances or vestibules in each of the two subunits can be clearly seen (arrows in Fig. 18A). The x-ray structure of the channel, however, represents a closed-state conformation. In the wild-type *Escherichia coli* ClC (EcClC) channel structure, residues from the N-termini of the D, F, and N α-helices constrict the channel, and two of these residues completely block the conduction pore, as shown in Fig. 18B. As we follow the EcClC pore from either the extracellular (top) or intracellular (bottom) opening toward the middle of the pore, it abruptly tapers and vanishes. Thus, before the crystal structure coordinates can be used to investigate the permeation of ions through the channel, a completely open-state structure needs to be created by using molecular dynamics or other means. Moreover, the ion conducting path of the EcClC channel, unlike that of the KcsA channel [3], takes a tortuous course through the protein, instead of being straight and perpendicular to the plane of the membrane. The meandering nature of the ClC pore complicates calculations of the force an ion experiences as it moves through the pore. Perhaps for these reasons, there have been only a few theoretical studies so far that attempt to relate the atomic structure of a ClC channel to the macroscopic properties.

The availability of the x-ray structure of the prokaryotic ClC Cl^- channel has prompted several theoretical investigations. Using a novel computational approach based on the Metropolis Monte Carlo method, Miloshevksy and Jordan [215] trace the ionic pathway through the protein as a Cl^- ion moves along several anion coordinating sites. They also calculate the electrostatic potential energy profiles before and after neutralizing the charges on several key amino acid residues. To carry out an all-atom molecular dynamics simulation of the prokaryotic ClC protein, Cohen and Schulten [216] built an open-state configuration of the channel by moving away the atoms that occlude the pore. They then constructed the profile of the potential of mean force along the ion conducting path.

Because the EcClC protein shares many signature sequence identities with the eukaryotic ClC channels, it is possible to build homology models of these channels based on the structural information provided by Dutzler et al. [5, 6]. With this aim in mind, Corry et al. [217, 218] first altered the x-ray structure of the prokaryotic ClC protein using molecular dynamics to create a set of open-state configurations. They expanded the EcClC conducting pathway to putative open-state by moving the atoms blocking the pore sideways. A set of cylinders of radius 5 Å were set at six points near the centerline of the pore containing the blocking atoms. Atoms inside the cylinders were pushed outward with a strong harmonic force, whereas all other atoms were held near to their initial positions using comparatively weak harmonic constraints. In this way, an open-state channel was molded with a minimum radius of 2.5 Å, which is the smallest radius that would allow passage of both Cl^- and NO_3^-, as illustrated in Fig. 18C. The open channel shows that the ClC pore is not axially symmetric like the well-known KcsA potassium channel, but meanders through the protein.

Assuming that the overall structure of the eukaryotic ClC channels is the same as that of EcClC, Corry et al. [217, 218] converted EcClC to ClC-0 and ClC-1. To do this, they first generated sequence alignments using the ClustalW algorithm, followed by a small amount of

manual adjustment, as detailed in Corry et al. [218]. Then, they replaced all nonconserved charged and dipole-containing and pore-lining amino acids to convert EcClC to ClC-0 or ClC-1. Incorporating these homology models into Brownian dynamics, they determined the current–voltage–concentration profiles of ClC-0 and ClC-1. These are illustrated in Fig. 19. The current–voltage relationship of ClC-0 is linear (A), whereas that of ClC-1 is asymmetrical and nonlinear (B). The experimental measurements obtained from ClC-0 [206] are superimposed (open circles) on the simulated data (filled circles). The chord conductance is 11.3 ± 0.5 pS, compared to the experimental value of 9.4 ± 0.1 pS. The conductance of ClC-1 obtained from Brownian dynamics simulations is 1.0 pS at -100 mV, much smaller than in ClC-0. At positive voltages, the current across ClC-1 is less than 0.01 pA, until the driving force is increased to +160 mV. There are several whole-cell current measurements demonstrating that the ClC-1 channel is inwardly rectifying [209, 219–221], as shown in the inset of Fig. 19B, but no single-channel current–voltage curve for ClC-1 is available in the literature. The estimated value of the ClC-1 conductance is about 1.2 pS [222, 223]. Thus, both the shape of the current–voltage relationship and the conductance determined from Brownian dynamics simulations are in accord with the experimental findings. In Fig. 19, the currents obtained from Brownian dynamics simulations in ClC-0 (C) and ClC-1 (D) by Corry et al. [217], under the applied potential of -80 mV, are plotted against the concentration of Cl^- ions in the reservoirs. The experimental data obtained from ClC-0 are shown in open circles in Fig. 19C and are fitted with a broken line. The solid lines fitted through the data points are calculated from the Michaelis–Menten equation. Again, there is a reasonable agreement between the simulated data and experimental measurements for ClC-0.

Brownian dynamics simulations also reveal the steps involved in ion permeation. As in the KscA K^+ channel, two resident Cl^- ions exist in a stable equilibrium in the ClC channels. When a third ion enters the pore from the intracellular space, this equilibrium is disrupted, and the outermost Cl^- ion is expelled to the extracellular space. The dwell histogram shown in Fig. 20, obtained by counting the average number of Cl^- ions in each thin layer, reveals the regions of the pore where ions preferentially dwell. In both ClC-0 (A) and ClC-1 (B), two ions dwell on the intracellular side of the channel. In the dwell histogram obtained from ClC-1 (Fig. 20B), there is also a third prominent peak, indicating that a third ion spends considerable time in the channel before a successful conduction event occurs.

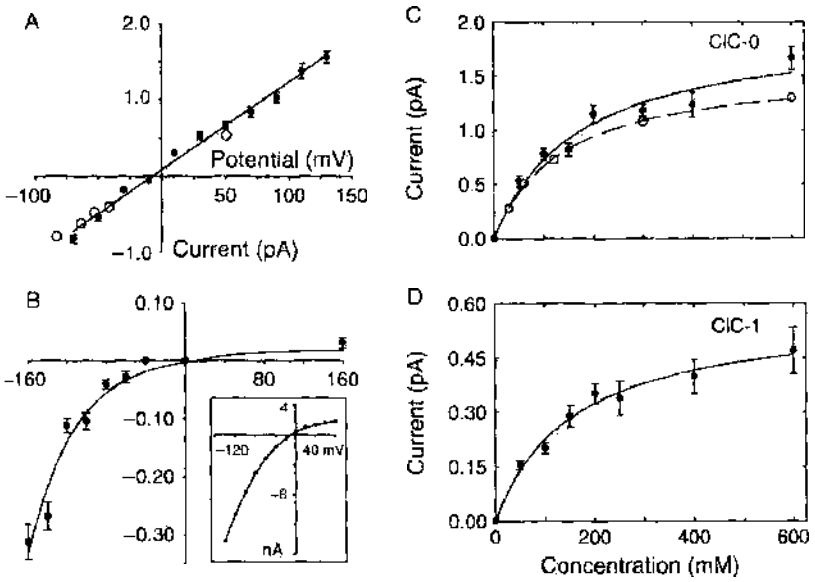


Figure 19. The current–voltage–concentration profiles. The current–voltage relationships for ClC-0 (A) and ClC-1 (B) obtained from Brownian dynamics simulations (filled circles) are compared with the experimental data of Miller [206] (open circles in A) and Bennetts et al. [219]. The outward currents (filled circles) of ClC-0 (C) and ClC-1 (D) are plotted against varying concentrations of NaCl in the reservoirs. The experimental measurements obtained by Tsung-Yu Chen are shown in open circles in (C). Reprinted with permission from [217], B. Corry et al., *Biophys. J.* 86, 846 (2004). © 2004, Biophysical Society.

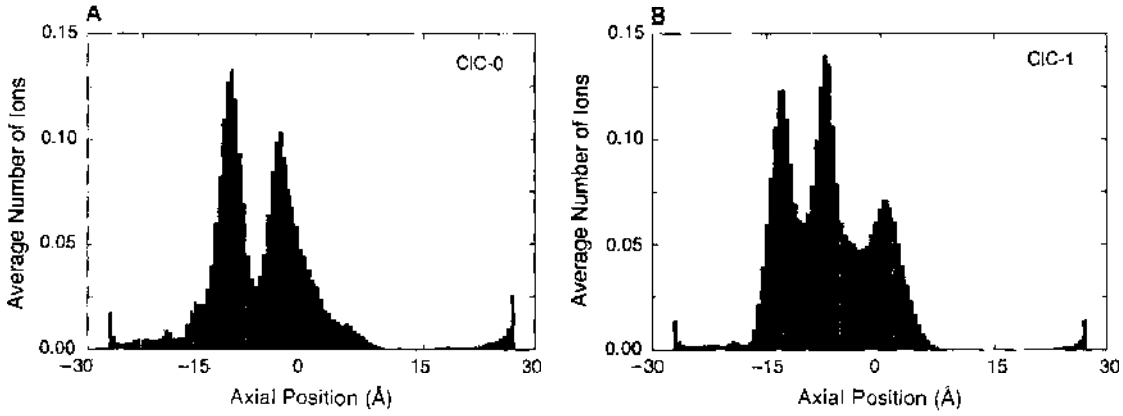


Figure 20. The dwell histograms of ClC-0 (A) and ClC-1 (B). The channel is divided into 100 thin sections, and the average number of ions in each section is calculated over a simulation period (0.4 μs) in the presence of an applied potential of −80 mV. Reprinted with permission from [217], B. Corry et al., *Biophys. J.* 86, 846 (2004). © 2004, Biophysical Society.

3.4. The L-Type Calcium Channel

Unlike the KcsA potassium and the bacterial ClC channels, the crystal structure of the calcium channel is not yet available. Thus, to model the calcium channel, one has to solve an inverse problem—in other words, a simplified channel model that captures the essential features of the full atomic details of the protein must be derived through a judicious use of clues from experimental data.

The calcium channel is ubiquitous in excitable cells. The channel is extremely selective against Na^+ ions and exploits a multi-ion Coulomb repulsion mechanism to achieve a high throughput of Ca^{2+} ions. It selects between ions of almost identical radius, the Pauling radii of Na^+ and Ca^{2+} ions being 0.95 and 0.99 Å, respectively. Moreover, the channel permits much larger ions to move across it, such as tetramethylammonium with a radius of about 2.8 Å [224]. Thus, a different mechanism of selectivity from that in the potassium channel must be at play, one that relies on the different charges on the ions. Monovalent ions can permeate the channel in the absence of calcium at much higher levels of conductance than can any divalent ions [225–228], but are blocked when the calcium concentration reaches only 1 μM [225, 229]. These properties of the calcium channel are largely determined by four glutamate residues that line the narrowest segment of the pore, or the selectivity filter. These residues are expected to be highly charged and to strongly bind Ca^{2+} ions in the channel, leading them to block the passage of Na^+ ions. Mutations of one or more of the glutamate residues lining the selectivity filter have provided many useful insights in selectivity in the calcium channel. The replacement of a glutamate residue with a neutral amino acid severely lowers the conductance of the channel for divalent ions and to a lesser extent for monovalent ions [230]. Also, the block of monovalent currents by divalent ions is severely hampered, only arising at much higher divalent concentrations [231].

Several attempts have been made in the past to construct theoretical models to explain permeation and selectivity in the calcium channel. The first type of proposed models is based on single-file reaction rate theory, in which ions sequentially hop from one site to another [232]. Because of the difficulty in obtaining both high selectivity and throughput with a single binding site [233], these models originally contained two binding sites in which repulsion between ions in neighboring sites increases transit rates [234, 226]. Because the two-site models could not accommodate the mutation data, Dang and McCleskey [235] put forward a new rate model where a single binding site is flanked by lower affinity sites to aid the exit of ions from the central site. Other mechanisms involving single sites have also been developed, such as competition between Ca^{2+} ions for binding sites [231, 236]. Although these models are purely phenomenological and do not take into account the channel shape or charged residues in the protein, they nevertheless have provided many useful insights as to how calcium channels may achieve their selectivity with a high throughput.

Nonner and Eisenberg [42] attempted to relate the observed properties of the calcium channel to its structure with the Poisson–Nernst–Planck theory. They modeled the selectivity filter of the channel as a 10 Å–long cylinder with a radius of 3 Å, to which a 20 Å–long conical vestibule is connected at each end. Four charged glutamate residues, placed along the wall of the cylinder, were allowed to be protonated, and other parameters, such as excess chemical potentials and diffusion coefficients of Na^+ and Ca^{2+}, were estimated analytically. The results obtained by solving their "extended" Poisson–Nernst–Planck theory accurately reproduce current–voltage relations and block of monovalent currents by divalent ions. The physical basis for the free parameters used to fit the channel data still needs to be elucidated.

Corry et al. [70] proposed a model of the L-type calcium channel, making use of the available information on its structure and conductance properties. Figure 21A shows the transverse section of the three-dimensional model, consisting of inner and outer vestibules and a selectivity filter. The selectivity filter is the most important part of the model and requires a careful design in order to reproduce the observed properties of calcium channels. Two crucial elements, namely, its size and charges on its walls, are determined from the experimental data. The diameter is set to 5.6 Å from the size of tetramethylammonium, the largest permeable ion [224]. The mutation data indicate presence of four negatively charged glutamate residues in the filter region [231]. They modeled the four glutamate residues as four fixed charges located in close proximity, which spread asymmetrically in a spiral pattern 1 Å behind the channel wall. To overcome the large image forces at the intracellular end of the channel, they placed four mouth dipoles, 5 Å in length, with their inner ends 1 Å inside the pore wall. The approximate locations of these charge are indicated in Fig. 21A. Brownian dynamics simulations performed with this model have been very successful in replicating many physiological properties of L-type calcium channels. These include current–voltage curves, saturation of conductance with concentration, selectivity against Na^+ ions, the anomalous mole fraction effect, attenuation of calcium current by external sodium ions, and the effect of mutating glutamate residues on blocking of sodium current [70].

An intuitive understanding of how these properties follow from the ion-channel interactions can be gleaned from a study of the potential energy profiles of ions in the channel. The ion-channel interaction has basically two components: a repulsive force due to the induced charges on the protein boundary and the electrostatic interaction of the ion with charge residues and dipoles in the channel wall. The simple Coulomb interaction between two ions is modified in the channel environment because they also interact via the surface charges induced by each other. Each of the four potential energy profiles shown in Fig. 21B is constructed by calculating the potential energy of the ion held at a fixed z position and then repeating these calculations at 1 Å steps. With no fixed charges on the protein wall, an ion entering the channel meets a steeply rising potential barrier (curve a for a Ca^{2+} ion,

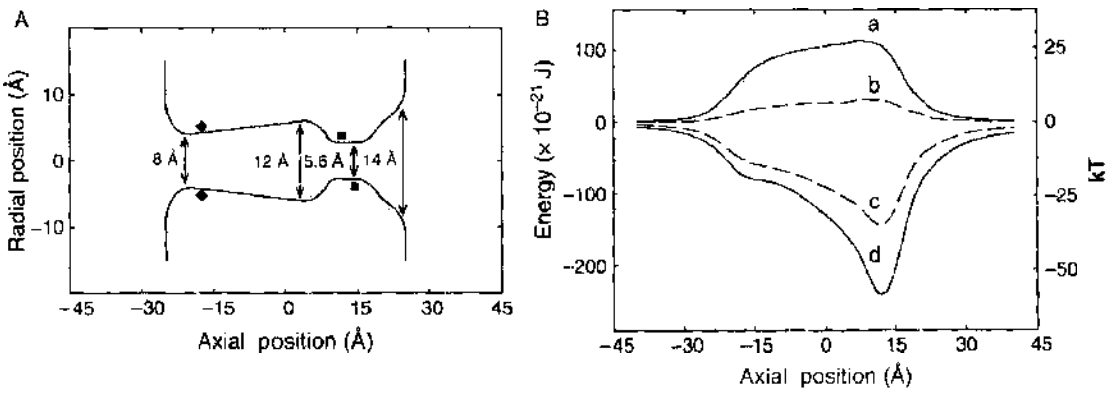


Figure 21. Model calcium channel and electrostatic energy profiles. (A) A three-dimensional channel model is generated by rotating the curves about the central axis by 180°. The positions of two of the four glutamate groups are shown by the squares, and the inner end of two of the four mouth dipoles by the diamonds. (B) Potential energy of an ion traversing the channel is plotted for a Ca^{2+} ion (a) and a Na^+ ion (b) in the absence of any fixed charges. With the glutamate groups and mouth dipoles included, the profiles are replotted for Ca^{2+} ion (d) and a Na^+ ion (c). Reprinted with permission from [70]. B. Corry et al., *Biophys. J.* 80, 195 (2001). © 2001, Biophysical Society.

curve **b** for a Na^+ ion). When the ring of four mouth dipoles and four glutamate charges are included in the model, this barrier is turned into a deep well. The depth of the well is 36 kT for a Na^+ ion (c) and 58 kT for a Ca^{2+} (d). The well is deep enough so that two Na^+ ions or one Ca^{2+} would be permanently trapped in the selectivity filter. When a second Ca^{2+} (or a third Na^+ ion) enters the channel from the extracellular space, two Ca^{2+} (or three Na^+) ions can coexist in the filter region in a semistable equilibrium, until the resident ion on the left climbs over a small energy barrier via thermal fluctuations and exits the channel.

Of the many properties of the calcium channels, the most interesting ones are associated with the blocking effects. In the anomalous mole fraction effect, the channel current vanishes at a certain range of Ca^{2+} concentrations in the presence of a fixed 150 mM Na^+, as shown in the inset of Fig. 22A. The Brownian dynamics results (Fig. 22A) indicate that the rapid drop and subsequent vanishing of the channel current is due to the blocking of Na^+ current by Ca^{2+} ions. Once the Ca^{2+} concentration is high enough to allow two Ca^{2+} ions in the filter, the channel starts conducting again but now Ca^{2+} ions instead of Na^+. Although sodium ions cannot block calcium, their presence in the vestibule can nevertheless retard entry of a second Ca^{2+} ion necessary for conduction. As illustrated in Fig. 22B, the predicted reduction in the channel current with increasing Na^+ concentration is in excellent agreement with the experimental data. A final example is the effect of mutating one of the glutamate residues to neutral glutamine on the blocking of Na^+ current (Fig. 22C). The mutation leads to a reduction in the depth of the potential well compared to the native case so that entry of a

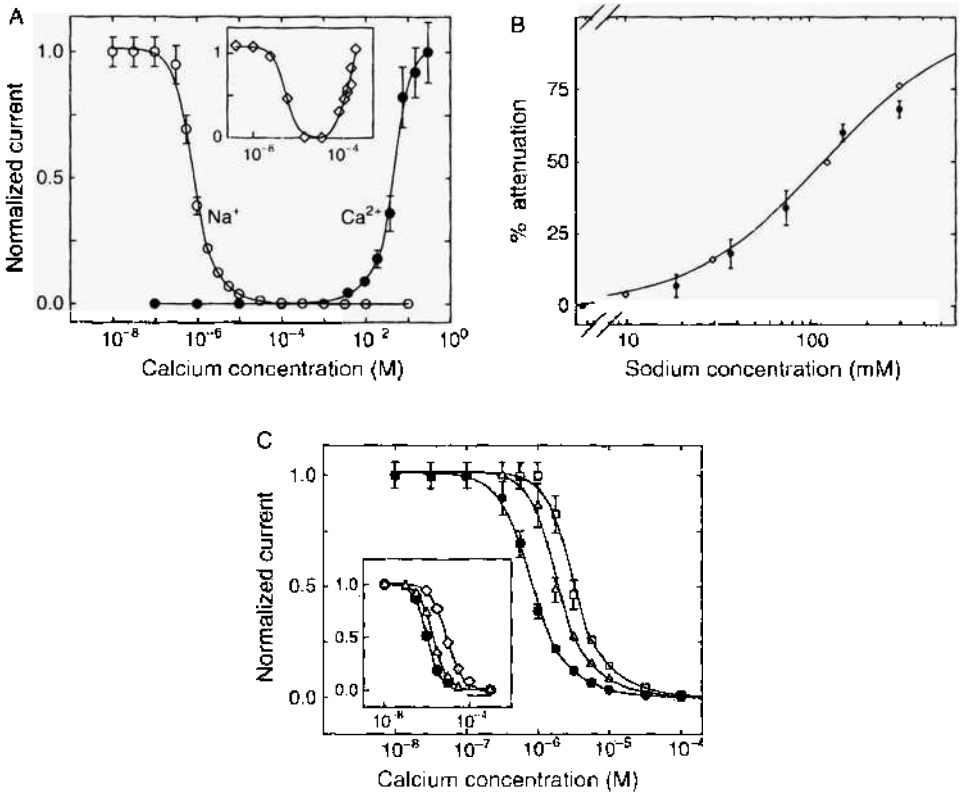


Figure 22. Brownian dynamics simulations of the L-type calcium channel. (A) Simulations accurately reproduce the mole fraction effect. Ca^{2+} (filled circles) and Na^+ (open circles) current passing through the channel normalized by the maximum value of each is plotted against the Ca^{2+} concentrations while keeping the Na^+ concentration fixed at 0.15 M. Experimental results [226] are shown in the inset. (B) Ca^{2+} current is attenuated by Na^+ ions. The percentage reduction in the channel current is plotted against Na^+ concentration while the Ca^{2+} concentration is fixed at 0.15 M (filled circles). The open diamonds and dotted line show the experimental data [276]. (C) The effect of removing glutamate charges on channel selectivity. The Na^+ current passing through the channel at different Ca^{2+} concentrations with all four glutamate charges in place (filled circles), the outermost glutamate removed (triangles), and the innermost glutamate removed (squares). Experimental data for wild type (filled circles) and for single glutamate to neutral glutamine mutations of two different residues (triangles and diamonds) are shown in the inset [231]. In all cases a driving potential of −0.2 V is applied. Reprinted with permission from [70]. B. Corry et al., *Biophys. J.* 80, 195 (2001). © 2001, Biophysical Society.

Ca^{2+} ion in the channel is delayed, and the blocking occurs at a higher Ca^{2+} concentration. Trends in the data (inset of C) are again reproduced by the Brownian dynamics simulations.

The level of agreement between theory and experiment obtained from Brownian dynamics studies of the L-type calcium channel is substantial and should encourage further applications of the Brownian dynamics method to modeling of other ion channels with limited structural information.

3.5. Other Biological Ion Channels

3.5.1. Mechanosensitive Channels

Several species of mechanosensitive channels have been identified, and the crystal structures of the large conductance mechanosensitive channel (MscL) from *Mycobacterium tuberculosis* at 3.5 Å and the small conductance mechanosensitive channel (MscS) from *E. coli* at 3.9 Å resoultion have been solved [237]. Both these channels have intriguing properties. The mechanosensitive MscS channel is gated both by tension and voltage and appears to have arginine residues exposed to the membrane that are important for voltage sensing [238]. This is particularly interesting in the context of voltage-gated sodium, calcium, and potassium channels that also contain highly conserved arginines [239]. The MscL channel is sensitive to tension only. It has such a large conductance in the open state it must undergo very substantial conformational changes from the crystal structure, which appears to be closed. The MscL channel is a pentameric protein consisting of two concentric rings of helices. The TM1 helices line the pore, the TM2 helices form the outer ring.

The mechanosensitve MscL channel has been the subject of a significant number of simulation and modeling studies aimed at understanding the gating mechanism, using the crystal structure or homology models of the *E. coli* MscL (Eco-MscL). The first simulations studied the crystal structure and several mutants to identify mobile and less mobile regions in the structure and to get hypotheses for further experimental studies [240, 241]. The most interesting question is perhaps how the structure changes under the influence of membrane tension. Several hypotheses have been put forward, falling in two broad categories. In the first, the TM1 and TM2 helices might form a larger "barrel." In this scenario, the TM2 helix might be the "tension sensor" and undergo a significant conformational change that eventually causes a change in TM1 as well to open the channel. Another hypothesis assumes that the TM1 and TM2 rings of helices might move together in an "iris-like" motion that opens the channel, analogous to the lens of some cameras, while keeping the direct interactions between TM1 and TM2 helices the same. The net effect in both cases is a sharp increase in the pore radius. This is illustrated in Fig. 23. The first possibility seems to be suggested by molecular dynamics simulations that apply a surface tension to the membrane or an explicit force on the protein directly [241–243], although the TM2 helix is very hydrophobic, and therefore not likely to line a large water-filled pore. The advantage of such simulations is that no preconceived notion of the mechanism is imposed, but a significant drawback is that such high surface tensions or forces are required to see structural changes on a nanosecond timescale that there is a considerable risk of simulation artifacts.

The second mechanism, the iris mechanism, has been investigated by molecular modeling and extensive mutagenesis experiments [244]. The mutagenesis experiments are consistent

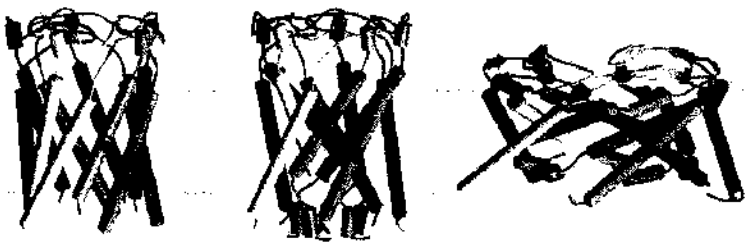


Figure 23. A hypothetical barrel-stave open conformation (left), a closed-state model of *E. coli* MscL built by homology to MscL (center), and a hypothetical tilted open conformation (right) with strong experimental support of Bentazos et al. [274] and Sukharev et al. [275]. Figures provided by Dr. Sukharev.

with this mechanism, and the resulting pore is larger than for the hypothetical barrel stave pore formed by parallel helices. Electron paramagnetic resonance experiments that look at the accessibility to hydrophilic and hydrophobic spin quenchers and the dynamics of every single residue in MscL also support the iris mechanism [245]. Finally, normal mode analyses of both the crystal structure and modeled intermediates also suggest this mechanism [246]. Eventually, a crystal structure of MscL trapped in the open state, which would probably be a mutant that might be based on modeling, might help to explain the different results and to hone further the battery of computational and experimental methods used to obtain models of states that are not directly accessible by high-resolution structural methods.

3.5.2. OmpF

Porins form large trimeric pores in the outer membrane of gram-negative bacteria, which passively transport small molecules down their concentration gradients. OmpF is the best known example of this family. It is a general diffusion pore from the outer membrane of *E. coli* that transports molecules up to ~650 Da. It shows gating behavior, but the molecular basis and the physiological relevance of this phenomenon are poorly understood [127]. OmpF is weakly cation selective, and its selectivity depends on the ionic strength of the solution. Porins have relatively long loops on the extracellular side and short turns on the intracellular side. The L3 loops fold back into the pore and form the so-called eyelet region or constriction zone, as shown in Fig. 24. The arrangement of oppositely charged residues on opposite walls of the narrowest region of the pore creates a strong transverse electrostatic field, which is expected to have a profound effect on the behavior of ions, water, and permeating molecules in this region.

OmpF has been extensively studied by electrophysiology methods, although not all of this data is straightforward to interpret in terms of the properties of a single protein. Nonetheless, OmpF is an attractive model pore for calculations because its high-resolution structure is known, as are structures of a range of mutants with altered electrophysiological properties. This combination of high-resolution structures and electrophysiological data allows systematic testing and calibrating of simulation methods.

OmpF has been the topic of a number of realistic molecular dynamics studies [77, 127, 247, 248] as well as Brownian dynamics [92, 249–251] and Poisson–Nernst–Planck calculations [92]. The crystal structure of OmpF embedded in a simulation model for molecular dynamics simulations, including lipids and solvent, is shown in Fig. 24. A molecular dynamics simulation of OmpF in a POPE bilayer [77] was the first simulation of a large protein channel in a full bilayer environment. The 1-ns simulation revealed complex, nonbulk properties of water within the transbilayer pore. Within the pore, water diffusion coefficients were reduced by up to a factor of 10, relative to bulk. The transverse electrostatic field in the pore resulted in a high degree of alignment of the water dipoles. In a more recent study, the

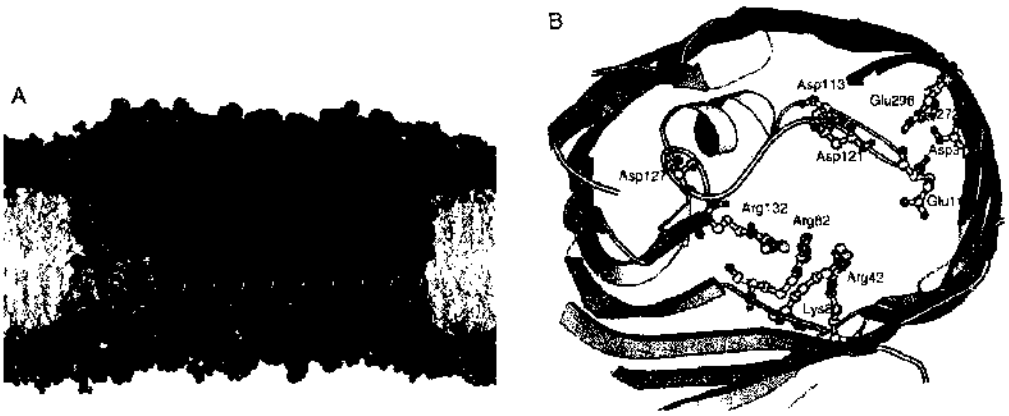


Figure 24. The OmpF porin. (A) In the simulation model, OmpF porin is embedded in a palmitoyl-oleoyl-phosphatidylethanolamine bilayer. The total system contains about 70,000 atoms. (B) In the closeup of the eyelet region of one of the OmpF monomers, there is a significant separation of positive charges and negative charges across the eyelet, resulting in a large local electric field.

orientation of permeating small dipolar molecules was calculated. Both alanine and glucose strongly oriented in the narrow part of the pore, but not appreciably outside this part [247]. Im and Roux [73] described in great detail the permeation of cations and anions obtained from atomistic molecular dynamics simulations of OmpF in a bilayer with 1 M KCl, based on a 5-ns simulation. They found different permeation paths for cations and anions in most of the channel.

This is consistent with a number of Brownian dynamics of ion flow through OmpF that showed that cations and anions follow distinct pathways with little overlap through the pore [92, 251]. It appears that anions require cations to permeate efficiently. OmpF is slightly selective for cations over anions, but ion pair formation counteracts this to some extent. Preliminary simulations of the same system with different salt concentrations and an applied electric field have also been described by Robertson and Tieleman [127].

Although OmpF is a wide pore with a very large conductance, the molecular dynamics calculations so far have not been long enough to calculate reliably a conductance. However, several continuum methods and Brownian dynamics simulations have also been applied to OmpF. A recent study [92] compared ion distributions obtained from the nonlinear Poisson–Boltzmann equation, Brownian dynamics, and molecular dynamics in OmpF. All three methods gave very similar results, with as most conspicuous feature the separation of cations and anions in two distinct sets of pathways through the channel that was observed in the earlier molecular dynamics study. This is interesting, because it shows that treating OmpF as a rigid protein and the solvent as a dielectric constant seems a reasonable approximation, probably because of the large size of the pore. The Poisson–Nernst–Planck theory and Brownian dynamics simulations were used to calculate the conductance and the reversal potential (a measure of the selectivity of the channel). The Poisson–Nernst–Planck and Brownian dynamics gave similar results, both close to experiment, for the reversal potential, but the former overestimated the conductance by about 50% [92].

OmpF porin is an interesting model system to study the interactions of permeating molecules with the channel. Experimentally, the noise in single-channel measurements of OmpF in the presence of permeating molecules can be analyzed to reveal a large amount of information about the interactions of such molecules. For example, the binding and unbinding of an antibiotic could be followed on a single-molecule level, and molecular modeling was used to suggest where and in which orientation this molecule was binding [252]. Robertson and Tieleman [247] have used biased molecular dynamics simulations in which alanine and methylglucose were pulled through OmpF. Although the pulling rate was significantly higher than the normal permeation rate, structural perturbation of the protein was minimal. During the permeation process, both alanine and methylglucose aligned strongly in the electric field in the eyelet region. Binding of the permeating dipolar molecules in the eyelet region was not observed, which could be because of the small size of the permeating molecules.

3.5.3. Alamethicin

Alamethicin (Alm) is a 20-residue Aib-rich channel-forming peptide, a member of the family of peptaibols. It has intensively been studied by experimental and computational methods (for reviews, see Refs. [14, 96]). There exist several variants, including covalently linked dimers of alamethicin [253] and peptides, in which the Aib residues have been replaced by leucine [254]. Alamethicin forms channels with well-defined conductance levels that correspond to channels formed by different numbers of peptides. Tieleman et al. [96, 255] have done a number of simulation studies in a palmitoyl-oleoyl-phosphatidylcholine lipid bilayer on these and related channel models to examine the conformational stability of N = 4 to 8 helix bundle models, and to attempt to link the models with experimental conductance data. These simulations are of duration comparable to the mean passage time of a single ion through such a channel (~5 ns is equivalent to an ionic conductance of 250 pS at 125 mV). The simulations suggest that the smallest observable conductance probably corresponds to an aggregate of five helices.

Because of its simplicity, alamethicin has also been a useful test system to investigate the effect of different simulation algorithms [96], and for one particular system (N6, a parallel hexameric bundle), a simulation has been extended to 100 ns [121], one the longest

simulations on a channel to date. In this fashion we obtained a validated "best guess" model for at least one conductance level of the Alm channel, which should prove useful as the basis for future, more indepth calculations of channel electrostatics and permeation models.

Starostin et al. [256] designed a variant of alamethicin, called K18. In this peptide, the glutamate or glutamine (both occur) in position 18 has been replaced by a lysine that points into the pore, and two alamethicin peptides have been covalently coupled. The resulting peptide shows preferential stabilization of half the channel levels of normal alamethicin, suggesting pairs of helices insert simultaneously and contribute to the channels. This makes it easier to determine the number of helices in a given measured conductance level. A second interesting property of K18 is that it forms channels with a pH-dependent selectivity [257]. Simulations of channel models formed by this peptide suggest why the experimentally determined selectivity for chloride over potassium is modest despite a high positive charge inside the channel. The channel is sufficiently wide for several chloride ions to be present, significantly decreasing the effective charge. This would lower the barrier for potassium ions. Because this channel is relatively simple, molecular modeling and additional simulations could be used to try to predict the effect of different modifications of the pore lining residues. Thus we could design a more anion-selective version of this channel, which could then be synthesized and experimentally tested.

3.5.4. Other Channels Formed by Peptides

In addition to the alamethicin channel already mentioned, simulations have studied a large number of simple channels formed by peptides. In many cases, these studies use only the predicted transmembrane fragment of larger proteins.

LS2 is a synthetic peptide consisting solely of serines and leucines that forms tetrameric proton-permeable helix bundles [258]. Models of LS2 have been built by several groups [259, 260], but critical comparison of these models with experimental conductance data has been difficult.

A second much-studied proton channel is Influenza A M2 (97 amino acids). This protein homo-tetramerizes to form a H^+-permeable channel activated by low pH [261]. M2 is an important system because it is the target of the anti-influenza drug amantadine. It is also important for computational studies because of the large amount of structural biology efforts directed at it, including solid-state nuclear magnetic resonance and site-specific labeling for Fourier-transform infrared (FTIR) spectroscopic measurements. It has been shown that a synthetic peptide corresponding to the single transmembrane helix of M2 will form ion channels in lipid bilayers, and there have been several structural studies of this peptide in lipid bilayers [262–264], as well as a significant number of modeling attempts [109, 265–267]. The pore is quite narrow, containing a number of water molecules, and is constricted toward its C-terminal mouth by the inward-facing His37 residues that are believed to contribute to gating of the channel [268]. Together, these simulations have shown a range of possible structures, but it remains difficult to decide which one is the most accurate structure. The structures vary in crossing angles of the four helices and in other details of the helical packing.

Simulations have also begun to consider proton transport directly, which will be an exciting direction in the future. The pH gating mechanism of M2 is controversial, with two main mechanisms postulated [269]. In one, pore-blocking histidines shuttle protons between two sites on a histidine. In the second, histidines become charged, move away from each other, and thus open the channel. Analyses of experimental data suggests that the first mechanism might be more likely [269], whereas several simulations have suggested the second [270]. No doubt this will receive further attention in the near future. Only very recently has actual proton transport been simulated in this channel, using the MS-EVB theory, but this has not yet included reactions with the histidines [270].

A third example of a peptide channel is the M2δ channel, a channel formed by five peptides with the sequence of the M2 transmembrane helix of the nicotinic acetylcholine receptor [271]. This peptide channel has the approximate conductance of the whole channel and has been studied by simulations [272]. As the structure of the nicotinic acetylcholine receptor is becoming increasingly accurate and now approaches atomistic resolution in the

transmembrane domain [273], it seems likely there will be an increase in simulation studies of the transmembrane domains of the nicotinic receptor and ultimately of the entire protein.

Several models have also been made of channels formed by peptides that are part of viral ion channels, such as VPU from HIV and NB from another strain of influenza. Several papers have been reviewed recently elsewhere [268]. A challenge remains as to how to validate models. Simulations typically show major instabilities but can be metastable if a structure is not correct. Probably the most powerful way of validating models would be a method to calculate the conductance of a model, which can be measured very accurately. Unfortunately, despite recent progress, particularly on gramicidin A, OmpF, and KcsA described above, this is not yet possible in a generally reliable way.

4. CONCLUSIONS

There has been an enormous stride in modeling and simulation of ion channels in the past 5 years. This progress has been brought about by a number of factors, including the availability of crystal structures of physiologically relevant ion channels, the obvious relevance of ion channels for biomedical and pharmaceutical research, the comparatively simple function and basic science of ion channels, the development of efficient and sophisticated simulation and modeling software, and the rapid increase in computer power available to an increasing number of researchers. In this review, we have attempted to summarize some of these recent advances in ion channel research. Of the three permeation models currently in use, the Poisson–Nernst–Planck theory can only be applied reliably for biological nanotubes whose radius is larger than 2 Debye lengths, or about 16 Å. The most appropriate tools for studying ionic channels that are selectively permeable to a specific ionic species are molecular and Brownian dynamics simulations. Brownian dynamics simulations enable calculation of conductance properties, whereas molecular dynamics can provide input and justification for Brownian dynamics as well as explaining finer details such as size-based selectivity. Because molecular dynamics simulations include all atomic detail and can deal with protein flexibility and conformational changes, they have successfully been used in a large number of studies to simulate local changes in structure and diffusion of water and ions, as well as to calculate potentials of mean force for ions in channels that can be used for Brownian dynamics simulations or kinetic theories.

Both methods have unjustified approximations and deficiencies that need to be better understood and improved in future work. For example, the use of continuum electrostatics in calculation of forces in Brownian dynamics simulations needs to be better validated by appealing to molecular dynamics. Here, water is treated as a continuum, and the effects of solvation and the structure of water are taken into account by frictional and random forces acting on ions. Inferences made from macroscopic electrostatics are valid only in the regions that are large compared to the diameters of water and ion molecules. In the narrow, constricted region of the channel, whose radius may be only ~2 Å, the representation of the dielectric as a continuous medium is a poor approximation. Molecular dynamics simulations are limited in timescale, system size, and the accuracy of the description of atomic interactions. It cannot predict channel conductance, which is the most reliably measured quantity. The timescale that is accessible depends mostly on the speed of computers, software, and algorithmic improvements, all of which will combine in the near future to allow simulations of several orders of magnitude longer than currently possible. Similarly, the force fields employed in molecular dynamics simulations need to be improved by including the polarization effects, perhaps using *ab initio* molecular dynamics as a guide in this process.

Because molecular dynamics simulations rely on atomic-detail structure of a channel, accurate structural information is essential for using this tool in structure–function studies. On the other hand, the simplified models used for Brownian dynamics, which may be construed as a simplified representation of the real microstructure of the channel, may adequately summarize the experimental observations and yield a set of testable predictions that can be compared with patch-clamp recordings. Thus, for this purpose, a detailed tertiary structure is not essential for Brownian dynamics—knowledge of the gross shape of the channel and

the approximate locations of the charged residues in the channel wall appear to be sufficient for this purpose. Future developments in the field will depend on the success of the protein chemists and structural biologists in resolving high-resolution structures of membrane proteins, as well as a further refinement of computational tools. Thus, one would like to see the development of a new computational tool that combines the two powerful computational techniques currently available—Brownian dynamics and molecular dynamics. In the wide region of the channel and the vestibules, the motion of charged particles can reliably be simulated with the semimicroscopic Brownian dynamics, whereas the motion of the atoms forming the channel wall and the interaction of the permeating ions with water molecules need to be elucidated by molecular dynamics calculations.

ACKNOWLEDGMENTS

S. H. C. is supported by grants from the Australian Research Council, the National Health & Medical Research Council of Australia. D. P. T is an Alberta Heritage Foundation for Medical Research Scholar. Work in D. P. T.'s laboratory is supported by the Canadian Institutes for Health Research and the National Science and Engineering Research Council. We thank our collaborators Toby Allen, Ben Corry, Scott Edwards, Matthew Hoyles, Serdar Kuyucak, Siu-Cheung Li, and Megan O'Mara, who have contributed to the Brownian dynamics and molecular dynamics studies of channels reported in this review, and Marian Zlomislic for her critical comments.

REFERENCES

1. E. Neher and B. Sakmann, *Nature* 260, 799 (1976).
2. O. P. Hamill, A. Marty, E. Neher, B. Sakmann, and F. J. Sigworth, *Pflügers Arch.* 391, 85 (1981).
3. D. A. Doyle, J. M. Cabral, R. A. Pfuetzner, A. Kuo, J. M. Gulbis, S. L. Cohen, B. T. Chait, and R. MacKinnon, *Science* 280, 69 (1998).
4. G. Chang, R. H. Spencer, A. T. Lee, M. T. Barclay, and D. C. Rees, *Science* 282, 2220 (1998).
5. R. Dutzler, E. B. Campbell, M. Cadene, B. T. Chait, and R. MacKinnon, *Nature* 415, 287 (2002).
6. R. Dutzler, E. B. Campbell, and R. MacKinnon, *Science* 300, 108 (2003).
7. R. B. Bass, P. Strope, M. Barclay, and D. C. Rees, *Science* 298, 1582 (2002).
8. B. Hille, "Ion Channels of Excitable Membranes," 3rd edn, Sinauer Associates, Sunderland, MA, 2001.
9. K. E. Cooper, E. Jakobsson, and P. Wolynes, *Prog. Biophys. Mol. Biol.* 46, 51 (1985).
10. M. B. Partenskii and P. C. Jordan, *Q. Rev. Biophys.* 25, 477 (1992).
11. B. Roux and M. Karplus, *Ann. Rev. Biophys. Biomol. Struct.* 23, 731 (1994).
12. R. S. Eisenberg, *J. Membr. Biol.* 150, 1 (1966).
13. R. S. Eisenberg, *J. Membr. Biol.* 171, 1 (1999).
14. D. P. Tieleman, P. C. Biggin, G. R. Smith, and M. S. P. Sansom, *Q. Rev. Biophys.* 34, 473 (2001).
15. S. Kuyucak, O. S. Andersen, and S. H. Chung, *Rep. Prog. Phys.* 64, 1427 (2001).
16. W. Nonner, D. P. Chen, and B. Eisenberg, *J. Gen. Physiol.* 113, 773 (1999).
17. E. W. McClesky, *J. Gen. Physiol.* 113, 765 (1999).
18. C. Miller, *J. Gen. Physiol.* 113, 783 (1999).
19. D. G. Levitt, *J. Gen. Physiol.* 113, 789 (1999).
20. D. G. Levitt, *Annu. Rev. Biophys. Biophys. Chem.* 15, 29 (1986).
21. J. O'M Bokris and A. K. N. Reddy, in "Modern Electrochemistry," Vol. 1, pp. 175–286, Plenum, New York, 1970.
22. B. Honig and A. Nicholls, *Science* 268, 1144 (1995).
23. D. G. Levitt, *Biophys. J.* 48, 19 (1985).
24. P. C. Jordan, R. J. Bacquet, J. A. McCammon, and P. Tran, *Biophys. J.* 55, 1041 (1989).
25. M. Cai and P. C. Jordan, *Biophys. J.* 57, 883 (1990).
26. R. Sankararamakrishnan, C. Adcock, and M. S. P. Sansom, *Biophys. J.* 71, 1659 (1996).
27. P. Weetman, S. Goldman, and C. G. Gray, *J. Phys. Chem.* 101, 6073 (1997).
28. C. Adcock, G. R. Smith, and M. S. P. Sansom, *Biophys. J.* 75, 1211 (1998).
29. W. Cheng, C. X. Wang, W. Z. Chen, Y. W Xu, and Y. Y. Shi, *Eur. Biophys. J.* 27, 105 (1998).
30. T. K. Rostovtseva, V. M. Aguilelaa, I. Vodyanoy, S. M. Bezrukov, and V. A. Parsegian, *Biophys. J.* 75, 1783 (1998).
31. M. S. P. Sansom, C. Adcock, and G. R. Smith, *J. Struct. Biol.* 121, 246 (1998).
32. G. G. Dieckmann, J. D. Lear, Q. Zhong, M. L. Klein, W. F. DeGrado, and K. A. Sharp, *Biophys. J.* 76, 618 (1999).
33. K. M. Ranatunga, C. Adcock, I. D. Kerr, G. R. Smith, and M. S. P. Sansom, *Theor. Chem. Acc.* 101, 97 (1999).
34. P. Smejtek, M. Mense, R. Word, and S. Wang, *J. Membr. Biol.* 167, 151 (1999).

35. B. Roux, B. Bernèche, and W. Im, *Biochemistry* 39, 13295 (2000).
36. P. Debye and J. F. Hückel, *Physik. Zeitschr.* 24, 185 (1923).
37. N. W. Ashcroft and N. D. Mermin, in "Solid State Physics," pp. 589–613. Holt, Rinehart and Winston, New York, 1976.
38. J. P. McKelvey, in "Solid State Semiconductor Physics," pp. 320–370. Harper & Row, NJ, 1966.
39. J. S. Blakemore, in "Solid State Physics," pp. 295–404. W. B. Saunders Co., London, 1974.
40. C. Croxton, in "Introduction to Liquid State Physics," pp. 247–262. John Wiley, New York, 1975.
41. D. Chen, J. Lear, and R. S. Eisenberg, *Biophys. J.* 72, 97 (1997).
42. W. Nonner and R. S. Eisenberg, *Biophys. J.* 75, 1287 (1998).
43. D. P. Chen, L. Xu, A. Tripathy, G. Meissner, and R. S. Eisenberg, *Biophys. J.* 76, 1346 (1999).
44. M. G. Kurnikova, R. D. Coalson, P. Graf, and A. Nitzan, *Biophys. J.* 76, 642 (1999).
45. A. E. Cardenas, R. D. Coalson, and M. G. Kurnikova, *Biophys. J.* 79, 80 (2000).
46. U. Hollerbach, D. P. Chen, D. D. Busath, and B. Eisenberg, *Langmuir* 16, 5509 (2000).
47. U. Hollerbach and R. S. Eisenberg, *Langmuir* 18, 3626 (2002).
48. B. Corry, S. Kuyucak, and S. H. Chung, *Biophys. J.* 78, 2364 (2000).
49. A. Syganow and E. von Kitzing, *J. Phys. Chem.* 99, 12030 (1995).
50. A. Syganow and E. von Kitzing, *Biophys. J.* 76, 768 (1999).
51. A. Syganow and E. von Kitzing, *Eur. Biophys. J.* 28, 393 (1999).
52. P. Graf, A. Nitzan, M. G. Kurnikova, and R. D. Coalson, *J. Phys. Chem. B* 104, 12324 (2000).
53. Z. Schuss, B. Nadler, and R. S. Eisenberg, *Phys. Rev. E* 64, 0361161 (2001).
54. A. Mamonov, R. D. Coalson, P. Graf, and M. G. Kurnikova, *Biophys. J.* 82, 209a (2002).
55. R. D. Coalson, M. G. Kurnikova, P. Graf, and A. Nitzan, *Biophys. J.* 82, 340a (2002).
56. B. Corry, S. Kuyucak, and S. H. Chung, *Biophys. J.* 84, 3594 (2003).
57. B. Nadler, U. Hollerbach, and R. S. Eisenberg, *Phys. Rev. E.* 68, 021905 (2003).
58. F. Reif, in "Fundamentals of Statistical and Thermal Physics," pp. 548–603. McGraw-Hill Kogakusha, Tokyo, 1965.
59. R. Zwanzig, in "Nonequilibrium Statistical Mechanics," pp. 3–29. Oxford University Press, Oxford, 2002.
60. W. F. van Gunsteren, H. J. C. Berendsen, and J. A. C. Rullmann, *Mol. Phys.* 44, 69 (1981).
61. W. F. van Gunsteren and H. J. C. Berendsen, *Mol. Phys.* 45, 637 (1982).
62. M. Hoyles, S. Kuyucak, and S. H. Chung, *Phys. Rev. E* 58, 3654 (1998).
63. L. G. Nilsson and J. A. Padró, *Mol. Phys.* 71, 355 (1990).
64. S. Z. Wan, C. X. Wang, and Y. Y. Shi, *Mol. Phys.* 93, 901 (1998).
65. E. Jakobsson and S. W. Chiu, *Biophys. J.* 52, 33 (1987).
66. S. Bek and E. Jakobsson, *Biophys. J.* 66, 1028 (1994).
67. S. C. Li, M. Hoyles, S. Kuyucak, and S. H. Chung, *Biophys. J.* 74, 37 (1998).
68. S. H. Chung, M. Hoyles, T. W. Allen, and S. Kuyucak, *Biophys. J.* 75, 793 (1998).
69. S. H. Chung, T. W. Allen, M. Hoyles, and S. Kuyucak, *Biophys. J.* 77, 2517 (1999).
70. B. Corry, T. W. Allen, S. Kuyucak, and S. H. Chung, *Biophys. J.* 80, 195 (2001).
71. S. H. Chung, T. W. Allen, and S. Kuyucak, *Biophys. J.* 82, 628 (2002).
72. S. H. Chung, T. W. Allen, and S. Kuyucak, *Biophys. J.* 83, 263 (2002).
73. W. Im, S. Seefeld, and B. Roux, *Biophys. J.* 79, 788 (2000).
74. B. Corry, M. Hoyles, T. W. Allen, M. Walker, S. Kuyucak, and S. H. Chung, *Biophys. J.* 82, 1975 (2002).
75. M. S. P. Sansom, G. R. Smith, C. Adcock, and P. C. Biggin, *Biophys. J.* 73, 2404 (1997).
76. M. E. Green and J. Lu, *J. Phys. Chem. B* 101, 6512 (1997).
77. D. P. Tieleman and H. J. C. Berendsen, *Biophys. J.* 74, 2786 (1998).
78. T. W. Allen, S. Kuyucak, and S. H. Chung, *J. Chem. Phys.* 111, 7985 (1999).
79. S. Garofoli and P. C. Jordan, *Biophys. J.* 84, 2814 (2003).
80. T. Hansson, C. Oostenbrink, and W. F. van Gunsteren, *Curr. Opin. Struct. Biol.* 12, 190 (2002).
81. S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner, *J. Am. Chem. Soc.* 106, 765 (1984).
82. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, *J. Comp. Chem.* 4, 187 (1983).
83. J. Hermans, H. J. C. Berendsen, W. F. van Gunsteren, and J. P. M. Postma, *Biopolymers* 23, 1513 (1984).
84. E. Lindahl, B. Hess, and D. van der Spoel, *J. Mol. Model.* 7, 306 (2001)
85. B. Roux, *Curr. Opin. Struct. Biol.* 12, 182 (2002).
86. C. E. Capener, I. H. Shrivastava, K. M. Ranatunga, L. R. Forrest, G. R. Smith, and M. S. P. Sansom, *Biophys. J.* 78, 2929 (2000).
87. C. E. Capener, H. J. Kim, Y. Arinaminpathy, and M. S. P. Sansom, *Human Mol. Genet.* 11, 2425 (2002).
88. M. A. Eriksson and B. Roux, *Biophys. J.* 83, 2595 (2002).
89. P. C. Biggin and M. S. Sansom, *Biophys. J.* 83, 1867 (2002).
90. D. P. Tieleman, L. R. Forrest, M. S. Sansom, and H. J. Berendsen, *Biochemistry* 37, 17554 (1998).
91. H. I. Petrache, D. M. Zuckerman, J. N. Sachs, J. A. Killian, R. E. Koeppe, and T. B. Woolf, *Langmuir* 18, 1340 (2002).
92. W. Im and B. Roux, *J. Mol. Biol.* 322, 851 (2002).
93. A. Burykin and A. Warshel, *Biophys. J.* 85, 3696 (2003).
94. O. Beckstein and M. S. Sansom, *Proc. Natl. Acad. Sci. USA* 100, 7063 (2003).
95. G. Hummer, J. C. Rasaiah, and J. P. Noworyta, *Nature* 414, 188 (2001).

96. D. P. Tieleman, B. Hess, and M. S. Sansom, *Biophys. J.* 83, 2393 (2002).
97. O. Beckstein, P. C. Biggin, and M. S. P. Sansom, *J. Phys. Chem. B* 105, 12902 (2001).
98. R. Allen, S. Melchionna, and J. P. Hansen, *Phys. Rev. Lett.* 89, 175502 (2002).
99. A. Burykin, C. N. Schutz, J. Villa, and A. Warshel, *Prot. Struct. Function Genet.* 47, 265 (2002).
100. P. C. Jordan, *Biophys. J.* 83, 1235 (2002).
101. J. Åqvist and V. Luzhkov, *Nature* 404, 881 (2000).
102. S. Berneche and B. Roux, *Nature* 414, 73 (2001).
103. Y. Zhou and R. MacKinnon, *J. Mol. Biol.* 333, 965 (2003).
104. T. W. Allen, O. S. Andersen, and B. Roux, *J. Am. Chem. Soc.* 125, 9868 (2003).
105. S. Berneche and B. Roux, *Proc. Natl. Acad. Sci. USA* 100, 8644 (2003).
106. D. P. Tieleman, M. S. P. Sansom, and G. A. Woolley, *Biophys. J.* 84, 1464 (2003).
107. C. H. Yu and R. Pomes, *J. Am. Chem. Soc.* 125, 13890 (2003).
108. F. H. Stillinger and C. W. David, *J. Chem. Phys.* 69, 1473 (1978).
109. A. M. Smondyrev and G. A. Voth, *Biophys. J.* 83, 1987 (2002).
110. Y. J. Wu and G. A. Voth, *Biophys. J.* 85, 864 (2003).
111. A. Warshel, *Acc. Chem. Res.* 35, 385 (2002).
112. M. A. Lill and V. Helms, *J. Chem. Phys.* 115, 7993 (2001).
113. B. L. de Groot, T. Frigato, V. Helms, and H. Grubmuller, *J. Mol. Biol.* 333, 279 (2003).
114. N. Chakrabarti, E. Tajkhorshid, B. Roux, and R. Pomes, *Structure (Cambridge)* 12, 65 (2004).
115. C. Dellago, M. M. Naor, and G. Hummer, *Phys. Rev. Lett.* 91, 105902 (2003).
116. P. S. Crozier, D. Henderson, R. L. Rowley, and D. D. Busath, *Biophys. J.* 81, 3077 (2001).
117. D. P. Tieleman, H. J. Berendsen, and M. S. Sansom, *Biophys. J.* 80, 331 (2001).
118. G. Ceve, *Biochim. Biophys. Acta* 1031, 311 (1990).
119. A. Kalra, S. Garde, and G. Hummer, *Proc. Natl. Acad. Sci. USA* 100, 10175 (2003).
120. Y. Y. Sham, Z. T. Chu, H. Tao, and A. Warshel, *Proteins* 39, 393 (2000).
121. D. P. Tieleman, in "Unsolved Problems in Noise and Fluctuations III," *AIP Proc.* 665, 298 (2002).
122. C. Sagui and T. A. Darden, *Annu. Rev. Biophys. Biomol. Struct.* 28, 155 (1999).
123. P. Carlon, U. Rothlisberger, and M. Parrinello, *Acc. Chem. Res.* 35, 455 (2002).
124. U. F. Rohrig, L. Guidoni, and U. Rothlisberger, *Biochemistry* 41, 10799 (2002).
125. L. Guidoni and P. Carloni, *Biochim. Biophys. Acta Biomembr.* 1563, 1 (2002).
126. B. Roux, *Acc. Chem. Res.* 35, 366 (2002).
127. K. M. Robertson and D. P. Tieleman, *Biochem. Cell Biol.* 80, 517 (2002).
128. E. Tajkhorshid, A. Aksimentiev, I. Balabi, M. Gao, B. Isralewitch, J. C. Phillips, F. Zhu, and K. Schulten, *Adv. Protein Chem.* 66, 195 (2003).
129. C. Domene and M. S. Sansom, *Biophys. J.* 85, 2787 (2003).
130. C. E. Capener and M. S. P. Sansom, *J. Phys. Chem. B* 106, 4543 (2002).
131. R. D. Hotchkiss and R. J. Dubos, *J. Biol. Chem.* 132, 791 (1940).
132. D. W. Urry, *Proc. Natl. Acad. Sci. USA* 68, 672 (1971).
133. S. B. Hladky and D. A. Haydon, *Nature* 225, 451 (1970).
134. S. B. Hladky and D. A. Haydon, *Biochim. Biophys. Acta* 274, 294 (1972).
135. O. S. Andersen, *Biophys. J.* 41, 119 (1983).
136. O. S. Andersen, *Biophys. J.* 41, 147 (1983).
137. A. Pullman, *Q. Rev. Biophys.* 20, 173 (1987).
138. O. S. Andersen and R. E. Koeppe, *Physiol. Rev.* 72, 89 (1992).
139. G. A. Woolley and B. A. Wallace, *J. Membr. Biol.* 129, 109 (1992).
140. D. D. Busath, *Annu. Rev. Physiol.* 55, 473 (1993).
141. R. E. Koeppe and O. S. Andersen, *Annu. Rev. Biophys. Biomol. Struct.* 25, 231 (1996).
142. B. A. Wallace, *J. Struct. Biol.* 121, 123 (1998).
143. D. G. Levitt, *Biophys. J.* 22, 209 (1978).
144. P. C. Jordan, *Biophys. J.* 39, 157 (1982).
145. H. Monoi, *Biophys. J.* 59, 786 (1991).
146. T. B. Woolf and B. Roux, *Proc. Natl. Acad. Sci. USA* 91, 11631 (1994).
147. T. B. Woolf and B. Roux, *Biophys. J.* 72, 1930 (1997).
148. S. W. Chiu, S. Subramaniam, and E. Jakobsson, *Biophys. J.* 76, 1929 (1999).
149. F. Tian, K. C. Lee, W. Hu, and T. A. Cross, *Biochemistry* 35, 11959 (1996).
150. F. Tian and T. A. Cross, *J. Mol. Biol.* 285, 1993 (1999).
151. S. Edwards, B. Corry, S. Kuyucak, and S. H. Chung, *Biophys. J.* 83, 1348 (2002).
152. J. Åqvist and A. Warshel, *Biophys. J.* 56, 171 (1989).
153. P. Jordan, *Biophys. J.* 58, 1133 (1990).
154. R. Elber, D. P. Chen, D. Rojewska, and R. Eisenberg, *Biophys. J.* 68, 906 (1995).
155. B. L. de Groot, D. P. Tieleman, P. Pohl, and H. Grubmuller, *Biophys. J.* 82, 2934 (2002).
156. O. S. Andersen, H. J. Apell, E. Bamberg, D. D. Busath, R. E. Koeppe, F. J. Sigworth, G. Szabo, D. W. Urry, and A. Woolley, *Nat. Struct. Biol.* 6, 609 (1999).
157. P. Tang and Y. Xu, *Proc. Natl. Acad. Sci. USA* 99, 16035 (2002).
158. B. Roux and M. Karplus, *J. Am. Chem. Soc.* 115, 3250 (1993).
159. T. W. Allen, T. Bastug, S. Kuyucak, and S. H. Chung, *Biophys. J.* 84, 2159 (2003).
160. T. W. Allen, O. S. Andersen, and B. Roux, *Proc. Natl. Acad. Sci. USA* 101, 117 (2004).

161. M. Sancho, M. B. Partenskii, V. Dorman, and P. C. Jordan, *Biophys. J.* 68, 427 (1995).
162. B. Roux, B. Prodhom, and M. Karplus, *Biophys. J.* 68, 876 (1995).
163. R. Ketcheim, B. Roux, and T. Cross, *Structure* 5, 1655 (1997).
164. L. E. Townsley, W. A. Tucker, S. Sham, and J. F. Hinton, *Biochemistry* 40, 11676 (2001).
165. R. MacKinnon, S. L. Cohen, A. Kuo, A. Lee, and B. T. Chait, *Science* 280, 106 (1998).
166. M. LeMasurier, L. Heginbotham, and C. Miller, *J. Gen. Physiol.* 118, 303 (2001).
167. Z. Lu, A. M. Klem, and Y. Ramu, *Nature* 413, 809 (2001).
168. J. H. Morais-Cabral, Y. Zhou, and R. MacKinnon, *Nature* 414, 37 (2001).
169. Y. Zhou, J. H. Morais-Cabral, A. Kaufman, and R. MacKinnon, *Nature* 414, 43 (2001).
170. Y. Jiang, A. Lee, J. Chen, M. Cadene, B. T. Chait, and R. MacKinnon, *Nature* 417, 523 (2000).
171. A. Kuo, J. M. Gulbis, J. F. Antcliff, T. Rahman, E. D. Lowe, J. Zimmer, J. Cuthbertson, F. M. Ashcroft, T. Ezaki, and D. A. Doyle, *Science* 300, 1922 (2003).
172. Y. Jiang, A. Lee, J. Chen, V. Ruta, M. Cadene, B. T. Chait, and R. MacKinnon, *Nature* 423, 33 (2003).
173. Y. Jiang, V. Ruta, J. Chen, A. Lee, and R. MacKinnon, *Nature* 423, 42 (2003).
174. B. E. Cohen, M. Grabe, and L. Y. Jan, *Neuron* 39, 395 (2003).
175. L. G. Cuello, J. G. Romero, D. M. Cortes, and E. Perozo, *Biochemistry* 37, 3229 (1998).
176. L. Heginbotham, M. LeMasurier, L. Kolmakova-Partensky, and C. Miller, *J. Gen. Physiol.* 114, 551 (1999).
177. E. Perozo, D. M. Cortes, and L. G. Cuello, *Science* 285, 73 (1999).
178. Y. S. Liu, P. Sompornpisut, and E. Perozo, *Nat. Struct. Biol.* 8, 883 (2001).
179. R. J. Mashl, Y. Tang, J. Schnitzer, and E. Jakobsson, *Biophys. J.* 81, 2473 (2001).
180. R. Coronado, R. L. Rosenberg, and C. Miller, *J. Gen. Physiol.* 76, 425 (1980).
181. D. Meuser, H. Splitt, R. Wagner, and H. Schrempf, *FEBS Lett.* 46, 447 (1999).
182. H. Schrempf, O. Schmidt, R. Kümmerlen, S. Hinnah, D. Müller, M. Betzler, T. Steinkamp, and R. Wagner, *EMBO J.* 14, 5170 (1995).
183. L. Heginbotham, T. Abramson, and R. MacKinnon, *Science* 258, 1152 (1992).
184. M. S. Sansom, I. H. Shrivastava, J. N. Bright, J. Tate, C. E. Capener, and P. C. Biggin, *Biochim. Biophys. Acta* 1565, 294 (2002).
185. T. W. Allen, S. Kuyucak, and S. H. Chung, *Biophys. J.* 77, 2502 (1999).
186. P. C. Biggin, G. R. Smith, I. Shrivastava, S. Choe, and M. S. Sansom, *Biochim. Biophys. Acta* 1510, 1 (2001).
187. L. Guidoni, V. Torre, and P. Carloni, *FEBS Lett.* 477, 37 (2000).
188. S. Berneche and B. Roux, *Biophys. J.* 78, 2900 (2000).
189. I. H. Shrivastava and M. S. P. Sansom, *Biophys. J.* 78, 557 (2000).
190. P. C. Biggin and M. S. Sansom, *Biophys. J.* 83, 1867 (2002).
191. Y. Liu, M. Holmgren, M. E. Jurman, and G. Yellen, *Neuron* 19, 175 (1997).
192. M. Zhou, J. H. Morais-Cabral, S. Mann, and R. MacKinnon, *Nature* 411, 657 (2001).
193. Y. Jiang, A. Lee, J. Chen, M. Cadene, B. T. Chait, and R. MacKinnon, *Nature* 417, 515 (2002).
194. L. Monticelli, K. M. Robertson, J. L. MacCallum, and D. P. Tieleman, *FEBS Lett.* 564, 325 (2004).
195. I. H. Shrivastava, D. P. Tieleman, P. C. Biggin, and M. S. P. Sansom, *Biophys. J.* 83, 633 (2002).
196. T. W. Allen, A. Bliznyuk, A. P. Rendell, S. Kuyucak, and S. H. Chung, *J. Chem. Phys.* 112, 8191 (2000).
197. B. Roux and S. Berneche, *Biophys. J.* 82, 1681 (2002).
198. A. Grossfield, P. Ren, and J. W. Ponder, *J. Am. Chem. Soc.* 125, 15671 (2003).
199. V. B. Luzhkov and J. Åqvist, *Biochim. Biophys. Acta* 1481, 360 (2000).
200. L. Guidoni, V. Torre, and P. Carloni, *Biochemistry* 38, 8599 (1999).
201. A. Accardi and C. Miller, *J. Gen. Physiol.* 122, 16a (2003).
202. A. Accardi and C. Miller, *Nature* 427, 803 (2004).
203. K. Steinmeyer, C. Lorenz, M. Pusch, M. Koch, and T. J. Jentsch, *EMBO J.* 13, 737 (1994).
204. N. Piwon, W. Gunther, M. Schwake, M. R. Bosl, and T. J. Jentsch, *Nature* 408, 369 (2000).
205. B. D. Simon, R. S. Bindra, T. A. Mansfield, C. Nelson-Williams, E. Mendonca, R. Stone, S. Schurman, A. Nayir, H. Alpay, A. Bakkaloglu, J. Rodri-Guez-Soriano, J. M. Morales, S. A. Sanjad, C. M. Taylor, D. Pilz, A. Brem, H. Trachtman, W. Griswold, G. A. Richard, E. John, and R. P. Lifton, *Nat. Genet.* 17, 171 (1997).
206. C. Miller, *Phil. Trans. Roy. Soc. Lond.* B288, 401 (1982).
207. T. J. Jentsch, T. Friedrich, A. Schriever, and H. Yamada, *Pflugers Arch.* 437, 783 (1999).
208. C. Lorenz, M. Pusch, and T. J. Jentsch, *Proc. Natl. Acad. Sci. USA* 93, 13362 (1996).
209. G. Y. Rychkov, M. Pusch, M. L. Roberts, and A. H. Bretag, *J. Physiol.* 530.3, 379 (2001).
210. M. Kawasaki, M. Suzuki, S. Uchida, S. Sasaki, and F. Marumo, *Neuron* 14, 1285 (1995).
211. K. Steinmeyer, B. Schwappach, M. Bens, A. Vandewalle, and T. J. Jentschh, *J. Biol. Chem.* 270, 31172 (1995).
212. D. Duan, S. Cowley, B. Horowitz, and J. R. Hume, *J. Gen. Physiol.* 113, 57 (1999).
213. T. Friedrich, T. Breiderhoff, and T. J. Jentsch, *J. Biol. Chem.* 274, 896 (1999).
214. C. G. Vonoye and A. L. George, Jr., *J. Physiol.* 539, 373 (2002).
215. G. V. Miloshevsky and P. C. Jordan, *Biophys. J.* 86, 825 (2004).
216. J. Cohen and K. Schulten, *Biophys. J.* 86, 836 (2004).
217. B. Corry, M. O'Mara, and S. H. Chung, *Biophys. J.* 86, 846 (2004).
218. B. Corry, M. O'Mara, and S. H. Chung, *J. Chem. Phys. Lett.* 386, 233 (2004).
219. B. Bennetts, M. L. Roberts, A. H. Bretag, and G. Y. Rychkov, *J. Physiol.* 535, 93 (2001).
220. C. Fahlke, R. Rüdel, N. Mitrovic, M. Zhou, and A. L. George, *Neuron* 15, 463 (1995).
221. C. Fahlke, C. L. Beck, and A. L. George, *Proc. Natl. Acad. Sci. USA* 94, 2729 (1997).
222. C. Saviane, F. Conti, and M. Pusch, *J. Gen. Physiol.* 113, 457 (1999).

223. F. Weinrich and T. J. Jentsch, *J. Biol. Chem.* 276, 2347 (2001).
224. E. W. McCleskey and W. Almers, *Proc. Natl. Acad. Sci. USA* 82, 7149 (1985).
225. P. G. Kostyuk, S. L. Mironov, and Y. M. Shuba, *J. Membr. Biol.* 76, 83 (1983).
226. W. Almers and E. W. McCleskey, *J. Physiol.* 353, 585 (1984a).
227. P. Hess, J. B. Lansman, and R. W. Tsien, *J. Gen. Physiol.* 88, 293 (1986).
228. C. C. Kuo and P. Hess, *J. Physiol.* 446, 629 (1993).
229. W. Almers, E. W. McCleskey, and P. T. Palade, *J. Physiol.* 353, 565 (1984).
230. A. Babinski, A. Yatani, G. Mikala, S. Tang, S. Yamamoto, and A. Schwartz, *Mol. Cell. Biochem.* 166, 125 (1997).
231. J. Yang, P. T. Ellinor, W. A. Sather, J. F. Zhang, and R. W. Tsien, *Nature* 366, 158 (1993).
232. R. W. Tsien, P. Hess, E. W. McCleskey, and R. L. Rosenberg, *Annu. Rev. Biophys. Chem.* 16, 265 (1987).
233. F. Benzanilla and C. M. Armstrong, *J. Gen. Physiol.* 60, 588 (1972).
234. P. Hess and R. W. Tsien, *Nature* 309, 453 (1984).
235. T. X. Dang and E. W. McCleskey, *J. Gen. Physiol.* 111, 185 (1998).
236. C. M. Armstrong and J. Neyton, *Ann. NY Acad. Sci.* 635, 18 (1991).
237. E. Perozo and D. C. Rees, *Curr. Opin. Struct. Biol.* 13, 432 (2003).
238. R. B. Bass, K. P. Locher, E. Borths, Y. Poon, P. Strop, A. Lee, and D. C. Rees, *FEBS Lett.* 555, 111 (2003).
239. F. Benzanilla and E. Perozo, *Adv. Protein Chem.* 63, 211 (2003).
240. D. E. Elmore and D. A. Dougherty, *Biophys. J.* 81, 1345 (2001).
241. J. Gullingsrud, D. Kosztin, and K. Schulten, *Biophys. J.* 80, 2074 (2001).
242. L. E. Bilston and K. Mylvaganam, *FEBS Lett.* 512, 185 (2002b).
243. G. Colombo, S. J. Marrink, and A. E. Mark, *Biophys. J.* 84, 2331 (2003b).
244. M. Betanzos, C. S. Chiang, H. R. Guy, and S. Sukharev, *Nat. Struct. Biol.* 9, 704 (2002).
245. E. Perozo, D. M. Cortes, P. Sompornpisut, A. Kloda, and B. Martinac, *Nature* 418, 942 (2002).
246. H. Valadie, J. J. Lacapere, Y. H. Sanejouand, and C. Etchebest, *J. Mol. Biol.* 332, 657 (2003).
247. K. M. Robertson and D. P. Tieleman, *FEBS Lett.* 528, 53 (2002).
248. W. Im and B. Roux, *J. Mol. Biol.* 319, 1177 (2002).
249. W. Im and B. Roux, *J. Chem. Phys.* 115, 4850 (2001).
250. P. S. Phale, A. Philippsen, C. Widmer, V. P. Phale, J. P. Rosenbusch, and T. Schirmer, *Biochemistry* 40, 6319 (2001).
251. T. Schirmer and P. S. Phale, *J. Mol. Biol.* 294, 1159 (1999).
252. E. M. Nestorovich, T. K. Rostovtseva, and S. M. Bezrukov, *Biophys. J.* 85, 3718 (2003).
253. G. A. Woolley, P. C. Biggin, A. Schultz, L. Lien, D. C. Jaikaran, J. Breed, K. Crowhurst, and M. S. Sansom, *Biophys. J.* 73, 770 (1997).
254. M. S. Sansom, *Q. Rev. Biophys.* 26, 365 (1993).
255. D. P. Tieleman, H. J. Berendsen, and M. S. Sansom, *Biophys. J.* 76, 1757 (1999).
256. A. V. Starostin, R. Butan, V. Borisenko, D. A. James, H. Wenschuh, M. S. Sansom, and G. A. Woolley, *Biochemistry* 38, 6144 (1999).
257. V. Borisenko, M. S. Sansom, and G. A. Woolley, *Biophys. J.* 78, 1335 (2000).
258. J. D. Lear, Z. R. Wasserman, and W. F. DeGrado, *Science* 240, 1177 (1988).
259. H. S. Randa, L. R. Forrest, G. A. Voth, and M. S. Sansom, *Biophys. J.* 77, 2400 (1999).
260. Q. Zhong, Q. Jiang, P. B. Moore, D. M. Newns, and M. L. Klein, *Biophys. J.* 74, 3 (1998).
261. L. J. Holsinger, D. Nichani, L. H. Pinto, and R. A. Lamb, *J. Virol.* 68, 1551 (1994).
262. K. C. Duff and R. H. Ashley, *Virology* 190, 485 (1992).
263. F. A. Kovacs and T. A. Cross, *Biophys. J.* 73, 2511 (1997).
264. A. Kukol, P. D. Adams, L. M. Rice, A. T. Brunger, and T. I. Arkin, *J. Mol. Biol.* 286, 951 (1999).
265. L. R. Forrest, W. F. DeGrado, G. R. Dieckmann, and M. S. Sansom, *Fold. Des.* 3, 443 (1998).
266. L. R. Forrest, A. Kukol, I. T. Arkin, D. P. Tieleman, and M. S. Sansom, *Biophys. J.* 78, 55 (2000).
267. K. J. Schweighofer and A. Pohorille, *Biophys. J.* 78, 150 (2000).
268. W. B. Fischer and M. S. Sansom, *Biochim. Biophys. Acta* 1561, 27 (2002).
269. J. D. Lear, *FEBS Lett.* 552, 17 (2003).
270. Y. J. Wu and G. A. Voth, *FEBS Lett.* 552, 23 (2003).
271. M. Montal and S. J. Opella, *Biochim. Biophys. Acta* 1565, 287 (2002).
272. R. Law, D. P. Tieleman, and M. S. P. Sansom, *Biophys. J.* 84, 14 (2003).
273. N. Unwin, *FEBS Lett.* 555, 91 (2003).
274. M. Bentazos, C. S. Chiang, H. R. Guy, and S. Sukharev, *Nat. Struct. Biol.* 9, 704 (2002).
275. S. Sukharev, S. R. Durell, and H. R. Guy, *Biophys. J.* 81, 917 (2001).
276. L. Polo-Parada and S. J. Korn, *J. Gen. Physiol.* 109, 693 (1997).

CHAPTER 13

Progress in Nanomedicine and Medical Nanorobotics

Robert A. Freitas, Jr.
Institute for Molecular Manufacturing, Palo Alto, California, USA

CONTENTS

1. NANOTECHNOLOGY AND NANOMEDICINE

"There is a growing sense in the scientific and technical community that we are about to enter a golden new era," announced Richard E. Smalley, founder of the Center for Nanoscale Science and Technology at Rice University in Texas and winner of the 1996 Nobel Prize in Chemistry. In his Congressional testimony [1] on June 22, 1999, Smalley spoke in support of a new National Nanotechnology Initiative before the Subcommittee on Basic Research of the U.S. House Science Committee in Washington, DC. "We are about to be able to

ISBN: 1-58883-048-9

Handbook of Theoretical and Computational Nanotechnology
Edited by Michael Rieth and Wolfram Schommers
Volume 6: Pages (619–672)

build things that work on the smallest possible length scales, atom by atom." Smalley said. "Over the past century we have learned about the workings of biological nanomachines to an incredible level of detail, and the benefits of this knowledge are beginning to be felt in medicine. In coming decades we will learn to modify and adapt this machinery to extend the quality and length of life. Twenty years from now, nanotechnology will have given us specially engineered drugs which are nanoscale cancer-seeking missiles, a molecular technology that specifically targets just the mutant cancer cells in the human body, and leaves everything else blissfully alone. To do this, these drug molecules will have to be big enough—thousands of atoms—so that we can code the information into them of where they should go and what they should kill. They will be examples of an exquisite, human-made nanotechnology of the future."

Following this testimony, the U.S. President, in his January 2000 State-of-the-Union speech, announced that he would seek $475 million for nanotechnology R&D via the National Nanotechnology Initiative, effectively doubling federal nanotech funding for Fiscal Year 2001. The President never referred to "nanotechnology" by name, but he gushed about its capabilities, marveling at a technology that will someday produce "molecular computers the size of a teardrop with the power of today's fastest supercomputers." Annual U.S. federal funding for nanotechnology research and development exceeded $500 million in 2002 [2], reaching $849 million in fiscal year 2004 [3] and could approach $1 billion in next year's budget; the European Commission has set aside 1.3 billion euros for nanotechnology research during 2003–2006 [4], with annual nanotechnology investment worldwide reaching approximately $3 billion in 2003. Private-sector analysts estimate that the worldwide market for nanoscale devices and molecular modeling should experience an average annual growth rate of 28% per year, rising from $406 million in 2002 to $1.37 billion in 2007, with a 35% per year growth rate in revenues from biomedical nanoscale devices [5].

In December 2002, the U.S. National Institutes of Health (NIH) announced a 4-year program for nanoscience and nanotechnology in medicine [4]. Burgeoning interest in the medical applications of nanotechnology has led to the emergence of a new field called nanomedicine [4, 6–13]. Most broadly, nanomedicine is the process of diagnosing [14], treating, and preventing disease and traumatic injury; of relieving pain; and of preserving and improving human health, using molecular tools and molecular knowledge of the human body [6]. The NIH Roadmap's new Nanomedicine Initiatives, first released in late 2003, "envision that this cutting-edge area of research will begin yielding medical benefits as early as 10 years from now" and will begin with "establishing a handful of Nanomedicine Centers... staffed by a highly interdisciplinary scientific crew including biologists, physicians, mathematicians, engineers and computer scientists... gathering extensive information about how molecular machines are built," who will also develop "a new kind of vocabulary—lexicon—to define biological parts and processes in engineering terms" [15]. Even state-funded programs have begun, such as New York's Alliance for Nanomedical Technologies [16].

It is useful to regard the development path of nanomedicine as a succession of three mutually overlapping and progressively more powerful technologies.

First, in the relatively near term, over the next 5 years, nanomedicine can address many important medical problems by using nanoscale-structured materials and simple nanodevices that can be manufactured today (Section 2). This includes the interaction of nanostructured materials with biological systems [8]—the first 12 Ph.D. candidates in "nanobiotechnology" began laboratory work at Cornell University in June 2000, and many other universities have started similar programs as state, federal, and international funding has soared.

Second, over the next 5–10 years, biotechnology will make possible even more remarkable advances in molecular medicine and biobotics (microbiological robots or engineered organisms), some of which are already on the drawing boards (Section 3).

Third, in the longer term, perhaps 10–20 years from today, the earliest molecular machine systems and nanorobots may join the medical armamentarium, finally giving physicians the most potent tools imaginable to conquer human disease, ill-health, and aging (Section 4). Issues relating to future Food and Drug Administration (FDA) approval of nanomedical materials, devices, and even nanorobots are beginning to be addressed by legal writers [17, 18].

2. MEDICAL NANOMATERIALS AND NANODEVICES

2.1. Nanopores

Perhaps one of the simplest medical nanomaterials is a surface perforated with holes, or nanopores. In 1997 Tejal Desai at Boston University and Mauro Ferrari at Ohio State University created what could be considered one of the earliest therapeutically useful nanomedical devices [19]. Along with collaborators at the Biomedical Microdevices Center at the University of California at Berkeley, Desai and Ferrari employed bulk micromachining to fabricate tiny cell-containing chambers within single crystalline silicon wafers. The chambers interface with the surrounding biological environment through polycrystalline silicon filter membranes that are micromachined to present a high density of uniform nanopores as small as 20 nanometers in diameter. These pores are large enough to allow small molecules such as oxygen, glucose, and insulin to pass but are small enough to impede the passage of much larger immune system molecules such as immunoglobulins and graft-borne virus particles. Safely ensconced behind this artificial barrier, immunoisolated encapsulated rat pancreatic cells may receive nutrients and remain healthy for weeks, happily secreting insulin back out through the pores, while the immune system remains blissfully unaware of the foreign cells which it would normally attack and reject.

Microcapsules containing replacement islets of Langerhans cells—most likely easily harvested piglet islet cells—could be implanted beneath the skin of some diabetes patients [20]. This could temporarily restore the body's delicate glucose control feedback loop without the need for powerful immunosuppressants that can leave the patient at serious risk for infection. Supplying encapsulated new cells to the body could also be a valuable way to treat other enzyme- or hormone-deficiency diseases, including encapsulated neurons that could be implanted in the brain and then electrically stimulated to release neurotransmitters, possibly as part of a future treatment for Alzheimer's or Parkinson's diseases. In conjunction with the biomedical company iMEDD, Desai has been active in continuing this work for immunoisolation [21], drug delivery [22, 23], and sensing [24], with U.S. patents in process. Neurotech (Paris) [25] and others are also working on encapsulated cell technologies [26] and similar techniques.

The flow of materials through nanopores can also be externally regulated [27]. The first artificial voltage-gated molecular nanosieve was fabricated by Charles R. Martin and colleagues [28] at Colorado State University in 1995. Martin's membrane contains an array of cylindrical gold nanotubules with inside diameters as small as 1.6 nanometers. When the tubules are positively charged, positive ions are excluded and only negative ions are transported through the membrane. When the membrane receives a negative voltage, only positive ions can pass. Future similar nanodevices may combine voltage gating with pore size, shape, and charge constraints to achieve precise control of ion transport with significant molecular specificity. Martin's recent efforts [29] have been directed at immobilizing biochemical molecular-recognition agents such as enzymes, antibodies, other proteins, and DNA inside the nanotubes as active biological nanosensors [30–32], to perform drug separations [33, 34], and to allow selected biocatalysis [34]. An exquisitely sensitive ion channel switch biosensor has also been built by an Australian research group [35]. The Australian scientists estimated that their device, currently being commercialized by Ambri Biosensor, could detect a minute change in chemical concentration equivalent to a single sugar cube tossed into Sidney harbor, or roughly one part in a billion billion ($\sim 10^{-18}$). Others are also investigating synthetic nanopore ion pumps [36] and voltage-gated nanopores embedded in artificial membranes [37], and molecular dynamics theoretical studies of viscosity [38] and diffusion [39] through nanopores are in progress.

Daniel Branton's team at Harvard University has conducted an ongoing series of experiments using an electric field to drive a variety of RNA and DNA polymers through the central nanopore of an alpha-hemolysin protein channel mounted in a lipid bilayer similar to the outer membrane of a living cell [40, 41]. As early as 1996, the researchers had determined that the individual nucleotides making up the polynucleotide strands must be passing single-file through the 2.6-nm-wide nanopore, and that changes in ionic current could be used to measure polymer length. By 1998, Branton had shown that the nanopore could

be used to rapidly discriminate between pyrimidine and purine segments (the two types of nucleotide bases) along a single RNA molecule. In 2000, the scientists demonstrated the ability to distinguish between DNA chains of similar length and composition that differ only in base pair sequence, and Branton continues to perfect this approach [42–46]. Current research is directed toward reliably fabricating pores with specific diameters and repeatable geometries at high precision [47–49], understanding the unzipping of double-stranded DNA as one strand is pulled through the pore [50], recognizing of folded DNA molecules passing through the pore [46], experimenting with new 3–10-nm silicon–nitride nanopores [46], and investigating the benefits of adding electrically conducting electrodes to pores to improve longitudinal resolution "possibly to the single-base level for DNA" [46]. It has been suggested that nanopore-based DNA-sequencing devices could allow per-pore read rates potentially up to 1000 bases per second [51]. Because nanopores can rapidly discriminate and characterize DNA polymers at low copy number, future refinements of this experimental approach may eventually provide a low-cost high-throughput method for very rapid genome sequencing.

2.2. Artificial Binding Sites and Molecular Imprinting

Another early goal of nanomedicine is to study how biological molecular receptors work and then to build artificial binding sites on a made-to-order basis to achieve specific medical results. Molecular imprinting [52–57] is an existing technique in which a cocktail of functionalized monomers interacts reversibly with a target molecule using only noncovalent forces. The complex is then cross-linked and polymerized in a casting procedure, leaving behind a polymer with recognition sites complementary to the target molecule in both shape and functionality. Each such site constitutes an induced molecular "memory" capable of selectively binding the target species. In one experiment involving an amino acid–derivative target, one artificial binding site per $(3.8\ \text{nm})^3$-polymer block was created. Chiral separations, enzymatic transition state activity, and high receptor affinities have been demonstrated.

Buddy D. Ratner at the University of Washington in Seattle has researched the engineering of polymer surfaces containing arrays of artificial receptors, which are created using a radiofrequency-plasma glow-discharge process to imprint a polysaccharide-like film with nanometer-sized pits in the shape of such biologically useful protein molecules as albumin (the most common blood protein), fibrinogen (a clotting protein), lysozyme and ribonuclease (two important enzymes), and immunoglobulin (antibodies) [52]. Each protein type sticks only to a pit with the shape of that protein. Such engineered surfaces could be used for quick biochemical separations and assays [56] and as recognition elements in biosensors and chemosensors [57], because such surfaces will selectively adsorb from solution only the specific protein whose complementary shape has been imprinted, and only at the specific place on the surface where the shape is imprinted. However, molecularly imprinted polymers have limitations, such as incomplete template removal, broad guest affinities and selectivities, and slow mass transfer; imprinting inside dendrimers (Section 2.7) may allow quantitative template removal, nearly homogeneous binding sites, solubility in common organic solvents, and amenability to the incorporation of other functional groups [55]. The RESIST Group at the Welsh School of Pharmacy at Cardiff University [53] and others [57] have looked at how molecularly imprinted polymers could be medically useful in clinical applications such as controlled drug release, drug monitoring devices, and biological and receptor mimics including artificial antibodies (plastibodies) or biomimicking enzymes (plastizymes) [57].

2.3. Quantum Dots and Nanocrystals

Fluorescent tags are commonplace in medicine and biology and are found in everything from human immunodeficiency virus (HIV) tests to experiments that image the inner functions of cells, but different dye molecules must be used for each color, color-matched lasers are needed to get each dye to fluoresce, and dye colors tend to bleed together and fade quickly after one use. "Quantum dot" nanocrystals have none of these shortcomings. These dots are tiny particles measuring only a few nanometers across, about the same size as a protein molecule or a short sequence of DNA. They come in a nearly unlimited palette of sharply defined colors that can be customized by changing particle size or composition. Particles

can be excited to fluorescence with white light, can be linked to biomolecules to form long-lived sensitive probes to identify specific compounds up to a thousand times brighter than conventional dyes used in many biological tests, and can track biological events by simultaneously tagging each biological component (e.g., different proteins or DNA sequences) with nanodots of a specific color.

Quantum Dot Corp. (QDC) [58], the manufacturer, believes this kind of flexibility could offer a cheap and easy way to screen a blood sample for the presence of a number of different viruses at the same time. It could also give physicians a fast diagnostic tool to detect, say, the presence of a particular set of proteins that strongly indicates a person is having a heart attack or to detect known cellular cancer markers [59]. On the research front, the ability to simultaneously tag multiple biomolecules both on and inside cells could allow scientists to watch the complex cellular changes and events associated with disease, providing valuable clues for the development of future pharmaceuticals and therapeutics. Quantum dots are useful for studying genes, proteins, and drug targets in single cells, tissue specimens, and living animals [60]. In August 2003, QDC, Matsushita, and SC BioSciences agreed to develop, manufacture, and market QDC's life science detection products [61]. "Qdot nanotechnology is revolutionizing biological detection," said Takao Kanamura, Senior Managing Director of Matsushita Kotobuki Electronics. "We project products developed under this agreement to generate revenue for QDC in excess of $100 million per year by 2007." Quantum dots are being investigated as chemical sensors [62] and for cancer cell detection [59], gene expression studies [63], gene mapping and DNA microarray analysis [64], immunocytochemical probes [65], intracellular organelle markers [66], live cell labeling [67, 68], medical diagnostics and drug screening [69], SNP (single nucleotide polymorphism) genotyping [70], vascular imaging [71], and many other applications [72, 73]. Quantum dot physics has been studied theoretically [74] and computationally using time-dependent density functional theory [75] and other methods [76–78].

Researchers from Northwestern University and Argonne National Laboratory have created a hybrid "nanodevice" composed of 4.5-nm nanocrystals of biocompatible titanium dioxide semiconductor covalently attached with snippets of oligonucleotide DNA [79]. Experiments showed that these nanocomposites not only retain the intrinsic photocatalytic capacity of TiO_2 and the bioactivity of the oligonucleotide DNA but, more important, also possess the unique property of a light-inducible nucleic acid endonuclease (separating when exposed to light or X-rays). For example, researchers would attach a strand of DNA that matches a defective gene within a cell to the semiconductor scaffolding, and then introduce the nanoparticle into the cell nucleus where the attached DNA binds with its defective complementary DNA strand, whereupon exposure of the bound nanoparticle to light or X-rays snips off the defective gene. "We call it a 'Swiss army knife' because, unlike today's drugs, we can inject 10 kinds of good genes all at once and target them in extremely specific or extremely broad ways," says researcher Tatjana Paunesku [80]. Other molecules besides oligonucleotides can be attached to the titanium dioxide scaffolding, such as navigational peptides or proteins, which, like viral vectors, can help the nanoparticles home in on the cell nucleus. This simple nanocrystal nanodevice might one day be used to target defective genes that play a role in cancer, neurological disease, and other conditions, though the work is still at a preliminary stage and testing in a laboratory model is at least 2 years away [80].

2.4. Fullerenes and Nanotubes

Soluble derivatives of fullerenes such as C_{60}—a soccer-ball-shaped arrangement of 60 carbon atoms per molecule—have shown great utility as pharmaceutical agents. These derivatives, many already in clinical trials, have good biocompatibility and low toxicity even at relatively high dosages. Fullerene compounds may serve as antiviral agents (most notably against HIV [81]; they have also been investigated computationally [82, 83]), antibacterial agents (*Escherichia coli* [84], *Streptococcus* [85], *Mycobacterium tuberculosis* [86], etc.), photodynamic antitumor [87, 88] and anticancer [89] therapies, antioxidants and antiapoptosis agents that may include treatments for amyotrophic lateral sclerosis (ALS or Lou Gehrig's disease) [90] and Parkinson's disease, and other applications—most being pursued by C Sixty [91], the

leading company in this area. C Sixty is also investigating possible drug-delivery "nanopills" consisting of two closed-end single-walled carbon nanotubes nested mouth-to-mouth, forming a capsule-like container [92].

Both single-walled [93, 94] and multiwalled [95–97] carbon nanotubes are also being investigated as biosensors; for example, to detect glucose [96, 98], ethanol [98], hydrogen peroxide [95], selected proteins such as immunoglobulins [94], and an electrochemical DNA hybridization biosensor [93]. NASA and the National Cancer Institute are developing a biosensor catheter to detect specific oligonucleotide sequences that serve as molecular signatures of cancer cells, with preliminary *in vitro* testing using tissue samples from patients with chronic myelogenous leukemia and acute promyelocytic leukemia, two neoplastic diseases for which molecular signatures have been well characterized [97].

2.5. Nanoshells and Magnetic Nanoprobes

Researchers Naomi Halas and Jennifer West at Rice University in Houston have developed a platform for nanoscale drug delivery called the nanoshell [99, 100]. Unlike carbon fullerenes, the slightly larger nanoshells are dielectric-metal nanospheres with a core of silica and a gold coating, whose optical resonance is a function of the relative size of the constituent layers. The nanoshells are embedded in a drug-containing tumor-targeted hydrogel polymer and injected into the body. The shells circulate through the body until they accumulate near tumor cells. When heated with an infrared laser, the nanoshells (each slightly larger than a polio virus) selectively absorb the infrared frequencies, melt the polymer, and release their drug payload at a specific site. Nanoshells offer advantages over traditional cancer treatments: earlier detection, more detailed imaging, fast noninvasive imaging, and integrated detection and treatment [101]. This technique could also prove useful in treating diabetes. Instead of taking an injection of insulin, a patient would use a ballpoint-pen-size infrared laser to heat the skin where the nanoshell polymer had been injected. The heat from nanoshells would cause the polymer to release a pulse of insulin. Unlike injections, which are taken several times a day, the nanoshell-polymer system could remain in the body for months.

In 2002, Rice University licensed its patented nanoshell technology to a private company started by Halas and West, Nanospectra Biosciences [102] of Houston, Texas, to develop commercial applications. Nanospectra is conducting animal studies at the M.D. Anderson Cancer Center at the University of Texas and specifically targets micrometastases, tiny aggregates of cancer cells too small for surgeons to find and remove with a scalpel. The company hopes to start clinical trials for the cancer treatment by 2004 and for the insulin-delivery system by 2006. Also in mid-2003, Rice researchers announced [103] the development of a point-of-care whole-blood immunoassay using antibody-nanoparticle conjugates of gold nanoshells [104]. By varying the thickness of the metal shell, researchers in Halas' group can precisely tune the color of light to which the nanoshells respond; near-infrared light penetrates whole blood very well, so it is an optimal wavelength for a whole blood immunoassay [103]. Successful detection of sub-nanogram-per-milliliter quantities of immunoglobulins was achieved in saline, serum, and whole blood in 10–30 minutes [104]. Groups at the University of Washington [105] and the Korea Advanced Institute of Science and Technology [106] are also investigating the use of gold nanoshells for biological applications.

An alternative approach pursued by Triton BioSystems [107] is to bond iron nanoparticles and monoclonal antibodies into nanobioprobes about 40 nanometers long. The chemically inert probes are injected and circulate inside the body, whereupon the antibodies selectively bind to tumor cell membranes. Once the tumor is covered with bioprobes after several hours, a magnetic field generated from a portable alternating magnetic field machine (similar to a miniaturized magnetic resonance imaging (MRI) machine) heats the iron particles to more than 170 degrees, killing the tumor cells in a few seconds [108]. Once the cells are destroyed, the body's excretion system kicks in and removes cellular residue and nanoparticles alike. Test subjects seem to feel no pain from the heat generated [108]. Triton BioSystems targets both visible tumors and micrometastases. Samuel Straface, chief executive at Triton BioSystems, explained that chemotherapy is analogous to napalm, killing large swaths of tissue in hopes

that all tumor cells are killed in the process, whereas the nanobioprobes function more like carefully planted explosives detonated by remote control [108]. "Chemotherapy is nasty stuff," he said. "The effects here are minimal. We expect virtually no collateral damage." Triton BioSystems began mouse trials in 2002, and tests show that the treatment itself brings no ill effects to the mice. By mid-2003, Straface was nearly finished with efficacy trials to see how the treatment actually works on tumors. Once those trials are done, the company will start designing human tests and ask the FDA for permission to begin human clinical trials in 2006.

Mirkin's group at Northwestern University uses magnetic microparticle probes coated with target protein-binding antibodies plus 13-nm nanoparticle probes with a similar coating but including a unique hybridized "bar-code" DNA sequence as an ultrasensitive method for detecting protein analytes such as prostate-specific antigen (PSA) [109]. After the target protein in the test sample is captured by the microparticles, magnetic separation of the complexed microparticle probes and PSA is followed by dehybridization of the bar-code oligonucleotides on the nanoparticle probe surface, allowing the determination of the presence of PSA by identifying the bar-code sequence released from the nanoparticle probe. Using polymerase chain reaction on the oligonucleotide bar codes allows PSA to be detected at 3 attomolar concentration, about a million times more sensitive than comparable clinically accepted conventional assays for detecting the same protein target.

2.6. Targeted Nanoparticles and Smart Drugs

Targeting drugs to particular organs or tissues, such as a cancer tumor, is one way to achieve some specificity of action. For example, an immunotoxin molecule [110] is an engineered hybrid of functional protein modules fabricated from two different types of proteins: a toxin and an antibody. Toxin proteins are normally produced and released by infectious bacteria. The protein binds to the surface of a host cell, penetrates it, and kills it. Toxin molecules are so potent that just a few of them can kill a cell. Antibodies are proteins produced by the immune system to recognize and bind to specific foreign materials. An immunotoxin molecule is made by fusing a part of the gene encoding a toxin with a part of the gene encoding an antibody that recognizes surface features on cancer cells. This creates a novel gene that can be used to express a new synthetic protein molecule. This new molecule will bind only to a cancer cell (via a module from the antibody protein), then penetrate it and kill it (via modules from the toxin protein). The first experiments with mice showed that these engineered proteins successfully eliminated certain tumors. Then, early in 2000, National Cancer Institute researchers confirmed that an immunotoxin made from a truncated form of *Pseudomonas* exotoxin was cytotoxic to malignant B-cells taken from patients with hairy cell leukemia [111]. Another set of clinic trials at the Universitaet zu Koeln in Germany that were completed in 2003 also found that a ricin-based immunotoxin had moderate efficacy against Hodgkin's lymphoma in some patients [112].

Multisegment gold/nickel nanorods are being explored by Leong's group at Johns Hopkins School of Medicine [113] as tissue-targeted carriers for gene delivery into cells that "can simultaneously bind compacted DNA plasmids and targeting ligands in a spatially defined manner" and allow "precise control of composition, size and multifunctionality of the gene-delivery system." The nanorods are electrodeposited into the cylindrical 100-nm-diameter pores of an alumina membrane, joining a 100-nm-length gold segment and a 100-nm-length nickel segment. After the alumina template is etched away, the nanorods are functionalized by attaching DNA plasmids to the nickel segments and attaching transferrin, a cell-targeting protein, to the gold segments, using molecular linkages that selectively bind to only one metal and thus impart biofunctionality to the nanorods in a spatially defined manner. Leong notes that extra segments could be added to the nanorods, for example, to bind additional biofunctionalities such as an endosomolytic agent, or magnetic segments could be added to allow manipulating the nanorods with an external magnetic field.

Targeted radioimmunotherapeutic agents [114] include the FDA-approved "cancer smart bombs" that deliver tumor-killing radioactive yttrium (Zevalin) or iodine (Bexxar) attached to a lymphoma-targeted (anti-CD20) antibody [115]. Other antibody-linked agents are being investigated such as the alpha-emitting actinium-based "nanogenerator" molecules that use

internalizing monoclonal antibodies to penetrate the cell and that have been shown, *in vitro*, to specifically kill leukemia, lymphoma, breast, ovarian, neuroblastoma, and prostate cancer cells at becquerel (picocurie) levels [116], with promising preliminary results against advanced ovarian cancer in mice [117]. However, drug specificity is still no better than the targeting accuracy of the chosen antibody, and there is significant mistargeting, leading to unwanted side effects.

Enzyme-activated drugs, first developed in the 1980s and still under active investigation [118], separate the targeting and activation functions. For instance, an antibody-directed enzyme-triggered prodrug cancer therapy is being developed by researchers at the University of Gottingen in Germany [119, 120]. This targeted drug molecule turns lethal only when it reaches cancer cells while remaining harmless inside healthy cells. In tests, mice previously implanted with human tumors are given an activating targeted enzyme that sticks only to human tumor cells, mostly ignoring healthy mouse cells. Then the antitumor molecule is injected. In its activated state, this fungal-derived antibiotic molecule is a highly strained ring of three carbon atoms that is apt to burst open, becoming a reactive molecule that wreaks havoc among the nucleic acid molecules essential for normal cell function. However, the molecule is injected as a prodrug—an antibiotic lacking the strained ring and with a sugar safety-catch. Once the sugar is clipped off by the previously positioned targeted enzyme, the drug molecule rearranges itself into a three-atom ring, becoming lethally active. Notes chemist Philip Ball [121]: "The selectivity of the damage still depends on antibody's ability to hook onto the right cells, and on the absence of other enzymes in the body that also activate the prodrug."

A further improvement in enzyme-activated drugs are "smart drugs" that become medically active only in specific circumstances and in an inherently localized manner. A good example is provided by Yoshihisa Suzuki at Kyoto University, who has designed a novel drug molecule that releases antibiotic only in the presence of an infection [122]. Suzuki started with the common antibiotic molecule gentamicin and bound it to a hydrogel using a newly developed peptide linker. The linker can be cleaved by a proteinase enzyme manufactured by *Pseudomonas aeruginosa*, a Gram-negative bacillus that causes inflammation and urinary tract infection, folliculitis, and otitis externa in humans. Tests on rats show that when the hydrogel is applied to a wound site, the antibiotic is not released if no *P. aeruginosa* bacteria are present, but if any bacteria of this type are present, then the proteolytic enzyme that the microbes naturally produce cleaves the linker and the gentamicin is released, killing the bacteria. "If the proteinase specific to each bacterium [species] can be used for the signal," wrote Suzuki, "different spectra of antibiotics could be released from the same dressing material, depending on the strain of bacterium." This specificity of action is highly desirable because the indiscriminate prophylactic use of antibiotics is associated with the emergence of strains of drug-resistant bacteria, and most antibiotics apparently have at least some toxicity for human fibroblasts. In subsequent work, an alternative antibiotic release system triggered by thrombin activity, which accompanies *Staphylococcus aureus* wound infections, was successfully tested as a high-specificity stimulus-responsive controlled drug release system [123].

Other stimulus-responsive "smart" hydrogels are being studied [124], including a hydrogel-composite membrane coloaded with insulin and glucose oxidase enzyme that exhibits a twofold increase in insulin release rate when immersed in glucose solution, demonstrating "chemically stimulated controlled release" and "the potential of such systems to function as a chemically-synthesized artificial pancreas" [125].

Nanoparticles with an even greater range of action are being developed by Raoul Kopelman's group at the University of Michigan. The first effort in this direction has produced PEBBLEs (probes encapsulated by biologically localized embedding) [126], dye-tagged nanoparticles constructed from a polyacrylamide matrix using bulk processes [127] that can be inserted into living cells as biosensors to monitor intracellular oxygen [128], calcium [126], zinc [129], and pH [126] levels; metabolism, or disease conditions. The next goal is the development of novel molecular nanodevices for the early detection and therapy of brain cancer, using silica-coated iron oxide nanoparticles with a biocompatible polyethylene glycol coating [130]. The miniscule size of the particles—20–200 nm—should allow them to penetrate

into areas o the brain that would otherwise be severely damaged by invasive surgery. The particles are attached to a cancer cell antibody or other tracer molecule that adheres to cancer cells and are affixed with a nanopacket of contrast agent that makes the particles highly visible during MRI. The particles enhance the killing effect during the subsequent laser irradiation of brain tissue. Nanoparticles allow MRI to see a few small brain tumor cells as smal as 50 μ—depending on the cancer type, tumor cells can range from 5 to 50 μ each and may grow in locations separate from the tumor site, and hence they are sometimes not visible to surgeons. Traditional chemotherapy and radiation kill cancerous cells but also destroy healthy cells; by using nanoparticles, the killing agents can be made to directly attack primarily the sick cells [131].

Fei Yan, a postdoctoral researcher in Kopelman's lab, is working on these nanodevices, called the Dynamic Nano-Platform (Fig. 1), in research originally funded by the Unconventional Innovations Program of the National Cancer Institute and now being commercialized as therapeutic "nanosomes" under license to Molecular Therapeutics [132]. According to the company, "the nanosome platform provides the core technology with interchangeable components that provide ultimate flexibility in targeting, imaging and treatment of cancer and cardiovascular disease indications."

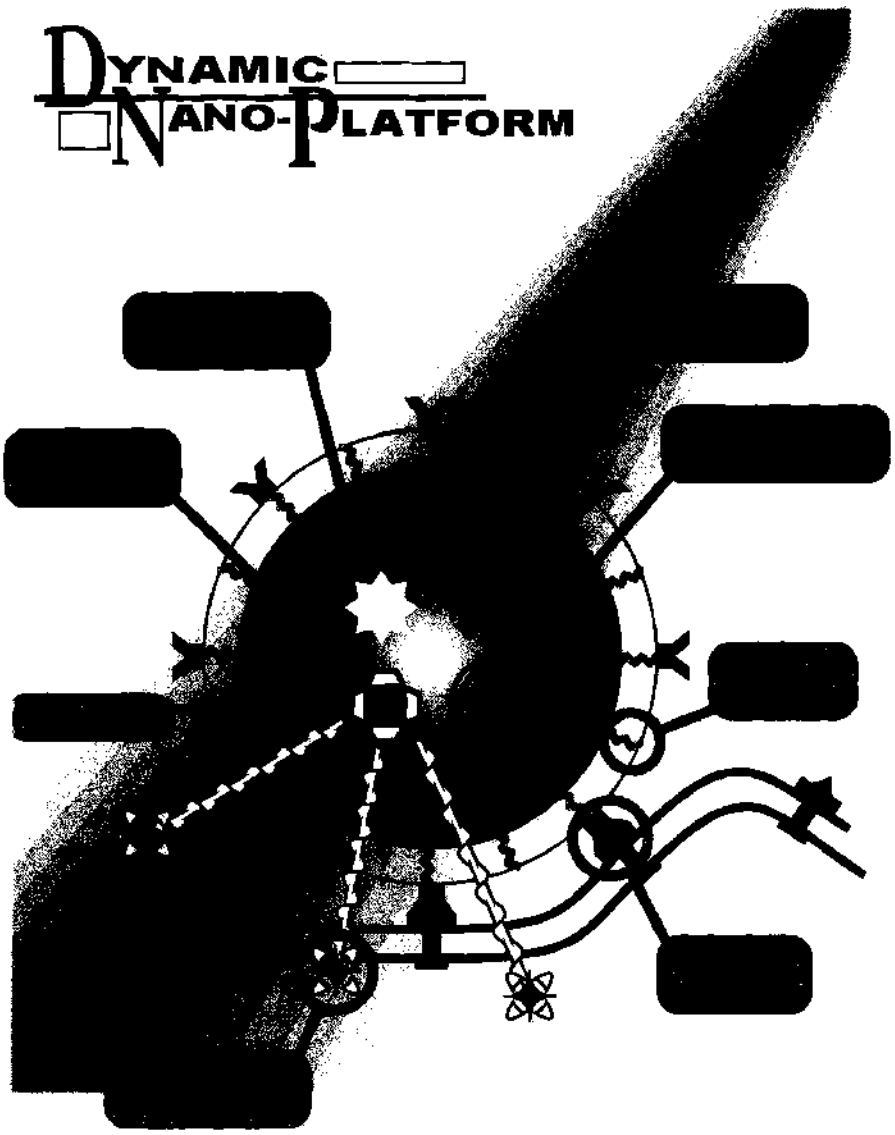


Figure 1. This illustration of the Dynamic Nano-Platform or "nanosome" shows proposed extensions of the technology, which may eventually incorporate magnetic and optical control and contrast elements to enable a number of function from biological sensing to targeted photo dynamic cancer therapy [131]. Image courtesy of Molecular Therapeutics, Inc. [132] and illustrator Eric E. Monson, who reserve all rights. © Molecular Therapeutics, Inc.

2.7. Dendrimers and Dendrimer-Based Devices

Dendrimers represent yet another nanostructured material that may soon find its way into medical therapeutics. Starburst dendrimers [133, 134] are tree-shaped synthetic molecules with a regular branching structure emanating outward from a core. Dendrimers form nanometer by nanometer, so the number of synthetic steps or "generations" dictates the exact size of the particles in a batch. Each molecule is typically a few nanometers wide, but some have been constructed up to 30 nm wide, incorporating more than 100,000 atoms. The peripheral layer of the dendrimer particle can be made to form a dense field of molecular groups that serve as hooks for attaching other useful molecules, such as DNA, which hunker down among the outermost branches. Computational studies have been done on some dendrimer-based nanoparticles [135].

In 1998, James R. Baker Jr. cofounded the Center for Biologic Nanotechnology at the University of Michigan to bring together doctors, medical researchers, chemists, and engineers to pursue the use of dendrimers [136] as a safer and more effective genetic therapy agent [134]. For Baker, these nanostructures are attractive because they can sneak DNA into cells while avoiding triggering an immune response, unlike the viral vectors commonly employed today for transfection. The dendrimer molecule is decorated with specific snippets of DNA, then injected into biological tissue. On encountering a living cell, dendrimers of a certain size trigger a process called endocytosis, in which the cell's outermost membrane deforms into a tiny bubble, or vesicle. The vesicle encloses the dendrimer, which is then admitted into the cell's interior. Once inside, the DNA is released and migrates to the nucleus, where it becomes part of the cell's genome. The technique has been tested on a variety of mammalian cell types [137] and in animal models [138, 139], though clinical human trials of dendrimer gene therapy remain to be done. Donald Tomalia at Central Michigan University, another cofounder of the Center for Biologic Nanotechnology and the original inventor of dendrimers, has reported using glycodendrimer "nanodecoys" to trap and deactivate some strains of influenza virus particles [140, 141]. The glycodendrimers present a surface that mimics the sialic acid groups normally found in the mammalian cell membrane, causing virus particles to adhere to the outer branches of the decoys instead of the natural cells. Tomalia's new company, Dendritic NanoTechnologies Inc. [142], sells a growing number of precision dendrimer molecules manufactured in bulk—at year-end 2003, the largest molecule for sale was a generation-10 ethylene diamine core PAMAM dendrimer with chemical formula $C_{40,942}H_{81,888}N_{16,378}O_{8,188}$. In July 2003, Starpharma was cleared by the FDA for human trials of their dendrimer-based anti-HIV microbicide [143]. Their product has been successful in preventing simian-HIV.

Baker's lab now has a more ambitious agenda, however: the synthesis of multicomponent nanodevices called tecto-dendrimers built up from a number of single-molecule dendrimer components [144–148]. Tecto-dendrimers are composed of a single core dendrimer, which may or may not contain a therapeutic agent, surrounded by additional dendrimer modules (Fig. 2). The additional dendrimer modules are of several types, with each type designed

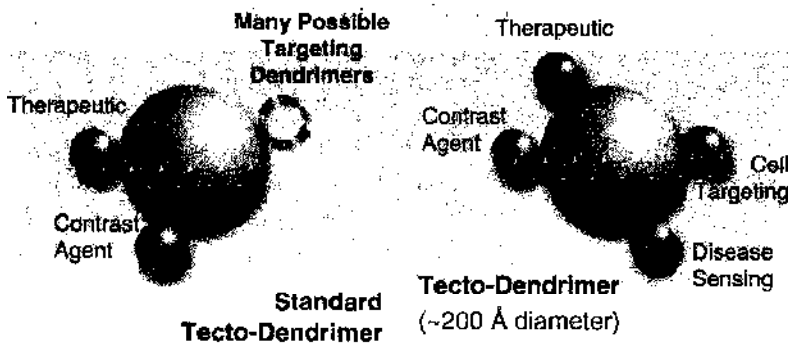


Figure 2. The standard tecto-dendrimer device, which may be composed of monitoring, sensing, therapeutic, and other useful functional modules [148]. Image courtesy of James Baker, University of Michigan © James Baker.

to perform a function necessary to a smart therapeutic nanodevice. Baker's group has built a library of dendrimeric components from which a combinatorially large number of nanodevices can be synthesized [148]. The initial library contains components that will perform the following tasks: diseased cell recognition, diagnosis of disease state, drug delivery, reporting location, and reporting outcome of therapy. By using this modular architecture, an array of smart therapeutic nanodevices can be created with little effort. For instance, once apoptosis-reporting, contrast-enhancing, and chemotherapeutic-releasing dendrimer modules are made and attached to the core dendrimer, it should be possible to make large quantities of this tecto-dendrimer as a starting material. This framework structure can be customized to fight a particular cancer simply by substituting any one of many possible distinct cancer recognition or "targeting" dendrimers, creating a nanodevice customized to destroy a specific cancer type and no other, while also sparing the healthy normal cells. In three nanodevices synthesized using an ethylenediamine core polyamidoamine dendrimer of generation 5, with folic acid, fluorescein and methotrexate covalently attached to the surface to provide targeting, imaging, and intracellular drug delivery capabilities, the "targeted delivery improved the cytotoxic response of the cells to methotrexate 100-fold over free drug" [147]. At least a half dozen cancer cell types have already been associated with at least one unique protein that targeting dendrimers could use to identify the cell as cancerous, and as the genomic revolution progresses it is likely that proteins unique to each kind of cancer will be identified, thus allowing Baker to design a recognition dendrimer for each type of cancer [148].

The tecto-dendrimer nanodevice platform is readily generalized to treat many diseases other than cancer. For instance, to cure viral infections, the body must kill virus-compromised cells, and infected cells are easy to recognize because they display nonhuman viral proteins on their surfaces. Baker envisions stocking a warehouse of viral protein-recognition dendrimers, making it possible to build any particular antiviral tecto-dendrimer by coupling the appropriate targeting dendrimer from the warehouse with the standard tecto-dendrimer. This same strategy could be applied against parasites, which also have unique nonhuman surface proteins, and even against parasites that hide inside human cells, such as malaria [148]. Tomalia's group at Central Michigan University is also pursuing tecto-dendrimer devices up to 5–100 nm in size [149]. Molecular modeling has been used to determine optimal dendrimer surface modifications for the function of tecto-dendrimer nanodevices and to suggest surface modifications that improve targeting [147].

NASA and the National Cancer Institute have funded Baker's lab to produce dendrimer-based nanodevices that can detect and report cellular damage resulting from radiation exposure in astronauts on long-term space missions [150]. By mid-2002, the lab had built a nanodevice that detects and reports the intracellular presence of caspase-3, one of the first enzymes that is released during cellular suicide or apoptosis (programmed cell death), which is one sign of a radiation-damaged cell. Baker's caspase-detecting dendrimer has two components. The first component fools white blood cells into identifying the dendrimer as a blood sugar so that the nanodevice is readily absorbed into the cell. The second component uses fluorescence resonance energy transfer (FRET) that employs two closely bound molecules. Before the cell undergoes apoptosis, the FRET system stays bound together and the interior of the white blood cell remains dark on illumination. Once apoptosis begins and caspase-3 is released, however, the bond is quickly broken and the white blood cell is awash in fluorescent light. A retinal scanning device is being devised to measure the amount of fluorescence inside an astronaut's body. If the level is above a certain baseline, then possible counteracting drugs can be taken.

2.8. Radio-Controlled Biomolecules

Although there are already many examples of nanocrystals attached to biological systems for biosensing purposes, the same nanoparticles are now being investigated as a means for directly controlling biological processes. Researchers Joseph Jacobson, Shuguang Zhang, and Kimberly Hamad-Schifferli, at the Massachusetts Institute of Technology, and their colleagues [151] have attached tiny radio frequency antennas—1.4-nm gold nanocrystals of less than 100 atoms—to DNA. When a ~1-GHz radio frequency magnetic field is transmitted

into the tiny antennas, the nanoparticles spin rapidly as alternating eddy currents are induced in the nanocrystals, producing inductive heating. The biological molecules to which the crystals are attached experience highly localized heating, causing the double-stranded DNA to separate into two strands in a matter of seconds in a reversible dehybridization process that leaves neighboring molecules untouched. When the magnetic field is removed, the strands rejoin immediately [151]. "Regulation of biomolecules using electronic radio frequency control represents a new dimension in biology," says Zhang. "So far, there are no tools that can do this. To be able to control one individual molecule in a crowd of molecules is very valuable."

The long-term goal is to apply the antennas to living systems and control DNA (e.g., gene expression, the ability to turn genes on or off) via remote electronic switching. To do that, the MIT researchers could attach gold nanoparticles to specific oligonucleotides that, when added to a sample of DNA, would bind to complementary gene sequences, blocking the activity of those genes and effectively turning them off. Applying the radio frequency magnetic field would then heat the gold particles, causing their attached DNA fragments to detach, turning the genes back on. Such a tool could give pharmaceutical researchers a way to simulate the effects of potential drugs that also turn genes on and off [152]. "It's clever to find a way to bridge two very different worlds—the biochemical world of nucleic acids and the physical world of electromagnetic waves," says biochemist Gerald Joyce of the Scripps Institute in La Jolla, California [153]. "You can even start to think of differential receivers—different radio receivers that respond differently to different frequencies. By dialing in the right frequency, you can turn on tags on one part of DNA but not other tags." Adds Jacobson, "Manipulation of DNA is interesting because it has been shown recently that it has potential as an actuator [Section 4.3]—a hard drive component—and can be used to perform computational operations." MIT has licensed the technology to a biotech startup, Waltham, Massachusetts-based engeneOS [154]. Von Kiedrowski's group at Ruhr-University Bochum in Germany is also developing biocompatible and thermostable gold cluster labels, tailored for nanotechnological, biomolecular, and nanomedical applications, that could be used for this purpose [155, 156].

The gold nanocrystals can be attached to proteins as well as DNA, opening up the possibility for future radio frequency biology to electronically control more complex biological processes such as enzymatic activity, protein folding and biomolecular assembly. In late 2002, Jacobson announced that his team had achieved electrical control over proteins as well [157]. The researchers separated an RNA-hydrolyzing enzyme called ribonuclease S into two pieces: a large protein segment made up of 104 amino acids and a small 18-amino-acid strand called the S-peptide. The RNAase enzyme is inactive unless the small strand sits in the mouth of the protein. Jacobson's group linked gold nanoparticles to the end of S-peptide strands and used the particles as a switch to turn the enzyme on and off: in the absence of the radio frequency field, the S-peptides adopt their usual conformation and the RNAase remains active, but with the external radio frequency field switched on, the rapidly spinning nanoparticles prevented the S-peptide from assembling with the larger protein, inactivating the enzyme. Eventually, Jacobson hopes that electronically controlled proteins and nucleotides will allow molecular biologists to cut and splice genetic information electronically, and perhaps eventually program computers to engineer new organisms [157].

Gregory Timp's group at the University of Illinois [158] is experimenting with 7-μ silicon-based microchips inserted into living cells to verify cell viability, as a precursor to testing GHz-frequency radio frequency microtransponders using carbon nanotube antennas inside cells.

Optically remote-controlled biomolecules are also being studied. For instance, researchers at the University of Washington [159] have added a reversible switch to endoglucanase (an enzyme that facilitates the breakdown of cellulose) by attaching two light-sensitive polymer chains next to the active sites. When exposed to visible light, one chain becomes hydrophilic, attracting water molecules and expanding, but when exposed to ultraviolet light, the chain becomes hydrophobic, expelling water molecules and contracting into a coil. The other chain works in reverse: under ultraviolet light it expands, and under visible light it contracts. Depending on the type of light applied, the enzyme's active site is either blocked or open;

hence enzyme activity is either off or on. The same group has also reported temperature-induced enzyme activity switching [160].

3. MICROSCALE BIOLOGICAL ROBOTS

Biotechnology originally contemplated the application of biological systems and organisms to technical and industrial processes, but in recent times the field has expanded to include genetic engineering and the emerging fields of genomics, proteomics, transcriptomics, gene chips, artificial chromosomes, tissue engineering, and even biobotics. Biotechnology now takes as its ultimate goal no less than the engineering of all biological systems, even completely designed organic living systems, using biological instrumentalities or "wet" nanotechnology. There are many good summaries of biotechnology [161–164] and its applications to future medicine [165–167] elsewhere, so here we focus solely on efforts to engineer natural nanomachine systems—microscale biological viruses (Section 3.1) and cells (Section 3.2)—to create new, artificial biological devices.

3.1. Engineered Viruses

Bacteriophage viruses, first employed therapeutically against bacteria by d'Herelle in 1922 [168], and other viruses have recently been used as self-replicating pharmaceutical agents [169]. During the last 10–15 years, bioengineered self-replicating viruses of various types [170, 171] and certain other vectors routinely were also being used in experimental genetic therapies as "devices" to target and penetrate certain cell populations, with the objective of inserting therapeutic DNA sequences into the nuclei of human target cells *in vivo*. Inserting new sequences into viral genomes, or combining components from two different viruses to make a new hybrid or chimeric virus [172, 173], is now routine, and virotherapies using oncolytic viruses that are replication competent in tumor but not in normal cells offer a novel approach for treating neoplastic diseases [174].

Efforts at purely rational virus design are underway but have not yet borne much fruit. For example, Endy et al. [175] computationally simulated the growth rates of bacteriophage T7 mutants with altered genetic element orders and found one new genome permutation that was predicted to allow the phage to grow 31% faster than wild type; unfortunately, experiments failed to confirm the predicted speedup. Better models are clearly needed [176, 177]. Nevertheless, combinatorial experiments on wild-type T7 by others [178–180] have produced new but immunologically indistinguishable T7 variants that have 12% of their genome deleted and that replicate twice as fast as wild type [180]. The Synthetic Biology Lab at MIT [181] seeks to build the next-generation T7, a bacteriophage with a genome size of about 40 Kbp and 56 genes. With DNA synthesis becoming cheap, "we wish to redesign and rebuild the entire genome, to create the next, and hopefully better, version of T7." Considerations in the redesign process include: "adding or removing restriction sites to allow for easy manipulation of various parts, reclaiming codon usage, and eliminating parts of the genome that have no apparent function." Synthesizing a phage from scratch "will allow us to better understand how Nature has designed the existing organism."

In a three-year project [182] culminating in 2002, the 7500-base polio virus was rationally manufactured "from scratch" in the laboratory by synthesizing the known viral genetic sequence in DNA, enzymatically creating an RNA copy of the artificial DNA strand, and then injecting the synthetic RNA into a cell-free broth containing a mixture of proteins taken from cells. The synthetic polio RNA then directed the synthesis of complete (and fully infectious) polio virion particles [182], allowing the researchers to claim that the virus was made without the use of any living cells. The rational design and synthesis of chimeric viral replicators is already possible today [183–186], but the rational design and synthesis of completely artificial viral sequences, leading to the manufacture of completely synthetic viral replicators, should eventually be possible.

Mark Young and Trevor Douglas at Montana State University have chemically modified the Cowpea chlorotic mottle virus (CCMV) viral protein cage surface to allow engineering of surface-exposed functional groups [187, 188]. This includes the addition of lamanin

peptide 11 (a docking site for lamanin-binding protein generously expressed on the surface of many types of breast cancer cells) to the viral coat, and the incorporation of 180 gadolinium atoms into each 28-nm viral capsid, allowing these tumor-targeting particles to serve as tumor-selective MRI contrast agents [189]. The researchers are now attempting to reengineer the artificial virion to make a complete tumor-killing nanodevice, exploiting a gating mechanism that results from reversible structural transitions in the virus [190]. Therapeutic anticancer compounds can be placed inside the viral capsule or even manufactured *in-situ*, using the capsid as a tiny reaction vessel. The natural viral gate of CCMV has been reengineered to allow control by redox potential (the oxidation state of a local environment, which influences the tendency of a molecule to lose or gain an electron). (Although CCMV, a plant virus, does not enter human cells, the final delivery vehicle could be a reconfigured human virus that does enter human cells.) Cellular interiors have a higher redox potential than blood, so viral capsids could be shut tight in transit but would open their redox-controlled gates after entering targeted cancer cells, releasing their payload of therapeutic agents. An alternative radiation-triggered switchable gate is also being developed. The team is exploring how the modified virus capsules work in a mouse model system and is encouraged by promising initial results. In principle, the four capabilities of the engineered capsids—high-sensitivity imaging, cell targeting, drug transport, and controlled delivery—represent a potentially powerful, yet minimally toxic, way to fight metastasized cancer [189].

Scientists from Osaka University have used a protein from the hepatitis B virus to synthesize 80-nm cagelike structures whose surface is modified to include a peptide that binds with a receptor on human liver cells. In one experiment, a fluorescent dye inserted into the cages reached cancerous human liver cells (both cultured in a lab dish and transplanted into mice) without affecting other cells [191]. Alterations in the surface peptides could allow the cages to be used as vehicles to deliver drugs or genes to other tissues as well.

3.2. Engineered Bacteria

Engineered bacteria were being pursued by Vion Pharmaceuticals in collaboration with Yale University. In their "Tumor Amplified Protein Expression Therapy" program [192], antibiotic-sensitive *Salmonella typhimurium* (food poisoning) bacteria were attenuated by removing the genes that produce purines vital to bacterial growth. The tamed strain could not survive very long in healthy tissue, but quickly multiplied 1000-fold inside tumors that were rich in purines. The engineered bacteria were available in multiple serotypes to avoid potential immune response in the host, and Phase I human clinical trials were underway in 2000 using clinical dosages. The next step would be to add genes to the bacterium to produce anticancer proteins that can shrink tumors, or to modify the bacteria to deliver various enzymes, genes, or prodrugs for tumor cell growth regulation.

In 1998, Glen Evans, then at the University of Texas Southwestern Medical Center, described the possible construction of synthetic genomes and artificial organisms. His proposed strategy involved determining or designing the DNA sequence for the genome, synthesizing and assembling the genome, then introducing the synthetic DNA into an enucleated pluripotent host cell to create an artificial organism. Genome engineers could modify an existing microbe by adding a biochemical pathway borrowed from other organisms, though this remains a difficult task because tailoring an existing system to match unique requirements demands detailed knowledge about the pathway. Ultimately, says Adam P. Arkin at Lawrence Berkeley National Laboratory, "we want to learn to program cells the same way we program computers." [193] Some genome engineers have started by building the biological equivalent of the most basic switch in a computer—a digital flip-flop. "Cells switch genes on and off all the time," observes MIT's Thomas F. Knight, Jr., who has pioneered some of this research. A cellular toggle switch, made of DNA and some well-characterized regulatory proteins, might be devised to turn on a specific gene when exposed to a particular chemical. These could be used in gene therapies: Implanted genes might be controlled with single doses of specially selected drugs, one to switch the gene on, another to switch it off [193].

Arcady Mushegian of Akkadix Corporation has looked at the genes present in the genomes of fully sequenced microbes to see which ones are always conserved in nature [194]. He concludes that as few as 300 genes are all that may be required for life, constituting the minimun possible genome for a functional microbe. An organism containing this minimal gene set would be able to perform the dozen or so functions required for life: manufacturing cellular biomolecules, generating energy, repairing damage, transporting salts and other molecules, responding to environmental chemical cues, and replicating. The minimal microbe—a basic cellular chassis—could be specified by a genome only 150,000 nucleotide bases in length. By 2001, Glen Evans could already produce made-to-order DNA strands that were 10,000 nucleotide bases in length [195] and was striving to increase this length by at least a factor of ten.

An engineered full-genome DNA, once synthesized, could be placed inside an empty cell membrane—most likely a living cell from which the nuclear material had been removed. Used in medicine, these artificial biorobots could be designed to produce useful vitamins, hormones, enzymes, or cytokines in which a patient's body was deficient; to selectively absorb and metabolize into harmless end products harmful substances such as poisons, toxins, or indigestible intracellular detritus; or even to perform useful mechanical tasks. One private company, engeneOS [196], was formed in 2000 to pursue the construction of these artificial biological devices, and in 2001 another company, Robiobotics LLC, put forward a business plan to pursue "whole genome engineering" and to begin seeking funding. Several other groups may be even further along in wetware engineering.

In November 2002, J. Craig Venter, of human genome–sequencing fame, and Hamilton O. Smith, a Nobel laureate, announced [197] their new company, Institute for Biological Energy Alternatives (IBEA), had received a $3 million, 3-year grant from the Energy Department to create a minimalist organism, starting with the *Mycoplasma genitalium* microorganism. Working with a research staff of 25 people, the scientists will remove all genetic material from the organism, then synthesize an artificial string of genetic material resembling a naturally occurring chromosome that they hope will contain the minimum number of *M. genitalum* genes needed to sustain life [197]. The artificial chromosome will be inserted into the hollowed-out cell, which will then be tested for its ability to survive and reproduce. To ensure safety, Smith and Venter said the cell will be deliberately hobbled to render it incapable of infecting people; it also will be strictly confined and designed to die if it does manage to escape into the environment [197].

In 2003, Glen Evans's new company Egea Biosciences [198] vaulted into the lead, receiving the first patent [199] "to include broad claims for the chemical synthesis of entire genes and networks of genes comprising a genome, the 'operating system' of living organisms." According to the company, Egea's proprietary GeneWriter and Protein Programming technology has been proven in extensive proof of concept studies and has produced libraries of more than 1,000,000 programmed proteins, produced over 200 synthetic genes and proteins, produced the largest gene ever chemically synthesized of over 16,000 bases, engineered proteins for novel functions, improved protein expression through codon optimization, and developed custom genes for protein manufacturing in specific host cells. Egea's software allows researchers to author new DNA sequences that the company's hardware can then manufacure to specification with a base-placement error of only $\sim 10^{-4}$, which Evans calls "word processing for DNA" [200].

According to Egea's patent [199], one "preferred embodiment of the invention" would include the synthesis of "a gene of 100,000 bp . . . from one thousand 100-mers. The overlap between 'pairs' of plus and minus oligonucleotides is 75 bases, leaving a 25 base pair overhang. In this method, a combinatorial approach is used where corresponding pairs of partially complementary oligonucleotides are hybridized in the first step. A second round of hybridization then is undertaken with appropriately complementary pairs of products from the first round. This process is repeated a total of 10 times, each round of hybridization reducing the number of products by half. Ligation of the products then is performed." The result would be a strand of DNA 100,000 base pairs in length, long enough to make a very simple bacterial genome [200]. Evans says his prototype machine can synthesize 10,000 bases in 2 day, and that 100,000-bp strands might require "a matter of weeks" to synthesize using

a future next-generation machine [200]. "Pretty soon, we won't have to store DNA in large refrigerators," says Tom Knight. "We'll just write it when we need it" [200].

4. MEDICAL NANOROBOTICS

The third major development pathway of nanomedicine—molecular nanotechnology (MNT), or nanorobotics [6, 8, 201]—takes as its purview the engineering of all complex mechanical medical systems constructed from the molecular level. Just as biotechnology extends the range and efficacy of treatment options available from nanomaterials, the advent of molecular nanotechnology will again expand enormously the effectiveness, comfort, and speed of future medical treatments while at the same time significantly reducing their risk, cost, and invasiveness. MNT will allow doctors to perform direct *in vivo* surgery on individual human cells. The ability to design, construct, and deploy large numbers of microscopic medical nanorobots will make this possible.

4.1. Early Thinking in Medical Nanorobotics

The first and most famous scientist to voice these possibilities was the late Nobel physicist Richard P. Feynman, who worked on the Manhattan Project at Los Alamos during World War II and later taught at CalTech for most of his professorial career. In his remarkably prescient 1959 talk "There's Plenty of Room at the Bottom," Feynman proposed employing machine tools to make smaller machine tools, with these to be used in turn to make still smaller machine tools, and so on all the way down to the atomic level [202]. Feynman prophetically concluded that this is "a development which I think cannot be avoided." Such nanomachine tools, nanorobots, and nanodevices could ultimately be used to develop a wide range of atomically precise microscopic instrumentation and manufacturing tools; that is, nanotechnology.

Feynman was clearly aware of the potential medical applications of the new technology he was proposing. After discussing his ideas with a colleague, Feynman offered [202] the first known proposal for a nanomedical procedure to cure heart disease: "A friend of mine (Albert R. Hibbs) suggests a very interesting possibility for relatively small machines. He says that, although it is a very wild idea, it would be interesting in surgery if you could swallow the surgeon. You put the mechanical surgeon inside the blood vessel and it goes into the heart and looks around. (Of course the information has to be fed out.) It finds out which valve is the faulty one and takes a little knife and slices it out. Other small machines might be permanently incorporated in the body to assist some inadequately functioning organ." Later in his historic lecture in 1959, Feynman urged us to consider the possibility, in connection with biological cells, "that we can manufacture an object that maneuvers at that level!"

The vision behind Feynman's remarks became a serious area of inquiry two decades later, when K. Eric Drexler, while still a graduate student at the Massachusetts Institute of Technology, published a technical paper [203] suggesting that it might be possible to construct, from biological parts, nanodevices that could inspect the cells of a living human being and carry on repairs within them. This was followed a decade later by Drexler's seminal technical book [201] laying the foundations for molecular machine systems and nanorobotics, and subsequently by Freitas's technical books [6, 8] on medical nanorobotics.

4.2. Nanorobot Parts and Components

Extending nanomedicine to molecular machine systems will probably require, among many other things, the ability to build precise structures, actuators, and motors that operate at the molecular level, thus enabling manipulation and locomotion. For example, in 1992 K. Eric Drexler of the Institute for Molecular Manufacturing theorized that an efficient nanomechanical bearing could be made by bending two graphite sheets into cylinders of different diameters, then inserting the smaller one into the larger one [201]. By 2000, John Cumings and Alex Zettl at U.C. Berkeley had demonstrated experimentally that nested carbon nanotubes do indeed make exceptionally low-friction nanobearings [204].

4.2.1. Nanobearings and Nanogears

To establish the foundation for molecular manufacturing, it is first necessary to create and to analyze possible designs for nanoscale mechanical parts that could, in principle, be manufactured. Because these components cannot yet be physically built in 2004, such designs cannot be subjected to rigorous experimental testing and validation. Designers are forced instead to rely on *ab initio* structural analysis and computer studies including molecular dynamics simulations. Notes Drexler [201]: "Our ability to model molecular machines (systems and devices) of specific kinds, designed in part for ease of modeling, has far outrun our ability to make them. Design calculations and computational experiments enable the theoretical studies of these devices, independent of the technologies needed to implement them."

In nanoscale design, building materials do not change continuously as they are cut and shaped, but rather must be treated as being formed from discrete atoms [205]. A nanoscale component is a supermolecule, not a finely divided solid. Any stray atoms or molecules within such a structure may act as dirt that can clog and disable the device, and the scaling of vibrations, electrical forces, thermal expansion, magnetic interaction, and surface tension with size lead to dramatically different phenomena as system size shrinks from the macroscale to the nanoscale [201].

Molecular bearings are perhaps the most convenient class of components to design because their structure and operation is fairly straightforward. One of the simplest examples is Drexler's overlap-repulsion bearing design [201], shown with end views and exploded views in Fig. 3 using both ball-and-stick and space-filling representations. This bearing has exactly 206 atoms including carbon, silicon, oxygen, and hydrogen, and it comprises a small shaft that rotates within a ring sleeve measuring 2.2 nm in diameter. The atoms of the shaft are arranged in a sixfold symmetry, and the ring has 14-fold symmetry, a combination that provides low-energy barriers to shaft rotation. At the atomic scale, the two opposing surfaces have periodic bumps and hollows, but the periods of these bumps are different for the two surfaces; that is, they are "incommensurate" [201, 207]. Two incommensurate surfaces cannot lock up in any particular position; hence, the barrier to free rotation is very low, on the order of ~0.001 kT (thermal noise at room temperature) [208]. Figure 4 shows an exploded view of a 2808-atom strained-shell sleeve bearing designed by Drexler and Merkle [201], using molecular mechanics force fields to ensure that bond lengths, bond angles, van der Waals distances, and strain energies are reasonable. Components of high rotational symmetry may consist of intrinsically curved, strained-shell, or special-case structures [201]. In the case of strained-shell structures, the bearing illustrated in Fig. 4 has bond strains of around ~10% (~38 zJ/atom), and similar hydrocarbon bearings have been designed with bond strains of ~5% (~11 zJ/atom) [6]. For comparison, strain energies [209, 210] are ≤3 zJ/atom for diamond lattice, ~25 zJ/atom in C_{240}, ~7–27 zJ/atom in the walls of infinite carbon nanotubes of diameter 0.7–1.3 nm, up to ~59 zJ (13% strain) for some bonds around a Lomer dislocation in diamond [211], ~70 zJ/atom in C_{60}, and at least ~80 zJ/atom for C_{36}. Fullerenes are among the most highly strained natural molecules ever isolated. For symmetrical diamondoid structures with negligible hoop stress, permissible bond strains may in theory be as large as ~140 zJ/atom producing a ~23% bond strain [201]; nanotube breaking strain is 20–30% for various chiral forms, and buckling strain is ~8% in axial compression. Bond strain in a simple strained-shell bearing can be lowered by making the bearing bigger, and thereby reducing the curvature. Thus, strained shell bearings are feasible, although in 2004 it remained unclear exactly how small they could be [212, 213] before becoming unstable. This 48-nm diameter bearing features an interlocking-groove interface that derives from a modified diamond (100) surface. Ridges on the shaft interlock with ridges on the sleeve, making a very stiff structure. Attempts to bob the shaft up or down, or rock it from side to side, or displace it in any direction (except axial rotation, wherein displacement is extremely smooth) encounter a very strong resistance [206]. Whether these bearings would have to be assembled in unitary fashion, or instead could be assembled by inserting one part into the otherwithout damaging either part, had not been extensively studied or modeled by 2004.

Molecular gears are another convenient component system for molecular manufacturing design-ahead. For example, Drexler and Merkle [201] designed a 3557-atom planetary gear, which is shown in side, end, and exploded views in Fig. 5. The entire assembly has 12

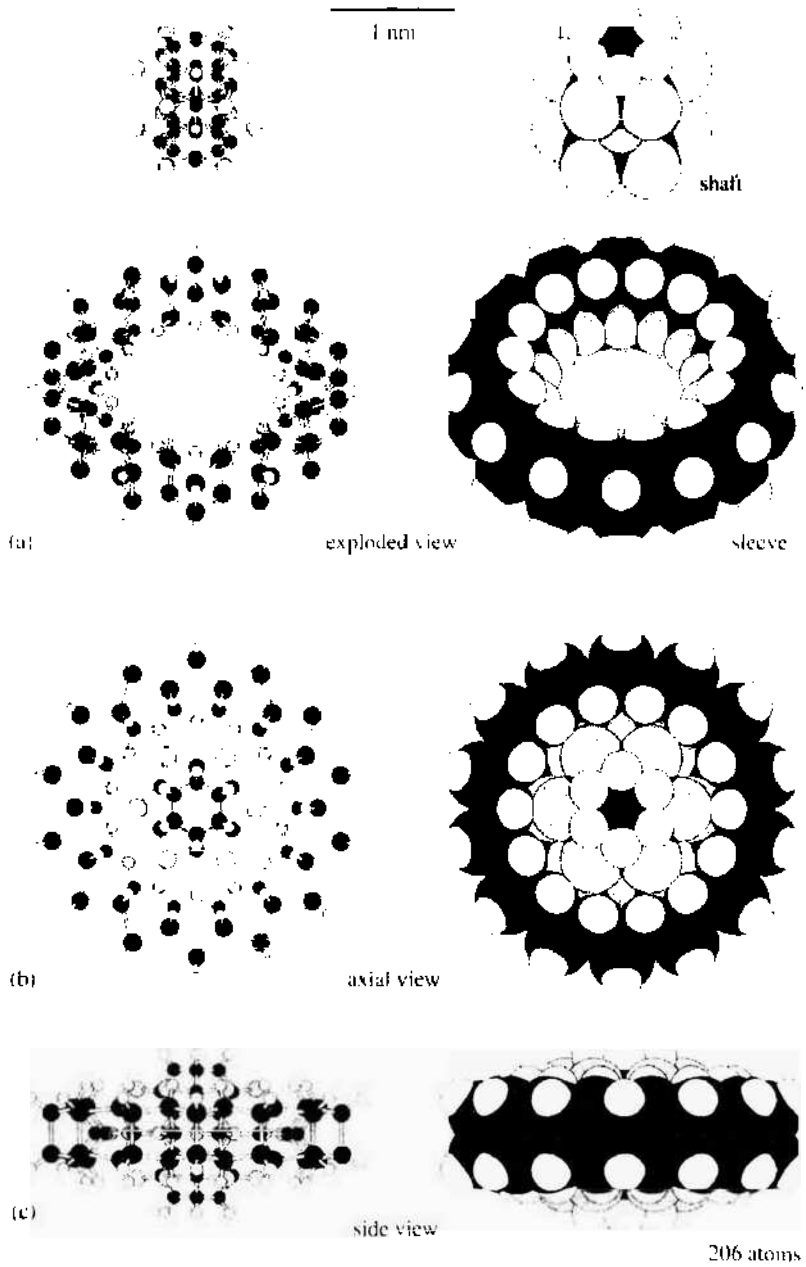


Figure 3. End views and exploded views of a 206-atom overlap-repulsion bearing [201]. Reprinted with permission from [201], K. E. Drexler, in "Nanosystems: Molecular Machinery, Manufacturing, and Computation," Wiley, New York, 1992. © 1992, John Wiley and Sons, Inc.

moving parts and is 4.3 nm in diameter and 4.4 nm in length, with a molecular weight of 51,009.844 Da and a molecular volume of 33.458 nm^3. An animation of the computer simulation shows the central shaft rotating rapidly and the peripheral output shaft rotating slowly. The small planetary gears rotate around the central shaft, and they are surrounded by a ring gear that both holds the planets in place and ensures that all of the components move in the proper fashion. The ring gear is a strained silicon shell with sulfur atom termination; the sun gear is a structure related to an oxygen-terminated diamond (100) surface; the planet gears resemble multiple hexasterane structures with oxygen rather than CH_2 bridges between the parallel rings; and the planet carrier is adapted from a Lomer dislocation [214] array created by R. Merkle and L. Balasubramaniam and is linked to the planet gears using C—C bonded bearings. The view in Fig. 5c retains the elastic deformations that are hidden in Fig. 5a—the gears are bowed. In the macroscale world, planetary gears are used in automobiles and other machines in which it is necessary to transform the speeds of rotating shafts.

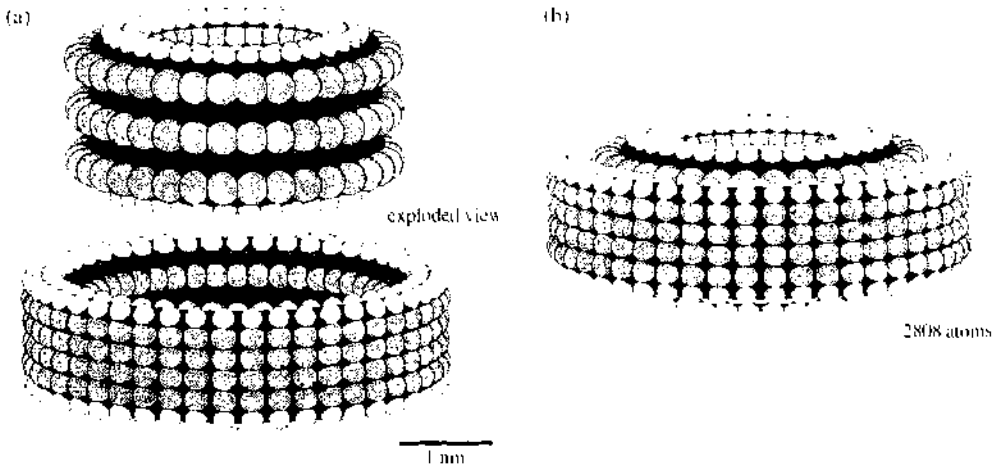


Figure 4. Exploded view of a 2808-atom strained-shell sleeve bearing [201]. Reprinted with permission from [201]. K. E. Drexler, in "Nanosystems: Molecular Machinery, Manufacturing, and Computation," Wiley, New York, 1992. © 1992, John Wiley and Sons, Inc.

Goddard and colleagues at CalTech [205, 215] performed a rotational impulse dynamics study of this "first-generation" planetary gear. At the normal operational rotation rates for which this component was designed (e.g., <1 GHz for <10 m/s interfacial velocities), the gear worked as intended and did not overheat [215]. Started from room temperature, the gear took a few cycles to engage, then rotated thermally stably at ~400 K. However, when the gear was driven to ~100 GHz, significant instabilities appeared although the device

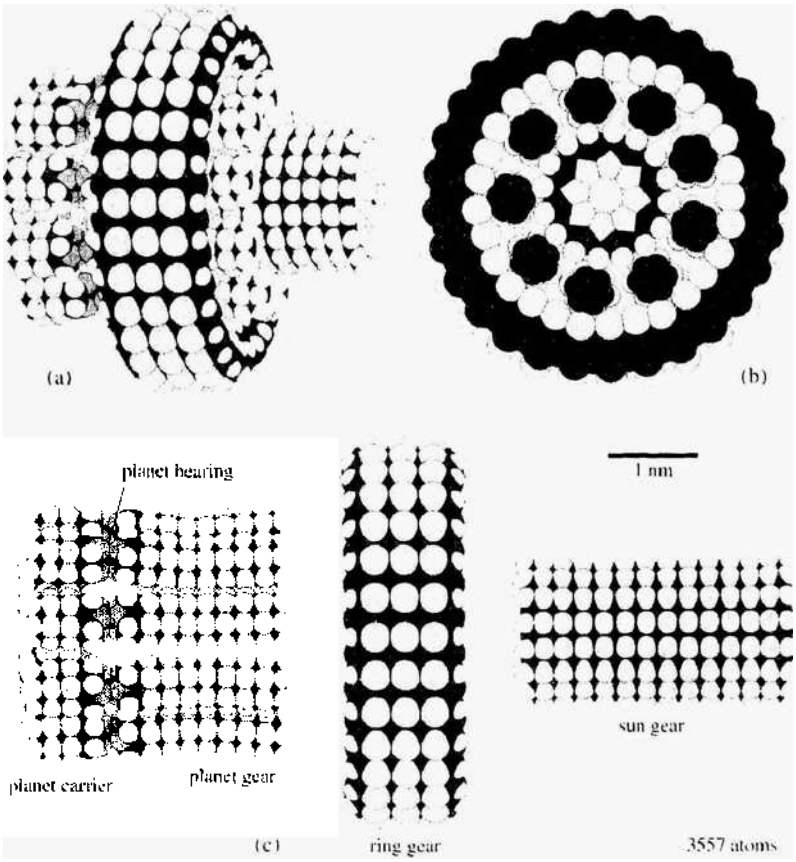


Figure 5. End-, side-, and exploded-view of a 3557-atom planetary gear [201]. Reprinted with permission from [201]. K. E. Drexler, in "Nanosystems: Molecular Machinery, Manufacturing, and Computation," Wiley, New York, 1992. © 1992, John Wiley and Sons, Inc.

still did not self-destruct [215]. One run at ~80 GHz showed excess kinetic energy causing the gear temperature to oscillate up to 450 K above baseline [205]. One animation of the simulation shows that the ring gear wiggles violently because it is rather thin. In an actual nanorobot incorporating numerous mechanical components, the ring gear would be part of a larger wall that would hold it solidly in place and would eliminate these convulsive motions that, in any case, are seen in the simulation only at unrealistically high operating frequencies.

Drexler and Merkle [216] later proposed a "second-generation" planetary gear design having 4235 atoms, a molecular weight of 72,491.947 Da, and a molecular volume of 47.586 nm^3. This new version was indeed more stable but still had too much slip at the highest frequencies. Commenting on the ongoing design effort, Goddard [205] suggested that an optimal configuration could have the functionality of a planetary gear but might have an appearance completely different from the macroscopic system, and offered an example: "Because a gear tooth in the xy plane cannot be atomically smooth in the z-direction, we may develop a Vee design so that the Vee shape of the gear tooth in the z-direction nestles within a Vee notch in the race to retain stability in the z-direction as the teeth contact in the xy plane. This design would make no sense for a macroscopic gear system since the gear could never be placed inside the race. However, for a molecular system one could imagine that the gear is constructed and that the race is constructed all except for a last joining unit. The parts could be assembled and then the final connections on the face made to complete the design."

4.2.2. Nanomotors and Power Sources

Another class of theoretical nanodevice that has been designed is a gas-powered molecular motor or pump [217]. The pump and chamber wall segment shown in Fig. 6 contain 6165 atoms with a molecular weight of 88,190.813 Da and a molecular volume of 63.984 nm^3. The device could serve either as a pump for neon gas atoms or (if run backward) as a motor to convert neon gas pressure into rotary power. The helical rotor has a grooved cylindrical bearing surface at each end, supporting a screw-threaded cylindrical segment in the middle. In operation, rotation of the shaft moves a helical groove past longitudinal grooves inside

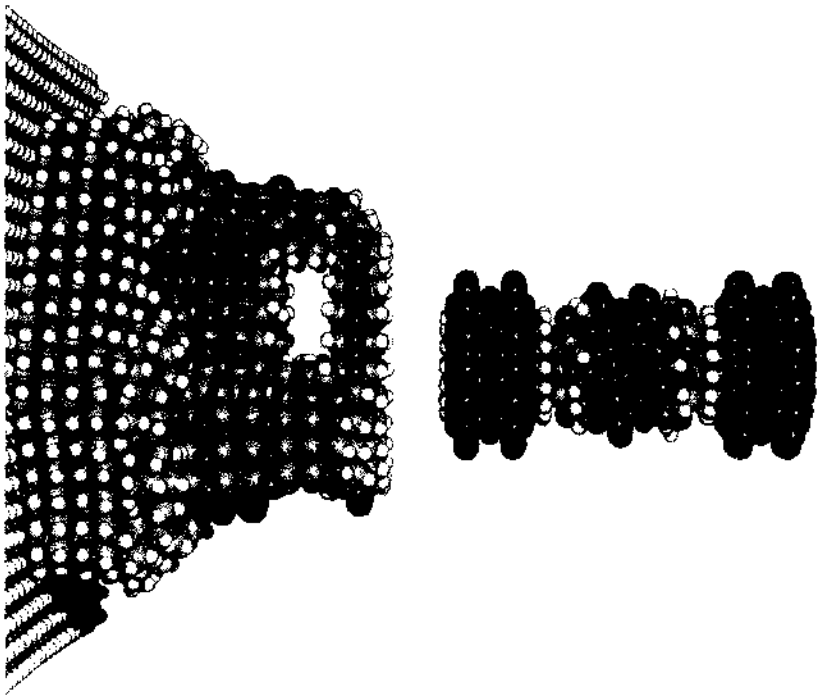


Figure 6. Side views of a 6165-atom neon gas pump/motor [217]. Image courtesy of K. Eric Drexler. Reprinted with permission from [217]. K. E. Drexler and R. C. Merkle, http://www.imm.org/Parts/Parts1.html or http://www.imm.org/Images/pumpApartC.jpg. © Institute for Molecular Manufacturing.

the pump housing. There is room enough for small gas molecules only where facing grooves cross, and these crossing points move from one side to the other as the shaft turns, moving the neon atoms along. Goddard [205] reported that preliminary molecular dynamics simulations of the device showed that it could indeed function as a pump, although "structural deformations of the rotor can cause instabilities at low and high rotational frequencies. The forced translations show that at very low perpendicular forces due to pump action, the total energy rises significantly and again the structure deforms." Merkle acknowledges that the pump moves neon atoms at an energy cost of 185 Kcal/mole-Å (12,900 zJ/atom-nm), which is not very energy efficient, but further refinement of this crude design is clearly warranted.

Almost all such design research in molecular nanotechnology is restricted to theory and computer simulation, which allows the design and testing of large structures or complete nanomachines and the compilation of growing libraries of molecular designs. The work is relatively inexpensive and does not require the support of a large team. Of course, calculations of many-body systems are notoriously difficult, with many computer packages making a number of simplifying assumptions (e.g., nuclei as point masses, electrons treated as a continuous charge distribution, and three-dimensional potential energy functions derived semiempirically from experimental data and treated as a classical field despite their true quantum mechanical character for ease of computation). Goddard [218] notes that future nanosystem simulations may require 1–100 million atoms to be considered explicitly, demanding further improvements in present-day molecular dynamics methodologies, which have only relatively recently entered the multimillion atom range [219].

Other researchers are pursuing experimental approaches to developing organic nanomotors for future nanorobots. Most notably, Carlo Montemagno, currently at the University of California at Los Angeles, has modified a natural biomotor to incorporate nonbiological parts, creating the first artificial hybrid nanomotor [220, 221]. Montemagno started with natural ATPase, a ubiquitous enzyme found in virtually every living organism and that helps to convert food into usable energy in living cells. The moving part of an ATPase molecule is a central protein shaft (or rotor, in electric motor terms) that rotates in response to electrochemical reactions with each of the molecule's three proton channels (comparable to the electromagnets in the stator coil of an electric motor). ATP (adenosine triphosphate) is the fuel that powers the molecular motor's motion.

Using the tools of genetic engineering, Montemagno added metal-binding amino acid residues to the ATPase. This allowed each motor molecule to bind tightly to nanoscale nickel pedestals prepared by electron beam lithography. Properly oriented motor molecules 12 nm in diameter were then attached to the pedestals with a precision approaching 15 nm, and a silicon nitride bar 100 nm long was bound to the rotor subunit of each motor molecule [222], all by self-assembly. In a microscopic video presentation, dozens of bars could be seen spinning like a field of tiny propellers. The group's first integrated molecular motor ran for 40 minutes at 3–4 revolutions per second. Subsequent motors have been operated for hours continuously by feeding them plenty of ATP. Montemagno has been measuring things like horsepower and motor efficiency, simple tests that would be familiar to any mechanical engineer inspecting a car engine. Montemagno is also trying to build a solar-powered, biomolecular motor-driven autonomous nanodevice, wherein light energy is converted into ATP, which then serves as a fuel source for the motor. "We think we'll be able to make autonomous devices that are powered by light on a scale of 1 μ or less, smaller than bacteria," he says.

Montemagno is developing a chemical means of switching his hybrid motors on and off reliably [223]. By engineering a secondary binding site tailored to a cell's signaling cascade, he plans to use the sensory system of the living cell to control nanodevices implanted within the cell [224]. Montemagno envisions tiny chemical factories operating inside living cells. He speculates that these nanofactories could be targeted to specific cells, such as those of tumors, where they would synthesize and deliver chemotherapy agents. Within a few years, he expects to have a motor assembled within a living cell, with the cell's physiology providing the energy to run it. "My 10-year goal is to make a device that harvests single molecules within a living cell, maybe a cellular pharmacy that produces a drug, stores it within the cell, and then based upon some signal, releases it," Montemagno said in 2000.

"For a technology that wasn't expected to produce a useful device before the year 2050, I think we've made a pretty good start. But we have a long way to go before it's safe to turn these little machines loose in the human body." Also following the bio-nano pathway is Mavroidis's group at Rutgers University, which in late 2003 received a $1 million 4-year National Science Foundation grant to produce a viral protein linear nanomotor prototype by 2007 that will "pave the way for development of complete nanorobotic assemblies" and later "make up the systems that travel the bloodstream or perform other unprecedented tasks in medicine and industry" [225].

Experimental nanomotor research is progressing in other laboratories as well. For instance, a 78-atom chemically powered rotating motor was synthesized in 1999 as a proof of principle by chemist T. Ross Kelly at Boston College [226], and chemically powered DNA-based nanomotors have also been designed and operated (Section 4.3.2). A chemically powered rotaxane-based linear motor exerting ~100 pN of force with a 1.9-nm throw and a ~250-s contraction cycle has been demonstrated by Stoddart's group at the University of California, Los Angeles [227, 228]. Ben Feringa at the University of Groningen in the Netherlands has built an artificial 58-atom motor molecule that spins when illuminated by solar energy [229]. Wong and Leigh at the University of Edinburgh [230] synthesized a motor consisting of a catenane made of a large ring with two smaller rings and four kinds of hydrogen-bonding sites. Pulsing the rings with long-wavelength ultraviolet light causes the small rings to complete a circuit along the large ring in about 70 min, though Leigh says faster ultraviolet laser cycling should make the rings spin at millions or billions of revolutions per minute if needed. White light and a bromine solution reset the motor, returning the rings to their original positions. Another potential nanorobot power source is a modified microbial fuel cell; laboratory demonstrations of such cells have used captive bacteria or immobilized enzymes [231] that, when fed organic material such as sugar [232], convert chemical energy into electricity that could be employed to power tiny electrical motors. In July 2003, U.S. patent 6,586,133 was awarded to Dale Teeters and colleagues at the University of Tulsa for their method of making micron-scale electrical batteries.

Also in 2003, the Zettl group [233] at the University of California, Berkeley, created an electrically powered 550-nm-wide nanomotor [234] by depositing a number of multiwalled nanotubes on the flat silicon oxide surface of a silicon wafer, then using an atomic force microscope to identify the best nanotubes from the pile. The researchers used electron beam lithography to simultaneously pattern a 110–300-nm gold rotor, nanotube anchors, and opposing stators around the chosen nanotubes. The researchers then annealed the rotor to the nanotubes, after which the surface was selectively etched to provide sufficient clearance for the rotor. A third stator was already buried under the silicon oxide surface. When the stators were alternately charged with 50 V of direct current, the gold rotor rocked back and forth up to 20 degrees, making a torsional oscillator. With a strong electrical jolt to the stators, the team could jerk the rotor and break the outer wall of the nested nanotubes, allowing the outer nanotube and attached rotor to spin freely around the inner nanotubes as a nearly frictionless bearing [204]; the carbon nanotube shaft to which the rotor is attached is only a few atoms across, perhaps 5–10 nm thick. The oscillating rotor might be used to generate microwave frequency oscillations possibly up to a few gigahertz, or the spinning rotor could be used to mix liquids in microfluidic devices. Zettl's team is taking friction and conductance measurements and plans to reduce the motor size by a factor of five in the future. Other groups are investigating nanotubes for use in mechanical [235] and nanoelectromechanical [236] systems.

Another well-known proposal is for nanorobotic devices to receive all power (and some control) signals acoustically [6, 201]. Externally generated ultrasonic pressure waves would travel through the aqueous environment to the nanodevice, whereupon a piston on the device would be driven back and forth in a well-defined manner, passing energy and information simultaneously into the device. Although an acoustically actuated nanoscale piston has not yet been demonstrated experimentally, microfluidic actuators are well known [237], and there are many reasons to expect that such small pistons will work as theory [201] predicts. For example, microscale pressure sensors have already been built using conventional microelectromechanical systems (MEMS) fabrication techniques (e.g., a piezoresistive pressure

microsensor diaphragm [238], a 250-μ medical pressure sensor that fits inside catheters [239], a 27-μ circular capacitive pressure sensor [240], and an optical pressure sensor 7 μ thick [241]. Micromachined flaps 45 μ in size have been raised from horizontal to vertical position by ultrasonic pulsing [242], demonstrating microscale acoustic actuation. Gas-filled 2–4μ micropores insonated at 1–2 MHz may exhibit "pistonlike" or "membranelike" vibration modes [243]. At the nanoscale, pressure applied, and then released, on carbon nanotubes causes fully reversible compression [244], and experiments have shown very low frictional resistance between nested nanotubes that are externally forced in and out like pistons [204]. Masako Yudasaka, who studies C_{60} molecules trapped inside carbon nanotubes or "peapods" at NEC (Nippon Electric Corp.), expects that "the buckyball can act like a piston" [245]. Computational simulations of acoustically driven nanopistons in a fluid environment would be useful in assessing the experimental utility of this potential nanorobot power source.

4.2.3. Nanocomputers

Truly effective medical nanorobots may require onboard computers to allow a physician to properly monitor and control their work. Molecular computing [246] has become one of the hottest research topics in nanotechnology. In 2000, a collaborative effort between the University of California, Los Angeles, and Hewlett Packard produced the first laboratory demonstration of completely reversible room-temperature molecular switches that could be employed in nanoscale memories, using mechanically interlinked ring molecules called catenanes [247], and there has been much recent progress with nanotube- and nanorod-based molecular electronics [248–250]. At least two independent companies—Molecular Electronics Corporation in Texas and California Molecular Electronics Corporation in California—have the explicit goal of building the first commercial molecular electronic devices including memories and other computational components of nanocomputers, possibly in the next few years, using techniques of self-assembly. There is also the possibility of low-speed biology-based digital nanocomputers, as briefly discussed in Section 4.3.4.

4.3. Self-Assembly and Directed Parts Assembly

4.3.1. Self-Assembly of Mechanical Parts

There is a wide range of different molecular systems that can self-assemble [251], and space does not permit more than a brief review here. Perhaps the best-known self-assembling molecular systems include those that form ordered monomolecular structures by the coordination of molecules to surfaces [252], called self-assembled monolayers (SAMs) [253], self-assembling thin films [254, 255] or Langmuir-Blodgett films [254, 256], self-assembling lipidic micelles and vesicles [257, 258], and self-organizing nanostructures [259, 260]. In many of these systems, a single layer of molecules affixed to a surface allows both thickness and composition in the vertical axis to be adjusted to 0.1 nm by controlling the structure of the molecules making up the monolayer, although control of in-plane dimensions to less than 100 nm is very difficult.

It is already known that self-assembling molecular systems can be made to self-replicate if the components have sufficient complexity. An example in biology is offered by the bacteriophages, viruses that infect and commandeer a bacterial cell's replication mechanism to synthesize their own component molecules such as nucleic acids and proteins. Spontaneous self-assembly of the component viral proteins then occurs, which produces hundreds of virus offspring in the host cell. Whereas new viruses are formed via self-assembly of randomly arranged and randomly moving component protein molecules, the various component molecules do not associate with one another at random during the assembly process but, rather, associate in a definite sequence. Biologists believe that conformational switches in protein molecules facilitate this randomized assembly of bacteriophages. In a protein molecule with several bond sites, a conformational switch causes the formation of a bond at one site to change the conformation of another bond site. As a result, a conformational change that occurs at one assembly step provides the essential substrate for assembly at the next step [261].

Several attempts have been made to achieve self-assembly of small mechanical parts to avoid direct parts grasping [262–265], and Saitou [266] gives a simple example of "sequential random bin picking," in which a process of sequential mating of a random pair of parts drawn from a parts bin which initially contains a random assortment of parts can produce the mating of a desired pair of parts. Griffith [267] complains that existing self-assembling systems are "essentially meso- or macro-scale versions of crystallization" and suggests expanding the toolbox of self-assembly by including dynamic components that emulate enzymatic allostery. Griffith has presented a simple "mechanical enzyme" analog: a two-bit mechanical state machine that programmatically self-assembles while floating at an interface between water and poly-fluorodecalin. The mechanical state machine has a mechanical flexure that acts as the "switch" in the state machine, making a mechanical allosteric enzyme. "The problem in designing self-assembling components is to avoid undesirable metastable states, and to make the desired assembled geometries the lowest energy conformations of the system," Griffith observes. For example, "you will observe that there are no straight edges on the components to avoid local energy minima on their collision." Griffith noted that he was "making progress on a 3-state system, however it is a problem of increasing difficulty, and as more component types are added to a system the challenge is to avoid any undesirable local energy minima," necessitating the development of energy versus orientation modeling tools [266].

The programming of engineered sequences of such conformational switches can allow the self-assembly of quite complicated mechanical structures. Saitou [267–271] has presented a model of self-assembling systems in which assembly instructions are written as conformational switches—local rules that specify conformational changes of a component. The model is a self-assembling automaton explicitly inspired [267] by the Penrose [272] self-replicating blocks and by Hosokawa's self-assembling triangular parts with embedded switches [273]. It is defined as a sequential rule-based machine that operates on one-dimensional strings of symbols. An algorithm is provided for constructing a self-assembling automaton that self-assembles a one-dimensional string of distinct symbols in a particular subassembly sequence. Classes of self-assembling automata are then defined on the basis of classes of subassembly sequences in which the components are self-assemble. For each class of subassembly sequence, the minimum number of conformations is provided that is necessary to encode subassembly sequences in the class. Finally, it is shown that three conformations for each component are enough to encode any subassembly sequence of a one-dimensional assembly of distinct components and having arbitrary length. Saitou claims [267] that his model of self-assembling automata can also be applied to self-assembly in two or three dimensions (in particular, to "the assembly of micro- to mesoscale components for microelectronic applications") an extension that might eventually permit the design of physical systems capable of self-replication via self-assembly.

Guided [274–277] or directed [278–280] self-assembly has become a growing research area. Yeh and Smith [263] have described a process of fluidic self-assembly of optoelectronic devices, Rothemund and Winfree [281] have described a tile assembly model for pseudocrystalline self-assembly, Breivik [282] has designed and patented a set of self-replicating physical polymers, and Gracias et al. [283] have impressed electrical circuits including light-emitting diodes on the surfaces of copper-polyimide truncated octahedra, each ~1 mm in diameter, and then induced these octahedra to self-assemble into specified three-dimensional electrical networks of up to 12 devices by comelting of opposing solder spots. (Gracias notes that hierarchical self-assembly [284] and shape-selective self-assembly using lock-and-key structures [285, 286] "offer more sophisticated strategies for the fabrication of asymmetrical networks incorporating more than one repeating unit.") Whitesides et al. [287, 288] first demonstrated capillary-force-driven assembly of a simple circuit and other structures from millimeter-scale components, as well as electrostatic self-assembly [289], and some of this work has since been extended to the fluidic self-assembly of microscale parts [263, 290–293], including "micro-origami" [294, 295] or "silicon origami" [296], as well as mesoscopic nucleic acid analogs [297]. The dynamics of Brownian self-assembly [298], the theory of designable self-assembling molecular machine structures [267, 299], and the computational modeling of self-assembly processes [300] are beginning to be addressed.

4.3.2. DNA-Directed Assembly

Working from the insight that DNA could serve as an assembly jig in solution phase, Smith and Krummenacker [301] devised a possible method for the assembly and covalent linkage of protein "parts" into specific orientations and arrangements, as determined by the hybridization of DNA attached to the proteins, called DNA-Guided Assembly of Proteins (DGAP). In this method, multiple DNA sequences would be attached to specific positions on the surface of each protein, and complementary sequences would bind, forcing protein building blocks (possibly including biomolecular motors, structural protein fibers, antibodies, enzymes, or other existing functional proteins) together in specific desired combinations and configurations, which would then be stabilized by covalent interprotein linkages. This technique could also be applied to nonprotein components that can be functionalized at multiple sites with site-specific DNA sequences, although proteins, at least initially, may be more convenient building blocks because of their size, their surface chemistry, the wide variety of functions and mechanical properties they can confer on the resulting assemblies, and the many existing techniques for introducing designed or artificially evolved modifications into natural proteins of known structure. A startup company, Molecubotics, Inc. [302], has put forward a business plan to pursue the DGAP method of molecular assembly and to begin seeking funding.

Although at this writing the experimental feasibility of the DGAP approach remains unknown, others are exploring methods for covalently attaching functional (fusion) proteins to a DNA backbone in a specified manner at ~8.5 nm (25 bp) intervals [303]; addressable protein targeting in macromolecular (e.g., nucleoprotein) assemblies [304–307]; attaching specific DNA tag sequences to protein molecules such as enzymes to a specific region on the surface of the target protein well away from the active site [308]; laying of silver wire [309, 310], gold wire [311], and other nanowires [312] using DNA as a molecular scaffolding or possible template for circuit construction; decorating DNA with fullerenes [313] and dendrimers [314]; assembling DNA-tagged nanometer-sized gold [315, 316], magnetic [317], or other particles into larger clusters, or into spontaneous lattices or magnetic chains [318]; silicon component aggregation on DNA-directed two-dimensional array [319]; and "protein stitchery" [320]. Drexler [321] notes that evolution has not maximized the stability of natural proteins, and that substantially greater stability may be engineered by various means (e.g., increasing folding stability by >100 Kcal/mole).

Early mechanical nanorobots might be made, at least in part, of DNA. The idea of using DNA to build nanoscale objects has been pioneered by Nadrian Seeman at New York University [322]. Two decades ago, Seeman recognized that a strand of DNA has many advantages as a construction material. First, it is a relatively stiff polymer. Its intermolecular interaction with other strands can be readily predicted and programmed as a result of the base-pair complementarity of nucleotides, the fundamental building blocks of genetic material. DNA also tends to self-assemble. Arbitrary sequences are readily manufactured using conventional biotechnological techniques, and DNA is readily manipulated and modified by a large number of enzymes. During the 1980s, Seeman worked to develop strands of DNA that would zip themselves up into more and more complex shapes—first tiny squares, then three-dimensional stick-figure cubes composed of 480 nucleotides each [323], then a truncated octahedron containing 2550 nucleotides [324]. By the mid-1990s, Seeman could fabricate nanoscale DNA stick figures of almost any regular geometric shape, by the billions per batch [325].

In 1999, Seeman reported yet another breakthrough—the construction of a mechanical DNA-based device that might serve as the basis for a nanoscale robotic actuator [326]. The mechanism has two rigid double-stranded DNA arms a few nanometers long that can be made to rotate between fixed positions by introducing a positively charged cobalt compound into the solution surrounding the molecules, causing the bridge region to be converted from the normal B-DNA structure to the unusual Z-DNA structure. The free ends of the arms shift position by 2–6 nm during this fully reversible structural conversion, like a hinge opening and closing. "It's a very simple nanomachine," admits Seeman, "but in the scheme of molecular devices it's huge because it generates more than four times the amount of movement produced by typical molecular devices." A large version of the device might function

as an elbow, whereas smaller devices could serve as finger joints. By 2002, Seeman's group had demonstrated a mechanical DNA-based rotary motor [327] and reported the design and construction of two-dimensional DNA arrays that might serve as templates for nanomechanical assembly [328–330]. Seeman is now collaborating with genetic engineers and computational chemists to achieve "the design and fabrication of practical nanoscale devices" and "to make rapid progress in demonstrating DNA based nanoscale devices," including "sequence-dependent devices [that] can provide the diversity of structures necessary for nanorobotics."

Other DNA-based actuation is known. For example, sequence-specific DNA hybridization is commonly used to bend silicon microcantilevers [331]. Alberti and Mergny [332] synthesized a sequence-dependent DNA "piston" composed of a 21-base oligonucleotide that displays a 5-nm, two-stroke, linear motor-type movement. Li and Tan [333] have made a single-DNA-molecule inchworm motor. Shu and Guo [334] synthesized a 30-nm-long chimeric pRNA (DNA-packaging) motor made from six strands of RNA surrounding a center strand of DNA; in the presence of ATP, the RNA strands push the DNA axle in succession, spinning it around and producing 50–60 pN of force.

Bernard Yurke at Bell Laboratories and Andrew Turberfield at the University of Oxford synthesized another DNA actuator using three single strands of artificial DNA that, when placed together, find their complementary partners and self-assemble to form a V-shaped structure [335]. The open mouth of this nanotweezer can be made to close by adding a special "fuel" strand that binds to the single-stranded DNA dangling from the ends of the arms of the tweezers and zips them closed, moving from a ~7-nm separation to a ~1-nm separation in ~13 s per cycle. A special "removal" strand, when added, binds to the fuel strand and pulls it away, opening the nanotweezers again. Yurke's original sequence-dependent DNA-based actuator was cycled seven times in sequence, and more recent work [336] has focused on a continuously running DNA nanomotor. Merkle [337] has theorized a similar chemically driven DNA-based "sliding strut" actuator that demonstrates that biotechnology and self-assembly can be used to make positional devices. Hao Yan and colleagues in Reif's lab at Duke University in Durham, North Carolina, have devised X-shaped DNA tiles that link up in a square grid with some of the strands consisting of sections of DNA that can lengthen and shorten by 6.8 nm, like tiny pistons, making a net whose mesh size can expand or contract under chemical control [338]. RNA polymerase has been used as a 15–20-pN positioning motor that "can be incrementally 'walked' or positioned along DNA with nanometer scale precision" [339].

Mirkin's group [340] has created two- and three-dimensional architectures with DNA and inorganic nanoscale building blocks, including using massively parallel tip arrays, to write patterns using a DNA-based "ink"; Mavroidis's group [341] is studying combining artificial DNA structures with carbon nanotubes; and von Kiedrowski's group [342] has used synthetic three-armed tris-oligonucleotidyl building blocks to assemble three-dimensional DNA nanoscaffolds such as supramolecular tetrahedra. Reif [343] has designed (but not yet built) "the first autonomous DNA nanomechanical devices that execute cycles of motion without external environmental changes. These DNA devices translate along a circular strand of ssDNA and rotate simultaneously." Reif's first DNA device, called "walking DNA," would achieve random bidirectional motion around a circular ssDNA strand by using DNA ligase and two restriction enzymes that consume ATP as their energy source. Reif's second DNA device, called "rolling DNA," would achieve random bidirectional motion by using DNA hybridization energy in trapped states with no DNA ligase or restriction enzymes. The emerging field of DNA-based molecular construction appears quite active.

4.3.3. Protein-Directed Assembly

After noting Merkle's [337] suggestion that the positional control of nanoscale building blocks would allow the extension of the normal reactions found in biological systems to include free-radical chemistry and more recent descriptions of similar reactive chemistries being employed by enzymes [344], Bradbury [345] proposed using artificial enzymes for nanoscale parts assembly. He envisions the engineering of artificial multifunctional proteins called "robozymes" having the following properties: first, unfolded, it grabs onto molecular

building blocks [346–348], carefully keeping them separate from each other to avoid nonspecific reactions; second, using specific enzyme catalytic sites near the bound building blocks, it "activates" the molecules (perhaps producing one or more free radicals); third, induced folding brings the building blocks into relatively precise alignment, allowing the desired chemical reaction or reactions to occur; and fourth, the protein is induced to unfold, releasing the final product. With protein folding forces in the tens to hundreds of pico-Newtons, such enzymes could also provide a means of threading one molecule through another molecule, mechanically producing interlinked structures such as rotaxane and catenane [349] without using self assembly (i.e., by adding hydrophobic ring structures to the molecular parts, which will then be attracted to each other in a polar solvent such as water), and thus allowing the positional assembly of very small (<1 kDa) molecular nanoscale parts.

Note that although most enzymes in cells are involved in manipulating small molecules (<0.25 kDa), there are several classes of enzymes involved in manufacturing complex covalently bound molecules such as vitamins, enzyme cofactors, antibiotics, and toxins with masses up to ~3 kDa. Molecules even larger than this are manipulated by tRNA-synthetase (a 40–100-kDa enzyme that manipulates ~30-kDa tRNAs); the spliceosome, the ribosome, the proteosome, and the DNA replication complex. (Many of these also involve "parts insertion" or "threading" maneuvers, such as the clamp and bridge helix mechanisms in RNA polymerase II that act as a translocation ratchet to feed DNA through the enzyme interior to produce mRNA [350].) By designing synthetic enzymes consisting of synthetic amino acids, we can envision grabbing molecular parts in a solution and then, as the enzyme folds, bringing them into proper alignment and causing them to react, which might be called "nanopart synthetases" or "protein-directed parts assembly." Of course, RNA-based ribozymes [351] may prove better suited than proteins for some reactions, so we are not limited to using enzymes to form the covalent bonds required in nanoparts.

Ratchet-action protein-based molecular motors are well known in biology [352], conformational cascades of a special genetic variant of yeast cell prions have already been used to assemble silver- and gold-particle-based nanowires [353], and the GTPase dynamin mechanoenzyme which self-assembles into rings or spirals, wrapping around the necks of budding vesicles and squeezing, pinching them off, during cellular endocytosis, is also well known [354]. Smith [355] has used methyltransferase-directed addressing of fusion proteins to DNA scaffolds to construct a molecular camshaft as a exemplar protein/nucleic acid biostructure. Bacland and Montemagno [221] have engineered a biomolecular motor constructed of ATPase protein [220] with an attached silicon nitride "propeller" arm [222] and a reversible on/off switch [223], and other task-optimized genetically engineered molecular motors have also been synthesized by others [356]. Protein-protein binding specificity has been used to bend silicon microcantilevers [357]. Finally, molecular chaperones are a group of proteins that assist in the folding of newly synthesized proteins or in the refolding of denatured proteins. Genetically engineered chaperonin protein templates (chaperone molecules) have been used to direct the assembly of gold (1.4, 5, or 10 nm) and CdSe semiconductor quantum dots (4.5 nm) into nanoscale arrays [358].

Immunoglobulin (Ig) or antibody molecules could be used first to recognize and bind to specific faces of crystalline nanoparts, and then as handles to allow attachment of the parts into arrays at known positions, or into even more complex assemblies. As reported by Freitas [1]: Kessler et al. [359] raised monoclonal antibodies (MAbs) specific for crystals of 1,4-dinitrobenzene having well-defined molecular-level structures. These antibodies were so specific they would not bind to the same molecule when it was conjugated to a protein carrier. IgG antibodies isolated from the serum of rabbits injected with crystals of monosodium urate monohydrate or magnesium urate octahydrate evidently bear in their binding sites an imprint of the crystal surface structure because they can act as nucleating templates for crystal formation in vitro with extremely low cross-reactivity, despite the similar molecular and structural characteristics of the two crystals [360]. Antibody binding to monosodium urate crystals has been known for decades [361], and viruses have been engineered with a specific recogniton moiety for ZnS nanocrystals used as quantum dots [362].

Like antigens with ordered multiple epitopes, crystals expose chemically and geometrically distinct surfaces, so different antibodies might recognize distinct faces of a crystal

(possibly including diamond crystal faces) in an interaction similar to that of antibodies for repetitive epitopes present on protein surfaces [347, 363]. For instance, one MAb to 1,4-dinitrobenzene crystals was shown to specifically interact with the molecularly flat, aromatic, and polar (101) face of these crystals, but not with other faces of the same crystal [348]. MAbs have also been elicited against cholesterol monohydrate crystals, one of which [364] was shown to specifically recognize the crystal's stepped (301) face. Here, the hydrophobic cholesterol hydrocarbon backbone is exposed on one side of the molecular steps while hydroxyl residues and water molecules are exclusively exposed on the other side. In both cases, crystal-specific antibodies were of the IgM idiotype [347]. This accords with the assumption that (unlike most commonly used antigens) crystals cannot be processed by the antigen-presenting cells, and hence antibodies must be induced through a T-cell-independent path [365]. Semiconductor-binding [278, 362] and calcite-binding [366] proteins are known that can discriminate among the various crystal faces of the given material and that can in some cases alter the pattern of crystal growth [367]. Sulfur-free gold-binding proteins recognize and noncovalently bind preferentially to the Au (111) crystal surface (gold-binding proteins use multiple repeats of 14–30-residue sequences to bind to this surface [368]).

Solubilized (derivatized) C_{60} and C_{70} fullerenes can induce the production of specific antibodies [346, 347, 369], usually by interaction with the combining sites of IgG [346]. It is speculated that highly hydrophobic pure fullerenes would be recognized by antibodies with hydrophobic amino acids in their binding sites [346, 370] or would interact with donor $-NH_2$ [371] and -SH [372] groups. There are several reports of antibodies being raised to single-walled carbon nanotubes [369, 370]. Computer simulations indicate that it may be possible to build antibodies that selectively bind to nanotubes of a specific diameter or chirality [370].

4.3.4. Microbe- and Virus-Directed Assembly

Artificial microbes might also be employed in molecular construction. A variety of biological molecular machines are already known that display linear motions; movements related to opening, closing, and translocation functions; rotary movements; and threading–dethreading movements [373]. Gerald J. Sussman at MIT claims that when computer parts are reduced to the size of single molecules, engineered microbes could be directed to lay down complex electronic circuits [193]. "Bacteria are like little workhorses for nanotechnology; they're wonderful at manipulating things in the chemical and ultramicroscopic worlds," he says. "You could train them to become electricians and plumbers, hire them with sugar and harness them to build structures for you." Regarding microbe-directed parts fabrication, one strain of bacteria (*Pseudomonas stutzeri AG259*) is known to fabricate single crystals of pure silver in specific geometric shapes, such as equilateral triangles and hexagons, up to 200 nm in size [374], and microorganisms can accumulate materials and synthesize inorganic structures composed of bismuth [375], CdS [376, 377], gold [378], magnetite [377, 379], silica [377], and silver [377].

As for microbe-directed parts assembly, Kondo et al. [380] used a grooved film (created by chemically precipitating cellulose tracks <1 nm apart onto a copper base) to train the bacterium *Acetobacter xylinum* to exude neat ribbons of cellulose along the prepared track at a rate of 4 μ/minute; the group is attempting to genetically modify the organism to secrete alternative sugar molecules that might better resist natural degradation. Natural fibroblasts in human tissue construct complex three-dimensional collagenous fiber networks of extracellular matrix (ECM) during wound healing, fibrillogenesis and fibroplasia. Although ECM strand positioning is stochastic in natural fibroblasts, cell functionality, and ECM network characteristics can be altered by chemotactic factors, contact guidance and orientation, hypoxia, and local mechanical stress. Fibroblasts can be genetically engineered, are capable of crosslinking collagen fibers (a "covalent parts joining" type of operation), and can apply ~100 pN forces while embedded in a three-dimensional collagen lattice [381].

To establish digital control over microorganisms, genetic circuits that can function as switches [382] or computational logic elements such as AND, NAND, and NOR gates (Ref. [6], Section 10.2.3.1) are under active investigation [383–386]. For example, in 2000, Gardner et al. [384] added a memory device to an *E. coli* bacterium using two inverters for which the output protein of each is the input protein of the other, and Elowitz and Leibler [383]

made an oscillator with three inverters connected in a loop: in one test of their system, "a fluorescent protein became active whenever one of the proteins was in its low state . . . the result was a population of gently twinkling cells like flashing holiday lights" [386]. By 2002, Weiss [385] had created a five-gene circuit in *E. coli* that could detect a specific chemical in its surroundings and turn on a fluorescent protein when the chemical concentration falls within preselected bounds [386].

The Synthetic Biology Lab at MIT is similarly trying to create cells that are "engineered genetic blinkers" [387] and that use light as a faster means of cellular input–output than chemical-mediated signals [388]. They are also creating a set of "BioBricks" [388, 389] that are "a [standardized] set of [building block] components that have been designed for use as logic functions within a cell. The members of this family are designed to be compatible, composable, interchangeable, and independent so that logic circuits may be constructed with little knowledge or concern for the origins, construction, or biological activities of the components."

Bostor University bioengineer Timothy Gardner explains [386] that the eventual goal "is to produce genetic 'applets', little programs you could download into a cell simply by sticking DNA into it, the way you download Java applets from the Internet." Bacterial memory has also been demonstrated: 150-base-long messages encoded as artificial DNA have been stored within the genomes of multiplying *E. coli* and *Deinococcus radiodurans* bacteria and then accurately retrieved [390]. Jacobson's team [151] has demonstrated remote electronic control over the hybridization behavior of DNA molecules by inductive coupling of a radio frequency magnetic field to a gold nanocrystal covalently linked to DNA (Section 2.8), offering the prospect of remote-controlled enzymes and "radio-controlled bacteria."

Bacteria can also be used as physical system components. For example, Kim et al. [391] are attempting to incorporate living bacteria into MEMS to form living cell motors for pumps and valves. The bacteria will be completely sealed inside the bioMEMS device. "When its flagellum is attached to a surface, the bacterium moves in a circular fashion, and always in the same direction," explains Tung. "A single bacterium can become a flagellar motor or pump, but a number of bacteria, all rotating in the same direction, could become a conveyor belt." Similarly, Linda Turner and colleagues at the Rowland Institute at Harvard have affixed a film of *Serratia marcescens* bacteria onto tiny beads, allowing the microbes' rotating appendages to carry the beads along. When the film is applied inside tiny tubes, the gyrating bacterial arms blend fluids twice as fast as diffusion alone [392]. Carlo Montemagno at the University of California, Los Angeles, has combined living cells with isolated MEMS structures to create cell-powered mechanical motors. In one experiment in 2003, a lithographically produced U-shaped structure 230 μ wide is attached to a cardiac muscle cell like a tiny prosthesis. When presented with glucose solution, the muscle cell contracts repeatedly, causing the mechanical structure to "walk" at a speed of ~46 μ/min with a repetition rate controlled by the spring constant of the MEMS structure [393]. Sequeira and Copik [394] also proposed using bacteria as power units for microscale mechanical systems.

Viral shells also provide useful templates for nanoscale assembly. Belcher [278, 362] employs virus capsid shells as scaffolds for the directed nanoassembly of nanoparticles such as quantum dots [362, 395] in a process she describes as the "biomimetic synthesis of nonbiological inorganic phases with novel electronic and magnetic properties directed by proteins and synthetic analogs." In one experiment [362], a genetically engineered M13 bacteriophage with a specific recognition moiety for zinc sulfide nanocrystals was used to assemble a ZnS-containing film having nanoscale ordering and 72-μ-sized domains. Viral coat proteins can be engineered by various techniques and have been used by others as scaffolds for nanomaterials synthesis [187] and self-assembly [396], including self-assembled monolayers [188].

4.4. Positional Assembly and Molecular Manufacturing

As machine structures become more complex, getting all the parts to spontaneously self-assemble in the right sequence will be increasingly difficult. To build such complex structures, it makes more sense to design a mechanism that can assemble a molecular structure by what is called positional assembly (i.e., picking and placing molecular parts). A device capable of

positional assembly would work much like the robot arms that manufacture cars on automobile assembly lines in Detroit, or that insert electronic components onto computer circuit boards with blinding speed in Silicon Valley. Using the positional assembly approach, the robot manipulator picks up a part, moves it to the workpiece, installs it, then repeats the procedure over and over with many different parts until the final product is fully assembled.

One of the leading proponents of positional assembly at the molecular scale is Zyvex Corp., a privately held molecular nanotechnology company headquartered in Richardson, Texas [397]. Zyvex is the first engineering firm to espouse an explicit goal of using positional assembly to manufacture atomically precise structures, or more specifically, "a user-controlled fabrication tool capable of creating molecularly precise structures with 3-dimensional capability in an economically viable manner." As a first step toward this goal, in 1998 Zyvex demonstrated the ability to use three independently controlled inch-long robotic arms to manipulate tiny carbon nanotubes in three dimensions under the watchful eye of a scanning electron microscope that can monitor objects and motions as small as 1 nm at near-video-scan rates. Zyvex has already demonstrated the ability to positionally assemble large numbers of MEMS-scale parts and still has a very long way to go before it can assemble nanoscale parts into useful machines, but its work is a step in the right direction, and the research continues today (Section 4.4.2).

Microscale devices could also be used to pick and place nanoscale parts. Agilent Laboratories has created an ultra-high-precision micromover platform [398] capable of providing linear two-dimensional movement in steps of 1.5 nm, the width of about nine bonded carbon atoms. The core of the micromover is a stepper actuator or linear motor that does not rotate, but instead steps right to left or front to back. The platform can travel a total of 30 μm in each direction in 2.5 ms; because each micrometer is made up of 1000 nanometers the micromover would take approximately 20,000 steps to traverse 30 micrometers, a distance about half the width of a single human hair. Another group led by Sylvain Martel, formerly at the Bio-Instrumentation Laboratory at MIT, is working on a similar nanopositioning device called the NanoWalker [399, 400].

Others have begun to explore the realm of submicron manipulation of objects. For example, Philip Kim and Charles Lieber at Harvard University created the first general-purpose nanotweezer [401]. Its working end is a pair of electrically controlled carbon nanotubes made from a bundle of multiwalled carbon nanotubes. To operate the tweezers, a voltage is applied across the electrodes, causing one nanotube arm to develop a positive electrostatic charge and the other to develop a negative charge. The attractive force can be increased or decreased by varying the applied voltage: 8.5 V completely closes the arms, whereas lower voltages give different degrees of grip. Using the tool, Kim and Lieber have successfully grasped 500-nm clusters of polystyrene spheres, or clusters about the same size scale as cellular substructures. Kim and Lieber were also able to remove a semiconductor wire 20 nm wide from a mass of entangled wires, using tweezer arms about 50 nm wide and 4 μ long. The early hope was that by growing single-walled nanotubes directly onto the electrodes the researchers could produce nanotweezers small enough to grab individual macromolecules.

The Kim–Lieber nanotweezer is very good at pinching and releasing objects, but the technique creates a large electric field at the tweezer tips that can alter the objects being manipulated, and the tweezers must be constructed one at a time, which makes the manipulation of large numbers of nano-objects a slow and tedious process. To try to improve on this, in 2001 a group led by Peter Boggild [402] of the Technical University of Denmark in Lyngby used standard micromachining processes to carve from a tiny slab of silicon an array of cantilevered micropliers that could be opened and closed electrically. Boggild then used an electron beam to grow a tiny carbon nanotweezer arm from the end of each cantilever, angled so that the tips were only 25 nm apart, making a better-controlled nanotweezer [403]. Other nanotube-based nanotweezers have since been reported by other groups [404, 405].

Precise covalent attachment of molecules to surfaces is also being pursued. Blackledge et al. [406] used a palladium-coated SFM (scanning force microscope) tip to chemically modify terminal functional groups on an organosiloxane-coated surface to create biotin-streptavidin assemblies in patterns with minimum 35-nm line widths. Díaz et al. [407] use redox probe microscopy (RPM), in which a SFM tip is modified with redox-active materials, whereupon

the interactions between tip and an adsorbate or between tip and a surface are modulated by the electrode potential. This system has also been used as a microtweezer to manipulate and position objects. Hla and Rieder [408, 409] have reviewed recent progress in using scanning tunneling microscopy (STM) to manipulate and synthesize individual molecules.

The ultimate goal of molecular nanotechnology is to develop a manufacturing technology able to inexpensively manufacture most arrangements of atoms that can be specified in molecular detail, including complex arrangements involving millions or billions of atoms per product object, as in the hypothesized medical nanorobots (Section 4.5). This will provide the ultimate manufacturing technology in terms of its precision, flexibility, and low cost. Two central mechanisms have been proposed to achieve these goals at the molecular scale: programmable positional assembly including fabrication of diamond structures using molecular feedstock (Section 4.4.1), and massive parallelism of all fabrication and assembly processes (Section 4.4.2).

4.4.1. Diamond Mechanosynthesis

Programmable positional assembly at the molecular scale is the central mechanism for achieving both great flexibility in manufacturing and the ultimate in precision. While ubiquitous at the scale of centimeters and meters, positional assembly at the molecular scale is still rudimentary, but its promise is immense. At the same time, a full analysis of how to use positional assembly to synthesize most arrangements of atoms permitted by physical law would be, at present, prohibitively complex. The 100 plus elements of the periodic table each have their own unique chemical properties, and the various combinations and permutations of these elements creates a combinatorial explosion of possibilities whose full analysis may occupy nanotechnologists worldwide for the rest of this century. A more manageable project is the analysis of a small set of positionally controlled tool tips that could be used in the mechanosynthesis of stiff hydrocarbons, and the analysis of a significant class of stiff hydrocarbons (in particular, diamond) that could be synthesized by the use of these tool tips.

Why diamond? There are four principle reasons. First, although the basic crystal structure of diamond was first elucidated in 1913 and the cleavage of diamond crystals has been seriously studied since at least 1928, it was not until the 1980s and especially the 1990s that the molecular surface characteristics of diamond were extensively investigated both theoretically and experimentally. Many practical questions about the molecular structure of diamond have now been resolved.

Second, the development since the 1950s of a significant experimental specialty known as adamantane chemistry now allows the convenient bulk synthesis of small molecules of pure diamond crystal in specific isomeric forms containing up to 50 atoms (including up to 22 carbon atoms) in size, and in some cases allows the rational regioselective functionalization of these molecules. Over 20,000 adamantane variants have been synthesized.

Third, an even more active experimental field (of similar vintage) is devoted to the synthesis, by means of chemical vapor deposition (CVD) [410], of macroscale thin films of diamond and related stiff hydrocarbon structures. These structures have practical, commercial applications today. The field of diamond CVD provides a wealth of understanding, both experimental and theoretical, of the myriad reaction mechanisms that can contribute to the growth of diamond.

Fourth, there is widespread interest in the exceptional properties of diamond. Among other desirable properties, it has extreme hardness, high thermal conductivity, low frictional coefficient, chemical inertness, and a wide bandgap. It is the strongest and stiffest material presently known at ordinary pressures. Recent investigations have been driven by the many emerging applications for diamond in MEMS mechanical and electromechanical devices [411, 412], optics, radiology, biochemical synthesis [413], and medicine [414–416], but most especially in various electronics devices [417–419]. A method for the precise manufacture of microscopic and nanoscale diamond structures would have tremendous utility in science and industry.

Useful conclusions can already be drawn from the CVD literature [410]. First, growth occurs on all major surfaces of diamond (albeit at widely differing rates, depending on reaction conditions). Second, even on a specific surface, growth occurs under a wide range

of experimental conditions and can use a wide range of different feedstocks because there is a great diversity of reaction mechanisms by which growth can occur. These two conclusions are important because they imply the availability of a wide range of synthetic methods by which diamond growth can take place. No single roadblock, nor any one theoretical or experimental obstacle, is likely to prevent progress because there are multiple alternative routes for achieving the synthesis of any specific stiff hydrocarbon. Thus, if any particular proposal for a specific tool tip, or any specific reaction mechanism should, on closer analysis, prove to be flawed or simply impossible, it can be discarded without fear of failure in the larger endeavor.

In contrast to high-pressure diamond synthesis [420] and low-pressure gas-phase diamond synthesis of diamond via CVD [410], positional mechanosynthesis has been proposed by Drexler [201] for the precise manufacture of diamond structures. Mechanosynthesis aims to achieve site-specific chemical synthesis by inducing chemical transformations controlled by positional systems operating with atomic-scale precision (e.g., the tip of a scanning probe microscope, or SPM [421]), thus enabling direct positional selection of reaction sites on the workpiece. STM has demonstrated an ability to manipulate surface structures atom by atom, and many proposed methods involve the use of STM to direct chemical reactions on the surface by: delivering an electric field to a subnanometer region of a surface to activate a chemical reaction, manipulating the chemistry of the tip to make it act as a catalyst that can then be introduced precisely into the region of desired reaction, or delivering mechanical energy from the tip to activate surface reactions. The reaction selectivity of all these methods relies on the exponential dependence of reaction rates on the activation barrier, which is lowered for surface reactions in a precisely defined area of the surface during mechanosynthesis [422].

Mechanosynthesis may be distinguished from simple piezochemistry, a general term that describes solution-phase chemical processes in which homogeneous, isotropic, slowly varying mechanical pressure (e.g., ~0.1–2 GPa in commercially available laboratory equipment) modifies chemical reactivity [423]. Under proper conditions, 2 GPa at room temperature might suffice to convert graphite to diamond [424]. Unlike forces resulting from hydrostatic pressure, forces applied by mechanosynthetic tool tips can be highly anisotropic and inhomogeneous on a molecular scale: Large loads (including compression, tension, shear, and torsion) can be applied to specific atoms and bonds in a controlled manner [201]. Stress is a scale-independent parameter [201], so tool tips for mechanosynthesis built of diamond could apply pressures equaling those in macroscale diamond-anvil pressure cells (e.g., >550 GPa [425]), with substantial effects on bonding (e.g., H_2 becomes metallic at ~150 GPa [426]). For example, a localized pressure of ~50 GPa applied to an activation volume of ~0.006 nm^3/C-atom in diamond yields a ~300 zJ/C-atom reduction in activation energy, slightly greater than the C—C covalent bond energy of a 278 zJ/C-atom (or 556 zJ/bond) in neopentane [427].

Undoped diamond normally consists of a rigid lattice of carbon atoms surface-passivated by hydrogen atoms, so a necessary aspect of diamond mechanosynthesis involves positionally controlled hydrogen abstraction [428–430] and donation [201, 431] at the surface of the diamond crystal lattice. The extensive theoretical analysis of the hydrogen abstraction tool has involved many people, including Donald W. Brenner [430, 432, 433], Richard J. Colton [432], K. Eric Drexler [201], William A. Goddard III [434], J. A. Harrison [433], Jason K. Perry [434], Ralph C. Merkle [431, 434], Charles B. Musgrave [434], Michael Page [430], O. A. Shenderova [433], Susan B. Sinnott [432, 433], and Carter T. White [432]. The institutions involved include the Materials and Process Simulation Center at Caltech; the Department of Materials Science and Engineering at North Carolina State University; the Institute for Molecular Manufacturing; the Department of Chemicals and Materials Engineering at the University of Kentucky; the Chemistry Department of the United States Naval Academy; the U.S. Naval Research Laboratories, Surface Chemistry Branch; and the Xerox Palo Alto Research Center.

Site-specific hydrogen abstraction from crystal surfaces has recently been achieved experimentally. For instance, Musgrave et al. [429] and Lyding et al. [435–437] have demonstrated the ability to abstract an individual hydrogen atom from a specific atomic position in a covalently bound hydrogen monolayer on a flat Si(100) surface, using electrically pulsed STM tip

in ultrahigh vacuum. Ho's group [438] has also demonstrated single-atom hydrogen abstraction experimentally, using STM. Hydrogen-donation tools have been described theoretically [201, 431] but have not yet been as extensively studied. However, there has been experimental demonstration of highly localized STM-catalyzed rehydrogenation of dehydrogenated hydrocarbon clusters adsorbed to the Pt(111) surface [439], and of hydrogen donation to a prepared azide-coated surface, producing highly localized amines, using a Pt-coated AFM (atomic force microscope) tip [440]. In the following analysis, we assume that site-specific hydrogen abstraction and donation tools for mechanosynthesis are available and can be used to abstract or donate hydrogen atoms at will on diamond surfaces.

The principal challenge in diamond mechanosynthesis is the controlled addition of carbon atoms to the growth surface of the diamond crystal lattice. The theoretical analysis of carbon atom insertion and carbon dimer placement on diamond has also involved many people including Tahir Cagin [218], K. Eric Drexler [201], Fedor N. Dzegilenko [441], Robert A. Freitas Jr. [442–444], William A. Goddard III [218], David J. Mann [444], Ralph C. Merkle [431, 442–445], Charles B. Musgrave [446], Jingping Peng [443, 444], Subhash Saini [441], Deepak Srivastava [441], and Stephen P. Walch [218, 445]. The institutions involved include the Materials Simulation Center at Caltech; ELORET; the Institute for Molecular Manufacturing; the IT Modeling and Simulation Group at NAS/MRJ, NASA Ames Research Center; Department of Chemical Engineering, Stanford University; the Thermosciences Institute, NASA Ames Research Center; the Xerox Palo Alto Research Center; and Zyvex Corporation.

The feasibility of precisely inserting individual carbon atoms, small hydrocarbon species, or small clusters of carbon atoms on a C(111) or C(100) diamond surface at specific sites was initially supported first by the computational work of Walch and Merkle [445]. Walch and Merkle analyzed several mechanosynthetic reactions, including placement of a carbon dimer onto a C(111) surface, insertion of a positionally controlled carbene into that dimer, and insertion of a positionally controlled carbene into a surface dimer on a C(100) surface using a nine-atom cluster to model the diamond surface. The latter insertion can take place with no barrier (according to computational results based on *ab initio* calculations using Gaussian with a 6-31G basis set and B3LYP density functional), provided the approach trajectory is appropriate. Subsequent removal of the mechanosynthetic tool tip using an appropriate withdrawal trajectory (e.g., including a 90° rotation of the tool to break the π bond of the double bond) is predicted to leave a single carbon atom in the bridged position on the dimer. Classical molecular dynamics simulations by Dzegilenko et al. [441] showed that a single weakly bonded carbon dimer could be selectively removed from the upper terrace of a reconstructed diamond C(100)–(2 × 1) surface by a carbon nanotube tip chemically modified with a C_2 carbene radical species strongly bonded to the end cap of the tip, but that this tool was not particularly useful for adding a carbon to the diamond surface. The carbene prefers to bond to a single radical (top) site on the C(111) diamond surface rather than at a bridged or fourfold hollow site, in agreement with the results of Walch and Merkle [445]. When planar C_6H_2 (methenylidene cyclopentene) is brought up to the C(100) surface, either a C_3H moiety with two lower C atoms of the tip initially deposited onto the fourfold locations forming bonds with C atoms of two neighboring surface dimers is attached, or else a C_4H_2 fragment is adsorbed atop two C atoms of neighboring surface dimers, with the reaction outcome depending critically on the initial tip–surface distances, the tip trajectory, and various allowed but undesired tip rearrangements [441]. Additional theoretical work by Musgrave's group [446] investigating the effects of highly localized electrical fields (such as might be generated by an STM tip) on the growth of diamond using tert-butyl and $C_{13}H_{22}$ cluster models indicates that although the isolated CH_2 radical is unstable (having two dangling bonds that force a ring opening on the cluster), the ethynyl (C_2H) and methyl (CH_3) radicals appear most promising for direct addition reactions.

In 2003, Merkle and Freitas [442] proposed a new family of mechanosynthetic tools intended to be employed for the placement of two carbon atoms—a CC dimer—onto a growing diamond surface at a specific site. Their analysis used density-functional theory with Gaussian 98 to focus on specific group IV–substituted biadamantane tool tip structures and to evaluate their stability and the strength of the bond they make with the CC dimer.

These tools should be stable in vacuum and should be able to hold and position a CC dimer in a manner suitable for positionally controlled diamond mechanosynthesis at room temperature. The function of a dimer placement tool is to position the dimer, then bond the dimer to a precisely chosen location on a growing molecular structure, and finally to withdraw the tool, leaving the dimer behind on the growing structure. To achieve this, the dimer is required to be both bonded relatively weakly to the tool and highly strained, and thus highly reactive, so it will easily bond to the growing molecular structure to which it is added.

There is a large combinatorial space of possible tools that might satisfy both requirements. Merkle and Freitas attempted to satisfy the two basic requirements by bonding the dimer to two group IV supporting atoms: carbon, silicon, germanium, tin, or lead. This series of elements forms progressively weaker bonds to carbon, so the proposed tools will likewise be progressively more weakly bound to the carbon-carbon (CC) dimer. The supporting group IV atoms are part of two substituted adamantane ($C_{10}H_{16}$) frameworks that position and orient them. The two substituted adamantane frameworks are rotated and fused together to make a biadamantane [447] structure (Fig. 7), creating very high angle strain in the bonds between the two supporting atoms and the dimer. This molecule, a bi-silaadamantane dicarbon, is only the tip of a complete tool. In a complete mechanosynthetic apparatus, a somewhat larger version of this molecule would likely be required so that the active tip could be held and positioned via a rigid handle structure. Subsequent molecular dynamics simulations of these tool tips [444] indicated a ~20% deposition success rate using the Ge tool at room temperatures, which may be an acceptable level of success for early experimental proof-of-concept demonstrations.

Although pick-and-place of individual carbon atoms or carbon dimers has not yet been demonstrated experimentally using scanning probe microscope tips, in 1985 Becker and Golovchencko [448] used voltage pulses on an STM tip to extract a single germanium atom from the (111) surface of a sample. STM has been used to bind silicon atoms to the tip, first pulling the atoms off the surface of a Si(111) crystal face and then reinserting them back into the crystal [449, 450], and segments of individual dimer rows of silicon atoms have been extracted from the Si(100) face to create structures with atomically straight edges and lateral features that are only 1 dimer in width [451]. Other researchers have moved clusters and single atoms of silicon across a silicon surface at room temperature [450–455]; mechanosynthesis of the Si(111) lattice has been studied theoretically by Herman [456–458] and appears feasible. A near-contact atomic force microscope operated at low temperature has been used for the vertical manipulation of selected single silicon atoms from the Si(111)–(7 × 7) surface, demonstrating removal of a selected silicon atom from its equilibrium position at the surface without otherwise perturbing the (7 × 7) unit cell, as well as the deposition of a single atom on a created vacancy at the surface [459]. The authors note, "These manipulation processes are purely mechanical, since neither bias voltage nor voltage pulse is applied between probe and sample."

As for carbon atoms, a STM tip has manipulated individual C_{60} molecules along terraces on a grooved copper plate [460], and the pick-and-place of C_{60} on silicon [461] surfaces has been demonstrated experimentally using an SPM. Other researchers have reported positioning carbon monoxide molecules on platinum surfaces [462]. More recently, Ho and Lee [463] have

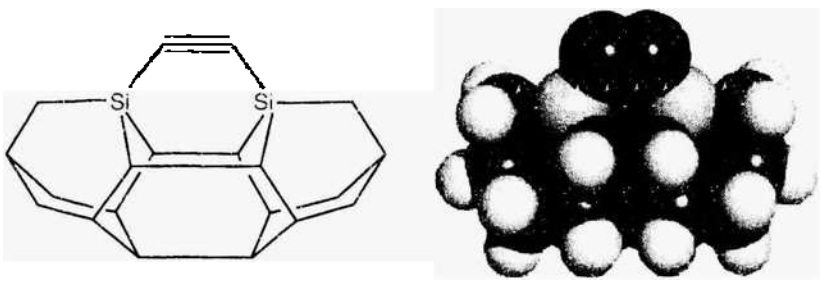


Figure 7. DCB6-Si dimer placement tool tip for diamond mechanosynthesis [442]. Reprinted with permission from [442], R. C. Merkle and R. A. Freitas, Jr., *J. Nanosci. Nanotechnol.* 3, 319 (2003). © 2003, American Scientific Publishers.

demonstrated the first repeatable site-specific mechanosynthetic covalent bonding operation of a diatomic carbon-containing molecule on a crystal surface, albeit electrically mediated. These researchers used an STM tip to locate two carbon monoxide molecules and one iron atom adsorbed on a silver surface in vacuum at 13 K. Next, they lowered the tip over one CO molecule and increased the voltage and current flow of the instrument to pick up the molecule; then they moved the tip-bound molecule over the surface-bound Fe atom and reversed the current flow, causing the CO molecule to covalently bond to the Fe atom, forming an iron carbonyl Fe(CO) molecule on the surface. Finally, the researchers repeated the procedure, returning to the exact site of the first Fe(CO) and adding a second CO molecule to the Fe(CO), forming a molecule of $Fe(CO)_2$, which in subsequent images of the surface appeared as a tiny "rabbit ears" structure, covalently bound to the silver surface [463].

These results, both experimental and theoretical, support the general feasibility of molecular positional operations that can modify a diamond workpiece, adding or removing small hydrocarbon clusters on that workpiece or even adding and removing single atoms or dimers under appropriate conditions. Repeated application of these basic operations should allow building up complex and atomically precise molecular structures. The mechanosynthetic strategy considered here is based on three principal assumptions:

1. Highly reactive tools: The reactions generally assumed to occur in the CVD synthesis of diamond involve highly reactive species. The molecular tools described here are likewise highly reactive, and for the same reason: synthesis of diamond structures is facilitated by the use of such tools.
2. Inert environment: Because the molecular tools can be highly reactive, they must be used in an inert environment. It is assumed that these tools will be used in vacuum, so that contact with solvent or gas molecules will not occur. More specifically, it is assumed that the mechanosynthetic work environment has no uncontrolled structures or molecules, and no free gas-phase molecules (e.g., contaminant molecules such as oxygen, nitrogen, hydrogen gas, or water, or undesired by-products of internal chemical reactions) that might react with the molecular tools or otherwise interfere with or disrupt mechanochemical operations. At the $\sim 10^{-12}$ atm pressure commonly achieved in laboratory UHV (ultra high vacuum) systems, a working volume smaller than ~4000 micron^3 would be more than 90% likely to contain no free gas-phase molecules.
3. Controlled trajectories: The molecular tools are positionally controlled at all times. They are not free to move at random. Their controlled trajectories are such that undesired encounters with the exterior surface of the tool tip holder or its housing, other molecular tools, or any other nonworkpiece structures are prevented. Undesired reactions between radicals and other physical structures is prevented either by keeping the radicals at a sufficient distance from surfaces with which they can react or by allowing radicals to contact with surfaces that are specifically designed to be inert and to resist attack by those radicals.

The assumption of positionally controlled highly reactive tools operating in an inert environment permits the use of novel and relatively simple reaction pathways. Although the ability to achieve an inert environment using present methods might lead to one possible implementation pathway, the primary purpose of this discussion has been to establish that, given a suitable vacuum environment, a relatively simple set of reactions and a relatively simple set of molecular tools should be sufficient to allow the manufacture of a wide range of nanoscale diamond structures with atomically precise features.

4.4.2. Massively Parallel Manufacturing

Complex objects assembled from simpler components may be manufactured either serially or in parallel. In serial assembly, objects are manufactured one at a time by a stepwise manufacturing process. Examples include handcrafted unique items such as an antique pocket watch, classical industrial "mass production" items such as automobiles that emerge only one by one at the end of an assembly line, or the traditional serial digital computer that executes instructions one by one in a linear sequence. In parallel assembly, objects are manufactured along many pathways simultaneously or at many different sites, such as polysomes

in living cells (multiple ribosomes translating a single mRNA strand simultaneously), mask lithography deposition of multiple circuits simultaneously on a single semiconductor wafer, or the modern parallel computer that at any moment is executing different instructions on thousands or even tens of thousands of independent processors in a highly parallel manner. Parallel manufacturing systems could have many possible control/configuration architectures, analogous to a SIMD (Single Instruction Multiple Data) approach; convergent [284, 464] or fractal [465, 466] assembly; agoric algorithms [467]; stigmergy [468], swarm [469, 470], or agent-based [467] approaches; or other manufacturing analogs taken from high-performance parallel computing.

Biology provides perhaps the best example of the power of massive parallelism in assembly. A single ribosome, able to make a single protein as directed by a single molecule of messenger RNA, is a marvelous manufacturing system. Yet, by itself, it would have little economic effect. However, billions of ribosomes operating together in each living cell can make all the proteins in a tree or—even more quickly—all the proteins in, for example, a rapidly growing kelp plant, which can literally grow 6 inches per day.

The difference between serial and parallel processing is equally crucial in molecular manufacturing, where the basic parts are very small. If a typical molecularly precise simple component is 1 nm^3 in volume, then to manufacture a 1 cm^3 volume of molecularly precise product requires the assembly of 1000 billion billion (10^{21}) individual simple molecular components. With serial manufacturing, just one molecular component is handled at a time—even at a 1 GHz operating frequency it would take many thousands of years, which is clearly not economically viable. With parallel manufacturing, however, vast numbers of molecular components can be processed simultaneously, reducing batch processing times to days, hours, or even less. Massively parallel assembly is the key to the economic viability of molecular manufacturing.

There are two principal pathways for achieving massively parallel assembly of molecularly precise physical structures: self-assembly and positional assembly. In commercial chemical synthesis, self-assembly usually takes place in fluid phase among mole ($\sim 10^{23}$) quantities of reactant molecules, which interact to produce mole quantities of product molecules. In Seeman's experiments (Section 4.3.2), producing DNA-based structures, and in other related experiments, involving supramolecular or biomolecular self-assembly, the number of product objects produced per batch is vastly less than mole quantities but is still very large by conventional standards in macroscale manufacturing. The inherent parallelism of self-assembly is the main advantage of this pathway over positional assembly in manufacturing.

To overcome this advantage and reap the full benefits of flexibility, precision, and quality in 21st century molecular manufacturing using positional assembly—also known (in the context of molecular manufacturing) as machine-phase nanotechnology—new techniques for massively parallel positional assembly must be developed. At least two such techniques for performing position assembly have already been clearly identified: massively parallel manipulator arrays and self-replicating systems.

Massively parallel manipulator arrays would use a very large array of independently actuated manipulation devices (e.g., scanning probe tips, robot arms, etc.) to process a very large number of molecular precise components simultaneously to build a larger product object. To produce large numbers of nanoparts and nanoassemblies, massively parallel scanning probe microscopes (SPM) arrays [471, 472], and microscale SPMs [473–476] would be most convenient. Force-sensing devices such as piezoelectric, piezoresistive, and capacitive microcantilevers made it possible to construct microscale AFMs on chips without an external deflection sensor. For example, in 1995 Itoh and colleagues [477] at the University of Tokyo fabricated an experimental piezoelectric ZnO_2-on-SiO_2 microcantilever array of 10 tips on a single silicon chip. Each cantilever tip lay ~70 μ from its neighbor and measured 150 μ long, 50 μ wide, and 3.5 μ thick, or ~26,000 μ^3/device, and each of the devices could be operated independently in the z-axis (e.g., vertically) up to near their mechanical resonance frequencies of 145–147 KHz at an actuation sensitivity of ~20 nm/volt (e.g., 0.3-nm resolution at 125 KHz).

Parallel probe scanning and lithography has been achieved by Quate's group at Stanford University, which has progressed from simple piezoresistive microcantilever arrays with five

tips spaced 100 μ apart and 0.04-nm resolution at 1 KHz but only one z-axis actuator for the whole array [478], to arrays with integrated sensors and actuators that allow parallel imaging and lithography with feedback and independent control of each of up to 16 tips, with scanning speeds up to 3 mm/s using a piezoresistive sensor [479]. By 1998, Quate's group had demonstrated [480] arrays of 50–100 independently controllable AFM probe tips mounted in two-dimensional patterns with 60 KHz resonances, including a 10 × 10 cantilevered tip array fabricated in closely spaced rows using throughwafer interconnects on a single chip; this work continues [481–483].

MacDonald's group at the Cornell Nanofabrication Facility pursued similar goals. In 1991, the team fabricated their first submicron stylus, driven in the xy plane using interdigitating MEMS comb drives [484], including the first opposable tip pair. By 1993, they had produced a 25-tip array on one xyz actuator, and by 1995 a complete working micro-STM (including xy comb drives) measuring 200 μ on an edge, and a micro-AFM measuring 2 mm on an edge including a 1-mm-long cantilever with a 20-nm-diameter integrated tip on a 6-μ high by 1-μ-diameter support shaft [485]. MacDonald's group demonstrated tip arrays with 5-μ spacings, exploiting the same process used to make the working micro-STM. With the same technology tips or small arrays of tips could be spaced 25–50 μ apart and integrated with individual z-axis microactuators, so that one xy-axis manipulator could support many tips, with each tip having a separate z actuator. By 1997, the group [486] had built and tested an array of micro-STMs on the surface of an ordinary silicon chip, with each tip on a cantilever 150 μ long with three-dimensional sensing and control. The largest prototype array has 144 probes, arranged in a square consisting of 12 rows of 12 probes each, with individual probe needles about 200 μ apart. Further development was to focus on increasing the range of movement and on fitting more and smaller probes into the same space.

Using conventional microlithography, researchers in the Millipede project at IBM's Zurich Research Laboratory [487] have fabricated scanning probe tip arrays of up to 1024 individual tips to achieve terabit-per-square-inch data storage densities [488]. Millipede project manager Peter Vettiger predicts low manufacturing costs and the ability to "build hundreds of these arrays on the same wafer" [489]. Simpler mechanical ciliary arrays consisting of 10,000 independent microactuators on a 1 cm^2 chip have been fabricated at the Cornell National Nanofabrication Laboratory for microscale parts transport applications [490], and other ciliary array systems for parts presentation in microscale manufacturing have been reported by the Fujita group [491]; Bohringer and colleagues [492]; Will's group [493]; Darling, Suh, and Kovacs [494]; and others. Microcantilever arrays for "electronic nose" [495] and related applications [496] have also been constructed, in at least one case having millions of interdigitated cantilevers on a single chip [497].

Active probe arrays for dip-pen nanolithography using DNA-based "ink" have been developed by Mirkin's group [498]. At a Materials Research Society meeting in late 2002, Mirkin reported that his group had constructed an array of 10,000 microscope tips, each capable of acting independently from the others: By using 10 tips in concert, they can draw essentially any desired shape. "The goal is to use dip pen nanolithography to generate [DNA-based] templates on surfaces that guide the assembly of nanoscale building blocks," Mirkin says. "It opens the door to placing electronic particles right where you want them. We think it's ultimately going to be a production tool.... This is not going to stop at 10,000. We can make arrays of arrays." The dip-pen lithography work of Li et al. [499] allows the direct on-surface fabrication of metal, semiconductor, and chemically diverse nanostructures made from both organic and inorganic materials.

Another approach is the use of independently mobile multiple manipulator platforms such as the NanoWalker system under investigation by Martel's group [500–503]. Martel envisions a fleet [504] of such wireless instrumented microrobots collectively forming a nanofactory system that could be used for positional nanomanufacturing operations.

Yet another alternative is Zyvex's patented Rotapod exponential assembly design concept [505], in which a single robotic arm on a wafer makes a second robotic arm on a facing surface by picking up micron-size lithographically produced parts, carefully laid out in advance in exactly the right locations so the tiny robotic arm can find them, and assembling them. The two robotic arms then make two more robotic arms, one on each of the two facing

surfaces. These four robotic arms, two on each surface, then make four more robotic arms. This process continues with the number of robotic arms steadily increasing in the pattern 1, 2, 4, 8, 16, 32, 64, and so forth until some manufacturing limit is reached (e.g., both surfaces are completely covered with tiny robotic arms). Thus, a single manipulator uses supplied parts to build a large manipulator array that can subsequently undertake the desired massively parallel manufacturing operations. However, the present Rotapod manipulator design is still under development, as it requires more precision to achieve flexible and molecularly precise fabrication. In 2001, Zyvex was awarded a $25 million, 5-year, National Institute of Standards and Technology (NIST) Advanced Technology Program government contract to develop prototype microscale assemblers using MEMS and nanoelectromechanical systems (NEMS) for prototype nanoscale assemblers [397].

Self-replicating systems would achieve massively parallel assembly first by fabricating copies of themselves, and then by allowing those copies to fabricate further copies, resulting in a rapid increase in the total number of systems. Once the population of replicated manipulator systems was deemed large enough, the manipulator population would be redirected to produce useful product objects, rather than more copies of itself. Self-replicating systems are widely found in natural biological systems but have not been pursued explicitly in macroscale manufacturing for at least two reasons: the widespread but erroneous perception of great technical difficulty, and the correct perception that such massive parallelism is unnecessary for traditional macroscale manufacturing. Nevertheless, ever since John von Neumann's theoretical studies of replicating systems in the 1940s and 1950s [506], and the well-known 1980 NASA engineering study of self-replicating lunar factories [507], manufacturing automation has been slowly progressing toward the goal of the fully self-replicating factory, including most notably Fujitsu Fanuc's nearly "unmanned" robot factory in Yamanashi Prefecture that uses robot arms to make robot arms. It is worth noting that self-replicating systems can be fully remote-controlled [508], fully autonomous [509], or various combinations in between [510, 511].

Note also that parallel manipulator arrays and individual replicators are both "self-replicating—the parallel arrays build more parallel arrays, after which the collective population of array manipulators builds useful products; similarly, the individual replicators build more individual replicators, then the collective population of replicators builds useful products. Each approach offers certain advantages and disadvantages—array systems can be more efficient due to positional specialization of workflow processing, while individual replicators are more tolerant of component failures—but both approaches exemplify "self-replication."

In the last few years there has been renewed research interest in the challenge of mechanical self-replicating systems [512], in part due to the realization that replication can be a fundamentally simple process. For example, Joseph Jacobson at the MIT Media Lab suggests that "to be useful for many applications, engineered systems must be able to manufacture multiple copies. Self-replicating systems may be useful in attaining that goal and it represents a new discipline in engineering." Today there are several ongoing university research programs, both theoretical and experimental, on mechanical (nonbiological) self-replicating machines [512]. The biotechnology and molecular engineering communities are just beginning to seriously study mechanical replicators operating in the nanoscale size domain. Current methods of self-assembly, although allowing massively parallel assembly, lack the flexibility, precision, and quality that are needed for twenty-first-century molecular manufacturing. Current methods of positional assembly, including massively parallel manipulator arrays and self-replicating systems, should allow molecularly precise massively parallel assembly, although further theoretical and experimental work will be required to fully realize this capability.

Note that for the "foreseeable future," it is likely that on-board storage of information will not be required by nanomechanical replicators. One example of an inherently safer and more flexible approach is the broadcast architecture [201, Section 16.3.2(a)] [513]. In this approach, information is broadcast by any of several means to the replicating component. The replicator can be built with an internal "dead man switch" that is automatically off unless activated by an encrypted broadcast signal [514]. The physical replicator becomes, in essence, a remote-controlled manipulator receiving instructions from the outside that guide

it, step by step, in assembling a second remote-controlled manipulator. After some number of repeat cycles, the result is a large number of identical remote-controlled manipulators. These manipulators can then be used to assemble large numbers of useful product objects by altering the stream of instructions sent to the population of replicated manipulator devices. Acoustic broadcast (mentioned earlier) can be used to combine both power and information transmission in one convenient mechanism.

Conceptual systems designs for molecular manufacturing have been offered by Drexler [201, 515, 516], Freitas [512], Hall [517], Merkle [208, 431, 464, 512, 513, 518, 519], and Phoenix [520], as extensively reviewed in Freitas and Merkle [512]. Though none of these conceptual designs has yet been subjected to rigorous computational simulation, as proposed by Goddard [218], several components such as gears and pumps have been simulated (Section 4.2) and have been found to perform largely as anticipated, though further design improvements are clearly needed.

4.5. Nanorobot Applications Designs and Scaling Studies

The idea of placing autonomous self-powered nanorobots inside of us might seem a bit odd, but actually the human body already teems with such nanodevices. For instance, more than 40 trillion single-celled microbes swim through our colon, outnumbering our tissue cells almost 10 to 1 [6]. Many bacteria move by whipping around a tiny tail, or flagellum, that is driven by a 30-nm biological ionic nanomotor powered by pH differences between the inside and the outside of the bacterial cell. Our bodies also maintain a population of more than a trillion motile biological nanodevices called fibroblasts and white cells such as neutrophils and lymphocytes, each measuring perhaps 10 μ in size [6]. These beneficial natural nanorobots are constantly crawling around inside of us, repairing damaged tissues, attacking invading microbes, and gathering up foreign particles and transporting them to various organs for disposal from the body [8].

There are ongoing attempts to build MEMS-based microrobots intended for *in vivo* use. For example, the "MR-Sub" project of the NanoRobotics Laboratory of Ecole Polytechnique in Montreal will use a MRI system as a means of propulsion for a microrobot in the blood vessels [521–523]. In this approach, a variable MRI magnetic field would generate a magnetic force on a robot containing ferromagnetic particles, providing a miniaturized system of propulsion able to develop sufficient power to direct a small device through the human body. Applications of the first-generation prototype might include targeted drug release, the reopening of blocked arteries, or taking biopsies. The project is currently gathering necessary information to define design rules for this type of microrobot, with a long-term goal "to further miniaturize the system and to create a robot made up of nanometric parts," making it "possible to carry out medical applications in the blood vessels which are still inaccessible." Other approaches to MEMS-based microrobots intended for *in vivo* use have been described in the literature [524, 525], including the magnetically controlled "cytobots" and "karyobots" proposed by Chrusch et al. [526] for performing wireless intracellular surgery.

There are even preliminary proposals for hybrid bionanorobots that could be constructed using currently foreseeable technologies. For example, Carlo Montemagno [527, 528] plans to use his modified ATPase motors (Section 4.2.2) to create a nanorobot that acts as a "pharmacy in a cell" by entering a cell, grabbing proteins produced by the cell that will not be used, and storing them until they are needed later by the patient. The device would consist of a tiny nickel drum, attached to the ATP-powered biological motor, that is coated with antibodies that adsorb the target molecules, whereupon an electric field pulls the molecules to a storage chamber and holds them in place.

The greatest power of nanomedicine will emerge in a decade or two, when we learn to design and construct complete artificial nanorobots using diamondoid nanometer-scale parts and subsystems including sensors, motors, manipulators, power plants, and molecular computers. If we make the reasonable assumption that we will someday be able to build these complex diamondoid medical nanorobots (Section 4.2), and to build them cheaply enough and in sufficiently large numbers to be useful therapeutically (Section 4.4), then what are the medical implications?

There are many possibilities [6–8, 529–535], but the development pathway will be long and arduous. First, theoretical scaling studies [530–535] are used to assess basic concept feasibility. These initial studies would then be followed by more detailed computational simulations of specific nanorobot components and assemblies, and ultimately full systems simulations, all thoroughly integrated with additional simulations of massively parallel manufacturing processes from start to finish, consistent with a design-for-assembly engineering philosophy. Once molecular manufacturing capabilities become available, experimental efforts may progress from component fabrication and testing to component assembly, and finally to prototypes and mass manufacture, ultimately leading to clinical trials. In 2004, progress in medical nanorobotics remains largely at the concept feasibility stage: Since 1998, the author has published four theoretical nanorobot scaling studies [530–535], summarized briefly below. Note that these studies are not intended to produce an actual engineering design for a future nanomedical product. Rather, the purpose is merely to examine a set of appropriate design constraints, scaling issues, and reference designs to assess whether or not the basic idea might be feasible, and to determine key limitations of such designs.

4.5.1. Respirocytes

The artificial mechanical red blood cell or "respirocyte" [530] is a bloodborne spherical 1-μ diamondoid 1000-atm pressure vessel (Fig. 8) with active pumping powered by endogenous serum glucose, able to deliver 236 times more oxygen to the tissues per unit volume than natural red cells and to manage carbonic acidity. The nanorobot is made of 18 billion atoms precisely arranged in a diamondoid pressure tank that can be pumped full of up to 3 billion oxygen (O_2) and carbon dioxide (CO_2) molecules. Later on, these gases can be released from the tank in a controlled manner, using the same molecular pumps. Respirocytes mimic the action of the natural hemoglobin-filled red blood cells. Gas concentration sensors on the outside of each device let the nanorobot know when it is time to load O_2 and unload CO_2 (at the lungs), or vice versa (at the tissues) (Fig. 9). An onboard nanocomputer and numerous chemical and pressure sensors enable complex device behaviors remotely reprogrammable by the physician via externally applied acoustic signals.

Each respirocyte can store and transport 236 times as much gas per unit volume as a natural red cell, so the injection of a 5-cc therapeutic dose of 50% respirocyte saline suspension, a total of 5 trillion individual nanorobots, into the human bloodstream can exactly replace the gas carrying capacity of the patient's entire 5.4 L of blood. If up to 1 L of respirocyte suspension could safely be added to the human bloodstream [8], this could keep a patient's tissues safely oxygenated for up to 4 h in the event a heart attack caused the heart to stop

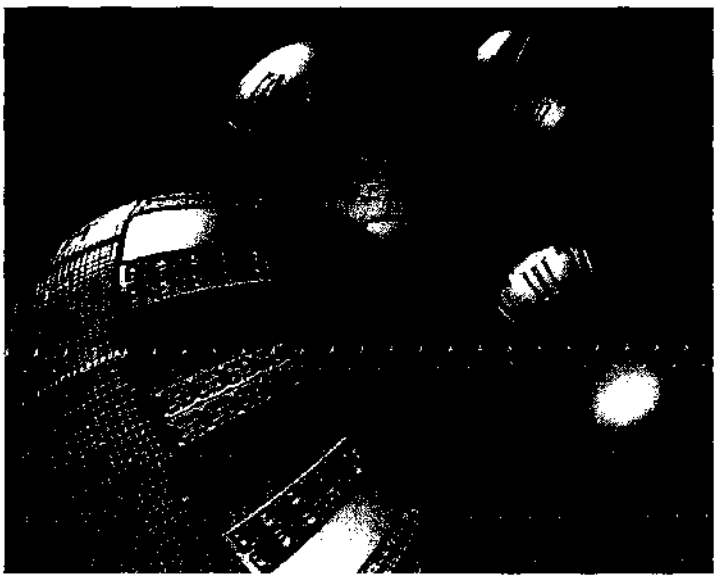


Figure 8. An artificial red cell: the respirocyte [530]. Reprinted with permission from [530], R. A. Freitas, Jr., *Artif. Cells, Blood Subst. Immobil. Biotech.* 26, 411 (1998), © 1998, Taylor and Francis.

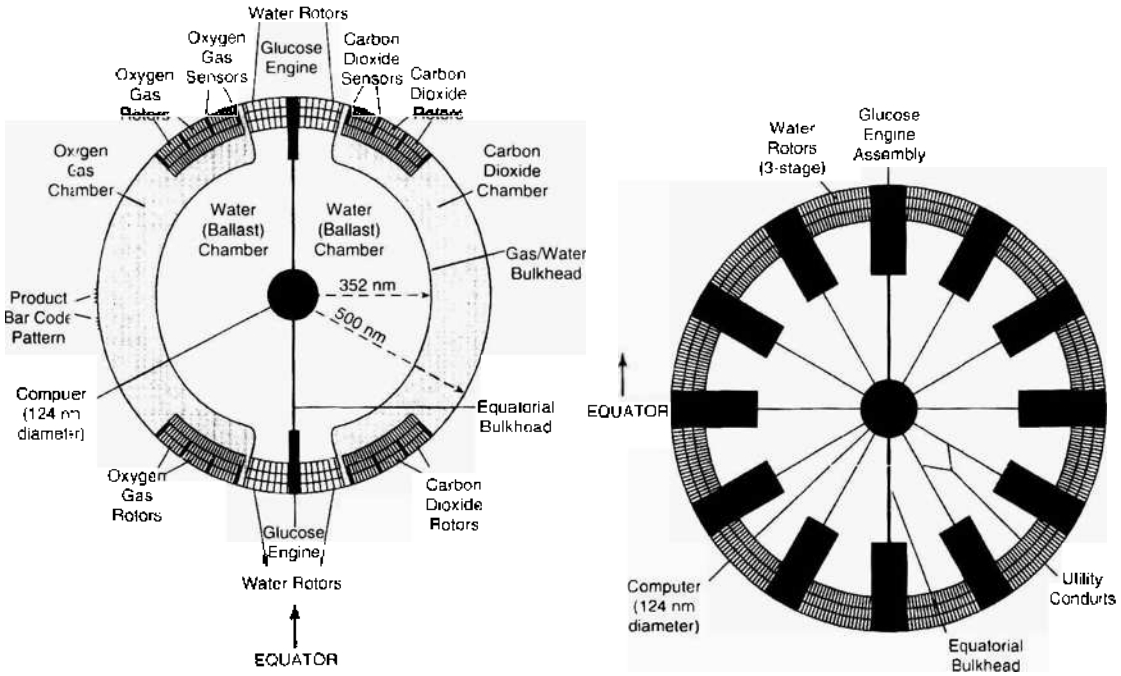


Figure 9. Internal cutaway view of respirocyte: equatorial (left) and polar (right) view [530]. Reprinted with permission from [530], R. A. Freitas, Jr., *Artif. Cells, Blood Subst. Immobil. Biotech.* 26, 411 (1998). © 1996, Taylor and Francis.

beating, or it would enable a healthy person to sit quietly at the bottom of a swimming pool for 4h, holding his breath, or to sprint at top speed for at least 15 min without breathing. Primary medical applications of respirocytes will include transfusable blood substitution; partial treatment for anemia, perinatal/neonatal, and lung disorders; enhancement of cardiovascular/neurovascular procedures, tumor therapies, and diagnostics; prevention of asphyxia; artificial breathing; and a variety of sports, veterinary, battlefield, and other uses.

4.5.2. Microbivores

An artificial mechanical white cell of microscopic size, called a "microbivore," has as its primary function destroying microbiological pathogens found in the human bloodstream, using a digest and discharge protocol [531, 532]. The microbivore is an oblate spheroidal nanomedical device (Fig. 10) measuring 3.4 μ in diameter along its major axis and 2.0 μ in diameter along its minor axis, consisting of 610 billion precisely arranged structural atoms in a gross geometric volume of 12.1 μ^3 and a dry mass of 12.2 pg. The device may consume up

Figure 10. An artificial white cell: the microbivore [531]. Designed by Robert A. Freitas Jr., illustrator Forrest Bishop. Reprinted with permission from [531], R. A. Freitas Jr., Zyvex preprint, 2001, http://www.rfreitas.com/Nano/Microbivores.htm, © 2001, Zyvex Corp.

to 200 pW of continuous power while completely digesting trapped microbes at a maximum throughput of 2 μ^3 of organic material per 30-s cycle, which is large enough to internalize a single microbe from virtually any major bacteremic species in a single gulp. The nanorobots would be ~80 times more efficient as phagocytic agents than macrophages in terms of volume/s digested per unit volume of phagocytic agent, and the nanorobots would have far larger maximum lifetime capacity for phagocytosis than natural white blood cells. Microbivores would fully eliminate septicemic infections in minutes to hours, whereas natural phagocytic defenses—even when aided by antibiotics—can often require weeks or months to achieve complete clearance of target bacteria from the bloodstream. Hence, microbivores appear to be up to ~1000 times faster-acting than either unaided natural or antibiotic-assisted biological phagocytic defenses and are able to extend the therapeutic competence of the physician to the entire range of potential bacterial threats, including locally dense infections.

During each cycle of nanorobot operation, the target bacterium is bound to the surface of the bloodborne microbivore like a fly on flypaper, via species-specific reversible binding sites [6]. Telescoping robotic grapples emerge from silos in the device surface, establish secure anchorage to the microbe's plasma membrane, and then transport the pathogen to the ingestion port at the front of the device, where the pathogen cell is internalized into a 2 μ^3 morcellation chamber. After sufficient mechanical mincing, the morcellated remains of the cell are pistoned into a separate 2 μ^3 digestion chamber, where a preprogrammed sequence of 40 engineered enzymes is successively injected and extracted six times, progressively reducing the morcellate ultimately to monoresidue amino acids, mononucleotides, glycerol, free fatty acids, and simple sugars. These simple molecules are then harmlessly discharged back into the bloodstream through an exhaust port at the rear of the device, completing the 30-s digestion cycle. This "digest and discharge" protocol [6] is conceptually similar to the internalization and digestion process practiced by natural phagocytes, except that the artificial process should be much faster and cleaner. For example, it is well known that macrophages release biologically active compounds during bacteriophagy [536], whereas well-designed microbivores need only release biologically inactive effluent.

4.5.3. Clottocytes

The artificial mechanical blood platelet or clottocyte [533] may allow complete hemostasis in as little as ~1 s, even in moderately large wounds. This response time is on the order of 100–1000 times faster than the natural system. The baseline clottocyte is conceived as a serum oxyglucose-powered spherical nanorobot ~2 μ in diameter (~4 μ^3 in volume) containing a compactly folded biodegradable fiber mesh. On command from its control computer, the device promptly unfurls its mesh packet in the immediate vicinity of an injured blood vessel following, say, a cut through the skin. Soluble thin films coating certain parts of the mesh dissolve on contact with plasma water, revealing sticky sections (e.g., sections complementary to blood group antigens unique to red cell surfaces) in desired patterns. Blood cells are immediately trapped in the overlapping artificial nettings released by multiple neighboring activated clottocytes, and bleeding halts at once.

Although up to 300 natural platelets might be broken and still be insufficient to initiate a self-perpetuating clotting cascade, even a single clottocyte, on reliably detecting a blood vessel break, can rapidly communicate this fact to its neighboring devices [6], immediately triggering a progressive, carefully controlled mesh-release cascade. Clottocytes may perform a clotting function that is equivalent in its essentials to that performed by biological platelets, but at only 0.01% of the bloodstream concentration of those cells or about 20 nanorobots per cubic millimeter of serum. Hence, clottocytes appear to be about 10,000 times more effective as clotting agents than an equal volume of natural platelets.

4.5.4. Vasculoid

Once a mature molecular nanotechnology becomes available, could blood be replaced with a single, complex robot? Purely as a design feasibility exercise, a robotic device that replaces and extends the entire human vascular system has been the subject of a preliminary scaling study [534, 535] and is properly called a "vasculoid," or vascular-like machine. Such a robot

would duplicate all essential thermal and biochemical transport functions of the blood, including circulation of respiratory gases, glucose, hormones, cytokines, waste products, and all necessary cellular components. The artificial vascular system would conform to the shape of existing blood vessels, replacing natural blood so thoroughly that the rest of the body would remain, at least physiochemically, essentially unaffected. The vasculoid appliance would be extremely complex, having ~500 trillion independent cooperating nanorobots. In simplest terms, the vasculoid is a watertight coating of nanomachinery distributed across the luminal surface of the entire human vascular tree. This nanomachinery constitutes a ~300 m^2 two-dimensional vascular-surface-conforming mosaic of conjoined square nanorobot plates, equipped with a mechanical ciliary array system that helps transport important nutrients and biological cells to the tissues in containerized form. Molecule-conveying docking bays make up 16% of all vasculoid plates, and nanotankers containing molecules for distribution can dock at these bays and load or unload their cargo. The entire appliance weighs ~2 kg and releases ~30 W of waste heat at a basal activity level and a maximum of ~200 W of power at peak activity level.

4.5.5. Chromosome Replacement and Antiaging Therapies

Medical nanorobots will also be able to intervene at the cellular level, performing *in vivo* cytosurgery. The most likely site of pathological function in the cell is the nucleus, and more specifically, the chromosomes. In one simple cytosurgical procedure called "chromosome replacement therapy," a nanorobot controlled by a physician would extract existing chromosomes from a diseased cell and insert new ones in their place [10, 529]. The replacement chromosomes will be manufactured to order, outside of the patient's body, in a laboratory benchtop production device that includes a molecular assembly line, using the patient's individual genome as the blueprint. The replacement chromosomes are appropriately demethylated, thus expressing only the appropriate exons that are active in the cell type to which the nanorobot has been targeted. If the patient chooses, inherited defective genes could be replaced with nondefective base-pair sequences, permanently curing a genetic disease. Given the speed with which nanorobots can be administered and their potential rapidity of action, it is possible that an entire whole-body procedure could be completed in 1 hr or less.

In the first half of the twenty-first century, nanomedicine should eliminate virtually all common diseases of the twentieth century, and virtually all medical pain and suffering as well. Because aging is believed to be the result of a number of interrelated molecular processes and malfunctions in cells, and because cellular malfunctions will be largely reversible, middle-aged and older people who gain access to an advanced nanomedicine can expect to have most of their youthful health and beauty restored. The end result of all these nanomedical advances will be to enable a process called "dechronification," which will first arrest biological aging, then reduce biological age by performing three kinds of procedures on each one of the 4 trillion tissue cells in the body.

First, a microbivore-class nanorobot device will be sent to enter every tissue cell, to remove accumulating metabolic toxins and undegradable material. Afterward, these toxins will continue to slowly reaccumulate, requiring a whole-body maintenance cleanout, perhaps annually, to reverse further aging.

Second, chromosome replacement therapy can be used to correct accumulated genetic damage and mutations in every one of the body's nucleus-bearing cells, with this therapy also perhaps repeated annually.

Third, persistent cellular structural damage that the cell cannot repair by itself such as enlarged or disabled mitochondria can be reversed as required, on a cell-by-cell basis, using cellular repair nanorobot devices.

By means of these annual checkups and cleanouts, and some occasional major repairs, biological age could be restored once a year to the more or less constant physiological age that the patient selects. Only conditions that involve a permanent loss of personality and memory information in the brain—such as an advanced case of Alzheimer's disease or a massive head trauma—may remain incurable in the nanomedical era.

It is a bright future that lies ahead for nanomedicine, but we shall all have to work very long and very hard to bring it to fruition.

ACKNOWLEDGMENT

I thank Zyvex Corporation for its financial support.

REFERENCES

1. U.S. House Testimony of Richard E. Smalley, 1999. http://www.house.gov/science/smalley_062299.htm.
2. National Nanotechnology Initiative: Research and Development FY, 2002. http://www.nano.gov 2002budget.html.
3. M. C. Roco, "National Nanotechnology Investment in the FY, 2004 Budget Request." AAAS Report XXVIII: Research & Development FY, 2004.
4. *Lancet* 362, 673 (2003).
5. "RB-162 Biomedical Applications of Nanoscale Devices," Business Communications Company, Inc., Norwalk, CT, 2003.
6. R. A. Freitas, Jr., "Nanomedicine: Basic Capabilities," Vol. I, Landes Bioscience, Georgetown, TX, 1999. http://www.nanomedicine.com.
7. R. A. Freitas, Jr., *J. Am. Dent. Assoc.* 131, 1559 (2000).
8. R. A. Freitas, Jr., "Nanomedicine: Biocompatibility," Vol. IIA, Landes Bioscience, Georgetown, TX, 2003. http://www.nanomedicine.com.
9. D. O. Weber, *Health Forum J.* 42, 32 (1999).
10. R. A. Freitas, Jr., *Stud. Health Technol. Inform.* 80, 45 (2002).
11. K. Bogunia-Kubik and M. Sugisaka, *Biosystems* 65, 123 (2002).
12. C. A. Haberzettl, *Nanotechnology* 13, R9 (2002).
13. D. F. Emerich and C. G. Thanos, *Expert Opin. Biol. Ther.* 3, 655 (2003).
14. K. K. Jain, *Expert Rev. Mol. Diagn.* 3, 153 (2003).
15. "NIH Roadmap: Nanomedicine," National Institutes of Health, 2003.
16. Alliance for Nanomedical Technologies. http://www.research.cornell.edu/anmt/.
17. F. A. Fiedler and G. H. Reynolds, *S. Cal. Interdisc. Law J.* 3, 593 (1994).
18. J. Miller, *Columb. Sci. Technol. Law Rev.* 4 (2002–2003).
19. T. A. Desai, W. H. Chu, J. K. Tu, G. M. Beattie, A. Hayek, and M. Ferrari, *Biotechnol. Bioeng.* 57, 118 (1998).
20. L. Leoni and T. A. Desai, *IEEE Trans. Biomed. Eng.* 48, 1335 (2001).
21. L. Leoni, A. Boiarski, and T. A. Desai, *Biomed. Microdev.* 4, 131 (2002).
22. S. L. Tao and T. A. Desai, *Adv. Drug Deliv. Rev.* 55, 315 (2003).
23. A. Ahmed, C. Bonner, and T. A. Desai, *J. Control Release* 81, 291 (2002).
24. L. Leoni, D. Attiah, and T. A. Desai, *Sensors* 2, 111 (2002).
25. Neurotech (Paris). http://www.neurotech.fr/product_tech.asp.
26. R. P. Lanza, J. L. Hayes, and W. L. Chick, *Nat. Biotechnol.* 14, 1107 (1996).
27. S. B. Lee and C. R. Martin, *J. Am. Chem. Soc.* 124, 11850 (2002).
28. M. Nishizawa, V. P. Menon, and C. R. Martin, *Science* 268, 700 (1995).
29. C. R. Martin and P. Kohli, *Nat. Rev. Drug Discov.* 2, 29 (2003).
30. H. Bayley and C. R. Martin, *Chem. Rev.* 100, 2575 (2000).
31. E. D. Steinle, D. T. Mitchell, M. Wirtz, S. B. Lee, V. Y. Young, and C. R. Martin, *Anal. Chem.* 74, 2416 (2002).
32. M. Wirtz, S. Yu, and C. R. Martin, *Analyst* 127, 871 (2002).
33. S. B. Lee, D. T. Mitchell, L. Trofin, T. K. Nevanen, H. Soderlund, and C. R. Martin, *Science* 296, 2198 (2002).
34. D. T. Mitchell, S. B. Lee, L. Trofin, N. Li, T. K. Nevanen, H. Soderlund, and C. R. Martin, *J. Am. Chem. Soc.* 124, 11864 (2002).
35. B. Cornell, V. Braach-Maksvytis, L. King, P. Osman, B. Raguse, L. Wieczorek, and R. Pace, *Nature* 387, 580 (1997).
36. Z. Siwy and A. Fulinski, *Phys. Rev. Lett.* 89, 198103 (2002).
37. J. Schmidt, "11th Foresight Conference on Molecular Nanotechnology," Palo Alto, CA, 2003. http://www.foresight.org/Conferences/MNT11/Abstracts/Schmidt/index.html.
38. L. A. Pozhar, E. P. Kontar, and M. Z. Hu, *J. Nanosci. Nanotechnol.* 2, 209 (2002).
39. S. K. Bhatia and D. Nicholson, *Phys. Rev. Lett.* 90, 016105 (2003).
40. A. Meller, L. Nivon, E. Brandin, J. Golovchenko, and D. Branton, *Proc. Natl. Acad. Sci.* 97, 1079 (2000).
41. A. Meller, L. Nivon, and D. Branton, *Phys. Rev. Lett.* 86, 3435 (2001).
42. A. Meller and D. Branton, *Electrophoresis* 23, 2583 (2002).
43. D. W. Deamer and D. Branton, *Acc. Chem. Res.* 35, 817 (2002).
44. D. Branton and A. Meller, in "Structure and Dynamics of Confined Polymers" (J. J. Kasianowicz, M. S. Z. Kellermayer, and D. W. Deamer, Eds.), p. 177, Kluwer Academic, Dordrecht, The Netherlands, 2002.
45. H. Wang and D. Branton, *Nat. Biotechnol.* 19, 622 (2001).
46. J. Li, M. Gershow, D. Stein, E. Brandin, and J. A. Golovchenko, *Nat. Mater.* 2, 611 (2003).
47. J. Li, D. Stein, C. McMullan, D. Branton, M. J. Aziz, and J. A. Golovchenko, *Nature* 412, 166 (2001).
48. D. Stein, J. Li, and J. A. Golovchenko, *Phys. Rev. Lett.* 89, 276106 (2002).
49. A. J. Storm, J. H. Chen, X. S. Ling, H. W. Zandbergen, and C. Dekker, *Nat. Mater.* 2, 537 (2003).
50. A. F. Sauer Budge, J. A. Nyamwanda, D. K. Lubensky, and D. Branton, *Phys. Rev. Lett.* 90, 238101 (2003).
51. D. W. Deamer and M. Akeson, *Trends Biotechnol.* 18, 147 (2000).

52. H. Shi and B. D. Ratner. *J. Biomed. Mater. Res.* 49, 1 (2000).
53. C. J. Allender, C. Richardson, B. Woodhouse, C. M. Heard, and K. R. Brain. *Int. J. Pharm.* 195, 39 (2000).
54. A. Bossi, S. A. Piletsky, E. V. Piletska, P. G. Righetti, and A. P. Turner. *Anal. Chem.* 73, 5281 (2001).
55. S. C. Zimmerman, M. S. Wendland, N. A. Rakow, I. Zharov, and K. S. Suslick. *Nature* 418, 399 (2002).
56. M. Yoshikawa. *Bioseparation* 10, 277 (2001).
57. O. Bruggemann. *Adv. Biochem. Eng. Biotechnol.* 76, 127 (2002).
58. Quantum Dot Corp., http: www.qdots.com .
59. X. Wu, H. Liu, J. Liu, K. N. Haley, J. A. Treadway, J. P. Larson, N. Ge, F. Peale, and M. P. Bruchez. *Nat. Biotechnol.* 21, 41 (2003).
60. X. Gao and S. Nie. *Trends Biotechnol.* 21, 371 (2003).
61. Quantum Dot Corporation, Matsushita (Panasonic), and Sc Biosciences Partner to Set the Standard for Biological Detection Using Quantum Dots. 2003. http://www.qdots.com/live/render/content.asp?id=15&press_id=13.
62. J. L. Medintz, A. R. Clapp, H. Mattoussi, E. R. Goldman, B. Fisher, and J. M. Mauro. *Nat. Mater.* 2, 630 (2003).
63. M. Han, X. Gao, J. Z. Su, and S. Nie. *Nat. Biotechnol.* 19, 631 (2001).
64. D. Gerion, W. J. Parak, S. C. Williams, D. Zanchet, C. M. Micheel, and A. P. Alivisatos. *J. Am. Chem. Soc.* 124, 7070 (2002).
65. F. Tokumasu and J. Dvorak. *J. Microsc.* 211, 256 (2003).
66. K. Hanaki, A. Momo, T. Oku, A. Komoto, S. Maenosono, Y. Yamaguchi, and K. Yamamoto. *Biochem. Biophys. Res. Commun.* 302, 496 (2003).
67. B. Dubertret, P. Skourides, D. J. Norris, V. Noireaux, A. H. Brivanlou, and A. Libchaber. *Science* 298, 1759 (2002).
68. J. K. Jaiswal, H. Mattoussi, J. M. Mauro, and S. M. Simon. *Nat. Biotechnol.* 21, 47 (2003).
69. X. Gao, W. C. Chan, and S. Nie. *J. Biomed. Opt.* 7, 532 (2002).
70. H. Xu, M. Y. Sha, E. Y. Wong, J. Uphoff, Y. Xu, J. A. Treadway, A. Truong, E. O'Brien, S. Asquith, M. Stubbins, N. K. Spurr, E. H. Lai, and W. Mahoney. *Nucleic Acids Res.* 31, e43 (2003).
71. Y. T. Lim, S. Kim, A. Nakayama, N. E. Stott, M. G. Bawendi, and J. V. Frangioni. *Mol. Imaging* 2, 50 (2003).
72. W. C. Chan, D. J. Maxwell, X. Gao, R. E. Bailey, M. Han, and S. Nie. *Curr. Opin. Biotechnol.* 13, 40 (2002).
73. A. Watson, X. Wu, and M. Bruchez. *Biotechniques* 34, 296 (2003).
74. M. Lannoo, C. Delerue, G. Allan, and Y. M. Niquet. *Philos. Trans. Ser. A Math. Phys. Eng. Sci.* 361, 259 (2003).
75. C. S. Garoufalis, A. D. Zdetsis, and S. Grimme. *Phys. Rev. Lett.* 87, 276402 (2001).
76. A. J. Williamson, J. C. Grossman, R. Q. Hood, A. Puzder, and G. Galli. *Phys. Rev. Lett.* 89, 196803 (2002).
77. C. S. Lent, B. Isaksen, and M. Lieberman. *J. Am. Chem. Soc.* 125, 1056 (2003).
78. S. Ogut, R. Burdick, Y. Saad, and J. R. Chelikowsky. *Phys. Rev. Lett.* 90, 127401 (2003).
79. T. Pauncsku, T. Rajh, G. Wiederrecht, J. Maser, S. Vogt, N. Stojicevic, M. Protic, B. Lai, J. Oryhon, M. Thurnauer, and G. Woloschak. *Nat. Mater.* 2, 343 (2003).
80. "Nanotechnology May Help Overcome Current Limitations of Gene Therapy." Northwestern University Press, 2003. http://www.scienceblog.com/community/article1507.html.
81. R. F. Schinazi, R. Sijbesma, G. Srdanov, C. L. Hill, and F. Wudl. *Antimicrob. Agents Chemother.* 37, 1707 (1993).
82. G. L. Marcorin, T. Da Ros, S. Castellano, G. Stefancich, I. Bonin, S. Miertus, and M. Prato. *Org. Lett.* 2, 3955 (2000).
83. Z. Zhu, D. I. Schuster, and M. E. Tuckerman. *Biochemistry* 42, 1326 (2003).
84. N. Tsao, P. P. Kanakamma, T. Y. Luh, C. K. Chou, and H. Y. Lei. *Antimicrob. Agents Chemother.* 43, 2273 (1999).
85. N. Tsao, T. Y. Luh, C. K. Chou, J. J. Wu, Y. S. Lin, and H. Y. Lei. *Antimicrob. Agents Chemother.* 45, 1788 (2001).
86. S. Bosi, T. Da Ros, S. Castellano, E. Banfi, and M. Prato. *Bioorg. Med. Chem. Lett.* 10, 1043 (2000).
87. Y. Tabata, Y. Murakami, and Y. Ikada. *Jpn. J. Cancer Res.* 88, 1108 (1997).
88. Y. Tabata, Y. Murakami, and Y. Ikada. *Fullerene Sci. Technol.* 5, 989 (1997).
89. N. Miyata and T. Yamakoshi, in "Fullerenes: Recent Advances in the Chemistry and Physics of Fullerenes and Related Materials" (K. M. Kadish and R. S. Ruoff, Eds.), Vol. 5, p. 345. Electrochemical Society, Pennington, NJ, 1997.
90. L. L. Dugan, E. Lovett, S. Cuddihy, B. Ma, T.-S. Lin, and D. W. Choi, in "Fullerenes: Chemistry, Physics, and Technology" (K. M. Kadish and R. S. Ruoff, Eds.), p. 467. Wiley, New York, 2000.
91. C Sixty, Inc., http://www.csixty.com.
92. A. N. Kirschner, D. H. Santosa, and S. R. Wilson, in "Fullerenes: Functionalized Fullerenes" (N. Martin, M. Maggini, and D. M. Guldi, Eds.), Vol. 9, p. 250. Electrochemical Society, Pennington, NJ, 2000.
93. H. Cai, X. Cao, Y. Jiang, P. He, and Y. Fang. *Anal. Bioanal. Chem.* 375, 287 (2003).
94. R. J. Chen, S. Bangsaruntip, K. A. Drouvalakis, N. W. Kam, M. Shim, Y. Li, W. Kim, P. J. Utz, and H. Dai. *Proc. Natl. Acad. Sci.* 100, 4984 (2003).
95. K. Yamamoto, G. Shi, T. Zhou, F. Xu, J. Xu, T. Kato, J. Y. Jin, and L. Jin. *Analyst* 128, 249 (2003).
96. S. Sotiropoulou and N. A. Chaniotakis. *Anal. Bioanal. Chem.* 375, 103 (2003).
97. M. Meyyappan, H. Partridge, and T. R. Govindan. "Nano and BioTechnology Research at NASA Ames." 2002.
98. J. Wang and M. Musameh. *Anal. Chem.* 75, 2075 (2003).
99. J. L. West and N. J. Halas. *Curr. Opin. Biotechnol.* 11, 215 (2000).
100. S. R. Sershen, S. L. Westcott, N. J. Halas, and J. L. West. *J. Biomed. Mater. Res.* 51, 293 (2000).

101. "Halas Wins Innovator Award in Fight Against Breast Cancer." Rice University Press, 2003. http://www.eurekalert.org/pub_releases/2003-07/ru-hwi070903.php.
102. Nanospectra Biosciences, Inc., http://www.nanospectra.com.
103. "Rice Makes First Rapid, Sensitive Whole-Blood Immunoassay." Rice University Press, 2003. http://www.eurekalert.org/pub_releases/2003-07/ru-rmf072203.php.
104. L. R. Hirsch, J. B. Jackson, A. Lee, N. J. Halas, and J. L. West, *Anal. Chem.* 75, 2377 (2003).
105. Y. Sun and Y. Xia, *Anal. Chem.* 74, 5297 (2002).
106. Y. T. Lim, O. O. Park, and H. T. Jung, *J. Colloid Interface Sci.* 263, 449 (2003).
107. Triton BioSystems, Inc., http://www.tritonbiosystems.com.
108. M. Kelly, "Startups Seek Perfect Particles to Search and Destroy Cancer," Small Times, 2003.
109. J.-M. Nam, C. S. Thaxton, and C. A. Mirkin, *Science* 301, 1884 (2003).
110. C. A. Pennell and H. A. Erickson, *Immunol. Res.* 25, 177 (2002).
111. D. H. Robbins, I. Margulies, M. Stetler-Stevenson, and R. J. Kreitman, *Clin. Cancer Res.* 6, 693 (2000).
112. R. Schnell, P. Borchmann, J. O. Staak, J. Schindler, V. Ghetie, E. S. Vitetta, and A. Engert, *Ann. Oncol.* 14, 729 (2003).
113. A. K. Salem, P. C. Searson, and K. W. Leong, *Nat. Mater.* 2, 668 (2003).
114. K. L. Dixon, *Nucl. Med. Commun.* 24, 951 (2003).
115. N. Pandit-Taskar, P. A. Hamlin, S. Reyes, S. M. Larson, and C. R. Divgi, *Curr. Oncol. Rep.* 5, 364 (2003).
116. M. R. McDevitt, D. Ma, L. T. Lai, J. Simon, P. Borchardt, R. K. Frank, K. Wu, V. Pellegrini, M. J. Curcio, M. Miederer, N. H. Bander, and D. A. Scheinberg, *Science* 294, 1537 (2001).
117. P. E. Borchardt, R. R. Yuan, M. Miederer, M. R. McDevitt, and D. A. Scheinberg, *Cancer Res.* 63, 5084 (2003).
118. G. Xu and H. L. McLeod, *Clin. Cancer Res.* 7, 3314 (2001).
119. L. F. Tietze, T. Feuerstein, A. Fecher, F. Haunert, O. Panknin, U. Borchers, I. Schuberth, and F. Alves, *Angew. Chem. Int. Ed. Engl.* 41, 759 (2002).
120. L. F. Tietze and T. Feuerstein, *Aust. J. Chem.* 56, 841 (2003).
121. P. Ball, *Nat. Sci.* Update, 2002.
122. Y. Suzuki, M. Tanihara, Y. Nishimura, K. Suzuki, Y. Kakimaru, and Y. Shimizu, *J. Biomed. Mater. Res.* 42, 112 (1998).
123. M. Tanihara, Y. Suzuki, Y. Nishimura, K. Suzuki, Y. Kakimaru, and Y. Fukunishi, *J. Pharm. Sci.* 88, 510 (1999).
124. K. S. Soppimath, T. M. Aminabhavi, A. M. Dave, S. G. Kumbar, and W. E. Rudzinski, *Drug Dev. Ind. Pharm.* 28, 957 (2002).
125. S. Bruhim, D. Narinesingh, and A. Guiseppi-Elie, *Biosens. Bioelectron.* 17, 973 (2002).
126. H. A. Clark, R. Kopelman, R. Tjalkens, and M. A. Philbert, *Anal. Chem.* 71, 4837 (1999).
127. H. A. Clark, M. Hoyer, M. A. Philbert, and R. Kopelman, *Anal. Chem.* 71, 4831 (1999).
128. H. Xu, J. W. Aylott, and R. Kopelman, *Analyst* 127, 1471 (2002).
129. J. P. Sumner, J. W. Aylott, E. Monson, and R. Kopelman, *Analyst* 127, 11 (2002).
130. H. Xu, F. Yan, E. E. Monson, and R. Kopelman, *J. Biomed. Mater. Res.* 66A, 870 (2003).
131. R. Clandos, Small Times, 2001. http://www.smalltimes.com/document_display.cfm?document_id=2293.
132. Molecular Therapeutics, Inc., http://www.moleculartherapeutics.com/tech.html.
133. J. F. Kukowska-Latallo, A. U. Bielinska, J. Johnson, R. Spindler, D. A. Tomalia, and J. R. Baker, Jr., *Proc. Natl. Acad. Sci.* 93, 4897 (1996).
134. J. D. Eichman, A. U. Bielinska, J. F. Kukowska-Latallo, and J. R. Baker, Jr., *Pharm. Sci. Technol. Today* 3, 232 (2000).
135. I. Lee, B. D. Athey, A. W. Wetzel, A. Kar, W. Meixner, and J. R. Baker, Jr., *Macromolecules* 35, 4510 (2002).
136. A. K. Patri, I. J. Majoros, and J. R. Baker, Jr., *Curr. Opin. Chem. Biol.* 6, 466 (2002).
137. J. F. Kukowska-Latallo, E. Raczka, A. Quintana, C. Chen, M. Rymaszewski, and J. R. Baker, Jr., *Hum. Gene Ther.* 11, 1385 (2000).
138. H. Nakanishi, O. Mazda, E. Satoh, H. Asada, H. Morioka, T. Kishida, M. Nakao, Y. Mizutani, A. Kawauchi, M. Kita, J. Imanishi, and T. Miki, *Gene Ther.* 10, 434 (2003).
139. L. Vincent, J. Varet, J. Y. Pille, H. Bompais, P. Opolon, A. Maksimenko, C. Malvy, M. Mirshahi, H. Lu, J. P. Vannier, C. Soria, and H. Li, *Int. J. Cancer* 105, 419 (2003).
140. J. D. Reuter, A. Myc, M. M. Hayes, Z. Gan, R. Roy, D. Qin, R. Yin, L. T. Piehler, R. Esfand, D. A. Tomalia, and J. R. Baker, Jr., *Bioconjug. Chem.* 10, 271 (1999).
141. J. J. Landers, Z. Cao, I. Lee, L. T. Piehler, P. P. Myc, A. Myc, T. Hamouda, A. T. Galecki, and J. R. Baker, Jr., *J. Infect. Dis.* 186, 1222 (2002).
142. Standard Products, Dendritic NanoTechnologies, Inc., 2003. http://www.dnanotech.com/store.asp#standard.
143. Starpharma Announces Success of U.S. FDA Application for Human Trials of Anti-HIV Product (VivaGel™), Starpharma, 2003.
144. J. R. Baker, Jr., A. Quintana, L. Piehler, M. Banazak-Holl, D. Tomalia, and E. Raczka, *Biomed. Microdevices* 3, 61 (2001).
145. T. A. Betley, J. A. Hessler, A. Meeke, M. M. Banaszak Holl, B. Orr, S. Uppuluri, D. A. Tomalia, and J. R. Baker, Jr., *Langmuir* 18, 3127 (2002).
146. D. A. Tomalia, H. M. Brothers 2nd, L. T. Piehler, H. D. Durst, and D. R. Swanson, *Proc. Natl. Acad. Sci.* 99, 5081 (2002).

147. A. Quintana, E. Raczka, L. Piehler, I. Lee, A. Myc, I. Majoros, A. K. Patri, T. Thomas, J. Mule, and J. R. Baker, Jr., *Pharm. Res.* 19, 1310 (2002).
148. "Dendrimers, Tecto-Dendrimers, 3-D Dendrimers," Center for Biologic Nanotechnology, University of Michigan, 2003. http://nano.med.umich.edu/Dendrimers.html.
149. D. A. Tomalia, D. R. Swanson, and B. Huang, "11th Foresight Conference on Molecular Nanotechnology," Palo Alto, CA, 2003. http://www.foresight.org/Conferences/MNT11/Abstracts/Tomalia/index.html.
150. H. Sparks, SPACE.com, 2002. http://www.space.com/businesstechnology/technology/radiation_nanobots_020717.html.
151. K. Hamad-Schifferli, J. J. Schwartz, A. T. Santos, S. Zhang, and J. M. Jacobson, *Nature* 415, 152 (2002).
152. A. Stikeman, *Technol. Rev.* 105 (2002).
153. C. Choi, United Press International, 2002. http://www.mail-archive.com/ctrl@listserv.aol.com/msg85249.html.
154. Biomolecular Functions, EngeneOS Press, 2002. http://www.prnewswire.com/cgi-bin/stories.pl?ACCT=104&STORY=/www/story/01-08-2002/0001644348.
155. W. M. Pankau, S. Mönninghoff, and G. von Kiedrowski, "Thermostable and Monoconjugable Gold Cluster by Gripping with a Dodekadentate Thioether Ligand," U.S. Patent Pending, 2003.
156. W. M. Pankau, K. Verbist, and G. von Kiedrowski, *Chem. Commun.* 2001, 519 (2001).
157. R. F. Service, *Science* 298, 2322 (2002).
158. G. Timp, Nanotechnology Industry Workshop, Center for Nanoscale Science and Technology, University of Illinois, 2003. http://www.cnst.uiuc.edu/NanoWorkshop2003/Abstracts%20and%20Posters-May0903.pdf.
159. T. Shimoboji, E. Larenas, T. Fowler, S. Kulkarni, A. S. Hoffman, and P. S. Stayton, *Proc. Natl. Acad. Sci.* 99, 16592 (2002).
160. T. Shimoboji, E. Larenas, T. Fowler, A. S. Hoffman, and P. S. Stayton, *Bioconjug. Chem.* 14, 517 (2003).
161. R. Kontermann and S. Dubel, Eds., "Antibody Engineering," Springer, New York, 2001.
162. A. Atala and R. P. Lanza, Eds., "Methods of Tissue Engineering," Academic Press, New York, 2001.
163. J. Sambrook and D. W. Russell, Eds., "Molecular Cloning: A Laboratory Manual," 3rd edn. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 2001.
164. I. S. Kohane, A. Kho, and A. J. Butte, "Microarrays for an Integrative Genomics," MIT Press, Cambridge, MA, 2002.
165. G. Stock, "Redesigning Humans: Our Inevitable Genetic Future," Houghton Mifflin, Boston, MA, 2002.
166. Y. M. Elcin, Ed., "Tissue Engineering, Stem Cells and Gene Therapies," Plenum Press, New York, 2003.
167. B. R. Glick and J. J. Pasternak, "Molecular Biotechnology: Principles and Applications of Recombinant DNA," 3rd edn. American Society for Microbiology, Washington, DC, 2003.
168. F. d'Herelle, "The Bacteriophage: Its Role in Immunity," Williams & Wilkins, Baltimore, MD, 1922.
169. R. J. Payne, D. Phil, and V. A. Jansen, *Clin. Pharmacol. Ther.* 68, 225 (2000).
170. C. A. MacHida, Ed., "Viral Vectors for Gene Therapy: Methods and Protocols," Humana Press, Totowa, NJ, 2002.
171. D. E. Post, F. R. Khuri, J. W. Simons, and E. G. Van Meir, *Hum. Gene Ther.* 14, 933 (2003).
172. S. El Mehdaoui, A. Touze, S. Laurent, P. Y. Sizaret, D. Rasschaert, and P. Coursaget, *J. Virol.* 74, 10332 (2000).
173. B. Dietzschold and M. J. Schnell, *Hybrid Hybridomics* 21, 129 (2002).
174. A. Kanerva, K. R. Zinn, T. R. Chaudhuri, J. T. Lam, K. Suzuki, T. G. Uil, T. Hakkarainen, G. J. Bauerschmitz, M. Wang, B. Liu, Z. Cao, R. D. Alvarez, D. T. Curiel, and A. Hemminki, *Mol. Ther.* 8, 449 (2003).
175. D. Endy, L. You, J. Yin, and I. J. Molineux, *Proc. Natl. Acad. Sci.* 97, 5375 (2000).
176. D. Endy and R. Brent, *Nature* 409, 391 (2001).
177. A. P. Arkin, *Curr. Opin. Biotechnol.* 12, 638 (2001).
178. F. W. Studier, *Virology* 95, 70 (1979).
179. J. Yin, *J. Bacteriol.* 175, 1272 (1993).
180. D. Kong and J. Yin, *Biotechnology (NY)* 13, 583 (1995).
181. "Synthetic Biology: Current Projects," 2003. http://syntheticbiology.org/projects/.
182. J. Cello, A. V. Paul, and E. Wimmer, *Science* 297, 1016 (2002).
183. D. Enshell-Seijffers, L. Smelyanski, and J. M. Gershoni, *Nucleic Acids Res.* 29, 1 (2001).
184. Y. G. Kim, M. Muralinath, T. Brandt, M. Pearcy, K. Hauns, K. Lowenhaupt, B. L. Jacobs, and A. Rich, *Proc. Natl. Acad. Sci.* 100, 6974 (2003).
185. K. J. Lee and J. C. de la Torre, *Curr. Top. Microbiol. Immunol.* 262, 175 (2002).
186. D. F. Hudson, C. Morrison, S. Ruchaud, and W. C. Earnshaw, *Trends Cell Biol.* 12, 281 (2002).
187. E. Gillitzer, D. Willits, M. Young, and T. Douglas, *Chem. Commun. (Camb.)* 20, 2390 (2002).
188. M. T. Klem, D. Willits, M. Young, and T. Douglas, *J. Am. Chem. Soc.* 125, 10806 (2003).
189. A. M. Rosenthal, "Viral Workhorses," Scientific American online, 2002.
190. T. Douglas and M. Young, *Nature* 393, 152 (1998).
191. T. Yamada, Y. Iwasaki, H. Tada, H. Iwabuki, M. K. L. Chuah, T. V. Driessche, H. Fukuda, A. Kondo, M. Ueda, M. Seno, K. Tanizawa, and S. Kuroda, *Nat. Biotechnol.* 21, 885 (2003).
192. D. Bermudes, B. Low, and J. Pawelek, *Adv. Exp. Med. Biol.* 465, 57 (2000).
193. K. Hopkin, *Sci. Am. Presents Extreme Eng.* 10, 78 (1999).
194. A. R. Mushegian, *Curr. Opin. Genet. Dev.* 9, 709 (1999).
195. "Researchers Build Huge DNA Chains," BBC, 2000. http://www.crystalinks.com/biology5.1.html.
196. EngeneOS, Inc., http://www.engeneos.com/.
197. J. Giles, Washington Post, 2002, p. A01. http://www.washingtonpost.com/ac2/wp-dyn?pagename=article&node=&contentId=A17496-2002Nov20.

198. Egea Biosciences, Inc., http://www.egeabiosciences.com.
199. G. A. Evans, U. S. Patent No. 6,521,427, 2003.
200. A. M. Goho, *Technol. Rev.* 106, 50 (2003).
201. K. E. Drexler, "Nanosystems: Molecular Machinery, Manufacturing, and Computation," Wiley, New York, 1992.
202. R. P. Feynman, *Eng. Sci.* (*CalTech*) 23, 22 (1960).
203. K. E. Drexler, *Proc. Nat. Acad. Sci.* 78, 5275 (1981).
204. J. Cumings and A. Zettl, *Science* 289, 602 (2000).
205. T. Cagin, A. Jaramillo-Botero, G. Gao, and W. A. Goddard III, *Nanotechnology* 9, 143 (1998).
206. K. E. Drexler, in "Prospects in Nanotechnology: Toward Molecular Manufacturing" (M. Krummenacker and J. Lewis, Eds.), p. 1, Wiley, New York, 1995.
207. G. He, M. H. Muser, and M. O. Robbins, *Science* 284, 1650 (1999).
208. R. C. Merkle, in "Prospects in Nanotechnology: Toward Molecular Manufacturing" (M. Krummenacker and J. Lewis, Eds.), p. 23, Wiley, New York, 1995.
209. R. C. Haddon, *Science* 261, 1545 (1993).
210. G. B. Adams, O. F. Sankey, J. B. Page, M. O'Keeffe, and D. A. Drabold, *Science* 256, 1792 (1992).
211. G. Leach, *Nanotechnology* 7, 197 (1996).
212. L. Peng, Z. L. Zhang, Z. Q. Xue, Q. D. Wu, Z. N. Gu, and D. G. Pettifor, *Phys. Rev. Lett.* 85, 3249 (2000).
213. O. Gulseren, T. Yildirim, and S. Ciraci, *Phys. Rev. B* 65, 153405 (2002).
214. W. M. Lomer, *Phil. Mag.* 42, 1327 (1951).
215. W. A. Goddard III, "Computational Chemistry and Nanotechnology," Presentation at the Fourth Foresight Conference on Molecular Nanotechnology, 1995.
216. K. E. Drexler, "Directions in Nanotechnology," Presentation at the Fourth Foresight Conference on Molecular Nanotechnology, 1995.
217. K. E. Drexler and R. C. Merkle, http://www.imm.org/Parts/Parts1.html or http://www.imm.org/Images/pumpApartC.jpg.
218. W. A. Goddard III, T. Cagin, and S. P. Walch, "Atomistic Design and Simulations of Nanoscale Machines and Assembly," Computational Nanotechnology Project, Materials Simulation Center, California Institute of Technology, 1996.
219. P. Walsh, A. Omelchenko, R. K. Kalia, A. Nakano, P. Vashishta, and S. Saini, *Appl. Phys. Lett.* 82, 118 (2003).
220. C. D. Montemagno and G. D. Bachand, *Nanotechnology* 10, 225 (1999).
221. G. D. Bachand and C. D. Montemagno, *Biomed. Microdevices* 2, 179 (2000).
222. R. K. Soong, G. D. Bachand, H. P. Neves, A. G. Olkhovets, H. G. Craighead, and C. D. Montemagno, *Science* 290, 1555 (2000).
223. H. Liu, J. J. Schmidt, G. D. Bachand, S. S. Rizk, L. L. Looger, H. W. Hellinga, and C. D. Montemagno, *Nat. Mater.* 1, 173 (2002).
224. J. J. Schmidt and C. D. Montemagno, *Drug Discovery Today* 7, 500 (2002).
225. Rutgers to Create Ultra-Tiny "Bio-Nano" Motors Under National Science Foundation Grant, Rutgers University, 2003. http://ur.rutgers.edu/medrel/viewArticle.html?ArticleID=3467.
226. T. R. Kelly, H. De Silva, and R. A. Silva, *Nature* 401, 150 (1999).
227. J. F. Stoddart, "11th Foresight Conference on Molecular Nanotechnology," Palo Alto, CA, 2003. http://www.foresight.org/Conferences/MNT11/Abstracts/Stoddart/index.html.
228. T. J. Huang, W. Lu, H.-R. Tseng, B. Brough, A. Flood, B.-D. Yu, J. F. Stoddart, and C.-M. Ho, "11th Foresight Conference on Molecular Nanotechnology," Palo Alto, CA, 2003. http://www.foresight.org/Conferences/MNT11/Abstracts/Huang/index.html.
229. N. Koumura, R. W. Zijlstra, R. A. van Delden, N. Harada, and B. L. Feringa, *Nature* 401, 152 (1999).
230. D. A. Leigh, J. K. Y. Wong, F. Dehez, and F. Zerbetto, *Nature* 424, 174 (2003).
231. S. Sasaki and I. Karube, *Trends Biotechnol.* 17, 50 (1999).
232. S. K. Chaudhuri and D. R. Lovley, *Nat. Biotechnol.* 21, 1229 (2003).
233. Zettl Research Group, http://www.physics.berkeley.edu/research/zettl/.
234. A. M. Fennimore, T. D. Yuzvinsky, Wei-Qiang Han, M. S. Fuhrer, J. Cumings, and A. Zettl, *Nature* 424, 408 (2003).
235. P. A. Williams, S. J. Papadakis, A. M. Patel, M. R. Falvo, S. Washburn, and R. Superfine, *Appl. Phys. Lett.* 82, 805 (2003).
236. S. Sapmaz, Y. M. Blanter, L. Gurevich, and H. S. J. van der Zant, *Phys. Rev. B* 67, 235414 (2003).
237. V. Kaajakari Abhijit and A. Lal, "11th International Conference on Solid-State Sensors and Actuators," Transducers '01 and Eurosensors XV, 2001. http://www.cae.wisc.edu/~sonicmem-research/actuation/transducers01_web.pdf
238. N. Shinde, R. Bashir, E. Groll, and G. Chiu, International Mechanical Engineering Congress and Exposition, American Society of Mechanical Engineers, Orlando, Florida, 2000.
239. P. Laitenberger, D. Pooley, and A. Stimpson, "Micro Electromechanical Systems: The Reality," The Generics Group Ltd., 2002.
240. "Capacitive Pressure Sensor Process," Zavracky Group, Microfabrication Laboratory at Northeastern University, http://www.ece.neu.edu/edsnu/zavracky/mfl/programs/soi/soi.html.
241. D. C. Abeysinghe, S. Dasgupta, J. T. Boyd, and H. E. Jackson, *Bull. Am. Phys. Soc.* 46 (2001).
242. V. Kaajakari and A. Lal, "10th International Conference on Solid-State Sensors and Actuators," 1999, p. 312.
243. D. L. Miller, *J. Acoust. Soc. Am.* 104, 2498 (1998).

244. S. A. Chesnokov, V. A. Nalimova, A. G. Rinzler, R. E. Smalley, and J. E. Fischer, *Phys. Rev. Lett.* 82, 343 (1999).
245. P. F. Schewe, B. Stein, and J. Riordon, *Am. Inst. Phys. Bull. Phys.* 531 (2001).
246. M. A. Reed and J. M. Tour, *Sci. Am.* 282, 86 (2000).
247. C. P. Collier, G. Mattersteig, E. W. Wong, Y. Luo, K. Beverly, J. Sampaio, F. M. Raymo, J. F. Stoddart, and J. R. Heath, *Science* 289, 1172 (2000).
248. P. G. Collins, M. S. Arnold, and P. Avouris, *Science* 292, 706 (2001).
249. J. M. Tour, "Molecular Electronics: Commercial Insights, Chemistry, Devices, Architecture and Programming," World Scientific, Singapore, 2003.
250. M. A. Reed and T. Lee, Eds., "Molecular Nanoelectronics," American Scientific, Stevenson Ranch, CA, 2003.
251. L. F. Lindoy and I. M. Atkinson, "Self-Assembly in Supramolecular Systems," Royal Society of Chemistry, 2000. http://www.rsc.org/is/books/selfassembly.htm.
252. W. Muller, H. Ringsdorf, E. Rump, G. Wildburg, X. Zhang, L. Angermaier, W. Knoll, M. Liley, and J. Spinke, *Science* 262, 1706 (1993).
253. G. M. Whitesides, J. P. Mathias, and C. Seto, *Science* 254, 1312 (1991).
254. A. Ulman, "An Introduction to Ultrathin Organic Films, from Langmuir-Blodgett to Self-Assembly," Academic Press, Boston, 1991.
255. R. Maoz, S. Matlis, E. DiMasi, B. M. Ocko, and J. Sagiv, *Nature* 384, 150 (1996).
256. J. A. Zasadzinski, R. Viswanathan, L. Madsen, J. Garnaes, and D. K. Schwartz, *Science* 263, 1726 (1994).
257. P. A. Bachmann, P. Walde, P. L. Luisi, and J. Lang, *J. Am. Chem. Soc.* 113, 8204 (1991).
258. H. Noguchi and M. Takasu, *Phys. Rev. E* 64, 41913 (2001).
259. W. P. Kirk and M. A. Reed, Eds., "Nanostructures and Mesoscopic Systems," Academic Press, New York, 1992.
260. S. I. Stupp, V. LeBonheur, K. Walker, L. S. Li, K. E. Huggins, M. Keser, and A. Amstutz, *Science* 276, 384 (1997).
261. J. D. Watson, N. H. Hopkins, J. W. Roberts, J. A. Steitz, and A. M. Weiner, "Molecular Biology of the Gene," 4th edn. Benjamin/Cummings, Menlo Park, CA, 1987.
262. P. H. Monceviez, M. J. Jakiela, and K. T. Ulrich, in "Proceedings of ASME Third Conference on Flexible Assembly Systems" (A. H. Soni, Ed.), Vol. DE-33, p. 41. American Society of Mechanical Engineering, New York, 1991.
263. H. J. Yeh and J. S. Smith, *Sens. Mater.* 6, 319 (1994).
264. K. Bohringer, K. Goldberg, M. Cohn, R. Howe, and A. Pisano, "Proceedings of the IEEE International Conference on Robotics and Automation," 1997.
265. A. Terfort and G. M. Whitesides, *Adv. Mater.* 10, 470 (1998).
266. K. Saitou, *IEEE Transactions of Robotics and Automation* 15, 510 (1999).
267. S. Griffith, 6.978 Final Paper, 2002. http://www.swiss.ai.mit.edu/projects/amorphous/6.978/finalpapers/saul-final.pdf.
268. K. Saitou and M. J. Jakiela, *Artif. Life* 2, 129 (1995).
269. K. Saitou and M. J. Jakiela, *Artif. Life* 2, 377 (1995).
270. K. Saitou and M. Jakiela, *Complex Systems* 10, 391 (1996).
271. O. Cetin, K. Saitou, H. Nishigaki, S. Nishiwaki, T. Amago, and N. Kikuchi, "Proceedings of the ASME International Mechanical Engineering Congress and Exposition," New York, 2001. http://www-personal.engin.umich.edu/~kazu/papers/imece-01.pdf.
272. L. S. Penrose, *Sci. Amer.* 200, 105 (1959).
273. K. Hosokawa, I. Shimoyama, and H. Miura, *Artif. Life* 1, 413 (1994).
274. J. Liu, T. Lee, B. L. Walsh, R. P. Andres, D. B. Janes, M. R. Melloch, J. M. Woodall, and R. Reifenberger, *Appl. Phys. Lett.* 77, 373 (2000).
275. X. M. Yang, R. D. Peters, P. F. Nealey, H. H. Solak, and F. Cerrina, *Macromolecules* 33, 9575 (2000).
276. Y. F. Gao and Z. Suo, *J. Appl. Phys.* 93, 4276 (2003).
277. J. C. Huie, *Smart Mater. Struct.* 12, 264 (2003).
278. S. R. Whaley, D. S. English, E. L. Hu, P. F. Barbara, and A. M. Belcher, *Nature* 405, 665 (2000).
279. S. J. Park, A. A. Lazarides, C. A. Mirkin, and R. L. Letsinger, *Angew. Chem. Int. Ed. Engl.* 40, 2909 (2001).
280. T. D. Clark, R. Ferrigno, J. Tien, K. E. Paul, and G. M. Whitesides, *J. Am. Chem. Soc.* 124, 5419 (2002).
281. P. W. K. Rothemund and E. Winfree, "Proceedings of the 32nd Annual ACM Symposium on Theory of Computing," p. 459. ACM, New York, 2000.
282. J. Breivik, *Entropy* 3, 273 (2001).
283. D. H. Gracias, J. Tien, T. L. Breen, C. Hsu, and G. M. Whitesides, *Science* 289, 1170 (2000).
284. I. S. Choi, N. Bowden, and G. M. Whitesides, *Angew. Chem. Int. Ed. Engl.* 38, 3078 (1999).
285. T. D. Clark, M. Boncheva, J. M. German, M. Weck, and G. M. Whitesides, *J. Am. Chem. Soc.* 124, 18 (2002).
286. P. W. K. Rothemund, *Proc. Natl. Acad. Sci.* 97, 984 (2000).
287. J. Tien, T. L. Breen, and G. M. Whitesides, *J. Am. Chem. Soc.* 120, 12670 (1998).
288. N. Bowden, F. Arias, T. Deng, and G. M. Whitesides, *Langmuir* 17, 1757 (2001).
289. J. Tien, A. Terfort, and G. M. Whitesides, *Langmuir* 13, 5349 (1997).
290. T. L. Breen, J. Tien, S. R. J. Oliver, T. Hadzic, and G. M. Whitesides, *Science* 284, 948 (1999).
291. U. Srinivasan, D. Liepmann, and R. T. Howe, *J. Microelectromech. Syst.* 10, 17 (2001).
292. T. D. Clark, J. Tien, D. C. Duffy, K. E. Paul, and G. M. Whitesides, *J. Am. Chem. Soc.* 123, 7677 (2001).
293. D. H. Gracias, V. Kavthekar, J. C. Love, K. E. Paul, and G. M. Whitesides, *Adv. Mater.* 14, 235 (2002).

294. E. Smela, O. Inganas, and I. Lundstrom, *Science* 268, 1735 (1995).
295. S. T Brittain, O. J. A. Schueller, H. Wu, S. Whitesides, and G. M. Whitesides, *J. Phys. Chem. B* 105, 347 (2001).
296. R. S. Fearing, in "Precision Sensors, Actuators and Systems" (H. S. Tzou and T. Fukuda, Eds.), p. 39. Kluwer Academic, Boston, MA, 1992.
297. M. Weck, I. S. Choi, N. L. Jeon, and G. M. Whitesides, *J. Am. Chem. Soc.* 122, 3546 (2000).
298. E. A. Rietman, "Molecular Engineering of Nanosystems," AIP Series, Springer, New York, 2001.
299. T. Hogg, *Nanotechnology* 10, 300 (1999).
300. S. Rasmussen and D. Pasqulini, "Poster Presentation, Second Astrobiology Science Conference," NASA/AMES Research Center, 2002, http://www.astrobiology.com/asc2002/abstract.html?ascid=337.
301. B. Smith and M. Krummenacker, "Paper Presented at the Albany Conference on Biomolecular Motors and Nanomachines," 1997, http://www.wadsworth.org/albcon97/abstract/krummena.htm.
302. Molecubotics, Inc., http://www.molecubotics.com/info.html.
303. S. S. Smith, L. Niu, D. J. Baker, J. A. Wendel, S. E. Kane, and D. S. Joy, *Proc. Natl. Acad. Sci.* 94, 2162 (1997).
304. N. C. Seeman, *DNA Cell Biol.* 10, 475 (1991).
305. J. A. Wendel and S. S. Smith, *Nanotechnology* 9, 297 (1998).
306. C. M. Niemeyer, B. Ceyhan, S. Gao, L. Chi, S. Peschel, and U. Simon, *Colloid Polym. Sci.* 279, 68 (2001).
307. S. S. Smith, in "Encyclopedia of Nanoscience and Nanotechnology" (H. S. Nalwa, Ed.), Vol. 10, p. 1. American Scientific, Stevenson Ranch, CA, 2003.
308. G. D. Meredith, H. Wu, and N. Allbritton, "11th Foresight Conference on Molecular Nanotechnology," Palo Alto, CA, 2003, http://www.foresight.org/Conferences/MNT11/Abstracts/Meredith/index.html.
309. E. Braun, Y. Eichen, U. Silvan, and G. Ben-Yoseph, *Nature* 391, 775 (1998).
310. H. Yan, S. H. Park, G. Finkelstein, J. H. Reif, and T. H. LaBean, *Science* 301, 1882 (2003).
311. J. K. N. Mbindyo, B. D. Reiss, B. R. Martin, D. J. Dermody, M. J. Natan, and T. E. Mallouk, *Adv. Mater.* 13, 249 (2001).
312. Y. Huang, X. Duan, Y. Cui, L. J. Lauhon, K.-H. Kim, and C. Lieber, *Science* 294, 1313 (2001).
313. A. M. Cassell, W. A. Scrivens, and J. M. Tour, *Angew. Chem. Int. Ed. Engl.* 37, 1528 (1998).
314. M. S. Shchepinov, K. U. Mir, J. K. Elder, M. D. Frank-Kamenetskii, and E. M. Southern, *Nucleic Acids Res.* 27, 3035 (1999).
315. C. A. Mirkin, R. L. Letsinger, R. C. Mucic, and J. J. A. Storhoff, *Nature* 382, 607 (1996).
316. C. J. Loweth, W. B. Caldwell, X. G. Peng, A. P. Alivisatos, and P. G. Schultz, *Angew. Chem. Intl. Ed.* 38, 1808 (1999).
317. J. M. Perez, T. O'Loughin, F. J. Simeone, R. Weissleder, and L. Josephson, *J. Am. Chem. Soc.* 124, 2856 (2002).
318. K. Tanaka, A. Tengeiji, T. Kato, N. Toyama, and M. Shionoya, *Science* 299, 1212 (2003).
319. R. Bashir, *Superlattices Microstruct.* 29, 1 (2001).
320. C. Park, J. 1. Camphell, and W. A. Goddard III, *Proc. Natl. Acad. Sci.* 89, 9094 (1992).
321. K. E. Drexler, *Annu. Rev. Biophys. Biomol. Struct.* 23, 377 (1994).
322. N. C. Seeman, *Trends Biotechnol.* 17, 437 (1999).
323. J. Chen and N. C. Seeman, *Nature* 350, 631 (1991).
324. Y. Zhang and N. C. Seeman, *J. Am. Chem. Soc.* 116, 1661 (1994).
325. N. C. Seeman, *Nanotechnology* 9, 257 (1998).
326. C. Mao, W. Sun, Z. Shen, and N. C. Seeman, *Nature* 397, 144 (1999).
327. H. Yan, X. Zhang, Z. Shen, and N. C. Seeman, *Nature* 415, 62 (2002).
328. E. Winfree, F. Liu, L. A. Wenzler, and N. C. Seeman, *Nature* 394, 539 (1998).
329. A. Carbone and N. C. Seeman, *Proc. Natl. Acad. Sci.* 99, 12577 (2002).
330. S. Xiao, F. Liu, A. E. Rosen, J. F. Hainfeld, N. C. Seeman, K. Musier-Forsyth, and R. A. Kiehl, *J. Nanoparticle Res.* 4, 313 (2002).
331. K. M. Hansen, H. F. Ji, G. Wu, R. Datar, R. Cote, A. Majumdar, and T. Thundat, *Anal. Chem.* 73, 1567 (2001).
332. P. Alberti and J.-L. Mergny, *Proc. Natl. Acad. Sci.* 100, 1569 (2003).
333. J. J. Li and W. Tan, *Nano Letters* 2, 315 (2002).
334. D. Shu and P. Guo, *J. Biol. Chem.* 278, 7119 (2003).
335. B. Yurke, A. J. Turberfield, A. P. Mills, Jr., F. C. Simmel, and J. L. Neumann, *Nature* 406, 605 (2000).
336. A. J. Turberfield, J. C. Mitchell, B. Yurke, A. P. Mills, Jr., M. I. Blakey, and F. C. Simmel, *Phys. Rev. Lett.* 90, 118102 (2003).
337. R. C. Merkle, *Trends Biotechnol.* 17, 271 (1999).
338. L. Feng, S. H. Park, J. H. Reif, and H. Yan, *Angew. Chem. Intl. Ed.* 42, 4342 (2003).
339. R. T. Pomerantz, M. Anikin, J. Zlatanova, and W. T. McAllister, "11th Foresight Conference on Molecular Nanotechnology," Palo Alto, CA, 2003, http://www.foresight.org/Conferences/MNT11/Abstracts/Pomerantz/index.html.
340. C. A. Mirkin, *Inorg. Chem.* 39, 2258 (2000).
341. C. Mavroidis, M. L. Yarmush, A. Dubey, A. Thornton, K. Nikitczuk, S. Tomassone, F. Papadimitrakopoulos, and B. Yurke, "Protein Based Nano-Machines for Space Applications," Final Report, Phase I NIAC Grant, Bio-Nano Robotics Lab, Rutgers University, 2002.
342. L. H. Eckardt, K. Naumann, W. M. Pankau, M. Rein, M. Schweitzer, N. Windhab, and G. von Kiedrowski, *Nature* 420, 286 (2002).
343. J. H. Reif, "Eighth International Meeting on DNA Based Computers (DNA 8)," Sapporo, Japan, Lecture Notes in Computer Science, New York, 2002.

344. A. Holzenburg and N. S. Scrutton, "Subcellular Biochemistry," Vol. 35, Plenum, New York, 2000.
345. R. J. Bradbury, "Protein Based Assembly of Nanoscale Parts," Aeiveos Corp., 2001.
346. B. X. Chen, S. R. Wilson, M. Das, D. J. Coughlin, and B. F. Erlanger, *Proc. Natl. Acad. Sci.* 95, 10809 (1998).
347. D. Izhaky and J. Pecht, *Proc. Natl. Acad. Sci.* 95, 11509 (1998).
348. N. Kessler, D. Perl-Treves, L. Addadi, and M. Eisenstein, *Proteins* 34, 383 (1999).
349. D. B. Amabilino and J. F. Stoddart, *Chem. Rev.* 95, 2725 (1995).
350. P. Cramer, D. A. Bushnell, and R. D. Kornberg, *Science* 292, 1863 (2001).
351. W. K. Johnston, P. J. Unrau, M. S. Lawrence, M. E. Glasner, and D. P. Bartel, *Science* 292, 1319 (2001).
352. M. Schliwa and G. Woehlke, *Nature* 422, 759 (2003).
353. T. Scheibel, R. Parthasarathy, G. Sawicki, X.-M. Lin, H. Jaeger, and S. Lindquist, *Proc. Natl. Acad. Sci.* 100, 4527 (2003).
354. M. A. McNiven, H. Cao, K. R. Pitts, and Y. Yoon, *Trends Biochem. Sci.* 25, 115 (2000).
355. S. S. Smith, *Nano Lett.* 1, 51 (2001).
356. W. Kliche, S. Fujita-Becker, M. Kollmar, D. J. Manstein, and F. J. Kull, *EMBO J.* 20, 40 (2001).
357. J. Fritz, M. K. Baller, H. P. Lang, H. Rothuizen, P. Vettiger, E. Meyer, H.-J. Guntherodt, C. Gerber, and J. K. Gimzewski, *Science* 288, 316 (2000).
358. R. A. McMillan, C. D. Paavola, J. Howard, S. L. Chan, N. J. Zaluzec, and J. D. Trent, *Nat. Mater.* 1, 247 (2002).
359. N. Kessler, D. Perl-Treves, and L. Addadi, *FASEB J.* 10, 1435 (1996).
360. M. Kam, D. Perl-Treves, R. Sfez, and L. Addadi, *J. Mol. Recog.* 7, 257 (1994).
361. P. Hasselbacher and H. R. Schumacher, *Arthritis Rheum.* 21, 353 (1978).
362. S.-W. Lee, C. Mao, C. E. Flynn, and A. M. Belcher, *Science* 296, 892 (2002).
363. M. Gross, *Nature* 373, 105 (1995).
364. D. Perl-Treves, N. Kessler, D. Izhaky, and L. Addadi, *Chem. Biol.* 3, 567 (1996).
365. J. J. Mond, A. Lees, and C. M. Snapper, *Annu. Rev. Immunol.* 13, 655 (1995).
366. D. A. Walters, B. L. Smith, A. M. Belcher, G. T. Paloczi, G. D. Stucky, D. E. Morse, and P. K. Hansma, *Biophys. J.* 72, 1425 (1997).
367. J. B. Thompson, G. T. Paloczi, J. H. Kindt, M. Michenfelder, B. L. Smith, G. Stucky, D. E. Morse, and P. K. Hansma, *Biophys. J.* 79, 3307 (2000).
368. S. Brown, M. Sarikaya, and E. Johnson, *J. Mol. Biol.* 299, 725 (2000).
369. B. C. Braden, F. A. Goldbaum, B. X. Chen, A. N. Kirschner, S. R. Wilson, and B. F. Erlanger, *Proc. Natl. Acad. Sci.* 97, 12193 (2000).
370. R. Dagani, *Chem. Eng. News* 78, 39 (2000).
371. A. Skiebe, A. Hirsch, H. Klos, and B. Gotschy, *Chem. Phys. Lett.* 220, 138 (1994).
372. R. Subramanian, P. Boulas, M. N. Vijayshee, F. D'Souzaj, M. T. Jones, and K. M. Kadish, *J. Chem. Soc. Chem. Commun.* 1994, 1847 (1994).
373. V. Balzani, A. Credi, and M. Venturi, "Molecular Devices and Machines: A Journey into the Nanoworld," VCH Verlagsgesellschaft MBH, Germany, 2003.
374. T. Klaus, R. Joerger, E. Olsson, and C. G. Granqvist, *Proc. Natl. Acad. Sci.* 96, 13611 (1999).
375. O. W. Nadeau, D. W. Gump, G. M. Hendricks, and D. H. Meyer, *Med. Microbiol. Immunol. (Berl.)* 181, 145 (1992).
376. M. Kowshik, N. Deshmukh, W. Vogel, J. Urban, S. K. Kulkarni, and K. M. Paknikar, *Biotechnol. Bioeng.* 78, 583 (2002).
377. R. R. Naik, S. J. Stringer, G. Agarwal, S. E. Jones, and M. O. Stone, *Nat. Mater.* 1, 169 (2002).
378. A. Ahmad, S. Senapati, M. I. Khan, R. Kumar, R. Ramani, V. Srinivas, and M. Sastry, *Nanotechnology* 14, 824 (2003).
379. D. Schuler and R. B. Frankel, *Appl. Microbiol. Biotechnol.* 52, 464 (1999).
380. T. Kondo, M. Nojiri, Y. Hishikawa, F. Togawa, D. Romanovicz, and R. M. Brown, Jr., *Proc. Natl. Acad. Sci.* 99, 14008 (2002).
381. M. Eastwood, D. A. McGrouther, and R. A. Brown, *Proc. Inst. Mech. Eng.* [H] 212, 85 (1998).
382. M. Ptashne, "A Genetic Switch: Phage (Lambda) and Higher Organisms," Blackwell Science, Boston, 1996.
383. M. B. Elowitz and S. Leibler, *Nature* 403, 335 (2000).
384. T. S. Gardner, C. R. Cantor, and J. J. Collins, *Nature* 403, 339 (2000).
385. Y. Yokobayashi, R. Weiss, and F. H. Arnold, *Proc. Natl. Acad. Sci.* 99, 16587 (2002).
386. E. Klarreich, *Sci. News* 163 (2003).
387. D. Endy, T. Knight, R. Rettberg, and G. J. Sussman, MIT Synthetic Biology Working Group, 2003. http://syntheticbiology.org/iap/.
388. MIT Synthetic Biology Working Group, 2003, http://syntheticbiology.org/projects.html.
389. MIT Synthetic Biology Working Group, 2003, http://biobricks.ai.mit.edu/.
390. P. C. Wong, K.-K. Wong, and H. Foote, *Commun. ACM* 46, 95 (2003).
391. J.-W. Kim, A. Malshe, and S. Tung, Institute of Biological Engineering (IBE) Annual Meeting, Athens, GA, 2003.
392. H. Pearson, Nature Science Update, 2003, http://www.nature.com/nsu/030707/030707-9.html.
393. C. D. Montemagno, "11th Foresight Conference on Molecular Nanotechnology," Feynman Prize Award Lecture, Palo Alto, CA, 2003.
394. F. Sequeira and A. J. Copik, Paper Presented at 8th Foresight Conference on Molecular Nanotechnology, 2000, http://www.foresight.org/Conferences/MNT8/Abstracts/Sequeira/.

395. C. Mao, C. E. Flynn, A. Hayhurst, R. Sweeney, J. Qi, G. Georgiou, B. Iverson, and A. M. Belcher, *Proc. Natl. Acad. Sci.* 100, 6946 (2003).
396. Q. Wang, T. Lin, L. Tang, J. E. Johnson, and M. G. Finn, *Angew. Chem. Int. Ed.* 41, 459 (2002).
397. Zyvex Corp., http://www.zyvex.com.
398. Agilent Technologies; http://www.labs.agilent.com/news/event/demos.html#6; *Technol. Rev.* (2003). http://www.technologyreview.com/articles/prototype0603.asp.
399. S. Martel, "Introduction to the NanoWalker: A Miniature Autonomous Robot Capable of Various Tasks at the Molecular and Atomic Scales," MIT Bioinstrumentation Laboratory, 2000.
400. S. Martel, L. C. Olague, J. B. C. Ferrando, S. Riebel, T. Koker, J. Suurkivi, T. Fofonoff, M. Sherwood, R. Dyer, and I. Hunter, "Proceedings of the SPIE: Microrobotics and Microassembly," SPIE, Newton, MA, 2001, Vol. 4568, p. 231.
401. P. Kim and C. M. Lieber, *Science* 286, 2148 (1999).
402. P. Boggild, T. M. Hansen, C. Tanasa, and F. Grey, *Nanotechnology* 12, 331 (2001).
403. R. F. Service, *Science* 294, 1448 (2001).
404. S. Akita, Y. Nakayama, S. Mizooka, T. Okawa, Y. Miyatake, S. Yamanaka, M. Tsuji, and T. Nosaka, *Appl. Phys. Lett.* 79, 1691 (2001).
405. H. Watanabe, C. Manabe, T. Shigematsu, K. Shimotani, and M. Shimizu, *Appl. Phys. Lett.* 79, 2462 (200).
406. C. Blackledge, D. A. Engebretson, and J. D. McDonald, *Langmuir* 16, 8317 (2000).
407. D. J. Diaz, J. E. Hudson, G. D. Storrier, H. D. Abruna, N. Sundararajan, and C. K. Ober, *Langmuir* 17, 5932 (2001).
408. S.-W. Hla and K.-H. Rieder, *Superlattices Microstruct.* 31, 63 (2002).
409. S.-W. Hla and K.-H. Rieder, *Ann. Rev. Phys. Chem.* 54, 307 (2003).
410. B. Dischler and C. Wild, Eds., "Low-Pressure Synthetic Diamond: Manufacturing and Applications," Springer, New York, 1998.
411. E. Kohn, M. Adamschik, P. Schmid, S. Ertl, and A. Floter, *Diam. Rel. Mater.* 10, 1684 (2001).
412. A. R. Krauss, O. Auciello, D. M. Gruen, A. Jayatissa, A. Sumant, J. Tucket, D. C. Mancini, N. Moldovan, A. Erdemir, D. Ersoy, M. N. Gardos, H. G. Busmann, E. M. Meyer, and M. Q. Ding, *Diam. Rel. Mater.* 10, 1952 (2001).
413. M. Adamschik, M. Hinz, C. Maier, P. Schmid, H. Seliger, E. P. Hofer, and E. Kohn, *Diam. Rel. Mater.* 10, 722 (2001).
414. I. Dion, C. Baquey, and J. R. Monties, *Int. J. Artif. Organs* 16, 623 (1993).
415. J. Franks and D. Finch, in "Nanotechnology in Medicine and the Biosciences" (R. R. H. Coombs and E. W. Robinson, Eds.), p. 133, Gordon and Breach, Amsterdam, The Netherlands, 1996.
416. D. P. Dowling, P. V. Kola, K. Donnelly, and T. C. Kelly, *Diam. Rel. Mat.* 6, 390 (1997).
417. S. Ertl, M. Adamschik, P. Schmid, P. Gluche, A. Floter, and E. Kohn, *Diam. Rel. Mater.* 9, 970 (2000).
418. H. Okushi, *Diam. Rel. Mater.* 10, 281 (2001).
419. J. Davidson, W. Kang, K. Holmes, A. Wisitsora-At, P. Taylor, V. Pulugurta, R. Venkatasubramanian, and F. Wells, *Diam. Rel. Mater.* 10, 1736 (2001).
420. F. P. Bundy, *J. Chem. Phys.* 38, 631 (1963).
421. H. H. Farrell and M. Levinson, *Phys. Rev. B* 31, 3593 (1985).
422. C. B. Musgrave, 2003, http://web.archive.org/web/20020603131403/http://chemeng.stanford.edu/html musgrave.html.
423. G. Jenner, *J. Chem. Soc. Faraday Trans. I* 81, 2437 (1985).
424. R. C. DeVries, *Annu. Rev. Mater. Sci.* 17, 161 (1987).
425. J. A. Xu, H. K. Mao, and P. M. Bell, *Science* 232, 1404 (1986).
426. R. J. Hemley and H. K. Mao, *Science* 249, 391 (1990).
427. D. F. McMillen and D. M. Golden, *Ann. Rev. Phys. Chem.* 33, 493 (1982).
428. X. Y. Chang, M. Perry, J. Peploski, D. L. Thompson, and L. M. Raff, *J. Chem. Phys.* 99, 4748 (1993).
429. A. Ricca, C. W. Bauschlicher, Jr., J. K. Kang, and C. B. Musgrave, *Surf. Sci.* 429, 199 (1999).
430. M. Page and D. W. Brenner, *J. Am. Chem. Soc.* 113, 3270 (1991).
431. R. C. Merkle, *Nanotechnology* 8, 149 (1997).
432. S. B. Sinnott, R. J. Colton, C. T. White, and D. W. Brenner, *Surf. Sci.* 316, L1055 (1994).
433. D. W. Brenner, S. B. Sinnott, J. A. Harrison, and O. A. Shenderova, *Nanotechnology* 7, 161 (1996).
434. C. B. Musgrave, J. K. Perry, R. C. Merkle, and W. A. Goddard III, *Nanotechnology* 2, 187 (1991).
435. J. W. Lyding, K. Hess, G. C. Abeln, D. S. Thompson, J. S. Moore, M. C. Hersam, E. T. Foley, J. Lee, Z. Chen, S. T. Hwang, H. Choi, P. H. Avouris, and I. C. Kizilyalli, *Appl. Surf. Sci.* 130, 221 (1998).
436. E. T. Foley, A. F. Kam, J. W. Lyding, and P. H. Avouris, *Phys. Rev. Lett.* 80, 1336 (1998).
437. M. C. Hersam, G. C. Abeln, and J. W. Lyding, *Microelectronic Eng.* 47, 235 (1999).
438. L. J. Lauhon and W. Ho, *J. Phys. Chem.* 105, 3987 (2000).
439. B. J. McIntyre, M. Salmeron, and G. A. Somorjai, *Science* 265, 1415 (1994).
440. W. T. Muller, D. L. Klein, T. Lee, J. Clarke, P. L. McEuen, and P. G. Schultz, *Science* 268, 272 (1995).
441. F. N. Dzegilenko, D. Srivastava, and S. Saini, *Nanotechnology* 9, 325 (1998).
442. R. C. Merkle and R. A. Freitas, Jr., *J. Nanosci. Nanotechnol.* 3, 319 (2003).
443. J. Peng, R. A. Freitas, Jr., and R. C. Merkle, *J. Comp. Theor. Nanosci.* 1, 62 (2004).
444. D. J. Mann, J. Peng, R. A. Freitas, Jr., and R. C. Merkle, *J. Comp. Theor. Nanosci.* 1, 71 (2004).
445. S. P. Walch and R. C. Merkle, *Nanotechnology* 9, 285 (1998).
446. C. B. Musgrave, 2001, http://chemeng.stanford.edu/html/stm.html.

447. H. F. Reinhardt, *J. Org. Chem.* 27, 3258 (1962).
448. R. S. Becker, J. A. Golovchenko, and B. S. Swartzentruber, *Nature* 325, 419 (1987).
449. C. J. Chen, "Introduction to Scanning Tunneling Microscopy," Oxford University Press, Oxford, UK, 1993.
450. D. Huang, H. Uchida, and M. Aono, *J. Vac. Sci. Technol. B* 12, 2429 (1994).
451. C. T. Salling and M. G. Lagally, *Science* 265, 502 (1994).
452. I.-W. Lyo and P. Avouris, *Science* 253, 173 (1991).
453. M. Aono, A. Kobayashi, F. Grey, H. Uchida, and D. H. Huang, *J. Appl. Phys.* 32, 1470 (1993).
454. P. Avouris, *Acc. Chem. Res.* 28, 95 (1995).
455. G. Meyer and K. H. Rieder, *Surf. Sci.* 377–9, 1087 (1997).
456. A. Herman, *Nanotechnology* 8, 132 (1997).
457. A. Herman, *Modelling Simul. Mater. Sci. Eng.* 7, 43 (1999).
458. A. Herman, *TASK Q* 1, 9 (1997).
459. N. Oyabu, O. Custance, I. Yi, Y. Sugawara, and S. Morita, *Phys. Rev. Lett.* 90, 176102 (2003).
460. M. T. Cuberes, R. B. Schlittler, and James K. Gimzewski, *Appl. Phys. Lett.* 69, 3016 (1996).
461. P. Moriarty, Y. R. Ma, M. D. Upward, and P. H. Beton, *Surf. Sci.* 407, 27 (1998).
462. D. M. Eigler and E. K. Schweizer, *Nature* 344, 524 (1990).
463. W. Ho and H. Lee, *Science* 286, 1719 (1999).
464. R. C. Merkle, *Nanotechnology* 8, 18 (1997).
465. R. Gordon and B. D. Aguda, in "Proceedings of the Annual International Conference IEEE Engineering in Medicine and Biology" (G. Harris and C. Walker, Eds.), p. 273, Part 1/4: Cardiology and Imaging, IEEE, New York, 1988.
466. A. Carbone and N. C. Seeman, *Nat. Comp.* 1, 469 (2002).
467. T. Toth-Fejel, *Nanotechnology* 11, 133 (2000).
468. E. Bonabeau, S. Guerin, D. Snyers, P. Kuntz, and G. Theraulaz, *BioSystems* 56, 13 (2000).
469. G. Theraulaz and E. Bonabeau, *J. Theor. Biol.* 177, 381 (1995).
470. E. Bonabeau, M. Dorigo, and G. Theraulaz, "Swarm Intelligence: From Natural to Artificial Systems," Santa Fe Institute Sciences of Complexity Series, Oxford University Press, New York, 1999.
471. K. E. Drexler, *J. Vac. Sci. Technol. B* 9, 1394 (1991).
472. A. A. G. Requicha, *Intl. J. Robotics Res.* 18, 344 (1999).
473. T. R. Albrecht, S. Akamine, M. J. Zdeblick, and C. F. Quate, *J. Vac. Sci. Technol. A* 8, 317 (1990).
474. M. I. Lutwyche and Y. Wada, *Appl. Phys. Lett.* 66, 2807 (1995).
475. D. A. Walters, J. P. Cleveland, N. H. Thomson, P. K. Hansma, M. A. Wendman, G. Gurley, and V. Elings, *Rev. Sci. Inst.* 67, 3583 (1996).
476. D. Lange, M. Zimmermann, C. Hagleitner, O. Brand, and H. Baltes, *Tech. Digest* 2, 1074 (2001).
477. T. Itoh, T. Ohashi, and T. Suga, Ninth International Conference on MEMS, 1996, p. 451.
478. S. C. Minne, P. Flueckiger, H. T. Soh, and C. F. Quate, *J. Vac. Sci. Technol. B* 13, 1380 (1995).
479. S. R. Manalis, S. C. Minne, and C. F. Quate, *Appl. Phys. Lett.* 68, 871 (1996).
480. S. C. Minne, G. Yaralioglu, S. R. Manalis, J. D. Adams, J. Zesch, A. Atalar, and C. F. Quate, *Appl. Phys. Lett.* 72, 2340 (1998).
481. K. Wilder, H. T. Soh, A. Atalar, and C. F. Quate, *Rev. Sci. Instr.* 70, 2822 (1999).
482. E. M. Chow, H. T. Soh, H. C. Lee, J. D. Adams, S. C. Minne, G. Yaralioglu, A. Atalar, C. F. Quate, and T. W. Kenny, *Sensors Actuators A* 83, 118 (2000).
483. T. Sulchek, R. J. Grow, G. G. Yaralioglu, S. C. Minne, C. F. Quate, S. R. Manalis, A. Kiraz, A. Aydiner, and A. Atalar, *Appl. Phys. Lett.* 78, 1787 (2001).
484. J. J. Yao, S. C. Arney, and N. C. MacDonald, *J. Microelectromech. Syst.* 1, 14 (1992).
485. Y. Xu, N. C. MacDonald, and S. A. Miller, *Appl. Phys. Lett.* 67, 2305 (1995).
486. S. A. Miller, K. L. Turner, and N. C. MacDonald, *Rev. Sci. Instrum.* 68, 4155 (1997).
487. P. Vettiger, G. Cross, M. Despont, U. Drechsler, U. Duerig, B. Gotsmann, W. Haeberle, M. Lantz, H. Rothuizen, R. Stutz, and G. Binnig, *IEEE Trans. Nanotechnol.* 1, 39 (2002).
488. M. I. Lutwyche, M. Despont, U. Drechsler, U. Durig, W. Haberle, H. Rothuizen, R. Stutz, R. Widmer, G. K. Binnig, and P. Vettiger, *Appl. Phys. Lett.* 77, 3299 (2000).
489. J. W. Toigo, *Sci. Am.* 282, 58 (2000).
490. D. Reznik, S. Brown, and J. Canny, "Proceedings of the IEEE International Conference on Robotics and Automation," IEEE Robotics and Automation Society, 1997, p. 1675, http://www.cs.berkeley.edu/~jfc/dreznik/Pubs/icra97.pdf.
491. Y. Mita, A. Kaiser, B. Stefanelli, P. Garda, M. Milgram, and H. Fujita, "IEEE International Conference on Systems, Man and Cybernetics (SMC)," Tokyo, Japan, 1999. http://www.ee.t.u-tokyo.ac.jp/mita/SMC99/.
492. K. F. Bohringer and H. Choset, Eds., "Distributed Manipulation," Kluwer Academic, Norwell, MA, 2000.
493. M. G. Coutinho and P. M. Will, "Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)," Leuven, Belgium, 1998.
494. R. B. Darling, J. W. Suh, and G. T. A. Kovacs, *Vac. Sci. Technol. A* 16, 1998 (1998).
495. F. M. Battiston, J. P. Ramseyer, H. P. Lang, M. K. Baller, Ch. Gerber, J. K. Gimzewski, E. Meyer, and H. J. Guntherodt, *Sensors Actuators B* 77, 122 (2001).
496. P. Vettiger, M. Despont, U. Drechsler, U. Durig, W. Haberle, M. I. Lutwyche, H. E. Rothuizen, R. Stutz, R. Widmer, and G. K. Binnig, *IBM J. Res. Dev.* 44, 323 (2000).
497. "Microcantilever Arrays," Protiveris Corp. (2003).

498. D. Bullen, S. Chung, X. Wang, J. Zou, C. Liu, and C. Mirkin, "Symposium LL: Rapid Prototyping Technologies," Materials Research Society Fall Meeting, Proceedings of the MRS, Vol. 758, Boston, MA, 2002. http://mass.micro.uiuc.edu/publications/papers/84.pdf.
499. Y. Li, B. W. Maynor, and J. J. Liu, *J. Am. Chem. Soc.* 123, 2105 (2001).
500. S. Martel, S. Riebel, T. Koker, M. Sherwood, and I. Hunter, "Proceedings of the Eighth IEEE International Conference on Emerging Technologies and Factory Automation," Special Session on Microrobotics in Manufacturing, Nice, France, 2001. http://www.nano.polymtl.ca.
501. S. Martel and I. Hunter, "Proceedings of the Third International Workshop on Microfactories," Minneapolis, MN, 2002, p. 97. http://www.nano.polymtl.ca.
502. S. Martel, A. Schindler, G. Baumann, S. Riebel, and T. Boitani, IEEE Automation, Taiwan, 2003. http://www.nano.polymtl.ca.
503. S. Martel and I. Hunter, *J. Micromechatronics* 2, 201 (2002).
504. S. Martel, T. Koker, S. Riebel, M. Sherwood, J. Suurkivi, and I. Hunter, "Proceedings of the SPIE: Microrobotics and Microassembly," Newton, MA, 2001, Vol. 4568, p. 221. http://www.nano.polymtl.ca.
505. G. D. Skidmore, E. Parker, M. Ellis, N. Sarkar, and R. Merkle, *Nanotechnology* 12, 316 (2001).
506. J. von Neumann, "Theory of Self-Reproducing Automata" (A. W. Burks, Ed.), University of Illinois Press, Urbana, IL, 1966.
507. R. A. Freitas, Jr. and W. P. Gilbreath, Eds., "Advanced Automation for Space Missions," NASA Conference Publication CP-2255, N83-15348, U.S. GPO, Washington, DC, 1982.
508. G. S. Chirikjian and J. Suthakorn, "Proceedings of the Eighth International Symposium on Experimental Robotics (ISER '02)," Italy, 2002. http://www.iser02.unisa.it/papers/67.pdf.
509. J. Suthakorn, A. B. Cushing, and G. S. Chirikjian, "Proceedings of the IEEE/ASME International Conference on Advanced Intelligence Mechatronics (AIM)," Kobe, Japan, 2003. http://custer.me.jhu.edu/publication/pdf/auton.pdf.
510. J. Suthakorn, Ph.D. Dissertation, Johns Hopkins University, 2003.
511. J. Suthakorn, Y. T. Kwon, and G. S. Chirikjian, "Proceedings of the IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA)," Kobe, Japan, 2003. http://custer.me.jhu.edu/publication/pdf/auton.pdf.
512. R. A. Freitas, Jr. and R. C. Merkle, "Kinematic Self-Replicating Machines," Landes Bioscience, Georgetown, TX, 2004. http://www.MolecularAssembler.com/KSRM.html.
513. R. C. Merkle, in "The Ultimate Limits of Fabrication and Measurement" (M. E. Welland and J. K. Gimzewski, Eds.), p. 25, Kluwer, Dordrecht, 1994.
514. "Foresight Guidelines on Molecular Nanotechnology," Rev. Draft Version 3.7, Foresight Institute, 2000.
515. K. E. Drexler, "Engines of Creation: The Coming Era of Nanotechnology," Anchor Press/Doubleday, New York, 1986.
516. K. E. Drexler, C. Peterson, and G. Pergamit, "Unbounding the Future: The Nanotechnology Revolution," William Morrow/Quill Books, New York, 1991.
517. J. S. Hall, *Nanotechnology* 10, 323 (1999).
518. R. C. Merkle, *J. Brit. Interplanet. Soc.* 45, 407 (1992).
519. R. C. Merkle, *Nanotechnology* 7, 210 (1996).
520. C. Phoenix, *J. Evol. Technol.* 13 (2003).
521. J. B. Mathieu, S. Martel, L. Yahia, G. Soulez, and G. Beaudoin, in "Proceedings of the Canadian Conference on Electrical and Computer Engineering (CCECE)," Montréal, Canada, 2003. http://www.nano.polymtl.ca.
522. J. B. Mathieu, S. Martel, L. Yahia, G. Soulez, and G. Beaudoin, in "Proceedings of the 25th Annual International Conference of IEEE Engineering in Medicine and Biology," Cancun, Mexico, 2003. http://www.nano.polymtl.ca.
523. "MR-Sub," NanoRobotics Laboratory, Ecole Polytechnique, Montreal, Canada, 2003.
524. K. Ishiyama, K. I. Arai, M. Sendoh, and A. Yamazaki, *J. Micromechatronics* 2, 77 (2003).
525. S. S. Flitman, "Paper Presented at Eighth Foresight Conference on Molecular Nanotechnology," 2000. http://www.foresight.org/Conferences/MNT8/Papers/Flitman/index.html.
526. D. D. Chrusch, B. W. Podaima, and R. Gordon, in "Conference Proceedings IEEE Canadian Conference on Electrical and Computer Engineering" (W. Kinsner and A. Sebak, Eds.), IEEE, Winnipeg, Canada, 2002. http://www.umanitoba.ca/faculties/medicine/radiology/Dick_Gordon_papers/Chrusch,Podaima,Gordon2002.
527. P. Sharke, *Mech. Eng.* 123, 42 (2001).
528. I. Sample, *New Sci.* 172, 30 (2001).
529. R. A. Freitas, Jr., *The Sciences* 40, 26 (2000).
530. R. A. Freitas, Jr., *Artif. Cells, Blood Subst., Immobil. Biotech.* 26, 411 (1998).
531. R. A. Freitas, Jr., Zyvex preprint, 2001. http://www.rfreitas.com/Nano/Microbivores.htm.
532. R. A. Freitas, Jr., *Foresight Update* 44, 11 (2001).
533. R. A. Freitas, Jr., *Foresight Update* 41, 9 (2000).
534. R. A. Freitas, Jr. and C. J. Phoenix, *J. Evol. Technol.* 11 (2002).
535. R. A. Freitas, Jr., *Foresight Update* 48, 10 (2002).
536. E. F. Fincher, L. Johannsen, L. Kapas, S. Takahashi, and J. M. Krueger, *Am. J. Physiol.* 271, R149 (1996).

CHAPTER 14

Molecular Engineering in Nanotechnology: Engineered Drug Delivery

István Majoros, Thommey Thomas, James R. Baker, Jr.
Center for Biologic Nanotechnology, University of Michigan, Ann Arbor, Michigan, USA

CONTENTS

1. INTRODUCTION

Molecular engineering on a nanoscale level has slowly become the wave of the future toward the realization of novel plateaus in a variety of scientific frontiers. The incorporation of nanotechnology into the design of novel materials has spread to encompass a host of disciplines, interrelating a diverse range of fields never previously thought to have the ability

ISBN: 1-58883-048-9

Handbook of Theoretical and Computational Nanotechnology
Edited by Michael Rieth and Wolfram Schommers
Volume 6: Pages (673–717)

to work together. The emergence of this technology has come at a time when scientific knowledge has grown vast enough to at least allow us a slight glimpse into the capabilities nanoscale molecular engineering possesses. The recent research and development in the past decade regarding molecular engineering on the nanoscale level, specifically within the field of drug delivery, has become of great importance to the disciplines of biomedical engineering, medicine, chemical engineering, and the like. A prime example of the importance of molecular engineering can be seen upon direct examination of the uses of synthetically fabricated macromolecules for application within biological systems. Engineered molecules such as these are under prominent investigation for use in chemotherapeutic drug design and delivery systems.

The use of nanoscale molecular engineering in the field of drug delivery possesses a great deal of advantages as compared to conventional technologies currently used today. These advantages stem primarily from carrier composition and delivery and release methods, which in combination allow for controlled drug release, resulting in fewer side effects and less toxicity to the body compared to conventional chemotherapeutic drugs available today. Secondary advantages of the use of this technology result from increased drug efficacy with use of lower concentrations. With increased drug efficacy, industrial/pharmaceutical companies will synthesize smaller drug batches, which in turn produce less biohazardous chemical waste, therefore making this technology more environmentally friendly than commercial conventional drug syntheses.

Drug delivery techniques have undergone a vast expansion in research and methodology related to carrier systems and delivery and release methods. During the past few decades, we have been fortunate enough to undergo a great revision in the ideologies and research approaches related to drug carrier systems and their ability to be delivered/incorporated into biological systems. By revision and refinement of current working knowledge related to nanoscale molecular engineering, scientists, engineers, and medical researchers alike have been able to devise devices using our technological capabilities to the fullest. By use of nanoscale molecular engineering techniques to produce and alter macromolecules that closely imitate the size and shape of biomolecules commonly found within human and animal systems, we possess the ability to engineer delivery of chemotherapeutic (as well as other) drugs into biological systems without the levels of toxicity and rejection by the body resulting from current cancer drug therapies on the market today.

The intentions of this chapter are to educate the reader on current strategies and methods used in the pursuit of the synthesis of an ideal chemotherapeutic drug delivery system. Each subtopic (carriers, imaging, drug delivery, targeting, apoptosis) provides a review of current research and innovations within the area under discussion as well as the application of this research toward the synthesis and production of our own multifunctional drug delivery device. The chapter as an entirety provides an objective view of the application of nanoscale molecular engineering toward the production of a multifunctional drug delivery device jointly possessing all of the functions (imaging, targeting, apoptosis sensing, drug delivery/release) mentioned within. It also serves to express the complexity associated with device synthesis, quality control during production, and characterization. The culmination of this research has resulted in the synthesis of a well-defined device capable of fulfilling multiple functions simultaneously.

2. CARRIERS

Nanotechnology in drug delivery represents one of the most rapidly advancing areas of interdisciplinary science in which chemist, chemical engineer, biologist, and physician contribute to human health care. Due to the complexity of the biochemical events and homeostatic regulation processes occurring within the cell, a variety of cellular mechanisms often prevent traditional drug therapies from working effectively. Many of these preventative mechanisms inhibit the activation or distribution of drug molecules to cancerous sites within the body. Nanoscale drug carriers, on the other hand, closely mimic the size and shape of biomolecules already present in the body and have the ability to suppress many of the preventative mechanisms activated during use of conventional cancer drug treatments.

In designing a drug carrier for use in biological systems, one must take under consideration (1) drug resistance due to physiological barriers, (2) drug resistance at the cellular level, (3) distribution and transformations the drug may undergo within the body, and (4) excretory properties of the drug from the body [1]. Controlled drug delivery systems offer numerous advantages compared to conventional treatment forms including improved efficacy, reduced toxicity (side effects), and improved patient compliance and convenience. Currently, researchers have turned much of their interest toward using polymers as vectors for drug delivery. Polymers are versatile materials capable of undertaking the dual role of both drug carrier and "programmable" (through use of molecular engineering) drug dispenser. Many inorganic and organic (natural and synthetic) polymeric drug carriers and delivery systems have been developed in the past decades and reviewed [2, 3]. Here, we review recently published polymeric and, more specifically, dendritic drug carrier systems from the aspect of molecular engineering in nanotechnology.

The relentless race toward the synthesis of the "perfect" drug carrier has led to the fabrication of a large variety of polymer materials exclusively for this purpose. Due to the voluminous quantity of polymer materials undergoing research or currently being used in drug carrier systems, an organizational scheme must be implemented in order to better characterize and distinguish similarities and differences between each of the polymer types and functions. A few simple classifications can be made in order to have a better understanding of the products currently in use or undergoing consideration for use as drug carriers. Table 1 demonstrates the different types of polymers frequently used in conjunction with current drug delivery techniques today. Following the scheme outlined in this table, the first major division is made in order to delineate between natural and synthetic polymers. A subsequent division, to differentiate further between synthetic polymers, separates these polymers into two classes: synthetic biodegradable polymers and synthetic nonbiodegradable polymers. This list provides only a sample of the most common carriers, as the unwieldy number of polymers and copolymers tested for use as drug carrier systems limits us from characterization of all of them in one text.

2.1. Natural Polymers

Natural or organic polymers, as their name implies, are not synthetically derived, but rather are naturally occurring materials. This attribute merely enhances their biocompatibility, lending many of these polymers compatibility in human systems. Recently studied natural polymers as drug delivery matrices include collagen, a protein-based carrier [4], and polysaccharide hydrogel [5] and alginate [6, 7], both of which can more generally be labeled as natural polysaccharide carriers as displayed in Table 1.

Table 1. Polymer carriers commonly used for drug delivery today.

Polymer types	Examples
Natural	
Protein based	Collagen, albumin, gelatin
Polysaccharides	Agarose, alginate, cyclodextrins, polysaccharide hydrogel
Synthetic	
Biodegradable	
Polyanhydrides	Poly(sebacic acid), poly(adipic acid), poly(terphthalic acid)
Polyesters	Poly(lactic acid), poly(glycolic acid), poly(hydroxy butyrate), poly(dioxanones)
Polyamides	Polyamino acids, poly(imino carbonates)
Nonbiodegradable	
Acrylic polymers	Polymethacrylates, poly(methyl methacrylate), poly hydroxy(ethyl methacrylate)
Silicones	Polydimethylsiloxane, colloidal silica
Cellulose derivatives	Cellulose acetate, hydroxypropyl methyl cellulose, carboxymethyl cellulose, ethyl cellulose
Others	Polyvinyl pyrrolidone, ethyl vinyl acetate
Dendrimers	Poly(amidoamine), Poly(propyleneimine), aromatic ester, and ether-type dendrimers

Collagen is a naturally occurring primary structural protein in the body, hence its excellent biocompatibility and weak antigenecity. Though it is necessary to mention collagen as a drug carrier due to its biomedical capabilities and the expansive research it has undergone, research has not concentrated on the use of collagen as a molecularly engineered material; therefore, we will not go into expansive detail regarding its characterization. This globular protein has a secondary structure consisting of ß-conformation sheets, which dictates its strength and structural stability due to extensive hydrogen bonding and hydrophobic interactions between the functional groups of the amino acid residues composing it. When placed in aqueous solution, it can be molded into a variety of engineered forms for use in drug, gene, and protein delivery. A variety of collagen drug delivery systems and nanoparticles are available for use on the market today [4]. Biodegradable collagen nanoparticles and nanospheres are thermally stable and can be taken up by the reticuloendothelial system. This advantage allows the nanoparticles and, consequently, the drugs carried within the particles to undergo enhanced uptake within a biological system.

Polysaccharide hydrogel, on the other hand, is an effective carrier that releases drug due to the effects of swelling. Hydrogel closely resembles living tissue due to its excellent water-absorbing capacity. It must be noted that there are also synthetic hydrogels derived in order to mimic the nonlinear response of biopolymers like proteins and DNA [3]. Many studies have tested the use of hydrogel for the release of insulin in response to glucose levels within the blood [2, 3, 8]. Hydrogel synthesized with glucose oxidase (which exhibits glucose-sensitive swelling) during polymerization can be used for glucose level–dependent insulin delivery. Other studies have indicated that hydrogel made of semi-interpenetrating networks of polyacrylamide have the ability to respond to antigens, which could lead to drug delivery to sites of immunological diseases [3]. Poly(lactic acid)-poly(ethylene oxide) (PLA–PEO) hydrogels with temperature dependent gel–sol transition have also been experimented with for use as injectable drug delivery systems. The gel–sol transition state is especially effective for injection of viscous liquids that solidify after delivery to the target location [2].

The third natural carrier to be discussed, alginate, has been used in numerous areas of biomedical science from drug delivery to use as macroporous scaffolding for growth of new tissue. Alginate is a polysaccharide immunoisolating substance extracted from various genera of brown seaweed. It consists of alternating molecules of the monosaccharides L-guluronic acid and D-mannuronic acid. Alginate beads with a high L-guluronic acid content have the largest pore sizes, ranging around 5–200 nm. These pores can be used to encapsulate various cells, which continue to sustain and undergo mitosis while enclosed. Due to this capability, alginate beads can be used to encapsulate inhibitory proteins for delivery to tumor sites [7].

Metastasis of a tumor involves growth of new blood vessels (angiogenesis) and proliferation of malignant cells. This process can be slowed by inhibitory proteins, which have the ability to considerably reduce the multiplication of malignant cells. Delivery of such proteins to the surface of and within a tumor is a considerably difficult process, but research has shown that alginate beads can be used for successful delivery of these anticancer agents. Alginate macroporous beads with interconnected pore structures have been implanted and studied in rats [7]. This alginate delivery system demonstrated the growth characteristics of the cells encapsulated within alginate beads (NHI3T3, 293, and BT4C) implanted into the brains of BD-IX rats. The macroporous alginate beads supported cell invasion as macrophages and fibroblasts were noted throughout the matrices. Morphology studies showed that encapsulated cells proliferated within the beads, growing denser as the increase in cell number continued. After 4 months, a substantial amount of living cells (close to 70%) were present within the alginate beads. A continuous delivery of tumor growth inhibitory proteins (via transfection of cells encapsulated within alginate beads by inducible vectors) in the tumor microenvironment could be of considerable therapeutic importance for the treatment of cancer or other diseases such as Parkinson's disease within the brain. It must be mentioned, however, that the research of alginate systems is also not of chief importance to molecular engineering. These carriers are by and large ill defined, owing to difficulties encountered during attempts to determine the number of molecules or cells within each system.

2.2. Synthetic Polymers

2.2.1. Dendritic Carriers

Dendrimers are perhaps some of the most complex polymers available for use as drug carriers today. These polymers are capable of working in biological systems with ease due to their size, shape, surface multifunctionality, and solubility. Taking advantage of nanotechnology through their nanometer-scale dimensions and intricate architecture, dendrimers form drug carriers able to travel throughout the body with ease due to their nanoscale size and generation-dependent biomolecule resemblance. These hyperbranched polymers are synthesized through reiterative reaction sequences and possess the ability to have specific atomic sites manipulated to attach drugs, dyes, and various sensors. Much research regarding dendrimers has focused on the engineering of these macromolecules into monofunctional as well as multifunctional devices. Due to their complex design, syntheses of these dendrimers must be highly controlled in order to achieve perfectly structured macromolecules. Precise reaction conditions are necessary to produce these drug carriers. Despite the explicit conditions necessary to produce these functional carriers, a large selection of dendrimers varying in generation, molecular weight, synthetic materials, and terminal groups is commercially available on the market in bulk quantities for research purposes.

Dendrimer size can be controlled through molecular engineering to resemble closely antibodies, enzymes, and globular proteins in size. Poly(amidoamine) (PAMAM) ammonia core dendrimer generations 3, 4, and 5 closely match in size and shape insulin (30 Å), cytochrome *C* (40 Å), and hemoglobin (55 Å), respectively [9]. Because of the close resemblance to these and other biomolecules, dendrimers are able to travel efficiently through the body. Figure 1 demonstrates the size and shape comparisons of the above-mentioned biomolecules as compared to differing dendrimer generations.

Dendrimers are composed of polymer materials ranging in molecular weight and terminal reactive groups. The most commonly researched dendrimers for drug delivery are poly(amidoamine) (PAMAM), "starburst dendrimers," modified poly(propyleneimine) (POPAM), and aromatic ether-type dendrimers. Dendrimers differ in construction from other polymers as they are synthesized from AB_n monomers (A and B signifying monomer units and n signifying a branching multiplicity of 2 or 3) but not AB monomers, which produce linear polymers. Each carrier is composed of an initiator core and secondary reagents (building blocks), which are attached through a reiterative sequence of steps to form generations. As the generation number of the dendrimer increases, dendrimer size increases as well. Repetition of these synthetic steps produces a highly branched macromolecule characterized by interior cavities capable of encasing therapeutic drugs or imaging units and reactive terminal groups capable of attaching drugs, dyes, and/or a variety of sensors. The use of molecular engineering techniques during the synthesis of a mono- or multifunctional dendritic drug carrier results in a product possessing a strikingly precise structure (a necessity for attachment of various secondary devices for visualization), cancer cell detection/targeting, and malignant cell destruction.

Traditionally, there are two general techniques used to synthesize dendrimers: convergent and divergent methods. The convergent method consists of synthesizing the dendrimer from the outside in, or from the periphery to the initiator core. This method is conducive to generating a considerably more homogeneous dendrimer as compared to those produced by divergent strategies. The divergent method, pioneered by Newkome and Tomalia, begins with the initiator core, and secondary reagents are added until the desired generation of dendrimer is produced. Though there are more defects present in divergently versus convergently synthesized dendrimers, the divergent method is often favored as it can produce dendrimers with up to 10 generations, whereas dendrimers produced by the convergent method can generate dendrimers with a maximum of 8 generations. Due to this shortcoming, most commercially available dendrimers are produced using the divergent method.

Although the production of each dendrimer requires a variety of differing reagents, the methodology behind each synthesis remains the same. PAMAM dendrimers are synthesized using an ethylenediamine (EDA) core with Michael addition of methyl acrylate (MA) and condensation (amidation) reactions of the resulting ester with large excesses of EDA.

Methanol (MeOH) is formed during the amidation reaction and must therefore be considered in all calculations regarding molecular weight. For the synthesis of POPAM dendrimers, a core of 1,4-diaminobutane is used, and acrylonitrile is added through exhaustive Michael addition followed by catalytic hydrogenation to form each generation. In all cases of dendrimer synthesis, each reaction is repeated to form new generations. A well-designed experiment followed with precision and accuracy can assure that a defined structure will ensue. The 10–200-nm resulting sizes of these dendritic carriers again closely resemble those of proteins and other biomolecules, allowing for easy passage within the body.

Due to the characteristic building-block nature of dendrimers (expressed through the formation of generation "layers"), it is possible to both theoretically and experimentally calculate the molecular weights and the number of terminal and tertiary groups that belong to each dendrimer. These numbers are initially generated theoretically via mathematical models that take into account degree of polymerization, molecular weights of each monomer, and molecular weights of each side product created by the multiple reaction steps. Experimental confirmation by UV absorption techniques and gel permeation chromatography, GPC, has verified the validity of these theoretical models for application to biological dendrimers. The mathematical descriptions of PAMAM, POPAM, and POMAM dendrimers provided

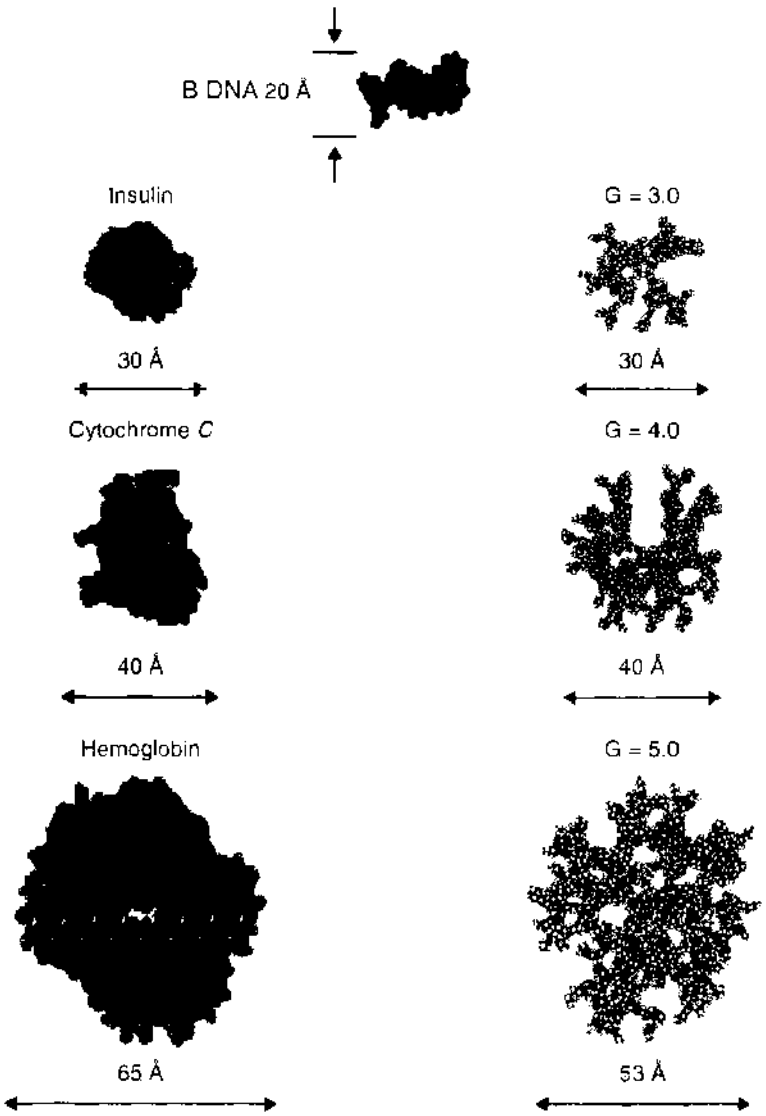


Figure 1. Size comparison between PAMAM dendrimers and biomolecules. Reproduced with permission from Y. Sayed-Sweet, D. M. Hedstrand, R. Spinder, and D. A. Tomalia, *J. Mater. Chem.* 7, 1199 (1997). © 1997, The Royal Society of Chemistry.

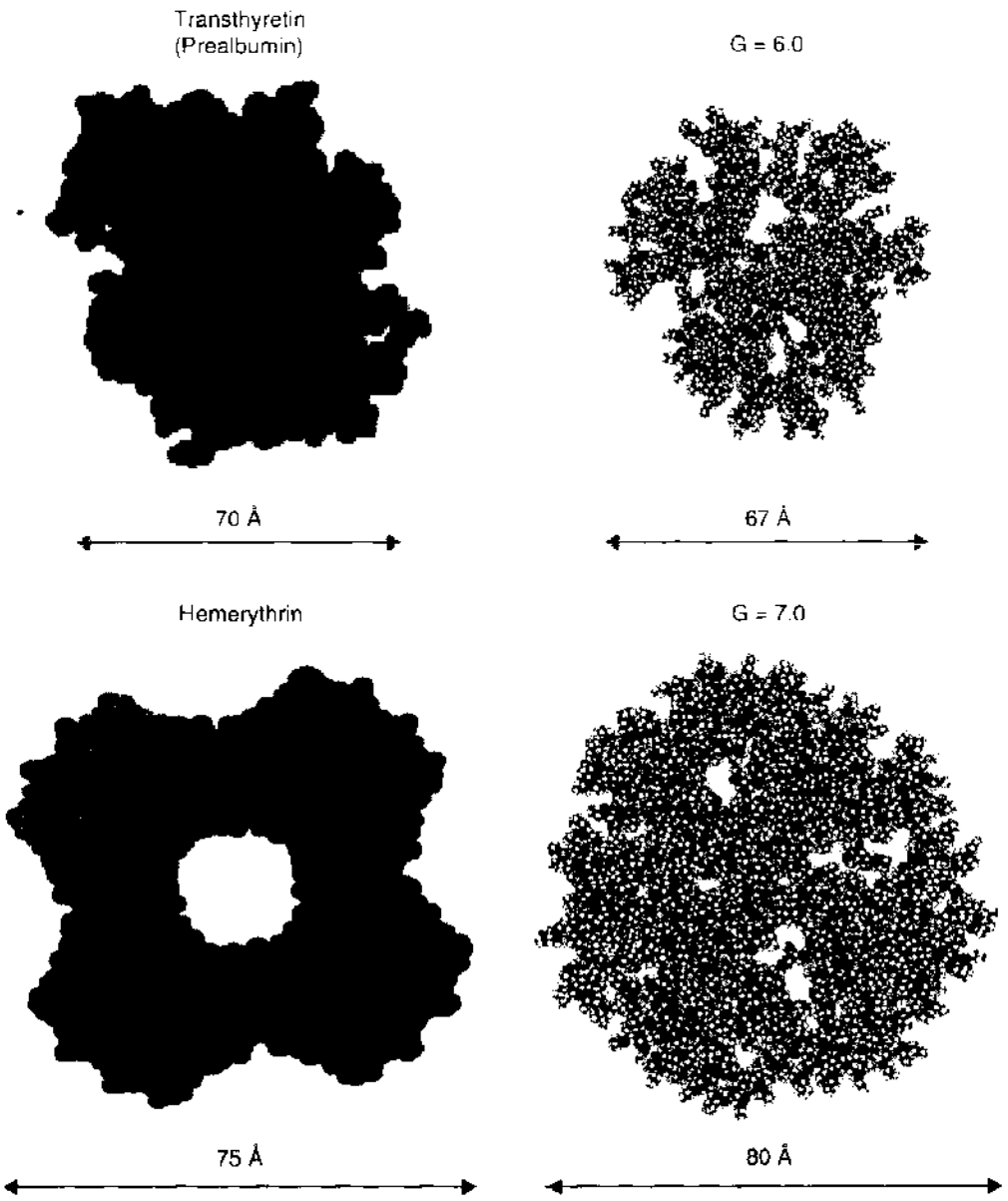


Figure 1. Continued.

in Ref. [10] can be used theoretically to determine the molecular weights, the degree of polymerization, and the number of terminal and tertiary amine groups for each of these types of dendrimers, depending on the generation. Examination of the molecular weights of the monomeric subunits that compose each dendrimer allows for the determination of the overall molecular weight through the use of the below given formulas. With this knowledge, one is able to characterize easily each of these dendrimers. As an example of how these formulas are used, the values to insert into the equations for PAMAM dendrimers are included below.

General notation for the formula of the molecular weight for each generation is given by

$$MW = MW_{\text{core}} + (\textit{Molecular weight of monomers}) F_c \left(\frac{F_r^{g+1} - 1}{F_r - 1} \right) \tag{1}$$

where F_c is core functionality (EDA core functionality = 4, NH_3 core functionality = 3), F_r is multiplicity of the repeating unit (MA + EDA's multiplicity is 2), and g is generation number.

The degree of polymerization (number of dendrimer repeating units) can be represented by the equation

$$\textit{Degree of polymerization} = F_c\left(\frac{F_r^{g+1}-1}{F_r-1}\right)\left(\overline{DP}_n = \frac{\overline{M}_n}{MW_{\text{monomer}}},\ MW_{\text{core}} \text{ is neglected}\right) \quad (2)$$

The number of terminal groups for a dendrimer can be determined through the use of the formula

$$\textit{Number of terminal groups, } Z = F_c F_r^g \quad (3)$$

The number of tertiary amines for a dendrimer can also be calculated using the following equation

$$\textit{Number of tertiary amines} = T_c + F_c\left(\frac{2^g-1}{2-1}\right) \quad (4)$$

where T_c is number of tertiary amines of the core (EDA has two tertiary amines).

The repetitive nature of dendrimer architecture allows us the ability to calculate these values using mathematical models. The recurrent synthetic steps produce generations of replicate structure without a great deal of side reactions to alter composition. It must be noted that no synthesis is perfect; therefore, these values must be treated only as theoretical approximations. These estimations, however, have been shown to be quite accurate in predicting the number of terminal groups, the degree of polymerization, and the molecular weights of each dendrimer generation [10]. Figure 2 presents the structure of a G = 3.0 PAMAM dendrimer. One must note the symmetrical molecule procured from the repetition of the synthetic process.

Some of the most important features of dendrimers are the terminal groups and the internal cavities found within generations. The terminal groups determine the surface functionality of the dendrimer: in other words, what functional groups can be attached to it. The types of terminal groups present and the size of the cavities are dependent on the reagents and reaction steps used during dendrimer synthesis.

The reactive terminal groups can be used to attach chemotherapeutic drugs, targeting ligands that attach to overexpressed receptors present on the cancer cell surfaces and to fluorophores for visualization of the location of the dendrimer within a system, as demonstrated by Fig. 3. Which terminal group is more likely to react is dependent on steric hindrance, hydrophobic interactions, and hydrogen bonding, as well as other factors. These carriers can be directed to cancer site locations if targeting agents are attached to the terminal groups of the dendrimer. Research by the University of Michigan Center for Biologic Nanotechnology has shown that dendrimers containing therapeutic agents administered "*in vivo*" can be targeted by antibodies to antigen-bearing tumors [11, 12].

Other studies have also shown that dendrimer terminal groups can be conjugated to fluorochromes and enter cells, acting as an apparatus to sense physiological changes within the cell [12]. Enzymatic cleavage of the bond between the carrier and a fluorophore is also used for visualization of carrier location as well as enzyme activation. It is necessary to mention, however, that there are issues regarding toxicity related to the reactivity of the terminal groups with blood, causing haemolysis. By utilizing dendrimers synthesized with terminal groups nonreactive to biological molecules in the blood, this problem can be prevented. The "*in vivo*" studies have also shown that intravenous injection of dendrimers with nonreactive terminal groups produces no evidence of toxicity.

Also important to drug delivery and imaging aspects of dendrimer carriers are the cavities present between generations. It has been shown that guest molecules can be entrapped into higher generation dendrimers when an amino acid derivative is used to cap each end group of the dendrimer [13]. The tested amino acid derivative cap was sterically crowded and bulky, therefore preventing guest molecules from diffusing out. Another encapsulation method makes use of hydrophobic interactions to entrap guest molecules. What can be contained within these cavities, however, is also determined by the interaction between host and guest

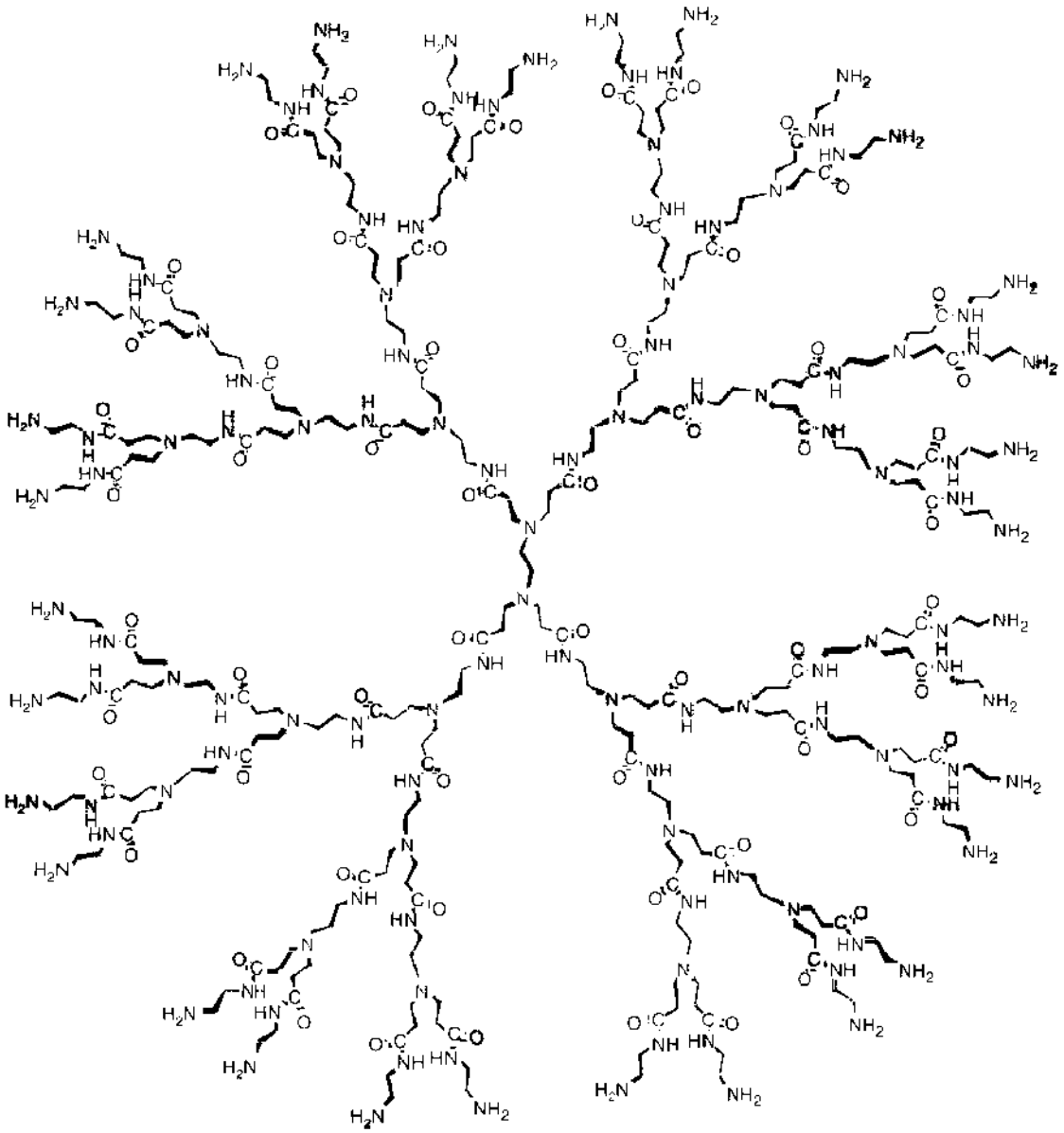


Figure 2. Chemical structure of PAMAM dendrimer (G = 3.0).

molecules. Polarity, size, and conformation of the guest molecule are determinants whether inclusion of the guest molecule can occur [14]. These void spaces allow for encapsulation of metal nanoparticles, for use as an imaging unit (for visualization of carrier location and enzyme activation), or for the encapsulation of therapeutic drugs, which can be released in a controlled manner with the rate ultimately dependent on dendrimer size, composition, and physiochemical properties.

Dendrimers can be synthesized purely as drug carriers or as a variety of other single-function devices, which is discussed elsewhere in this chapter. In the past, researchers have conjugated each of these units to dendritic carriers separately, but much research is currently focusing on developing a dendritic carrier with a therapeutic drug, targeting unit, fluorophore, and apoptotic sensor attached. Conjugation of all of these units to one carrier creates a multifunctional device capable of simultaneously delivering medication to a patient in targeted locations, sensing the location of the device in the body, as well as dispersing and activating of the drug that was delivered. More information regarding newly synthesized dendritic multifunctional devices can be found in Refs. [9, 11–13, 15–39]. The fact that these dendrimers have the ability to be used for a variety of functions simultaneously sets these

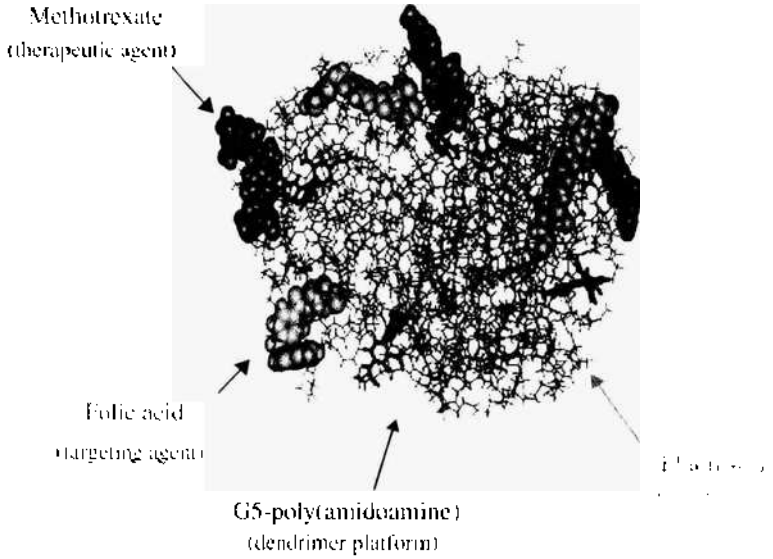


Figure 3. Capped terminal groups with conjugated guest molecules (3D model: with targeting, imaging, and drug).

polymers apart from other carrier systems as one of the most complex drug delivery systems available for use today.

Though the numerous benefits of using dendrimers as drug carriers far outweigh the negative aspects, it is necessary to note the concentration and sometimes generation-dependent (for PAMAM dendrimers: Fig. 4) hemolysis present with use of dendrimers terminated with $-NH_2$ groups. Testing on cell lines *in vitro* has determined that dendrimers terminated by carboxyl, hydroxyl, and acetyl groups, however, do not exhibit hemolytic or cytotoxic properties [38]. Accumulation of dendrimers (generation dependent) in areas of the body involved in excretion (the liver and kidneys) has also somewhat slowed clinical trials for these carriers. Areas in the body associated with filtration and excretory functions—more specifically, the kidneys and liver—are also affected by use of these carriers, as dendrimers can build up in either of these locations dependent on the generation (size) used. Continued research in this field will certainly speed the process toward the modification and approval of these carriers for clinical use in the near future.

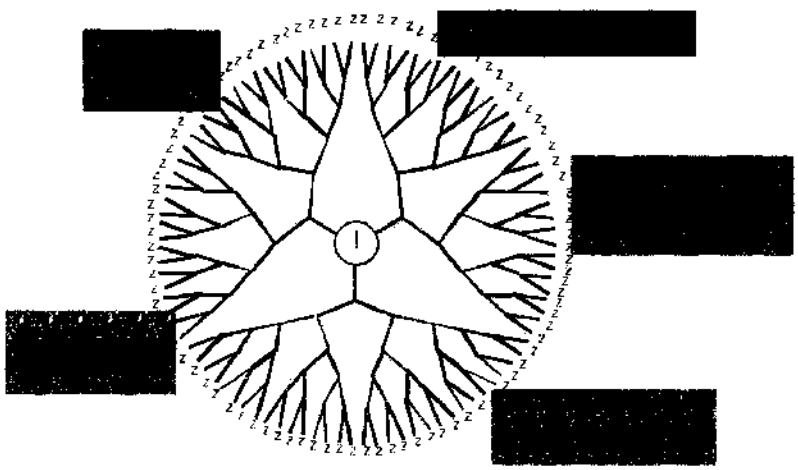


Figure 4. PAMAM dendrimer conjugated with various functions.

2.2.2. Polymer Nanoparticles

Polymer nanoparticles can be synthesized to recognize certain biomolecules, therefore enhancing their drug delivery capabilities. A molecular imprinting technique common to chromatography allows for this capability. Use of a molecular template during the polymerization process and removal of this molecule after polymer formation yields a nanoparticle left with nanopores or micropores (in the shape of the template molecule) capable of recognizing the molecule previously used during polymer formation. This system can be used to create carriers that will only release a drug upon identification of the template molecule within the biological system. This ability allows scientists and physicians to deliver medication releasable only in specific conditions. This type of system is applicable to release of insulin in response to glucose deficiency in the blood or release of other drugs in response to an overabundance of malignant cells [8].

2.2.3. Magnetic Polymeric Composites

Polymer matrices combined with powdered drugs and magnetic beads can be used to deliver drugs by differing the rates of diffusion through pores in the polymer matrix altered by the presence of a magnetic field. Research by Langer and coworkers [40] has shown that without the presence of a magnetic field, drug diffusion is slow through matrix pores, yet upon introduction of an external oscillating magnetic field, drug release is increased as pores of the matrix are compressed in such a manner that larger amounts of drug are available for release within a shorter period of time. Rates of drug release were recorded to be 30 times higher with the presence of a magnetic field versus without, and it was shown that alterations in the strength and frequency of the magnetic field allowed for precise control and modification of release rate. Possible uses of these magnetic polymeric composites are endless. These composites may be used for long-term storage and release of insulin for diabetics or for storage and release of other long-term, rate-release–dependent medications. It would be necessary, however, for devices to be worn to activate the carriers to release medications at given times. Patient compliance might be an issue with carriers of this type, but biocompatibility as demonstrated in Ref. [40] shows great promise for these devices.

2.2.4. Hydrophobic Polymers

Hydrophobic polymers like poly(ethylene-vinyl acetate) or poly(lactic-glycolic acid) copolymers can be dissolved in solvents and mixed with proteins or drugs ranging in size for slow release during a controlled number of days. The principle behind the idea of using hydrophobic polymers stems from the fact that biological systems where drug carriers are used are surrounded by aqueous mediums, and by using a polymer carrier in which the exterior is hydrophobic and the interior cavities are hydrophilic, these cavities or pores can be used to encase drugs or proteins for controlled delivery and release. By controlling the implant geometry, the release rate can be altered to maintain a constant rate [40]. In order for a carrier to erode heterogeneously, it is best for it to be hydrophobic but to possess also water-sensitive linkages [2]. Poly(anhydrides) fall into this category, as their anhydride bonds are sensitive to water and cleavage.

It has been shown that, upon cutting thin cross-sections of an ethylene-vinyl acetate copolymer with a cryomicrotome, no pores (where a drug or protein awaiting release resides) were found when there was an absence of a drug or protein within the polymer. However, upon placement of either a drug or protein within the polymer matrix, pores were formed due to phase separation that took place. Alteration in pore size is key to achieving different release rates. By controlling molecular weight or composition or the size or concentration of the particles added for diffusion by the polymer, one has the ability to determine pore size and to control release rate. Determination of pore size is necessary, as pore size dictates the rate of diffusion of the protein or drug into the system [40].

Poly(anhydride) carriers have been used for years as delivery agents for proteins and low-molecular-weight chemotherapeutic drugs. Poly(anhydrides) can be prepared through melt-condensation, ring opening polymerization, interfacial condensation, dehydrochlorination, and by use of dehydrative coupling agents [41]. Use of poly(anhydride) carrier systems

allows for controlled degradation of the polymer while maintaining the drug or protein in the interior of itself. This is due to the water-sensitive anhydride bonds on the surface of the polymer as well as to the general hydrophobic nature of the polymer, which prevents water from entering the interior pores or cavities. The rate of degradation can be altered by variations in composition. The more hydrophobic the monomer, the higher the stability of the anhydride bond to hydrolysis. This combination, again, allows for controlled heterogeneous degradation of the carrier, analogous to the degradation of a bar of soap [2].

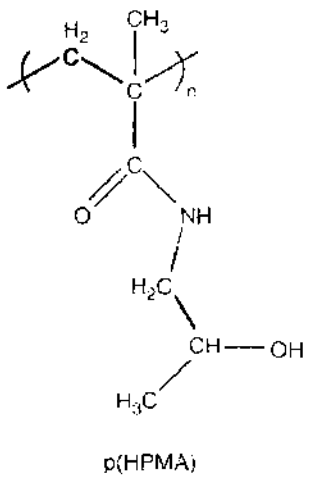
2.2.5. Homopolymers, Cross-linked Polymers

A significant number of research activities have been reported using a variety of homopolymers and cross-linked polymers. Most prominent under study are cross-linked hydrogels, polypyrrole (PPy), poly-*N*-(2-hydroxypropyl) methacrylamide (HPMA), poly(ethylene glycol) (PEG) conjugates, and synthetic cyclodextrin-based drug delivery systems. Poly(lactic acid) (PLA) and poly(glycolic acid) (PGA) have also been used as effective homopolymer drug carrier and delivery systems as has the homopolymer polycaprolactone, which can be conjugated as a paste to Taxol for use as an anticancer therapeutic agent [42]. These polymers have been coupled with different types of drugs, adenosine triphosphate (ATP), lysozyme, and growth factors. These systems are biocompatible, water soluble, stable at pH = 7.4, and capable of releasing the carried substrates.

The general properties of cross-linked hydrogel synthesized by propoxylation of lactitol and esterification with chlorinated poly(ethylene glycol) bis(carboxymethyl)ether was described [43]. The swelling ratios decreased with increasing cross-linking and were sensitive to the temperature and to concentrations of salt and glucose. The release rate of acetylsalicylic acid (the model drug) was found to be dependent on the time and the cross-linking ratio. Cross-linked hydrogel can also be used as a barrier for diffusion of a drug within a system. Again, the release rate of a drug is dependent on the time and the ratio of cross-linking—the more the cross-linking (higher rigidity), the less the swelling of the hydrogel.

Polypyrrole has the ability to act as an electroactive membrane, which is useful for drug release and medium sensing. This conducting electroactive polymer (CEP) has been tested for use as a drug delivery agent due to its unique redox properties. These properties allow controlled ionic transport through the polypyrrole membrane. Adenosine triphosphate (ATP) as a model drug was electrostatically entrapped into a polypyrrole (PPy) membrane during electrochemical synthesis [44]. Release occurs during reduction and has been induced by a chemical trigger, dithiothreitol, and by pH increase (0.1 M NaOH/0.1 M NaCl solution). In the latter, a fast release process was observed due to the deprotonation of the PPy.

Poly(HPMA) (Fig. 5) has effectively been researched and used as a drug carrier in biological systems. The affectivity of conjugates of poly(HPMA) and doxorubicin (DOX) in a



p(HPMA)

Figure 5. Poly-*N*-(2-hydroxypropyl methacrylamide) [Poly(HPMA)].

slightly cross-linked or in a star-like shape (both coupled by an ATG antibody) were studied [45]. HPMA conjugated with DOX displayed a tumor-selective deposition of the drug. DOX has shown to be internalized via endocytosis for lysosomal delivery of the active drug to the selected area [45]. There was no significant difference observed in the binding activity of the device configurations compared to the free antibody binding to its T-cell surface antigens. The *in vitro* cytotoxic effect of conjugates was tested on mouse T-splenocytes and EL4 T-cell lymphoma cells. The conjugates almost completely inhibited proliferation of stimulated cells. The effect depended on the structure of the peptide spacer and the molecular weight of the device. The *in vivo* effect of drug-polymer conjugates differed in the configuration of Phe in the oligopeptide spacer on the growth of EL4 lymphoma and in the survival time of mice (D < L < DL configurations).

Conjugation of platinates to polymeric carriers increases system solubility, reduces toxicity, and increases drug delivery to a localized tumor area due to the enhanced permeability and retention (EPR) effect. Research has shown that HPMA copolymer platinates with a molecular weight of 25,000–31,000 Da and a platinum content of 3–7 wt% do not exhibit cytotoxic effects on B16F10 cell lines. Use of these conjugates allows for delivery of a cytotoxic drug to a specified area through endocytosis without toxic effects to the rest of the body. Preliminary clinical trials on animals and man have also proven this. One problem encountered with the use of HPMA-platinates, or platinates in general, is the increased resistance tumors acquire toward these conjugates. Research has shown that tumor cells typically experience a 3- to 5-fold decrease in sensitivity toward the use of platinum anticancer agents, therefore use of these carriers may be considered ineffective in the long run [45].

Cleavage studies [46] of PEG–DOX conjugates reported the mechanism and viability of these conjugates as therapeutics. The release reaction mechanism involved enzymatic ester cleavage of PEG + spacer and amide cleavage of the spacer-drug by lactonization. The most effective device against ovarian adenocarcinoma tumors (SKOV-3) was found to be daunorubicin amide-bonded to 3-(2′-PEG-alaninyl-4′,6′-dimethylphenyl)-3,3-dimethylpropionic acid.

A second study involving amino-containing daunorubicin (DNR) showed DNR covalently attached to PEG through a series of linkers via a variety of chemical groups [47]. Half-life time measurement in fresh rat plasma indicated that the ester bond covalently attaching the PEG linker to the carrier was cleaved by enzymatic hydrolysis 10–20 times faster than other bonds. The cleavage process took place in two steps: first, PEG carrier cleavage by enzymatic hydrolysis followed by a rapid 1,4- or 1,6-benzyl elimination reaction, which released the free drug. Subsequent *in vivo* testing (M109, SKOV-3) demonstrated superior tumor growth inhibition via use of this carrier as compared to equimolar amounts of the native drug.

Another study [39] has shown that linkage of folic acid (the targeting unit) using a PEG spacer arm to cyclodextrin creates a bioconjugate-targeted drug delivery system to areas of cancer cells, which possess folate receptors. Linkage of folic acid and cyclodextrins using PEG provides space in which the folic acid can react with the folate receptor while also preventing steric interference from the cyclodextrins. Cyclodextrins are carbohydrate macrocycles that have the ability to complex with a variety of hydrophobic molecules. Natural cyclodextrins have been known to produce toxic side effects such as hemolysis. More specifically, research has shown that ß-cyclodextrins are the most toxic of the natural cyclodextrins. Due to this, hydroxyl substitution using hydroxy-alkyl or sulfo-alkyl groups has been used to produce semisynthetic cyclodextrins in order to avoid the bulk of the side effects produced through the use of natural cyclodextrins. This carrier is characterized by a low molecular weight, allowing for easy passage into tumors and possibly greater drug delivery ability as compared to higher molecular weight delivery systems. The use of this bioconjugate carrier as a drug delivery agent is especially effective due to the inclusion capacity of cyclodextrins, the solubility and biocompatibility of PEG, and the ability to target folate receptors in areas of cancer due to the use of folic acid as a targeting unit [39].

Homopolymers PLA and PGA occur in D and L (or *R* and *S*, respectively) forms [2]. Both PLA and PGA (Fig. 6) are derived from monomers that are also natural metabolites in the human body; therefore, degradation of either produces monomers nontoxic to biological systems. Breakdown products are important to consider when assessing which carrier to use in a biological system. It is necessary to consider this as some polymers, though stable and

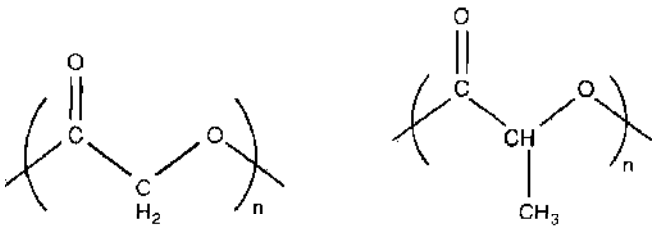


Figure 6. PLA and PGA polymer structures.

nontoxic as macromolecules, break down into monomers that are damaging and toxic to biological systems.

Therefore, it is safe to say that the degradation products of each drug carrier must be monitored in order to determine whether or not it can be considered biocompatible. In its natural state, lactic acid occurs in the L form; therefore, the L state of PLA (known as PLLA) is considered to be more biologically compatible. In comparison to one another, PGA homopolymers degrade faster than PLA homopolymers due to crystallinity factors as well as the pendent methyl group on PLA acting as a steric inhibitor to hydrolytic attack. Research by Uhrich et al. [2] has also suggested that PLA-derived microparticles will degrade faster than PLA-derived nanoparticles.

2.2.6. Inorganic Porous Silica Carriers

Porous silica-based matrices derived from sol gel have been used for delivery and controlled release of proteins as substrates for bone growth and ibuprofen (30%) [48]. In the latter, the drug was released within 3 days under "*in vitro*"-simulated conditions. In the case of bone growth, the "*in vitro*" release of a trypsin inhibitor (a model protein of similar size to growth factors with documented effect in bone) was dosage- and time-dependent. The release pattern included an initially slow release, with a further release driven by diffusion. Though this carrier is mentioned because of its special usages as well as to complete the discussion regarding the variety of carriers available on the market, it must be noted that inorganic porous silica carriers make no use of nanotechnology and therefore will not be described in further detail.

2.2.7. Block Copolymer

Block copolymers are being researched as a potentially extremely useful means of drug delivery, as they have the capability of attaching two or more drugs simultaneously. Due to block copolymers' amphiphilicity, they have the ability to self-assemble into aggregates with hydrophobic interiors and hydrophilic exteriors. Hydrophobic interior blocks are encased by hydrophilic exterior blocks, allowing for encapsulation of a polar drug and stabilization between the exterior and aqueous medium by the hydrophilic shell [42]. Block copolymers are valuable carriers due to their ability to mimic biological transport systems, lipoproteins, or viruses [3]. Recent block copolymer carrier studies have been based on lactic and glycolic acid and ethylene glycol. One study produced nanoparticles from poly(D,L-lactic acid)-*co*-poly(ethylene glycol) (PLA–PEG) block copolymers for use as a site-specific drug delivery system. A second study has reviewed the use and biocompatibility of poly(ethylene glycol) (PEG) block copolymers.

The vascular endothelial growth factor (VEGF) and bovine serum albumin have been co-encapsulated into microspheres prepared from poly-(D,L-lactic acid-*co*-glycolic acid)/poly(ethylene glycol) (PLGA/PEG) [49]. The ability to determine and limit the rate of release from PLGA microspheres is derived from the understanding of end-group effects [2]. The degradation of microspheres was monitored during 28 days (the hydrolytic products are nontoxic lactic acid and glycolic acid), and the release rate was found to be constant after the sixth day.

Amorphous copolymers, that is, lactic and glycolic acid copolymers (Fig. 7), can be formed with CO_2 into a three-dimensional matrix with porosity up to 95% [50]. Sheridan et al. have shown that angiogenic factors (e.g., VEGF) can be incorporated into block copolymer matrices during the fabrication process and provide controlled release. VEGF released in this manner was found to retain more than 90% of its bioactivity.

PLA–PEG systems are promising drug delivery systems due to their long circulation times and their potential ability for high drug-loading. Researchers have shown that varying the amount of PEG present in the block copolymer can alter the degradation rate of the system, which in turn can be used to control the rate of drug delivery [51]. Nanoparticle size, uniformity, hydrodynamic radius, and aggregation behavior of site-specific drug delivery poly(D,L-lactic acid)-*co*-poly(ethylene glycol) (PLA–PEG) (Fig. 8) polyester-polyether block copolymers as a function of the PLA segment have been studied [52]. The aggregation number and nanoparticle size depend not only on the copolymer composition but also on the concentration of copolymer in the organic phase during preparation.

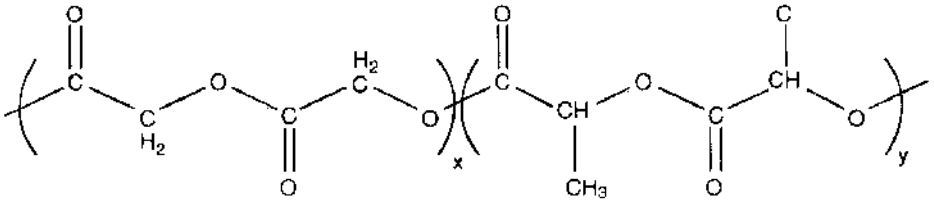
PEG block copolymers have widely been noted as biocompatible carriers with the ability to be excreted at the rate of 98% in humans, per a molecule with the molecular weight of 4000 amu [2]. Figure 9 displays the chemical structure of PEG. PEG is hydrophilic, which leads to a resistance to protein adsorption due to water molecules that hydrogen bond tightly to the chain. This protein resistance allows for increased circulation time within the blood *in vivo* as compared to poly(lactic acid) PLA particles. The inclusion of PEG into block copolymer systems therefore allows for longer protein circulation in the body due to prolonged endocytosis, phagocytosis, liver uptake and clearance, as well as other adsorptive processes [2].

2.2.8. Graft Copolymer

A new class of hetero-grafted amphiphilic copolymers has been synthesized [53]. To poly-L-lysine and poly-L-ornithine backbones, palmitoyl has been attached covalently and methoxy-poly(ethylene glycol) (mPEG) has been added as a block segment to form PLP and POP copolymers, respectively. These copolymers assembled into polymeric vesicles (400–600 nm in size, measured by photon correlation spectroscopy) in the presence of cholesterol. The vesicles were able to encapsulate drugs such as doxorubicin HCl with a capacity of more than 10 mg/g of polymer and hence could serve as drug carriers. Additionally, these polymer vesicles have been evaluated *in vitro* as nonviral gene delivery systems and were found to be effective at gene expression with low toxicity *in vivo*. Vesicle-DNA complexes are about one to two orders of magnitude less cytotoxic than the parent polymer-DNA complexes, although more hemolytic than the parent polymer-DNA complexes, and showed improved gene transfer to human tumor cell lines despite the absence of receptor-specific ligands and lymosomotropic agents such as chloroquine.

2.2.9. Linear, Random Copolymer

Three monomers, *N*-isopropylacrylamide (NiPAM), protein-reactive *N*-acryloxy-succinimide (NASI), and hydrophobic alkylmethacrylate (AMA), were employed to synthesize linear random copolymers [54] suitable for making thermoreversible structures with proteins.



PLGA

Figure 7. Poly(lactic acid-*co*-glycolic acid) (PLGA): PLA–PGA copolymer.

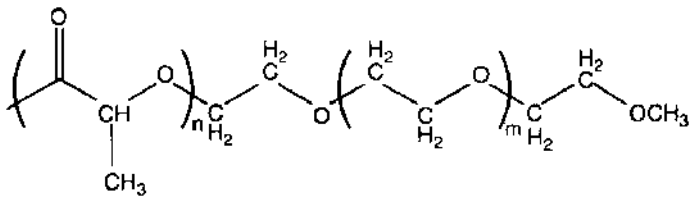


Figure 8. PLA–PEG copolymer.

These thermoreversible polymers are so called due to their temperature-dependent solubility. This characteristic enables one to induce a phase separation by increasing the temperature to higher than the solubility limit of the polymers, which is useful for attaching proteins without the use of cross-linking methods, which can alter protein structure. This capability was demonstrated by localizing a therapeutic protein, recombinant human bone morphogenetic protein-2 (rhBMP-2), with high effectivity *in vivo* in a rat intramuscular experiment. These thermoreversible polymers can be applied for delivery of therapeutic proteins to improve their therapeutic efficacy.

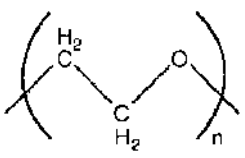
2.2.10. Oligonucleotides

The methodology of nano-encapsulation has been developed for delivery of oligonucleotides (ODNs) or antisense oligonucleotides into the intracellular compartment [55–59]. Antisense oligonucleotides with base sequences complementary to specific RNA have the ability to modulate gene expression after binding to intracellular mRNA. However, they have poor stability in biological fluids and low intracellular penetration; therefore, it is necessary to encapsulate them to ensure deliver without having undergone damage. Single-stranded ODNs can also be complexed noncovalently into the intracellular compartment or be delivered into the nuclei of cells [57].

Nanocapsules have been prepared by interfacial polymerization of isobutylcyano-acrylate (IBCA) in a water/oil emulsion (in the presence of ODNs in the water phase) with a size of ~350 nm [55]. Zeta potential measurements and fluorescence-quenching assays confirmed localization of ODNs in the water core of nanocapsules. These nanocapsules were incubated in the presence of serum and then delivered into biological systems where they were efficiently protected from degradation by nucleases.

Other researchers have found the benefits of complexation of ODNs with porphyrins, which also can act as carriers [57]. A variety of meso-substituted, water-soluble cationic porphyrins (meso tetra (*N*-methyl-4-pyridyl)-porphyrin tosylate [TMPy(4)], meso tetra (4-*N*, *N*,*N*-trimethylanilinium)-porphyrin chloride [TAPP], and meso tetra (4-sulfonatophenyl)-porphyrin dihydrochloride [TPPS]) have been used for ODN complexation. These porphyrins have been shown to protect the ODNs from nuclease degradation. Researchers have determined that their delivery capacity was dependent on the charge ratio and the type of porphyrin used as well as the chemical substituents of the ODN.

A new class of fluoropyrimidines (FdUMP[10]), oligodeoxynucleotides (ODNs) synthesized from 5-fluoro-2′deoxyurindine-5′monophosphate (FdUMP) nucleotides, have also been developed [58]. Folic acid was attached to the FdUMP[10] oligomer via a phosphodiester linkage (FA-FdUMP[10]), and it showed improved cytotoxicity toward human colorectal tumor cells (H630) and 5-fluorouracil (5-FU)-resistant colorectal tumor cells (H630-10).



PEG

Figure 9. Polyethylene glycol) structure (PEG).

The enhanced cytotoxicity of FA-FdUMP[10] was 183,000-fold greater relative to 5-FU in H630-10 cells and 25-fold better than unconjugated FdUMP[10] toward H630-10 cells in a folate-free medium.

3. IMAGING

In vivo imaging tools using noninvasive techniques have rapidly been increasing in quality as technology progresses. These tools can be used to detect and monitor tumor growth as well as other cancer-related physiological events. Imaging can also be used to detect molecular markers specific to the presence of tumors. The majority of imaging techniques rely on the interaction of energy with tissues in the body [60]. Magnetic resonance imaging (MRI) is one of these techniques. Other methods, however, rely on use of molecular probes for imaging. The strength of these chemical sensors lies in their minimal interference with the cells or sample in question [61]. The latest imaging techniques have included the use of dendrimer nanocomposites or nanoparticles as contrast agents. The nontoxicity of the poly(amidoamine) (PAMAM) dendrimer has made it safe for use in both human and animal systems. Different sizes of these dendrimers can be used for imaging different areas of the body. Scientists and researchers have been taking advantage of these versatile polymer–metal conjugates for use in visualizing cancer and cancer markers. Of course, other methods are also still widely in use for imaging; these methods include positron emission tomography (PET) scans, computer tomography (CT) scans, as well as a host of others. The costs and efficiencies of these technologies vary, but use of MRI or nanocomposite imaging allows for the best resolution as compared to other current technologies available. Table 2 describes current methods in use for imaging of human and animal systems. The methods are listed in order of resolution quality, beginning with the highest. General descriptors are given for the particular uses of each imaging unit, as not all will be detailed within this review.

3.1. Nanocomposite- and Nanoparticle-Based Imaging

Dendrimers containing metal integrated into their structures have a wide variety of properties and functions. Many methods of imaging have taken advantage of the unique structural characteristics dendrimers possess. Due to the repetitive branching, the interior cavities present between generations, and the reactive terminal groups, these polymers make

Table 2. Available imaging systems, resolution qualities, and uses.

Method	Resolution	Notes on use
Magnetic resonance imaging (MRI)	10–100 μm	Anatomical and physiological imaging, high contrast, phenotyping of tumors, cell tracking
X-ray computed tomography (CT) imaging	50 μm	Lung, bone tumor imaging, need iodine contrast dye to differentiate between soft tissues
Ultrasound imaging	50 μm	Vascular, interventional, and embryo imaging using acoustic echoes
Intravital microscopy	1 μm	Confocal, multiphoton, imaging of all parts at high resolutions but at limited depths
Positron Emission Tomography (PET)	1–2 mm	Imaging of metabolic levels and patterns of molecules (sugars, amino acids, enzymes, neurotransmitters, etc.)
Fluorescence-mediated tomography (FMT)	1–2 mm	Imaging of targeted fluorochromes in deep tumors
Fluorescence reflectance imaging (FRI)	1–2 mm	Imaging of surface tumor molecular events
Single photon emission tomography (SPECT)	1–2 mm	Imaging of molecular probes and markers
Bioluminescence imaging (BLI)	Several mm	Cell tracking and gene expression
Optical imaging	–	Physiological and molecular imaging of body

perfect candidates for use as imaging devices. Nontoxic metals like gold, platinum, and iron can be conjugated or loaded into the interior of dendrimers in a variety of ways for use in biological systems with minimal to no toxicity. Though procedures for synthesizing these dendrimers vary based on how the metal interacts with the dendrimer in question, the resulting products are used in similar ways. Use of poly amido(amine) (PAMAM) dendrimers is common, as these have been well characterized and are known to be safe for use in biological systems. In regard to imaging, these dendrimers have properties that allow them to be used for biosensing and biolabeling, but it must also be noted that these dendrimer–metal conjugates, or nanocomposites, retain properties to serve other functions in an array of additional fields.

Dendrimers containing metal nanoparticles within their structures can be categorized into three groups. Inclusion of metal nanoparticles within the dendrimer structure by any of the following methods affords the dendrimer in question a variety of functions. The first category of dendrimers contains dendrimers known as dendrimer nanocomposites. These dendrimers rely on metal nanoparticles as integral parts of their structure. The second class consists of dendrimers with reactive groups that can bind metal ions to them. The third group is composed of dendrimers containing encapsulated metal nanoparticles within their structure [62].

Advances in technology have made it possible to photosensitize organic–inorganic hybrid polymer composites for use as photoconductive materials. The most commonly synthesized hybrid polymer composites are called dendrimer nanocomposites (DNCs). DNCs are composed of both inorganic and organic nanosized particles characterized by unique physical and chemical properties. The properties of these nanoparticles are distinctive due to the dispersion of inorganic guests at an atomic/molecular level and due as well to the properties of their organic hosts [63]. Techniques involving the use of DNCs for imaging continue to undergo rapid development.

It is also interesting to note that, due to the enhancement of the localized electric field surrounding metal nanoparticles, it is possible to alter the laser-induced optical breakdown threshold of a material. Lowering the breakdown threshold allows one to selectively target DNCs to release encapsulated drugs while avoiding tissue damage of the surrounding areas. Because the enhanced localized electric field is confined to a range of approximately 1 nm around the nanoparticle, it is also possible to use laser irradiation for breakdown of organisms within a cell through the nanoheating effect [64]. Use of an amino acid for stabilization of a gold nanocomposite allows for the ability to freely use and distribute the nanocomposite in water. This development allows for the use of nanocomposites in organic environments for biosensing and biolabeling.

Other methods of detection of nanocomposite dendrimers within a biological system include biomolecular detection through the use of light scattering spectroscopy, as subwavelength-sized nanoparticles show pronounced resonance for visible light in their scattering spectrum. Variations in a nanoparticle's size, composition, and/or shape as well as nanoparticle functionalization have the ability to alter nanoparticle plasmon resonance (NPP). Environmental (external) factors, including dielectric properties of the immediate environment of the nanoparticle as well as alterations in charge distribution, can also influence the NPP resonance position [65].

Gold nanoparticle clusters specifically have received a great deal of attention for their optical capabilities. By combining two commercially available chemicals, hydrogen tetrachloroaurate and sodium citrate, gold nanoparticles are easily produced [66]. Small gold clusters display strong UV absorption that decays exponentially as well as a broad superimposed band at ~500 nm (SP band), which decreases in energy and intensity as the size of the gold clusters decreases [67]. The size of the dendrimer greatly influences the structure of dendrimer–gold nanocomposites. Research by Satoh et al. has found that higher generations of poly(amidoamine) (PAMAM) dendrimers (G5.5) tend to have gold present in both the interior and exterior of the dendrimer structure, whereas dendrimers of G1.5 tend to have the presence of gold nanoparticles in the exterior of the dendrimer structure [68]. Gold dendrimer nanocomposites are nontoxic and therefore have the ability to be used in biological

systems. Iron and platinum are also commonly used in the synthesis of metal nanocomposite dendrimers and are safe for use in biological systems as well.

The use of silver metallic particles formed from silver salts in nanocomposite dendrimers has also been a focus of interest for researchers. Synthesis of silver particles and PAMAM dendrimers has been reported to occur by way of direct complexation as well as by substitution of copper in the preformed copper–PAMAM nanocomposite [63]. These nanocomposites are stable in polar solvents and, like gold nanocomposites, vary in primary structure dependent on characteristics of the dendrimer with which it has complexed. It must be noted that the presence of silver alters dendrimer structure as well. Further details regarding silver's structural effects on dendrimers can be found elsewhere [63]. Unlike gold nanocomposites, silver nanocomposites are toxic to biological systems and cannot be used within humans or animals.

The second class of metal-containing dendrimers consists of dendrimers with reactive groups that have the ability to attach metal ions. One of the more commonly used metals for biosensing and imaging in biological systems in this class is covalently bonded gold. The size of the gold (or any metal) clusters attached in this manner can be controlled by reaction conditions, as the kinetics of cluster growth are dependent on the amount of surface coverage of the dendrimer [69]. Binding of metal ions can also be electrostatic as well as coordinative.

The third class of dendrimers used for imaging consists of dendrimers containing encapsulated metal nanoparticles. Dendrimer structure entails the presence of large enclosed spaces at higher generations and, due to the prevalence of these interior spaces, it is possible to use them to enclose nanoparticles for either imaging or drug delivery. Metal particles encapsulated by a dendrimer can be delivered through the bloodstream or locally in the body for imaging of the localized area. It has successfully been demonstrated that dendrimers can be used to encapsulate nanoclusters containing only a few particles of metal [70]. One method of preparing dendrimers with encapsulated metal nanoclusters is by direct reduction. This method is only valid, however, if the corresponding metal ions possess the ability to be drawn into the interior spaces of the dendrimer and if the metal in question has a corresponding strongly complexed precursor ion. For this reason, Ag^{+} cannot be encapsulated by a dendrimer through use of the direct reduction method [71]. Metal nanoparticles encapsulated within dendrimers also do not undergo agglomeration due to the dendrimer's encasing them. It is also important to note that solubility differences between the dendrimer and the metal ions in solution help to drive the encapsulation of the ions into the interior of the dendrimer, as depicted in Fig. 10 [72].

Encapsulated gold nanoparticles have the ability to be used as optical agents within biological systems. Gold particles possess large polarizability in the optical region; therefore, steps may be taken to visualize gold particle aggregates within a system. Two methods may be used for visualization of these particles: plasmon polariton of Rayleigh resonance (RR) or surface-enhanced Raman scattering (SERS) [73]. Similar to dendrimer encapsulated nanoparticles, lipid-encapsulated nanoparticles have also successfully been used as ultrasonic contrast agents; they are stable in the bloodstream and circulate for about 1 h [74].

3.2. Optical Imaging

Optical imaging makes use of the variance in optical properties possessed by different tissues and fluids within the body. Optical imaging allows for physiological and molecular imaging of mounted/prepared cells and tissue samples as well as *in vivo* imaging of biological systems. The drawback to using an optical imaging method is the inability of the majority of these methods to penetrate tissues to any substantial depth [75]. Because most of these imaging methods are to be used *in vivo*, the thickness or depth of the tissue is of concern, as thickness prevents fluorescence from being easily observed.

However, due to the flexible properties of the fluorochromes used in optical imaging, it is possible to either use these fluorochromes alone or to link these imaging units with antibodies or other ligands to make their detection abilities especially versatile. Given that tumors and other malfunctioning areas of the body overexpress different receptors, it is possible to use these directed fluorochromes to target toward and detect areas of pathological injury.

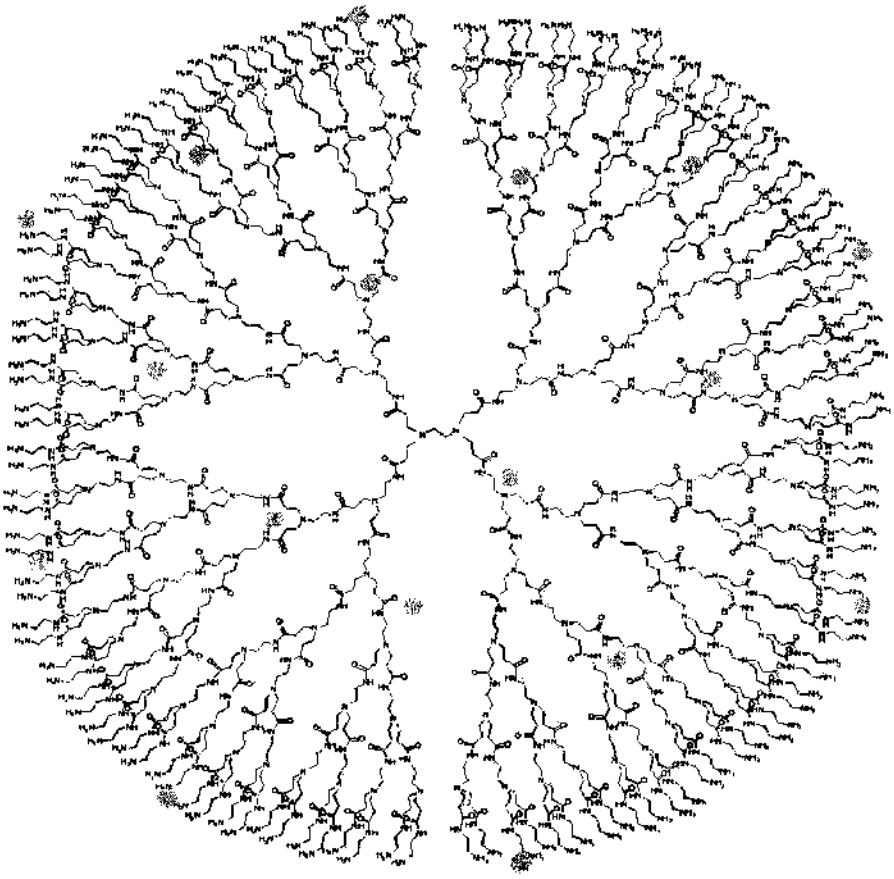


Figure 10. Image of encapsulated metal particles by dendrimer to demonstrate how the metal particles are enclosed.

Having stated this, there are two main types of fluorochromes used for optical imaging: they are termed nonspecific and specific (or targeted) fluorochromes.

To observe optimally fluorescence through the use of optical imaging, fluorochromes in the near-infrared range (NIR) of 650–900 nm must be used [75]. These wavelengths provide the best imaging capabilities, because light in this range can cross through tissue very efficiently, and absorption of these waves is very low in hemoglobin and water, therefore allowing for clear visualization of fluorophores in tissues. Little absorption of these wavelengths by hemoglobin and water also prevents background (auto) fluorescence.

Nonspecific fluorochromes, as their name indicates, are used to image nonspecific areas of tissue. These fluorochromes are not targeted toward any specific receptors; therefore, they are used only to generally image tissues. Nonetheless, certain conditions must be met for these fluorochromes to be visible within a biological system. Foremost, as mentioned before, the most successfully imaged fluorochromes have excitation and emission maxima that fall within the NIR spectrum of 650–900 nm. It is also necessary for these probes to be hydrophilic, to possess low toxicity, and to have a high extinction coefficient, high biocompatibility, and a strong fluorescent signal. To avoid imaging problems or clarity between tissue density/types, it is best to use a fluorochrome that exhibits no photobleaching or photosensitizing effects.

Cyanine dyes are the most commonly used nonspecific fluorochromes [75]. There are many variations of these dyes, each best functioning in different areas of the body. For example, indocyanine green (ICG) is best used for imaging of the hepatic, ophthalmic, and cardiac regions of the body. Nonspecific fluorochromes have also proven to be beneficial in tumor detection as compared to the use of MRI with a gadolinium contrast agent.

The second class of optical contrast agents possessing targeting capabilities are labeled "specific" or "targeted" fluorochromes. These fluorochromes become visible after cleavage of a linker between the ligand (which binds to the overexpressed receptor) and fluorochrome.

They allow for imaging of tumors and also possess the ability to detect tumor grading. Due to the binding properties of the ligands (attached to the fluorochromes) with the receptors (overexpressed by tumors), it follows accordingly that the more progressed a tumor becomes (higher grade), the larger the number of receptors, therefore leading to an increased number of ligands binding and, accordingly, more fluorescence given off. This enhanced fluorescence allows for differentiation between tissue types as well.

Unlike other imaging methods, which do not have the ability to differentiate between molecules within a system, targeted optical contrast agents can be used to detect enzymatic activity and enzyme concentration differences in tissues. Activatable fluorochromes are designed such that a change in their molecular conformation (altered by enzymatic activity) activates their fluorescent capabilities. Different enzymes are involved in a range of pathologies [75].

Conformational change, hence "activation" of a fluorochrome by one of these said enzymes, acts as an indicator of the presence of pathology within the body. Activation occurs via cleavage of the fluorochrome from the ligand with which it was joined. These fluorochromes do not fluoresce in their native state due to quenching that occurs between fluorochromes via energy resonance transfer. Upon cleavage, however, this resonance energy transfer no longer occurs, allowing the fluorochromes to fluoresce within the system [76].

Extensive research has gone into the creation of a cathepsin-B–sensitive optical probe. Cathepsin-B has been shown to be up-regulated in areas of necrosis, inflammation, and angiogenesis [76]. Therefore, the creation of a probe that senses protease over expression can be used to indicate the presence of a cancerous growth. It has been shown that protease imaging has been successfully applied to imaging of autoimmune disease activity in a model of inducible rheumatoid arthritis. Inflammatory cathepsins were used to detect the inflammatory response of arthritis by injecting wild-type mice with an immunogenic serum. Near-infrared fluorescent (NIRF)-reflectance imaging using a cathepsin-B–sensitive probe allowed for detection of inflammatory activity before any outward clinical manifestation of the disease [75]. The ability to detect protease overexpression before outward symptoms of pathology occur allows for the possibility to alter greatly the course of one's clinical outcome if one can be treated accordingly in time.

A subsequent division of targeted fluorochromes includes injectable fluorescent probes and photoproteins. This class of proteins gives off photons following oxidation or oxidative conversion of a substrate such as luciferin by luciferase. This process usually requires ATP, oxygen, and catalysis, but due to the fact that luciferase is not present naturally in mammalian systems, for example, use of this substrate for detection allows for imaging without background fluorescence (noise) [75]. Photons released by this process can be detected using sensitive charge-coupled device technology. Bioluminescence imaging systems are also used for detection of luciferase-released photons *in vivo*. Luciferase is detected in biological systems by light emittance, which is visible wherever luciferase is expressed within a system.

The development of optical tomographic imaging techniques offers the prospect of providing a lower detection limit for fluorochromes and a higher spatial resolution, but a second imaging modality is still required to be able to image an anatomical background [77].

3.3. Magnetic Resonance Imaging

Nanotechnology has greatly influenced the capabilities of imaging systems today. Magnetic resonance imaging (MRI) has particularly benefited from the abilities nanotechnology has to offer. Molecular engineering has allowed for the creation of targeted nanoparticle contrast agents to help distinguish between abnormal and normal tissue and cell types. These contrast agents can also be conjugated with a chemotherapeutic agent, which allows for treatment and visualization to occur simultaneously. Use of nanoparticles as contrast dyes in MR imaging has been on the upsurge and will eventually become equal in use or surpass today's conventionally used contrast agents.

Traditionally, magnetic resonance imaging has relied on gadolinium-based dyes for the ability to improve contrast and differentiate between normal tissue and cancerous masses, aneurysms, blocked vessels within the body, and so forth. The problem, however, with

gadolinium-based contrast agents is that though they have the ability to provide contrast between normal tissues and otherwise, these agents also have the tendency to equilibrate through interstitial space very rapidly, therefore removing a great deal of the precision magnetic resonance (MR) imaging has the ability to provide [78]. MR imaging, regardless, provides high spatial and temporal resolution images (up to 100 μm) using a noninvasive method. There are a variety of contrast dyes available for use in MR imaging; many use dendrimers as well as superparamagnetic nanoparticles. These dyes are imaged via energy transfer between the tissues in question and the magnetic resonance imaging unit.

Superparamagnetic nanoparticles have the ability to be used as contrast agents for imaging in biological systems [79]. These nanoparticles are named such due to the magnetic properties they possess when placed within a magnetic field. However, they do not possess any magnetic properties when removed from a magnetic environment. These particles can be used in MR imaging for detecting irregularities (malignant or benign growths) in the liver, spleen, and lymph nodes. Accumulation of magnetic nanoparticles occurs in target tissues, and degradation of the particles occurs over a period of days. A darkening effect is observed on T2-weighted images of these systems due to cells within the system having undergone phagocytosis [80]. Use of these nanoparticles allows visualization of metastases as small as 1 mm within lymph nodes.

More specifically, use of iron oxide nanoparticles for MR imaging allows for a more distinct delineation between normal and injured areas than with the use of gadolinium as a contrast dye. Iron oxide particles have a long half-life within blood, but, unlike gadolinium-based contrast agents, these particles will not travel to extravascular areas. Though the injection of these nanoparticles does not help differentiate normal and ischemic tissues using T1-weighted images, use of T2-weighted imaging presents a very clear contrast between normal and ischemic tissues in animal studies as well as in human trials.

Further enhancing the capabilities of nanoparticle contrast agents are targeted nanoparticle contrast agents. These agents have been developed to attach to specific tissues or cells for detection of atypical attributes. The ability of these contrast agents to bind to particular sites allows researchers the utility of a contrast agent that not only allows them to image a localized area but also to quantify the amount of injury present. An example of these targeted nanoparticle contrast agents are tissue factor (TF)-targeted nanoparticles. These nanoparticles bind to smooth muscle cell membranes *in vivo* and have the ability to detect differentiation in these membranes due to restenosis, for example, where vascular smooth muscle cells proliferate and move from the media to the intima of a vessel wall [81]. Though traditional contrast agents allow for visualization of a localized area, targeted nanoparticle contrast agents allow for the distinct visualization of specific tissues or cells.

MR contrast agents containing PAMAM dendrimer cores are widely being developed for use in vascular imaging due to their ability to stay in blood circulation for extended periods of time as well as due to their greatly enhanced relaxivities as compared to gadolinium [82]. Often coupled to these dendrimer cores are perfluorocarbon nanoparticles. These nanoparticles are attached to dendrimers or antibodies and then delivered locally or through the bloodstream to tissue sites for imaging by MRI. When coupled to antibodies for delivery, these nanoparticles are then activated at injured tissue sites for imaging purposes.

Dendrimer size determines the area in the body in which it can be used for imaging. These dimensions also determine which area of the body will be affected most by excretion of these polymers from the system. Generation number distinguishes dendrimer size—the larger the generation number, the larger the dendrimer. MR contrast agents with a G6 PAMAM dendrimer core have been shown to act in the blood pool and have the capability of imaging blood vessels of mice at 100-μm resolution while exhibiting no renal excretion. However, these large generation dendrimers have also shown to accumulate in the liver [82]. Lower generation dendrimers (smaller in size), on the other hand, accumulate in the kidneys instead of the liver, unlike higher generation dendrimers and gadolinium. To distinguish between the uses of each polymer based on its size, larger generations of PAMAM dendrimers are used for visualization of blood vessels, whereas smaller generations of PAMAM dendrimers are used as contrast agents for kidney tubules, and so forth [82].

Research has ensued during the past 20 years to make advances in magnetic resonance imaging techniques [83–95]. There have been, however, some roadblocks. Physical properties of the dendrimers used as delivery devices for the nanoparticles can limit the use of these polymer conjugates in clinical practice. Though the imaging capabilities of nanoparticles as contrast agents are comparable and better than currently used dyes, use of higher generation dendrimers as the delivery device for these nanoparticles has not always been the best policy in human systems, as they collect in various areas and take a substantial amount of time to degrade in the system. This problem can lead to toxicity in the system in amounts larger than currently experienced with dyes used in standard practice today.

4. TARGETING

A cell becomes cancerous through the disruption of its normal growth and differentiation, initiated by genetic damage that leads to unregulated DNA replication and cell proliferation. Currently employed cancer therapy strategies exploit drug- or radiation-induced interruption of cell proliferation *via* arresting cell division through inhibition of one of the biochemical pathways of cell division. The commonly used drugs either inhibit DNA replication or interact with microtubule assembly that regulates cell division. The selectivity of these conventional cancer therapies simply relies on the property of a malignant cell to replicate faster than a normal cell. Obviously, this is a poor strategy because all dividing cells will be destroyed by these treatments, especially the relatively rapidly dividing normal cells such as the hair follicle, gastrointestinal epithelial cells, and bone marrow cells. The destruction of these normal cells may lead to toxicity, depending on the severity of normal cell damage. In addition, some of these drugs have organ-specific toxicity such as cardiotoxicity. Another drawback is that the chemotherapeutic drug has to be maintained in circulation for sufficient time and at adequate concentration to catch all the tumor cells at a specific point of their cell cycle at which the drug interacts. Furthermore, although cancer cells initially respond to the chemotherapeutic agent, relapse often occurs because the cells that survive the treatment become drug-resistant, further aggravating the cancer, owing to their lack of response to multiple drugs. Despite these severe drawbacks, some chemotherapeutic regimens have successfully been employed to cure certain types of childhood leukemias [96].

Conventional cancer therapy has never achieved complete cures in all patients in any type of cancer. Therefore, targeted drug release has been a subject of extensive research in the past two decades. This fast-developing oncological therapeutic approach is to target specifically the tumor cells and minimize any collateral damage by sparing healthy cells, acting similarly to a guided missile. This is achieved by identifying specific receptors in cancer cells that are qualitatively or quantitatively different from normal cells and targeting these molecules with a targeting agent linked to a drug carrier and therapeutic drug specifically designed to disrupt the cell cycle, thereby killing the cell. Apart from the specificity, targeted drug delivery overcomes the dilution effect of the drug in the body by effective homing of the drug into the tumor and maximizing the amount released into the cancer cell. In addition, because of the high specificity of the targeting agent and controlled release of the drug, the effective dose of the drug can be reduced by several orders of magnitude, preventing toxic side effects and lessening drug resistance.

There are two main types of cancer targeting. In the first type, intervention of an unregulated growth-promoting activity of the cell by a drug or an antibody inhibits the cancer cell growth. There are 50 or more identified cancer genes that have altered DNA sequences, the protein products of which can be possible targets by direct inhibition of their function or by blocking another downstream signal transduction pathway that was initiated by this altered protein. An example of this type of targeting is the usage of a tyrosine kinase inhibitor or an antibody to inhibit the growth of cancers caused by increased expression of epidermal growth factor (EGF) receptor and tyrosine kinase activity [94] or Abl protein [72].

In the second type of cancer targeting, a cell surface receptor or antigen, for example, expressed specifically or in excess, is targeted by the corresponding ligand or antibody. The ligand is linked to a chemotherapeutic drug with or without using a carrier platform and delivers the drug specifically into the cancer cell to induce cytotoxicity. Use of a carrier

platform allows the fabrication of a multifunctional device that can also carry other devices such as an apoptosis sensor and an imaging unit. The drug can be any cytotoxic drug that initiates programmed cell death, and the specificity of the drug effect depends on the specificity of the cell surface target. In the next paragraph, we discuss the second type of targeting.

There are several criteria for proper selection of a targeted drug delivery system (DDS). The targeting component of the DDS relies specifically on characteristics of the cancerous tissue to be targeted. The targeting component of the system must target a specific cancer signature on the tumor. The DDS must be soluble in aqueous body fluids such as the blood and urine and must maintain its stability in blood while remaining nontoxic and nonimmunogenic. The DDS must be able to penetrate through tissue barriers, including vasculature, and should have the ability to migrate into the interstitial fluid to reach the cancer cells. Along with these characteristics, the DDS must also be able to be endocytosed into the cancer cell and release the drug into the appropriate cell compartments to initiate apoptosis. The exogenous components of the system must also have the ability to be excreted from the body with or without undergoing biodegradation, and the monomer components of each macromolecule must be nontoxic so that, if the macromolecule undergoes biodegradation, the monomeric units will not cause toxicity to the system. Although a "perfect" system that meets all the above criteria has not been yet fashioned for any type of tumor targeting, several agents that meet many of the above criteria have been described. The different types of platforms used for drug targeting have been given elsewhere in this chapter. Here we describe the cancer-specific targeting molecules that have been tested. These can be categorized into (1) a small molecule such as folic acid (FA), (2) a protein such as the EGF, and (3) an antibody such as Herceptin [human EGF receptor 2 (HER2)-binding antibody].

4.1. Folate Receptor: Small Molecule Targeting

The vitamin folic acid serves as a one-carbon source for synthesis of several intermediary metabolites, most importantly for the synthesis of nucleotides required for DNA synthesis and cell growth. The FA is taken up into a cell either through a low-affinity reduced folate carrier (RFC) or through a high-affinity FA-receptor (FAR). The RFC is ubiquitously expressed in all cells and carries the naturally occurring reduced folates. FAR is expressed in low levels in normal cells and preferentially carries the oxidized FA into cells. Because of the high affinity and the cancer cell specificity, FA is the more suitable targeting agent as compared to reduced folates. There are three types of FAR: the α, β, and γ isoforms. The α and β isoforms are glycosylphosphatidylinositol (GPC)-anchored membrane proteins, and the γ isoform is a soluble protein. The α and β isoforms are membrane receptors respectively identified in epithelial and nonepithelial cells. The α isoform is overexpressed in several human carcinomas, even up to a 100-fold [96–100] containing FA. This provides striking selectivity for targeting carcinomas using drug conjugates containing FA.

In addition to the tumor specificity of FAR-α in human carcinomas, there are other advantages for the selection of FA as a targeting agent. Expression of FAR in transformed epithelial cells is relatively enhanced on the basolateral surface ("blood side") of cancer cells as compared to the predominant apical ("tissue side") expression in normal cells, which complements the cancer cell specificity of FA [101]. The small molecular size of FA allows easy tumor penetration and favorable pharmacokinetics. FA is easily available and is inexpensive, and the chemical conjugation of FA to a polymer such as a dendrimer and the purification of the FA–conjugate complex is relatively simple. FA binds to its receptor with high affinity ($K_d = 0.1$ to 1 nM) and is largely nonimmunogenic. The tumor cell density of FA receptors tends to be elevated during tumor progression, providing increased therapeutic potential during later stages of the cancer. Because of these advantages, FA has widely been studied for the targeting of bioactive agents such as protein toxins, oligonucleotides, plasmids, liposome-entrapped drugs, radiopharmaceutical agents, and MRI agents [100, 102–113].

FA is internalized into cells through receptor-mediated endocytosis of the FA–FAR complex into endosomes, trafficking and releasing of FA into the cytosolic compartments, and recycling part of the endocytosed FAR back into the plasma membrane [114–116]. FA can be conjugated at its γ and α carboxyl group without losing the ability to bind to a FA-receptor

and to internalize into cells [117]. A variety of folate-conjugates have been synthesized and shown to be endocytosed into FA-receptor expressing cells [101, 102, 118–121]. Although the trafficking pathway of FA has been investigated in multiple studies, a precise mechanism has not been elucidated. Trafficking through both clathrin-coated pits and caveolae has been reported. The trafficking mechanisms of any of the FA–drug conjugates and the releasing of folate and other moieties from the conjugates are not known. Nevertheless, FA–drug conjugates have been used for *in vivo* targeting. ^{99}Tc conjugated to FA specifically concentrates in FAR-expressing xenograft tumors [112, 121]. In clinical studies, the radioactive folate conjugates containing ^{111}In diaethylenetriamine penta-acetic acid, (DTPA)–FA, has been shown to concentrate into malignant ovarian tumors with only minimum uptake of the conjugate into benign tumors [101].

FA can be directly conjugated to several agents such as a radiopharmaceutical, a chemotherapeutic drug, or protein toxins for delivery into the cells [112]. The direct conjugation of FA to a biological molecule has the limitation of not being able to conjugate other functions (e.g., an apoptosis-detecting agent) onto the bifunctional conjugate. In addition, direct conjugation may lead to the loss of biological activity of the attached molecule [122]. Therefore, a polymer carrier onto which several functions such as a fluorescent imaging agent and an apoptosis-detecting agent have been added in addition to FA and a chemotherapeutic drug will have the advantages of being able to deliver several functions simultaneously.

An FA-targeted liposomal drug delivery system has been employed specifically to deliver drugs into cancer cells *in vitro* and *in vivo*, increasing the therapeutic index of the drug and overcoming MDR [107, 123]. FA-conjugated liposomes of a diameter of 60 nm endocytose into FAR-expressing KB cells [106]. The liposomal targeting system has the disadvantage of being removed by the mononuclear phagocyte system (MPS). Liposome recognition by MPS can be prevented by surface modification of the liposomes. However, such modification may result in reduced affinity of the FA conjugated to the liposomes. Moreover, *in vivo*, FA-targeted liposomes possess poor target specificity, as nontargeted liposomes are also taken up by cells.

In addition to liposomes, other polymers have also been used as platforms for FA-targeting. FA covalently linked to PEG-coated poly(amino-ethylene glycol)cyanoacrylate-*co*-hexadecyl cyanoacrylate showed increased avidity for FAR [110]. FA conjugated to the water-soluble polymer *N*(-2 hydroxypropyl) methacrylamide (HMPA) is also taken up into FAR-expressing cells [101]. The drug doxorubicin, covalently linked to HPMA through a peptide spacer, is specifically hydrolyzed by lysosomal enzymes and is released into the cytosol to induce cytotoxity.

PAMAM dendrimers with well-defined and highly branched structures contain multiple surface primary amino groups ideal for conjugation of several molecules of FA and multiple other functions onto the surface amino groups of the dendrimer. FA-conjugated PAMAM dendrimers bind and internalize into FA-receptor–expressing KB cells. Competition assay with free FA shows that the PAMAM dendrimer device has a similar affinity for the folate receptor as compared to free FA. Therefore, the affinity of FA is not lost by conjugation with PAMAM dendrimer. The binding of the dendrimer device is relatively acid-resistant as compared to the binding of free FA. This is probably due to the increased avidity caused by multiple anchoring of the device to two or more adjacent receptor binding sites. Multivalent interactions of FA moieties have also been demonstrated in FA linked to PEGylated cynoacrylate nanoparticles [110]. Multiple anchoring does not seem to influence internalization of the dendrimer, as confocal microscopic analysis indicated it has cytosolic and nuclear localization [111].

In vivo studies have shown that G5-FA–6TAMRA conjugate concentrates into xenograft KB cell tumors in nude mice with the dendrimer (Fig. 11). As given above, FA has several advantages as a tumor-targeting agent. FA-mediated targeting systems have great clinical potential because of the overexpression of FAR in a multitude of human cancers, and FA-based nanoparticles may serve as beneficial agents for diagnosis and treatment of cancer.

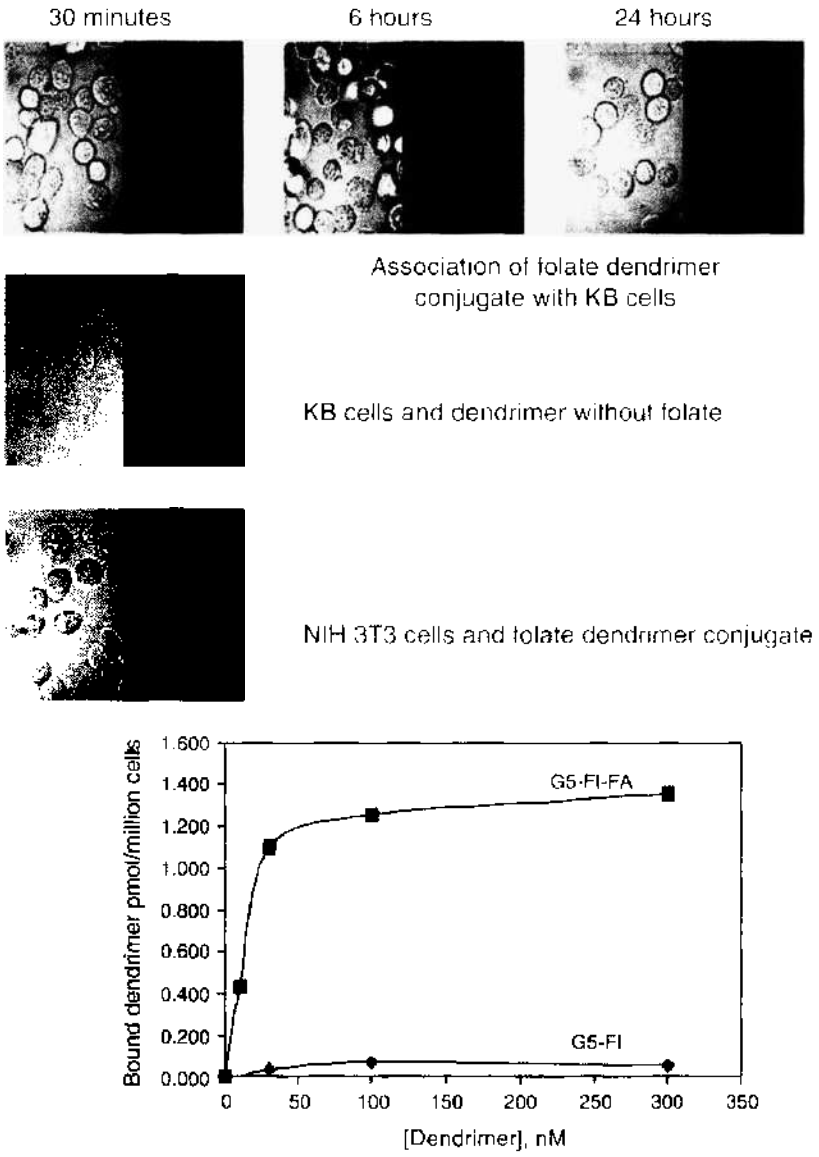


Figure 11. Association of folate dendrimer conjugate with KB cells and imaged by fluorescein. (Quantitation of bound dendrimer on KB cells by two-photon fluorescence measurements through optical fibers. KB cells were trypsinized and rinsed with folate-free media in the absence of serum. Aliquots of 5 million cells suspended in 0.5 ml media were incubated with different concentrations of the dendrimers, rinsed 2× with media, and the cells were spun down in a microfuge. Two-photon optical fiber was inserted into the cell pellet, and the number of photons emitted was quantified using a photon counter. The values obtained were converted to pmols from standard curves generated for the two dendrimers.)

4.2. Protein- and Peptide-Based Targeting

Several growth factor receptors are known to be relatively overexpressed in cancer cells as compared to control cells, making them suitable for targeting using the corresponding ligand as the targeting agent. These include cell membrane receptors for ligands such as epidermal growth factor [123], vascular endothelial growth factor [124], and transferrin [125]. Nonetheless, as many of these receptors are also expressed in normal cells and are involved in normal cell growth regulation, targeting these receptors offers poor cancer cell specificity and results in inhibition of normal cell growth. Moreover, conjugation with the intact protein could result in alteration of the tertiary structure of the protein, leading to decreased affinity for the receptor.

Tumor neovascularization, the formation of new blood vessels, is crucial for the progression of the growth of a tumor exceeding a volume of 1 mm^3 due to the otherwise lack of nutrients and oxygen available to the growth. There are several advantages for targeting

newly forming vasculature in a tumor as compared to targeting the cancer cell itself. The vasculature is directly accessible from the blood, and, by limiting the targeting agent from having to travel within the positive pressure of the tumor interstitial fluid, the drug treatment can often be more effective. As a multitude of tumor cells depend on a single blood vessel for nutrition and oxygen, shutting down the vessel can kill all of the cells dependent on the vessel to receive nutrients. As compared to the rapidly transforming tumor epithelial cells, vascular endothelial cells are genetically stable and therefore are less prone to drug resistance. Moreover, as newly formed vessels can be specifically targeted, the toxicity for existing vasculature is minimal.

Several proteins have been identified to regulate the neovascularization process. These proteins are therefore possible target molecules for targeted drug therapies. These proteins include endothelial specific growth factors, antigens, and so forth [126, 127]. Proteins classified as integrins have been identified as crucial for mediating neovascularization [128]. The growth factor VEGF is known to regulate the initiation of neovascularization and seems to have no identified functional role in normal vascular cells. Due to these characteristics, this growth factor has been an attractive target molecule in drug therapies combating newly forming tumor vasculature [124]. One drawback of targeting the vascular cells is that, as the cancer cells are not directly targeted and killed and only made to be deprived of nutrients, the effect on the cancer cell may be more cytostatic rather than cytotoxic. In addition, tumor cells have the ability to stay dormant and alive under hypoxic and nutrient-deprived conditions by down-regulating energy consumption and by up-regulating enzymes such as mitochondrial hexokinase.

Targeting of cancer cell–specific receptors with peptide fragments (with <100 amino acids) has several advantages. Peptides are smaller in size compared to intact proteins; therefore, peptide-based targeting produces better tumor penetration and better pharmacokinetics such as faster blood clearance, resulting in increased tumor specificity. Peptides are generally nonimmunogenic and can be synthesized and conjugated to a polymer easily without undergoing denaturation. Because of their smaller size, multiple peptide molecules can be conjugated onto a dendrimer, resulting in multivalence and increased avidity of the conjugate. In certain types of cancers, the receptors for some of the naturally occurring peptides such as somatostatin, vasoactive intestinal peptide, bombesin, and leutenizing hormone-releasing hormone (LH-RH) are overexpressed, making these suitable for tumor targeting [129–132]. These naturally occurring peptide targets nonetheless will have low tumor specificity, and therefore identification of tumor specific peptides is important for tumor selectivity.

One strategy to identify specific tumor receptors on cancer cells is by using random peptide phage libraries. In this procedure, random peptide phage libraries are allowed to internalize into a cancer cell *in vitro* by repeated cycling steps of cellular internalization, washing of surface bound phages, and amplification of the internalized phage in bacteria. The phage DNA purified from the final bacterial preparation is sequenced and the corresponding peptide identified. A recent study using such a phage display technique identified a peptide with a major core motif with the amino acid sequence LTVXPWY, which binds to several cancer cell lines [133]. Another peptide similarly identified to bind to the ErbB-2 receptor is KCCYSL [134].

Once a tumor-specific peptide has been identified, it has to be determined if the peptide is stable in the blood and is not readily hydrolyzed by peptidases. If the peptides are susceptible to plasma peptidases, the peptides have to be modified to inhibit the degradation by peptidases. The modification procedures include "end-capping" to inhibit hydrolysis by exopeptidases, substitution of an L-amino acid with a D-amino acid or an imino group to prevent hydrolysis by endopeptidases [135–137]. Such procedures should not result in losing the binding characteristics of the peptide on the receptor. As the flexibility of the amino acids may be important for the binding of the peptide to the receptors, introduction of a spacer (such as an additional small peptide) between the active peptide and the polymer is important. As in the case of FA, proper internalization of the peptide–conjugate is important for intracellular drug delivery. Therefore, *in vitro* studies have to be conducted to monitor the binding and internalization of the peptide–conjugate.

As the molecular size cutoff for kidney filtration through the glomeruli is 50 kDa, conjugates below this size are filtered through the kidney for elimination. However, filtered peptides and proteins are quantitatively reabsorbed back in the proximal tubules through a process called pinocytosis. If the peptide- and protein-conjugates are also reabsorbed into the proximal tubules in a similar fashion, this could result in accumulation and cytotoxicity within the kidney. *In vitro* and *in vivo* studies are needed to verify this. Peptide reabsorption has been shown to be reversed by administering a pharmacological dose of a basic amino acid [138, 139].

Protein- and peptide-based targeting has emerged as a new strategy for tumor-specific drug delivery. Improving the stability of proteins through modifications and identification of tumor-specific peptides will enable the development of a new class of anticancer therapeutic drug conjugates.

4.3. Antibody Targeting

As several cell-surface antigens have been identified to be overexpressed in cancers, antibody-based targeting has extensively been pursued in the past two decades [140, 141]. The most commonly used antibody for targeting is the IgG class, having a molecular weight of 155,000 Da. Several tumor-selective antigens have been used for targeting by the antibody either directly or by use of antibody-drug conjugates. These include antibody conjugates targeted against human EGF receptor 2 (HER2)-neu [142], carcinoembryonic antigen (CEA) [143], mutated p53 [144, 145], and cell adhesion molecules such as integrins and selectins [146–148], mucins [149], and several other extracellular and intracellular proteins [141]. Although *in vitro* studies have proven to be promising in several of these studies, the extension of these observations to application *in vivo* resulted only in limited targeting efficiency.

Antibody targeting sometimes has the advantage of inhibiting cell growth by two independent mechanisms. Certain antibodies such as Herceptin are growth-inhibitory to the cells, caused by the inhibition of receptor-mediated signal transduction [97]. In addition, targeting of certain antibodies to the tumor cells causes these cells to be recognized by killer cells. Therefore, conjugation of a therapeutic drug to such antibodies has the potential of yielding the dual therapeutic actions of direct growth-inhibition and apoptosis induced by the delivered drug.

There are several drawbacks in selecting an intact antibody as the targeting molecule. The tumor-to-blood ratio of an antibody is normally very low. The Fc region is the primary recognition site for the immunogenicity of the antibody and for uptake into the liver, which lead to a host immune reaction and removal of the antibody conjugate in the liver. Antibodies are large molecules, and further increasing the size of the drug delivery device by conjugation to a polymer results in poor tissue penetration and cellular endocytosis. Although the tumor vasculature is sometimes leaky enough to allow entry of a large molecule such as the antibody–polymer complex, the degree of leakiness varies broadly and at several regions may not allow the complex to be uptaken passively. The immunogenicity can be overcome by "humanizing" the antibody by recombining the immunogenic sites of the animal-derived antibody through genetic engineering technology. The immunogenicity can also be reversed by coupling with polymers such as PEG. These procedures will allow augmenting the circulation time of the complex. Nonetheless, such protein modification may affect the antigen recognition site. One alternative to using an intact antibody is using antibody fragments such as the divalent $F(ab')_2$, the monovalent Fab, and the single-chain Fv segment. These smaller-sized fragments, in addition to being smaller and providing better pharmacokinetics, will have reduced immunogenicity due to the removal of the Fc region. Immunoliposomes conjugated with EGFR-binding antibody fragments have recently been used to target EGFR-expressing cells [150].

One limitation to antibody targeting is that the antigen is generally heterogeneously expressed in tumor cells with multiple genetic variants of the protein and with a different glycosylation pattern, leading to a lack of response of these tumor cells to therapy. Moreover, during tumor progression, the antigen is released from the cells and becomes nonresponsive to the antibody-based targeting. Antibody targeting of tumor vasculature in xenograft

human tumors in animals can pose experimental difficulties as the endothelial cells in these tumors are derived from the host and may not cross-react with a human antibody [151]. The conjugation of an antibody to a polymer is achieved through attaching the polymer to the functional groups of the antibody such as lysine-NH_2, with or without using a linker. Substitution of functional groups at the antigen-binding site of the antibody can result in a substantial decrease in the affinity of the antibody onto the cell-surface antigen. Moreover, the conjugation of the polymer to the antibody can result in a loss of antibody solubility and aggregation.

Another application of the antibody targeting is the antibody directed enzyme prodrug therapy, or ADEPT. In this approach, an inactive precursor drug (prodrug) and an enzyme that can cleave the prodrug to the active free drug is identified. The enzyme is conjugated to an antibody that in turn is capable of targeting a corresponding antigen on the tumor cell. The antibody–enzyme conjugate is administered systemically and allowed to bind specifically to the tumor. When the unbound circulating conjugates are cleared after a waiting period, the prodrug is administered to bind specifically to the catalytic site of the enzyme that is located in the tumor, to be cleaved to the active drug. Several enzyme/prodrug combinations, using the ADEPT strategy, have been subjected to *in vivo* and clinical trials to achieve the increased therapeutic index (the ratio of the dose that gives maximum tumor volume reduction to the dose that produces toxicity) as compared to the drug alone [141].

Recent identification and approval of several antibodies such as Herceptin and Ritoxan [97] relatively specific for certain types of tumors supports the need for further investigations on designing novel antibody-based drug delivery systems. Identifying tumor-specific antigens and optimizing the pharmacokinetics by using appropriate antibody fragments will improve the therapeutic index of a targeted drug.

5. DRUG DELIVERY

The *in vitro* and *in vivo* applications of drug delivery systems have expanded exponentially in the past decade. Multiple polymers with defined structures simulating a natural biological macromolecule are currently being tested as platforms for spatially and temporally controlled drug delivery. The structure and synthesis of these polymers have been discussed in a previous section, and here we describe the utility in delivering drugs into cancer cells.

In cancer therapy, a DDS is delivered either locally at the tumor site or systemically into the circulation. In the local delivery protocol, a DDS such as a drug-loaded microsphere is implanted at the primary tumor site, and the drug is delivered in a controlled fashion. As the drug is delivered locally, there is no requirement for a specific targeting agent. The DDS can be injected directly into the tumor or infused into the tumor vasculature, using an implanted or external pump, so that the polymer is released at the tumor site in a controlled manner. Such local drug delivery reduces systemic toxicity and provides high local drug levels without undergoing degradation in other organs. Locally delivered drug-loaded polymers show increased efficacy to regress tumor as compared to free drugs [152–154]. A magnetic targeted system in which the drug is absorbed onto iron magnets and released at the tumor site under the control of an applied magnetic field has been subjected to clinical trials [155]. The local delivery of DDS has the limitations of difficult tumor site accessing and not being able to deliver the drug to malignant cells that might have migrated beyond the delivery site.

In systemic delivery, the DDS is delivered through the circulating blood to all tumor cells, both at the primary tumor site and metastases. There are two different categories of systemic delivery systems. In the first category, the drug is delivered in a polymer to achieve slow and sustained release of the drug into the tumor. An example of such slow-releasing systems is Zoladex, which is effectively applied for the local release of the hormone LH-RH and has been shown to delay progression of prostate cancer [156]. Although such sustained-release delivery systems can be synthesized easily, they do not have any tumor specificity and therefore can lead to side effects from the parent drug that is felt throughout the system.

In the second type of systemic drug delivery, the DDS carrying the drug is injected intravenously for delivery into the tumor cells. In order to target systemically a drug into a tumor cell, there are several structural and physiological abnormalities of tumors that must be

considered. One primary structural defect is presented by the tumor vasculature itself, which has a variety of irregularities differentiating it from normal vasculature. The endothelial cells that compose tumor vasculature do not adhere tightly enough to prevent passive leakage of blood components. The cutoff size of these tumor vasculature pores has been estimated to be around 400 nm. Tumor vasculature is also not well developed and often lacks an endothelial lining and a basement membrane. The passive diffusion of the interstitial fluid in the tumor is greatly compromised because of the lack of a proper lymphatic network. For these reasons, a nontargeted polymer–drug conjugate can also enter the cancer cells and get trapped inside, a process called "enhanced permeability and retention," or the EPR effect. Clinically, these tumor abnormalities serve as an advantage because drugs can be delivered passively without the requirement for a specific targeting agent. Cytokines such as the tumor necrosis factor μ and interleukins have been used to increase the vascular permeability and tumor diffusion of a macromolecule [156–158]. Nonetheless, drug delivery through EPR is inefficient because of tumor heterogeneity and the absence of the EPR effect in large portions of the tumor.

Systemically administered DDS containing a tumor-specific targeting molecule can complement the EPR effect to deliver the drug into the tumor cells through binding and internalization of the DDS. Even in the absence of an EPR effect, the drug can be preferentially delivered by the tumor vasculature through tight binding on the tumor cell receptor of the targeting molecule. Nonetheless, one major constraint that needs to be overcome is the increased interstitial pressure of the tumor and the consequent lowering of the diffusion of the polymer along the interstitial space to the tumor cell. This becomes further aggravated with the increase in size of the drug delivery agent. For this reason, smaller-sized polymer conjugates are preferred for drug delivery, and a peptide fragment of a protein or an antibody is preferred over use of an intact macromolecule as the targeting molecule.

When systemically administered, nanoparticles are taken up by the mononuclear phagocyte system (MPS) of tissues such as the liver, spleen, lungs, bone marrow, and blood unless the nanoparticle is surface-modified to prevent recognition by the MPS. The degree of clearing of a drug-conjugated nanoparticle by the MPS depends on the physical and chemical properties of the polymer such as the size, surface charge, hydrophobicity, and biodegradability. The macrophage uptake of a nanoparticle is reduced when the size of the nanoparticle is smaller (<100 nm) and when it is free of surface hydrophobicity. Several different types of nanoparticles designed to bypass the MPS have been used in the past decade for drug delivery and to induce cancer cell death.

A cancer cell can be induced to die through two methods: necrosis or apoptosis. Necrosis is a violent type of cell death induced by physiological conditions such as heat, extreme pH, or osmotic swelling, which leads the cells to release intracellular components into the surrounding microenvironment, leading an inflammatory response. Apoptosis, or programmed cell death, on the other hand initiates a cascade of biochemical events within the cell that, in a defined fashion, silently kill the cell and allow for the remnants to be preferentially phagocytosed and cleared by macrophages. The preferential method of inducing cancerous cell death is through apoptosis, which avoids the inflammatory response generated by necrosis. Several chemotherapeutic drugs have been used in the past to induce tumor cell apoptosis.

Although there are more than 100 FDA-approved cancer therapeutic drugs, not more than 20 have been used for *in vivo* and clinical studies. Some of the commonly used chemotherapeutic drugs include doxorubicin (intercalates with DNA and stops DNA replication), methotrexate (inhibits the enzyme dihydrofolate reductase and stops nucleotide and DNA synthesis), taxol (promotes microtubule assembly and causes cell-cycle arrest), cisplatin, vincristine (inhibits microtubule assembly and causes cell-cycle arrest), and 5-fluorouracil (inhibits thymidylate synthase). The choice of these drugs has been based not only on their properties to induce cell growth arrest and programmed cell death (apoptosis), but also on their ability to achieve a high therapeutic index. The choice of a free drug for therapy is also based on its effect on inducing drug resistance. Although the pharmacokinetics, pharmacodynamics, and toxicity of the free drug are known, determination of these parameters as well as the therapeutic index and the drug resistance properties of the polymer–drug conjugate is important for their applicability in the clinic.

Despite the availability of several identified chemical toxins such as azide, mercury salts, colchicines, and so forth, which are known to induce cellular apoptosis, these compounds are not commonly selected for polymer conjugation. This is because the targeting agents identified so far are only *relatively* specific and these toxins, if collaterally delivered to normal cells, will cause potent toxic side effects. Nonetheless, identification of highly specific targets and usage of these toxins below their threshold toxicity levels may be beneficial for future targeted delivery. Moreover, identification of other drugs that promote apoptosis by interacting with one of the intermediary steps of the apoptotic pathway may serve as possible agents for delivery. An example of such a biochemical step is the binding of the oncoprotein MDM2 to p53 that prevents p53-mediated apoptosis [159]. An inhibitor of this interaction may have possible use as a drug in certain cancers with lowered p53 function. Similarly, delivery of a proapoptotic function such as the activator of the cystolic Bax protein may also serve as a possible target. The delivering of inhibitors of certain biochemical events in the apoptotic cascade has a great deal of potential for these agents to serve as possible drugs in certain types of cancers.

Although chemotherapeutic drugs induce cell death in the first rounds of administration, cancer cells eventually become resistant to the drug. This is because the cancer cells that survive the inital chemotherapeutic intervention "learn" to evade further attack by the drug through several biochemical modifications. There are several altered cellular biochemical pathways that lead to drug resistance such as the activity and expression of cystolic enzymes and membrane proteins. One major aberrant biochemical mechanism leading to multidrug resistance (MDR) is the over expression of the membrane P-glycoprotein (P-gp). P-gp is a protein that mediates the extrusion of a wide variety of drugs from the cell [160]. *In vitro* studies using drug-resistant cell lines have shown that drug encapsulation within a nanoparticle can overcome drug resistance [161, 162]. Drugs delivered in liposomes have been shown to overcome multidrug resistance [163]. Although the precise mechanism of the protection of drug resistance when delivered in a polymer is not known, it is possible that the highly localized release of the drug from the polymer into a specific subcellular compartment prevents it from active extrusion by P-gp.

Liposomes are one of the most extensively investigated platforms for drug delivery. As described previously, liposomes contain hydrophilic head groups and a hydrocarbon tail of 14 to 18 carbon atoms. In water, the hydrophobic chains wrap around themselves and form vesicles by entrapment of water between them. A chemotherapeutic drug can therefore be entrapped within this vesicular space for delivery into cells. One obstacle in using liposomes is their removal by the MPS. Although the clearance of liposomes by MPS can be exploited for targeting macrophages or the liver, the liposomes must be surface-modified and protected from MPS in order to target other tissues. Liposomes can be coated to make "stealth liposomes," which are stable in circulation for more than a day. The FDA-approved drugs Doxil and DaunoXome are drug-encapsulated stealth liposomes used to treat AIDS-related Kaposi's sarcoma [164]. Stealth liposomes linked to a specific targeting molecule have been tested for specific drug delivery. Folic acid conjugated to liposomes through a PEG linker and encapsulated with doxorubicin can deliver the drug into FA-receptor–expressing cells [107]. Liposomes conjugated with streptavidin bind to tumors that had been pretargeted with antibody–biotin complex with elevated tumor association as compared to nontargeted tumors [165].

To overcome the tissue stability constraints encountered in liposomal drug delivery, many investigators have shifted their strategy to develop other synthetic nanoparticles as drug delivery platforms [166]. Drug delivery methods researched and currently under development include encapsulation of the drug within the interior space of the nanoparticle, adsorption onto the cell surface, or covalent conjugation to the device. A nanoparticle in which a delivery agent is encapsulated into the cargo space of the particle through noncovalent linkage is termed a nanocomposite. Several biodegradable nanocomposites, surface modified for long circulation, have been shown to outperform the respective free drug in improving the efficacy for tumor regression *in vivo* [167–169]. The drug mitoxantrone encapsulated in or adsorbed on poly(butylcyanoacrylate) nanoparticles coated with poloxamine show significant decrease in tumor volume as compared to the free drug [170]. In addition to the delivery

of chemotherapeutic drugs, nanocomposites have been employed to deliver oligonucleotides [166]. These studies using encapsulated drugs were confounded by the lack of stability of the drug in the polymer in biological fluids and the presence of a large amount of free drug in the preparation.

There are several advantages for using a stable multifunctional DDS onto which several functions are covalently linked. In such a system, a drug is linked through covalent ester bonding and is released in cells through endogenous esterases. Multiple therapeutic modalities, attacking the cancer cells or their bystanders through multiple mechanisms by way of combinatorial therapy, could effectively kill a tumor mass. Targeting more than one target in a neoplastic cell or vasculature is more effective than targeting of a single receptor. For drug delivery, in addition to an apoptosis-inducing agent, the addition of a specific inhibitor of an altered tyrosine kinase or PI3 kinase or a "bioreductive prodrug" (which is converted to active drug specifically under the hypoxic tumor conditions) would greatly enhance the therapeutic index. Such multiheaded therapeutic devices can only be designed if several free functional moieties are present on the surface of the nanoparticle.

PAMAM dendrimer nanodevices are suitable platforms for multifunctional drug delivery because they meet several criteria crucial for biological application and compatibility. These dendrimers are available commercially in large quantities and are under good manufacturing process (GMP). As described previously, the molecular size of these polymers ranges from 1 to 20 nm in diameter, similar to biological macromolecules, and they also possess a large number of free primary amino groups for conjugation with multiple functions. For example, the generation 5 (G5, 5 nm in diameter) PAMAM dendrimer has approximately 120 amino groups on its surface available for conjugation with multiple functions. In the past decade, dendrimers have been widely used for conjugation with several types of bioactive materials [171–175].

We have generated a three-functional dendrimer nanodevice containing FA as the targeting molecule, methotrexate (MTX) as the chemotherapeutic drug, and fluorescein isothiocyanate (FITC) as the fluorescent detecting dye (see Fig. 3). These individual components have been covalently conjugated to the surface amino group of the G5 PAMAM dendrimer. The device is taken up in FAR-expressing KB cells and induces cytotoxicity in a dose-dependent manner. *In vivo* studies in mice comparing the three-functional device to a free drug show a significantly increased therapeutic index by the nanodevice (manuscript under preparation). These results indicate that PAMAM dendrimer conjugates are promising models for further clinical investigation of targeted cancer therapies.

6. APOPTOSIS

6.1. Apoptosis as a Biological Event

Kerr et al. originally described two forms of cell death, necrosis and apoptosis, which may occur in the absence of pathological manifestations [176]. Apoptosis, or programmed cell death, is a normal physiologic process that occurs during embryonic development as well as in maintenance of tissue homeostasis. The overabundance or deficiency of apoptosis in cells can lead to recognizable disease. Many illnesses can be commonly associated with a defect in apoptosis. Alzheimer's disease and stroke can be attributed to an excess of apoptosis; the excessive cell death is what leaves these patients in a deteriorated physical and mental state. Cancer and autoimmune diseases, however, can be credited to a lack of apoptosis taking place. Cell growth becomes out of control and proliferation of these malignant cells continues, increasing the ratio of cancerous to healthy cells and eventually leading to death if left uncontrolled [177]. On reading this section, it must be made clear the numerous number of research articles that have been made available during the past few years. Figure 12 presents the drastic increase in articles published in the past few years, which has left us with an enormous amount of research to sift through in order to present the clearest version of the course of apoptosis.

The term apoptosis, from the Greek word for "falling off" of leaves from a tree, is used to describe a process in which a cell actively participates in its own destructive processes.

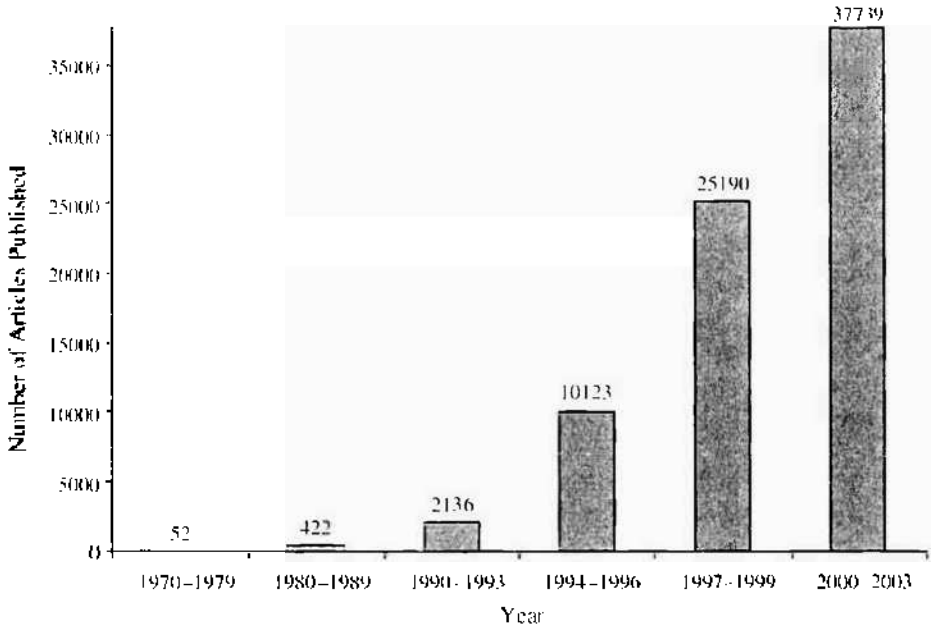


Figure 12. Chart of increase of journal articles published related to apoptosis.

Certain morphological, biochemical, and molecular changes characterize the apoptotic cascade. The duration of this destructive process differs by cell type and can also be influenced by the presence of inducing or inhibiting agents. Events can be characterized to occur during the early, middle, or late phases of the apoptotic process and are detected through the use of flow cytometry and fluorescence microscopy and through newly developed atomic force microscopy (AFM). There are two pathways for induction of apoptosis via death signals: the mitochondria pathway and the death receptor pathway. Both pathways use caspases for disassembly of cell structures, leading to cell death.

Though the process of apoptosis is possibly reversible if detected in its earliest stages, once caspase activity has begun, the process becomes irreversible. Caspases, a structurally related group of cysteine aspartate-specific proteases, cleave peptide bonds following specific recognition sequences. They play a central role in activating apoptosis of vertebrate cells. Activation of the caspase family is one of the earliest markers of an apoptotic event. However, the final phases of apoptosis require the activation of cysteine proteases of the caspase family as well. Early apoptotic events also consist of changes in the plasma membrane, including increased permeability, a loss of membrane symmetry, and construction of membrane-bound apoptotic bodies. Externalization of phosphatidyl serine from the inner to the outer plasma membrane leaflet also occurs (via caspase activation) and can be used as an indicator of the initialization of apoptosis [178].

Additional events of apoptosis include the condensation of the cytoplasm and nucleus due to lysosomal rupture and cell shrinkage, degradation of cellular proteins, membrane blebbing, condensation of nuclear chromatin, and internucleosomal cleavage of DNA. In contrast to necrotic cell death, cell shrinkage is an elemental step in apoptosis for the instigation of death enzymes. Cells that have undergone shrinkage show an increased presence of caspase-3 like activity and have fragmented DNA [178].

In the past few years, great emphasis has been placed on mitochondria as important components to the apoptotic process. Though there has been some uncertainty regarding the order in which these events take place (caspase activation, then mitochondrial activity or vice versa) or whether the mitochondria even play a vital role in the apoptotic process [177], there are at least three general mechanisms occurring in the apoptotic cascade that involve cell mitochondria: (i) disruption of the electron transport chain, which means effectively loss of cell metabolism and ATP production; (ii) release of caspase activating proteins; and (iii) alteration of redox potential [179].

Alteration of cell metabolism, largely due to a decrease in mitochondrial respiration, is a key step toward cell death [180]. Disruptions of the mitochondria begin with a decrease in membrane potential. Research by Heiskanen et al. supports that a decrease in membrane potential does not occur partially in all mitochondria within a cell, but occurs *fully* (full depolarization) within a finite proportion of mitochondria within a cell [181]. Opening of the mitochondrial permeability transition pore (PT pore) occurs as a result of this and allows for the passage of small molecules and ions, equilibrating the ion concentrations between the matrix and the intermembrane space of the mitochondria. This equilibrium causes destruction of the respiratory chain and rupturing of the mitochondrial outer membrane. Uncoupling of the respiratory chain leads to a decrease of ATP production in the mitochondria [182]. The mitochondrial outer membrane ruptures due to the expansion of the matrix space (as a result of the equilibration of ion concentrations via the PT pore, the matrix experiences hyperosmolarity, causing it to swell) inside the inner membrane. The folded cristae of the inner membrane allow for the inner membrane to have a much larger surface area than the outer membrane; therefore, the expansion of the inner membrane causes the outer membrane to rupture, leading to the release of cytochrome *C* and other caspase-activating proteins from the mitochondria into the cytosol [179, 181]. The release of cytochrome C helps activate caspase-9, which cleaves procaspase-3 into its active form, caspase-3, which is largely responsible for the biochemical and morphological changes by which we so commonly characterize the apoptotic pathway. The final stage of apoptosis is characterized by the dying cells, fragmenting into "apoptotic bodies" that are rapidly eliminated by phagocytic cells without eliciting significant inflammatory damage to surrounding cells.

Apoptosis can be initiated by a variety of biological events. Irradiation (UV or γ), growth factor withdrawal, signal transduction after engagement of cell surface receptors ("death receptors" Fas/Cd95/Apo-1, and so forth), as well as chemotherapeutic drugs can all induce cell destruction via the apoptotic pathway [183]. Therapeutic drugs intended to either circumvent or transduce apoptotic signals must be designed with these events in mind in order to inhibit or promote apoptosis. Caspase activators and inhibitors are some of the commonly researched ways to alter apoptosis events.

A variety of methods are used to detect whether or not apoptosis has occurred or is occurring within a cell. Double-dye fluorescence resonance energy transfer (FRET) detection as well as single-dye detection using rhodamine derivatives, rhodamine 110, (which is used to detect caspase activity in cells), rhodamine 123, and tetramethylrhodamine methyl ester (TMRM) and tetramethylrhodamine ethyl ester (TMRE) (which are used to measure membrane potential in mitochondria) are all commonly used as apoptosis detectors.

Necrosis, which typically occurs as a result of cell injury or exposure to cytotoxic chemicals, is distinct from apoptosis in both morphological and biochemical respects. Necrosis is caused be external damage to a cell, usually caused by destruction of the plasma membrane. Necrotic cell death begins with swelling of the cell and its mitochondrial contents, followed by rupture of the cell membrane [184]. Soluble cytosolic proteins are then able to leak out from the lysed membrane. Necrosis often occurs secondarily to the apoptotic process and can be considered as degenerative cell death [181]. In contrast to apoptosis, necrosis can trigger an inflammatory reaction in the surrounding tissue as a result of the release of its cytoplasmic contents, much of which are proteolytic enzymes, which attract inflammatory cells and trigger tissue destruction characteristic of inflammation. The tissue then undergoes repair and scarring, leading to permanent damage and changes in tissue structure.

6.2. Introduction to Apoptosis Detection

Apoptosis detection methods are employed to determine if the apoptosis process has occurred or is in the process of occurring within cells. Apoptosis detection is most often performed using single- or double-dye detection methods. Though there are a large variety of techniques in use, the following sections review the most commonly practiced. However, some methods are more valuable than others, as they are able to detect apoptosis *in vivo*. Most techniques nonetheless rely on *in situ* evaluation, as phagocytic tissues and cells quickly

remove apoptotic cells left *in vivo*. Other detection methods require the cells to be fixed in order to detect the occurrence of apoptosis. Detection methods used to study apoptosis in fixed tissues are difficult to evaluate due to the structural complexity of cells and tissues; therefore, electron microscopy and two-photon microscopy are promising methods under investigation for further study of apoptosis in live tissues [177].

Detectors are commonly made with one or two dyes, and by staining with these dyes or by coupling these dye reagents to a dendrimer or other delivery device, these dyes are delivered into the cell or attach to cell surfaces in order to determine whether apoptosis has occurred. Detection methods can be categorized based on the phase of apoptosis they detect. Some techniques detect apoptotic activity based on membrane permeability and DNA fragmentation, whereas other methods detect activity based on chromatin condensation or caspase activity. Flow cytometry is widely used to quantify dye measurements as apoptotic, necrotic, and viable cells all exhibit different light-scattering patterns. Most often, apoptosis detection methods are used in combination with one another for better evaluation of data. The majority of the methods of apoptosis detection are expensive as well as time-consuming and are subjected to a great deal of human error in handling and interpretation of test results; therefore, it is in an experimenter's best interest to use fully the methods available for proper evaluation of data [82]. Table 3 lists apoptotic events separated by location in the cell that are detectable or made visible by imaging methods currently in use. These events are most commonly made detectable by surface staining dyes, flow cytometry, or by the other methods listed in italics. Research into imaging apoptosis is directed by the events below.

6.2.1. Single-Dye Apoptosis Sensor

Single-dye detection methods are often the less complex choice for apoptosis sensing. There are a wide variety of techniques in practice that use single-dye detection. These include detection by coumarin-based dyes; rhodamine-derivative dye; terminal deoxynucleotidyl transferase (TdT)-mediated end-labeling of DNA strand breaks; (TUNEL method); use of annexin V for the detection of phosphatidyl serine on cell membranes and also in combination with propidium iodide (PI) for differentiation between apoptotic and necrotic cells; use of DNA-binding dyes such as the Hoechst dye and DAPI; use of green fluorescent protein (GFP) as a marker; and staining of cell surface and intracellular structures, which are measured quantitatively by flow cytometry [185].

Detection of caspase-3 activity is often used to determine whether the apoptotic process is occurring. Coumarin-based fluorogenic substrates such as Ac-DEVD–AFC and

Table 3. Detectable changes within a cell that denote the occurrence of an apoptotic event.[a]

DNA cleavage and nuclear events	Biochemical events	Mitochondrial events	Changes in surface morphology and composition
Segmentation in chromatin, nuclei	Caspase activity, *FRET detection, rhodamine derivatives*, various *fluorophores*	Permeability transition, detection by *vital dyes*	Time lapse characterization of surface morphology
Chromatin condensation	Detection of caspase cleavage products	Mitochondrial antigens	Phospholipid externalization, *annexin V* binding
DNA cleavage *in situ* by *TUNEL method*, detection of DNA fragmentation, strand breaks	Transglutaminase activity	Cytochrome *C* release and alterations	Changes in membrane permeability, *DAPI, Hoechst*
Anti-single-stranded DNA antibody	PARP activity	Metabolic activity	
Hairpin oligos to detected double-stranded breaks	Death antigens		

[a] Dyes and methods used for visualization of these events are italicized.

Z-DEVD–AMC are often used to detect caspase-3 activity in cellular lysates; however, more effort has been placed on use of rhodamine-based derivatives for caspase-3 detection, as coumarin-based fluorogenic substrates have low extinction coefficients. It has been reported by Liu et al. that the $(Z\text{-DEVD})_2$–Rh 110 exhibits a much higher turnover rate than the coumarin-based Ac-DEVD–AFC substrate and is at least 10-fold more sensitive than Z-DEVD–AMC under their assay conditions [186]. The products released from coumarin-based substrates also have short excitation and emission wavelengths, which limit assay sensitivity, and are cell impermeable; therefore, they only work with cellular extracts [187]. Residual fluorescence of the substrates as well as poor wavelength separation of their products due to cellular auto fluorescence gives way to high background interference in microscopic assays [186].

Rhodamine derivatives including rhodamine 110, rhodamine 123, and others are often used to detect caspase activity as well as changes in mitochondrial membrane potential. There are many advantages to using rhodamine-based substrates over other dyes. They have longer excitation and emission wavelengths, which reduces interference from testing compounds; the uncleaved substrate does not fluoresce and therefore has a very low background signal, which makes it very clear whether apoptosis has occurred or not; and finally, once cleaved, rhodamine has very strong fluorescent properties. Rhodamine and its derivatives are also cell-permeable, which allows for flow cytometry to measure the amount of fluorescence emitted, and therefore clearly verify the occurrence of the apoptotic process [188].

Though it is reportedly difficult to couple rhodamine 110 to the tetrapeptide (a chain of four amino acids, D-E-V-D, which is cleaved by caspase-3 after the second D) in order to form the fluorogenic substrates, sequential coupling of the amino acids to rhodamine 110 has been the most suitable way of preparing the substrate [186]. Use of two DEVD blocking groups to prevent fluorescence of the substrate leads to the necessity of two hydrolysis reactions for cleavage of both the DEVD sites in $(Z\text{-DEVD})_2$–Rh 110. This limits the linear dynamic range of the substrate, and the creation of a Rhodamine 110–based substrate with only one hydrolyzable amide group (one DEVD site) would allow for more efficiency in the use of the rhodamine 110 derivative as a fluorogenic substrate.

A second method used for apoptosis detection is known as terminal deoxynucleotidyl transferase mediated digoxigenin nick end labeling, (i.e. the TUNEL method). DNA fragmented by activated endonucleases into segments 180–200 bp in length are labeled by dUTP, which is later detected by light or fluorescence microscopy [189]. This method is performed *in situ* for observation of apoptosis at the single cell level, where it is often difficult to distinguish between the occurrence of apoptosis or necrosis in a cell. Though considered the standard for detection of DNA fragmentation, a hallmark of the apoptotic process, the TUNEL method often fails to distinguish between apoptosis and necrosis, as cells in late phases of necrosis undergo DNA damage, which would stain as a false positive for apoptosis. The TUNEL method detects all DNA damage/fragmentation, so artificial DNA breakage would be stained as well, also giving a false positive identification for apoptosis. Another drawback of this method is that it is unfeasible for use in live cells. Tissues must be frozen, then sectioned, mounted, and stained with the anti-digoxigenin antibody and conjugated with a fluorescent dye such as fluorescein [190].

An early marker of the apoptotic process is indicated by the externalization of phosphatidyl serine from the inner to the outer plasma membrane leaflet. The Ca^{2+}-dependent protein annexin V can be used to detect the presence of phosphatidyl serine on cell membranes. The externalization of phosphatidyl serine occurs via caspase activation; therefore, its presence, indicated by the binding of annexin V to the phosphatidyl serine, acts as a good marker of the initialization of apoptosis [185]. Due to the morphological differences that characterize apoptosis and necrosis, use of annexin V in combination with use of propidium iodide (PI) staining makes it possible to differentiate between apoptotic and necrotic cells. As discussed previously, a fundamental difference between the two types of cell death is characterized by cell membrane permeability. Cells experiencing apoptotic cell death possess cell membranes that are impermeable until the final stages of the death process, whereas cells undergoing necrotic death possess membranes that are permeable. Due to the large molecular weight of PI and the impermeability of the membranes of cells undergoing apoptosis, cells

that stain annexin V positive and PI negative are considered apoptotic, whereas cells that stain both annexin V and PI positive are labeled necrotic [191]. In later stages of apoptosis, however, apoptotic cells may stain both annexin V and PI positive due to loss of membrane integrity, therefore making it difficult or near impossible to differentiate between apoptosis and necrosis during these stages [192].

Another method of visualization can be achieved by the use of dyes that bind internally to DNA fragments. Use of the various Hoechst dyes (33342, 33258) is particularly prevalent; Hoechst 33342 is a fluorescent dye that permanently stains DNA for visualization of chromatin condensation, changes in chromatin conformation, and nuclear fragmentation, for qualitative determination of apoptosis. Hoechst 33342 is a simple method for detection of early and late events in the apoptotic process *in vivo* [189]. It also has the ability to detect early changes in membrane permeability, as it can internally label DNA within apoptotic cells [193]. Other variations of Hoechst dyes work in similar ways.

Chromatin condensation and fragmentation and DEVD-like caspase activity can be analyzed by staining with DAPI and PhiPhiLux–G_1D_2, respectively. PhiPhiLux–G_1D_2 is a fluorogenic substrate that is cleaved in a DEVD-dependent manner to produce rhodamine molecules, which fluoresce red under a G2A filter, whereas DAPI stains nuclei (apoptotic or viable) blue under a DAPI filter [194]. PhiPhiLux, which is used to measure caspase activation via FACS, poorly penetrates the cell membrane and needs a longer incubation time [187]. Using this technique, it is not possible to measure apoptotic events shorter than the 1-h staining time necessary to produce visible results using this method.

Green fluorescent protein is a versatile protein that has the ability to be used in a variety of circumstances. Pertaining to apoptosis detection, GFPs have the ability to differentiate between necrotic cells and cells undergoing apoptosis. GFPs also permit the user the ability of monitoring the morphological changes occurring within cells undergoing apoptosis. In addition, these proteins allow for the "real-time" monitoring of the apoptotic or necrotic processes without damage or alteration to the cell by way of measuring the decrease in GFP fluorescence intensity. The lower the GFP fluorescence present, the greater the loss of cell viability [195]. Green fluorescent protein affects the same cells stained by PI and annexin V staining; however, in accordance with research by Strebel et al. [185], use of GFP methodology requires minimal steps toward achieving the same results as other dye detection techniques while minimizing handling errors. GFP readout can be performed by flow cytometry as well as in a fluorescence plate reader and is the ideal assay for large-scale screening [185].

Flow cytometry is used to quantify the amount of apoptosis occurring, based on the fact that light-scattering patterns of apoptotic cells are differentiated from light-scattering patterns of viable cells as well as from cells undergoing necrotic cell death. Apoptotic cells give lower forward and higher side scatter values than viable cells due to their smaller size and differences in cytoplasm and nucleus consistency. Flow cytometry analysis is based on the detection of DNA fragmentation and loss and morphological changes as well as changes in membrane permeability [192]. Various fluorogenic substrates as described previously can be used in combination with this detection method in order to quantitatively determine if apoptosis is taking place.

6.2.2. Double-Dye FRET Reagent-Based Apoptosis Sensor

Fluorescence resonance energy transfer detection is used to examine structural and dynamic characteristics of biological molecules in aqueous solution [196]. Macromolecules to undergo examination are labeled with covalently linked donor and acceptor fluorophores. After a specific molecule has been chosen to undergo FRET, the donor fluorophore of that molecule is excited by continuous laser illumination. The fluorophore donor transfers energy to the acceptor fluorophore, where it is re-emitted as fluorescence.

FRET detection measures noncovalent bonding events in biological and macromolecular systems. Due to the presence of caspase-3, which cleaves certain cellular substrates during apoptosis, and the effects of caspase-3 on fluorescence resonance energy transfer, FRET detection can be used to determine whether apoptosis has occurred. FRET detection can also be used to monitor the spontaneous folding and unfolding of molecules due to

denaturing agents in solution as well as to map qualitatively protein–protein interactions. There are two ways most generally used to measure the FRET effect with steady-state microscopy. The first method of detection of the FRET effect measures the decrease of donor fluorophore emission and the increase in acceptor fluorophore emission in a two-dye system. The second method of detection of the FRET effect is called acceptor photobleaching. In this method, the donor is excited, the acceptor is photobleached, and donor emission is measured before and after acceptor photobleaching [197]. However, this method cannot be used *in vivo*.

In order for the FRET effect to occur, it is necessary for the fluorescence emission band of the donor fluorophore molecule to overlap with the excitation band of the acceptor molecule within 20–80 Å of the donor, as displayed in Fig. 13 [198].

It is assumed that due to the peptide conformation in aqueous solution, the donor and acceptor molecules are in close proximity to one another (10–100 Å), which allows for the energy transference from the donor fluorophore to the acceptor fluorophore.

The transfer of energy due to the FRET effect can be detected by the reduction of fluorescence from the donor fluorophore and an increase in the intensity of the fluorescent emissions from the acceptor fluorophore. The temporal increase in fluorescent intensity by the acceptor is called "acceptor in-growth" [198]. Donor fluorophore colors are restricted to blue or cyan, which have limited fluorescent capabilities and also undergo autofluorescence and are paired with green- or yellow-colored acceptor fluorophores.

As stated previously, the FRET effect is used to determine the conformation of biomolecules as well as to monitor biological processes such as apoptosis within a cell. FRET has great importance in the field of apoptosis detection and has widely been implemented in the study of programmed cell death. The detection of apoptosis in cells is again made possible by the activation of caspase-3. If the linker between the donor and acceptor fluorophores is synthesized to contain a cleavable substrate (containing the amino acid sequence D-E-V-D), then caspase-3, which actively cleaves cellular substrates possessing the sequence D-E-V-D, can be used to cleave the linker between the donor and acceptor fluorophores in the molecular probe [199].

The presence of caspase-3, which is only active during apoptosis, is detected by the elimination of the FRET effect. This allows for the determination whether or not apoptosis is occurring.

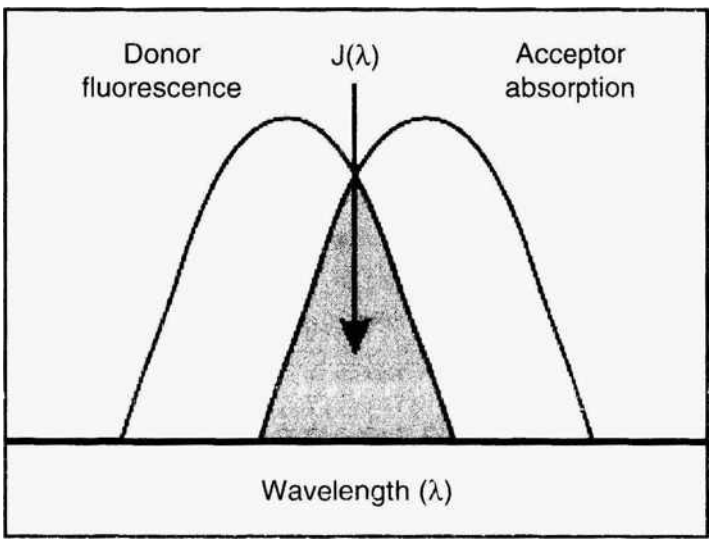


Figure 13. Image of overlapping acceptor:donor spectra where $J(\lambda)$ = spectral overlap integral of the absorption spectrum of the acceptor and fluorescence emission spectrum of the donor = integral$[(\varepsilon_A(\lambda) * F_D(\lambda) * \lambda^4\, d\lambda$ $cm^3\, M^{-1}$, where ε_A = extinction coefficient of acceptor, and F_D = fluorescence emission intensity of donor as a fraction of the total integrated intensity. From Fluorescence resonance energy transfer, available at www.probes.com/handbook/boxes/0422.html. Reprinted with permission. © Molecular Probes, Inc.

Resulting cleavage of the peptide, by caspase-3, between valine and aspartic acid in the recognition sequence D-E-V-D results in the elimination of the FRET effect as the donor and acceptor fluorophores are no longer joined. Flow cytometry is used to quantify the amount of fluorescence present. By observing the intensity shift between the emissions of the donor and acceptor fluorophores, it is possible to determine the change in the FRET effect as a function of the cleavage of the linker by the enzyme caspase-3 [199]. Figure 14 demonstrates an example of a peptide with a cleavable D-E-V-D site present. Cleavage of this site by caspase-3 results in a shift of low to high fluorescence of the donor.

There are, however, a few problems with the reliance on caspase-3 to cleave the D-E-V-D substrate within the linker between the donor and acceptor fluorophores. These problems lie within the fact that the linker might not be fully accessible for cleavage by the enzyme for various structural reasons, which would hinder the detection abilities of the FRET effect on apoptosis. The linker, for example, may be hindered from cleavage due to orientation of the donor and acceptor fluorophores to one another as well as due to the distance between the two fluorophores [199]. This method, nonetheless, has shown to be able to detect apoptosis exceptionally well in living, intact cells.

6.2.2.1. Synthesis of FRET-Apoptosis Device Figure 15 demonstrates the synthetic steps necessary to produce a nanodevice for apoptosis detection.

Evaluation of the FRET effect necessitates the calculation of the transfer efficiency of the photons between the donor and acceptor. The number of photons emitted from the donor and acceptor are counted within a certain time frame, and the ratio of these numbers is called the energy-transfer efficiency. This ratio is dependent on the distance between the donor and acceptor fluorophores. The distance between the two fluorophores must be calculated in order to determine the dynamic and conformational changes of the macromolecule under consideration [200]. The efficiency of the energy transfer can be calculated using the following formula,

$$E(r) = [1 + (r/R_0)^6]^{-1} \tag{5}$$

where $E(r)$ is the efficiency of the Förster energy transfer between the fluorophore donor and acceptor, r is the distance between the donor and acceptor fluorophores, and R_0 is the Förster radius, (i.e., the distance at which the efficiency is 1/2).

The larger the efficiency $E(r)$ of the energy transfer, the shorter the distance (r) between the fluorophores. This formula is valid due to the angular averaging of the dipole–dipole interaction, which occurs due to the fast movement of the fluorophores. It must be noted, however, that the value r fluctuates due to the dynamics of the macromolecule to which it is attached, as the molecule under observation will change shape and undergo folding and unfolding of its structure and other various conformational changes [200]. If the time

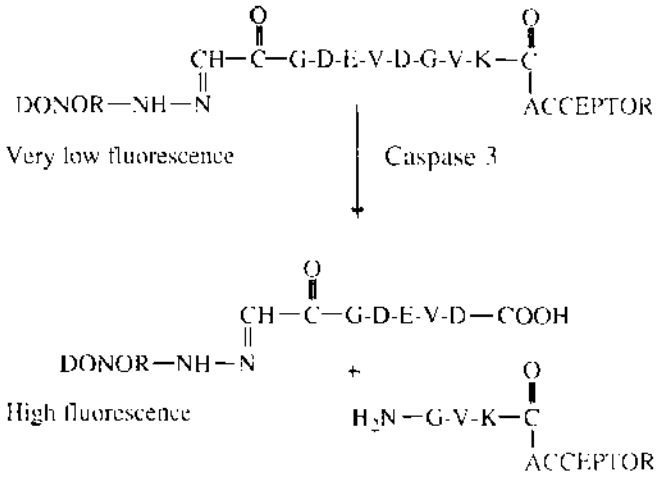


Figure 14. Peptide with cleavable valine–aspartic acid site marked.

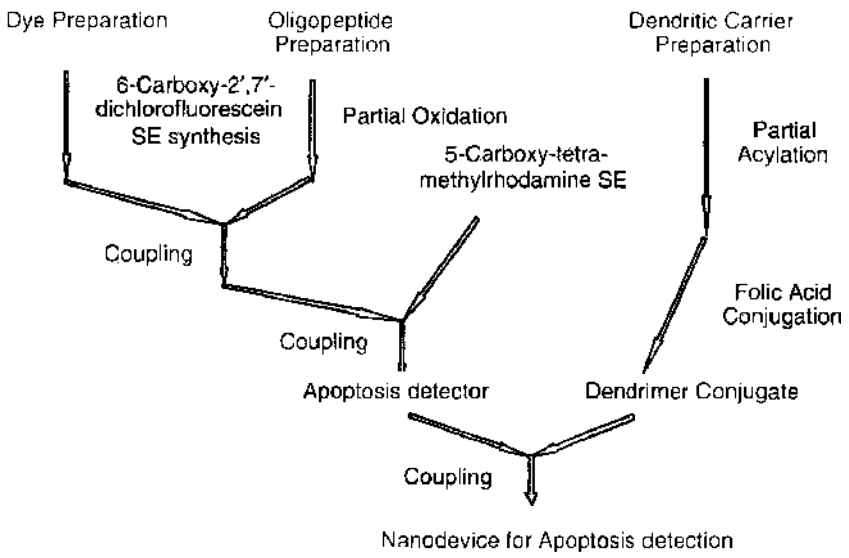


Figure 15. Strategy for FRET reagent and dendritic device synthesis.

window of observation is such that enough photons have been emitted for calculation and the distance between the donor and acceptor fluorophores does not change, then the efficency can be turned into a useful calculation of distance r.

It must also be noted that at a constant concentration of free and associated FRET pairs, the emission of the FRET donor is inversely proportional to the mole fraction of associated molecules [198].

Though use of double-dye (fluorophore) FRET detection is the most commonly used FRET detection technique, use of triple- and possible quadruple-dye detection has also been under study [196]. Research by Liu and Lu has shown that it is possible to label biomolecules that possess many arms with three or even four dyes for maximal detection of conformational changes as a result of fluorescent resonance energy transfer [196]. The labeling of the radiating arms of dendrimers independently of one another using fluorophores of different colors allows for determination of temporal and spatial locations of each pair of arms as well as for the assessment of overall changes in structural dynamics. One problem, however, encountered with using four dyes for detection is the possible nonoverlaps of spectra between donor and acceptor fluorophores, a prerequisite for the FRET effect, thuspreventing energy transfer between that particular pair of fluorophores [196]. This, however, (although a small inefficiency) does not eliminate the capabilities that using four dye can have, as the detection capabilities with this number of dyes can have great positive influence on determining the structural changes within a biomolecule.

7. CONCLUSIONS

Nanotechnology has emerged on the scientific forefront just within the past few decdes, but within this short time, researchers using this technology have already pioneered numerous advancements in a variety of fields as never possible before. More specifically to the topics contained within this chapter, molecular engineering on the nanoscale has allowed scientists and researchers alike the capability of synthesizing and modifying both bioleical and synthetic molecules on the molecular level for use in animal as well as human syams. The advancement of technology is, however, always occurring at a rapid-fire place, so that has been reviewed and discussed here represents a mere fraction of what is currently going on in the scientific world. The material presented within must be assessed with this mnd. The application of molecular engineering has allowed for the creation of a multifunctonal dendritic nano-biomolecule capable of delivering targeted therapeutic drugs while alsoaving the ability to fluoresce and detect cell death. Synthesis of a device such as thishas been made possible by through molecular engineering techniques on the nanoscale level,

which in turn enable scientists to perform the precise synthetic steps necessary to produce such an implicate macromolecule. Continued research and advancements within the fields of molecular engineering and cancer research are sure to bring even more complex and exciting results in the near future.

REFERENCES

1. I. Brigger, C. Dubernet, and P. Couvreur, *Adv. Drug Del. Rev.* 54, 631 (2002).
2. K. E. Uhrich, K. M. Shakesheff, S. M. Cannizzaro, and R. S. Larger, *Chem. Rev.* 99, 3181 (1999).
3. O. Pillai and R. Panchagnula, *Curr. Opin. Chem. Bio.* 5, 447 (2001).
4. C. H. Lee, A. Singla, and Y. Lee, *Int. J. Pharm.* 221, 1 (2001).
5. K. Y. Lee, C. M. Peteres, W. K. Anderson, and J. D. Mooney, *Nature* 408, 998 (2000).
6. P. Eiselt, J. Yeh, K. R. Latvala, D. L. Shea, and J. D. Mooney, *Biomaterials* 21, 1921 (2000).
7. T. A. Read, V. Stensvaag, H. Vindenes, E. Ulvestad, R. Bjerkvig, and T. F. Frits, *Int. J. Dev. Neurosci.* 17, 653 (1999).
8. C. M. Henry, *Chem. Eng. News* 80, 39 (2002).
9. R. Esfand and D. A. Tomalia, *Research Focus, Reviews* 6, 427 (2001).
10. I. J. Majoros, C. B. Mehta, and J. R. Baker, Jr., *J. Comp. Theor. Nanosci.* in press (2004).
11. L. Balogh, A. Bielinska, J. D. Eichman, R. Valluzzi, I. Lee, J. R. Baker, Jr., T. S. Lawrence, and M. K. Khan, *Chimica Oggi/Chemistry Today* 5, 35 (2002).
12. J. R. Baker, Jr., A. Quintana, L. Piehler, M. Banaszak-Holl, D. A. Tomalia, and E. Raczka, *Biomed. Microdev.* 3, 61 (2001).
13. K. Kono, M. Liu, and J. M. Fréchet, *Bioconjug. Chem.* 10, 1115 (1999).
14. R. Esfand, D. A. Tomalia, A. E. Beezer, J. C. Mitchell, M. Hardy, and C. Orford, *Polym. Prep.* 41, 1324 (2000).
15. J. Li, D. Qin, J. R. Baker, and D. A. Tomalia, *Polym. Prep.* 41, 1446 (2000).
16. J. Li, D. R. Swanson, D. Qin, H. M. Brothers, L. T. Piehler, D. A. Tomalia, and D. J. Meier, *Langmuir* 15, 7347 (1999).
17. J. Li, L. T. Piehler, D. Qin, J. R. Baker, Jr., D. A. Tomalia, and D. J. Meier, *Langmuir* 16, 5613 (2000).
18. T. A. Betley, M. M. Banaszak-Holl, B. G. Orr, D. R. Swanson, D. A. Tomalia, and J. R. Baker, Jr., *Langmuir* 17, 2768 (2001).
19. I. Lee, B. D. Athey, A. W. Wetzel, A. Kar, W. Meixner, and J. R. Baker, Jr., *Macromolecules* 35, 4510 (2002).
20. T. A. Betley, J. A. Hessler, A. Mecke, M. M. Banaszak-Holl, B. G. Orr, S. Uppuluri, D. A. Tomalia, and J. R. Baker, Jr., *Langmuir* 18, 3127 (2002).
21. A. W. Bosman, H. M. Janssen, and E. W. Meijer, *Chem. Rev.* 99, 1665 (1999).
22. A. Nourse, D. B. Millar, and A. P. Minton, *Biopolymers* 53, 316 (2000).
23. G. Sui, M. Micic, Q. Huo, and R. M. Leblanc, *Langmuir* 16, 7847 (2000).
24. C. Kojima, K. Kono, K. Maruyama, and T. Takagishi, *Bioconjug. Chem.* 11, 910 (2000).
25. S. C. Zimmerman, M. S. Wendland, N. A. Rakow, I. Zharov, and K. S. Suslick, *Nature* 418, 399 (2002).
26. P. Singh, *Bioconjug. Chem.* 9, 54 (1998).
27. D. A. Tomalia, D. M. Hedstrand, and L. R. Wilson, "Encyclopedia of Polymer Science and Engineering," Vol. 2, p. 46. J. Wiley & Sons, New York, 1990.
28. D. A. Tomalia and P. R. Dvornic, "Polymeric Materials Encyclopedia," Vol. 3, p. 1814. CRC Press, New York, 1996.
29. G. Pistolis, A. Malliaris, C. M. Paleos, and D. Tsiourvas, *Langmuir* 13, 5870 (1997).
30. S. Uppuluri, S. E. Keinath, D. A. Tomalia, and P. R. Dvornic, *Macromolecules* 31, 4498 (1998).
31. G. Teobaldi and F. Zerbetto, *J. Am. Chem. Soc.* 125, 7388 (2003).
32. A. Schmitzer, E. Perez, J. Rico-Lattes, A. Lattes, and S. Rosca, *Langmuir* 15, 4397 (1999).
33. J. Ruiz, G. Laduente, S. Mareen, C. Omelas, S. Lazare, E. Cloutet, J. C. Blais, and D. Astrue, *J. Am. Chem. Soc.* 125, 7250 (2003).
34. D. A. Tomalia, H. Baker, J. Dewald, M. Hall, G. Kallos, S. Martin, J. Roeck, J. Ryder, and P. Smith, *Polym. J.* 17, 117 (1985).
35. D. A. Tomalia, A. M. Naylor, and W. A. Goddard III, *Angewandte Chemie* 29, 138 (1990).
36. L. J. Hobson and W. J. Feast, *Polym. J.* 40, 1279 (1999).
37. Y. Zhao, X. Shuai, C. Chen, and F. Xi, *Chem. Mater.* 15, 2836 (2003).
38. R. Duncan, *Polym. Mater. Sci. Eng.* 84, 214 (2001).
39. P. Caliceti, S. Salmaso, A. Semerizato, T. Carofiglio, R. Fornasier, M. Fermeglia, M. Ferrone, and S. Prici, *Bioconjug. Chem.* 14, 899 (2003).
40. R. Langer, *Acc. Chem. Res.* 33, 94 (2000).
41. N. Kumar, R. S. Langer, and A. J. Domb, *Adv. Drug Del. Rev.* 54, 889 (2002).
42. C. Allen, Y. Yu, D. Maysinger, and A. Eisenburg, *Bioconjug. Chem.* 9, 564 (1998).
43. H. J. Han, M. J. Krochta, J. M. Kurth, and Y. L. Hsieh, *J. Agr. Food Chem.* 48, 5278 (2000).
44. J. M. Pernaut and R. J. Reynolds, *J. Phys. Chem. B.* 104, 4080 (2000).
45. E. Gianasi, M. Wasil, E. G. Evagorou, A. Keddle, G. Wilson, and R. Duncan, *Euro. J. Cancer* 35, 994 (1999).
46. B. R. Greenwald, H. Y. Choe, D. C. Conover, K. Shum, D. Wu, and M. Royzen, *J. Med. Chem.* 43, 475 (2000).

47. B. R. Greenwald, A. Pendri, D. C. Conover, H. Zhao, H. Y. Choe, A. Martinez, K. Shum, and S. Guan, *J. Med. Chem.* 42, 3657 (1999).
48. M. E. Santos, S. Radin, and P. Ducheyne, *Biomaterials* 20, 1695 (1999).
49. W. T. King and W. Patrick, Jr., *J. Biomed. Mater. Res.* 51, 383 (2000).
50. M. H. Sheridan, L. D. Shea, M. C. Peters, and D. J. Mooney, *J. Controlled Release* 64, 91 (2000).
51. S. A. Hagan, G. A. Coombes, M. C. Garnett, S. E. Dunn, M. C. Davies, L. Illum, and S. S. Davis, *Langmuir* 12, 2153 (1996).
52. T. Riley, S. Stolnik, C. R. Heald, C. D. Xiong, M. C. Garnett, L. Illum, and S. S. Davis, *Langmuir* 17, 3168 (2001).
53. M. D. Brown, A. Schatzlein, A. Brownlie, V. Jack, W. Wang, L. Tetley, A. I. Gray, and I. F. Uchegbu, *Bioconjug. Chem.* 11, 880 (2000).
54. H. Uludag, B. Norrie, N. Kousinioris, and T. Gao, *Biotech. Bioeng.* 73, 510 (2001).
55. G. Lambert, E. Fattal, H. Pinto-Alphandary, A. Gulik, and P. Couvreur, *Int. J. Pharm.* 214, 13 (2001).
56. G. Lambert, E. Fattal, H. Pinto-Alphandary, A. Gulik, and P. Couvreur, *Pharm. Res.* 17, 707 (2000).
57. S. M. Flynn, S. T. George, L. White, W. Devonish, and G. B. Takle, *Biotechnology* 26, 736 (1999).
58. J. Liu, C. Kolar, A. T. Lawson, and H. W. Gmeiner, *J. Org. Chem.* 66, 5655 (2001).
59. I. J. Majoros, B. Keszler, S. Woehler, T. Bull, and J. R. Baker, Jr., *Macromolecules* 36, 5526 (2003).
60. R. Weissleder, *Nat. Rev. Cancer* 2, 1 (2002).
61. M. Brasuel, R. Kopelman, T. J. Miller, R. Tjalkens, and M. A. Philbert, *Anal. Chem.* 73, 2221 (2001).
62. K. R. Gopidas, J. K. Whitesell, and M. A. Fox, *J. Am. Chem. Soc.* 125, 6491 (2003).
63. M. F. Ottaviani, R. Valluzzi, and L. Balogh, *Macromolecules* 35, 5105 (2002).
64. J. Y. Ye, L. Balogh, and T. Norris, *App. Phys. Lett.* 80, 1713 (2002).
65. G. Raschke, S. Kowarik, T. Franzl, C. Sonnichsen, T. A. Klar, J. Feldmann, A. Nichtl, and K. Kurzinger, *Nano Lett.* 3, 935 (2003).
66. C. Mirkin, *Inorg. Chem.* 39, 2258 (2000).
67. M. J. Hostetler, J. E. Wingate, C. J. Zhong, J. E. Harris, R. W. Vachet, M. R. Clark, J. D. Londono, S. J. Green, J. J. Stokes, G. D. Wignall, G. L. Glish, M. D. Porter, N. D. Evans, and R. W. Murray, *Langmuir* 14, 17 (1998).
68. K. Satoh, T. Yoshinaga, and K. Esumi, *J. Colloid Interface Sci.* 255, 312 (2002).
69. M. Brust, M. Walker, D. Bethell, D. J. Schiffrin, and R. Whyman, *J. Chem. Soc., Chem. Commun.* 116, 801 (1994).
70. M. E. Garcia, L. A Baker, and R. M Crooks, *Anal. Chem.* 71, 256 (1999).
71. R. M. Crooks, M. Zhao, L. Sun, V. Chechik, and L. K. Yeung, *Accts. Chem. Res.* 34, 181 (2001).
72. Y. Niu and R. M. Crooks, *Chem. Mater.* 15, 3463 (2003).
73. B. Dragnea, C. Chen, E. S. Kwak, B. Stein, and C. C. Kao, *J. Am. Chem. Soc.* 125, 6374 (2003).
74. C. S. Hall, J. N. Marshall, M. J. Scott, P. J. Gaffney, S. A. Wickline, and G. M. Lanza, *J. Acous. Soc. Am.* 108, 3049 (2000).
75. C. Bremer, V. Ntziachristos, and R. Weissleder, *Eur. Radiol.* 13, 231 (2003).
76. K. Marten, C. Bremer, S. M. Khazaie, B. Sloane, C. H. Tung, and R. Weissleder, *Gastroenterology* 122, 406 (2002).
77. L. Josephson, M. F. Kircher, U. Mahmood, Y. Tang, and R. Weissleder, *Bioconjug. Chem.* 13, 554 (2002).
78. C. Chapon, F. Franconi, L. Lemaire, L. Marescaux, P. Legras, J. P. Saint-Andre, B. Denizot, and J. J. Le Jeune, *Invest. Radiol.* 38, 141 (2003).
79. L. Josephson, M. F. Kircher, U. Mahmood, Y. Tang, and R. Weissleder, *Bioconjug. Chem.* 13, 554 (2002).
80. A. R. Padhani, *Eur. J. Cancer* 38, 2116 (2002).
81. G. M. Lanza, X. Yu, P. M. Winter, D. R. Abendschein, K. K. Karukstis, M. J. Scott, L. K. Chinen, R. W. Fuhrhop, D. E. Scherrer, and S. A. Wickline, *Circulation* 106, 2842 (2002).
82. H. Kobayashi, S. Kawamoto, S. K. Jo, N. Sato, T. Saga, A. Hiraga, J. Konishi, S. Hu, K. Togashi, M. W. Brechbiel, and R. A. Star, *Kidney Int.* 61, 1980 (2002).
83. L. Kostakoglu, H. Agress, Jr., and S. J. Goldsmith, *Radiographics* 23, 315 (2003).
84. D. Artemov, M. Solaiyappan, and Z. M. Bhujwalla, *Cancer Res.* 61, 3039 (2001).
85. D. Hogemann, V. Ntziachristos, L. Josephson, and R. Weissleder, *Bioconjug. Chem.* 13, 116 (2002).
86. A. Hemminki, K. R. Zinn, B. Liu, T. R. Chaudhuri, R. A. Desmond, B. E. Rogers, M. N. Barnes, R. D. Alvarez, and D. T. Curiel, *J. Natl. Cancer Inst.* 94, 741 (2002).
87. J. L. Evelhoch, R. J. Gillies, G. S. Karczmar, J. A. Koutcher, R. J. Maxwell, O. Nalcioglu, N. Raghunand, S. M. Rosen, B. D. Ross, and H. M Swartz, *Neoplasia* 2, 152 (2000).
88. B. Laxman, D. E. Hall, M. S. Bhojani, D. A. Hamstra, T. L. Chenevert, B. D. Ross, and A. Rehemtulla, *Proc. Natl. Acad. Sci.* 99, 16551 (2002).
89. S. A. Anderson, R. K. Rader, W. E. Westlin, C. Null, D. Jackson, G. M. Lanza, S. A. Wickline, and J. J. Kotyk, *Magn. Reson. Med.* 44, 433 (2000).
90. X. Yu, S. K. Song, J. Chen, M. J. Scott, R. J. Fuhrhop, C. S. Hall, P. J Gaffney, S. A. Wickline, and G. M. Lanza, *Magn. Reson. Med.* 44, 867 (2000).
91. S. Flacke, S. Fischer, M. Scott, R. Fuhrhop, J. Allen, M. Mclean, P. Winter, G. Sicard, P. Gaffney, S. Wickline, and G. M. Lanza, *Circulation* 104, 1280 (2001).
92. C. Hall, J. Marsh, M. Scott, P. Gaffney, S. Wickline, and G. M. Lanza, *J. Acous. Soc. Am.* 110, 1677 (2001).
93. C. Chapon, F. Franconi, L. Lemaire, L. Marescaux, P. Legras, J. P. Saint-Andre, B. Denizot, and J. J. Le Jeune, *Invest. Radiol.* 38, 141 (2003).

94. L. Josephson, M. F. Kircher, U. Mahmood, Y. Tang, and R. Weissleder, *Bioconjug. Chem.* 13, 554 (2002).
95. H. Kobayashi, S. Kawamoto, S. K. Jo, N. Sato, T. Saga, A. Hiraga, J. Konishi, S. Hu, K. Togashi, M. W. Brechbiel, and R. A. Star, *Kidney Int.* 61, 1980 (2002).
96. D. Weitman, R. H. Lark, L. R. Coney, D. W. Fort, V. Frasca, V. R. Surawski, and B. A. Kamen, *Cancer Res.* 52, 3396 (1992).
97. J. F. Ross, P. K. Chaudhuri, and M. Ratnam, *Cancer* 73, 2432 (1994).
98. G. Toffoli, *Int. J. Cancer* 74, 193 (1997).
99. M. Wu, J. Fan, W. Gunning, and M. Ratnam, *Cancer Epidemiol. Biomarkers Prev.* 8, 775 (1999).
100. H. Wang, X. Zheng, F. Behm, and M. Ratnam, *Blood* 15, 3529 (2000).
101. Y. Lu and P. S. Low, *Adv. Drug Del. Rev.* 54, 675 (2002).
102. C. P. Leamon, I. Pastan, and P. S. Low, *J. Biol. Chem.* 268, 24847 (1993).
103. S. Li, H. M. Deshmukh, and L. Huang, *Pharm Res.* 15, 1540 (1998).
104. A. Bielinska, J. F. Kukowska-Latallo, J. Johnson, D. A. Tomalia, and J. Baker, Jr., *Nuc. Acid Res.* 24, 2176 (1996).
105. J. F. Kukowska-Latallo, A. Bielinska, J. Johnson, R. Spindler, D. A. Tomalia, and J. Baker, Jr., *Proc. Natl. Acad. Sci. U.S.A* 93, 4897 (1996).
106. R. J. Lee and P. S. Low, *J. Biol. Chem.* 269, 3198 (1994).
107. R. J. Lee and P. S. Low, *Biochim. Biophys. Acta* 1233, 134 (1995).
108. A. Gabizon, *Bioconjug. Chem.* 10, 289 (1999).
109. D. M. Kranz, T. A. Patrick, K. E. Brigle, M. J. Spinella, and E. J. Roy, *Proc. Natl. Acad. Sci. U.S.A* 92, 9057 (1995).
110. B. Stella, S. Arpicco, M. T. Peracchia, D. Desmaele, J. Hoebeke, M. Renoir, J. d'Angelo, L. Cattel, and P. Couvreur, *J. Pharm. Sci.* 89, 1452 (2000).
111. Y. Ravindranath, *Curr. Opin. Oncol.* 15, 23 (2003).
112. C. P. Leamon, P. A. Parker, I. R. Vlahov, L. C. Xu, J. A. Reddy, M. Vetzel, and N. Douglas, *Bioconjug. Chem.* 13, 1200 (2002).
113. S. D. Konda, S. Wang, M. Brechbiel, and E. C. Wiener, *Invest. Radiol.* 37, 199 (2002).
114. S. Mayor, K. G. Rothberg, and F. R. Maxfield, *Science* 264, 1948 (1994).
115. R. G. Anderson, B. A. Kamen, K. G. Rothberg, and S. W. Lacey, *Science* 255, 410 (1992).
116. M. Wu, J. Fan, W. Gunning, and M. Ratnam, *J. Membr. Biol.* 159, 137 (1997).
117. C. P. Leamon, R. B. DePrince, and R. W. Hendren, *J. Drug Target.* 7, 157 (1999).
118. J. J. Turek, C. P. Leamon, and P. S. Low, *J. Cell Sci.* 106, 423 (1993).
119. J. T. Douglas, B. E. Rogers, M. E. Rosenfeld, S. I. Michael, M. Feng, and D. T. Curiel, *Nat. Biotechnol.* 14, 1574 (1996).
120. P. Ginobbi, T. A. Geiser, G. Ombres, and G. Citro, *Anticancer Res.* 17, 29 (1997).
121. J. Sudimack and R. J. Lee, *Adv. Drug Deliv. Rev.* 41, 147 (2000).
122. S. F. Atkinson, T. Bettinger, L. W. Seymour, J. P. Behr, and C. M. Ward, *J. Biol. Chem.* 276, 27930 (2001).
123. D. Goren, A. T. Horowitz, D. Tzemach, M. Tarshish, S. Zalipsky, and A. Gabizon, *Clin. Cancer Res.* 6, 1949 (2000).
124. M. J. Backer and J. M. Backer, *Bioconjug. Chem.* 12, 1066 (2001).
125. Z. M. Qian, H. Li, H. Sun, and K. Ho, *Pharm. Rev.* 54, 561 (2002).
126. R. A. Brekken, C. Li, and S. Kumar, *Int. J. Cancer* 100, 123 (2002).
127. L. Sepp-Lorenzino and K. A. Thomas, *Exp. Op. Invest. Drugs* 11, 1447 (2002).
128. R. Pasqualini, E. Koivunen, and E. Ruoslahti, *Nat. Biotechnol.* 15, 542 (1997).
129. I. Virgolini, M. Raderer, A. Kurtaran, P. Angerberger, S. Banyai, Q. Yang, S. Li, M. Banyai, J. Pidlich, and B. Niederle, *N. Engl. J. Med.* 331, 1116 (1994).
130. J. C. Q. Reubi, *J. Nucl. Med.* 41, 63 (1997).
131. M. Langer and A. G. Beck-Sickinger, *Curr. Med. Chem-Anti Cancer Agents* 1, 71 (2001).
132. A. V. Schally and A. Nagy, *Life Sci.* 72, 2305 (2003).
133. M. Shadidi and M. Sioud, *FASEB J.* 17, 256 (2003).
134. N. Karasseva, V. Glinsky, N. Chen, R. Komatireddy, and T. Quinn, *J. Protein Chem.* 21, 287 (2002).
135. W. A. Breeman, W. H. Baker, M. De Jong, L. J. Hofland, D. J. Kwekkeboom, P. P. Kooij, T. J. Visser, and E. P. O. Krenning, *J. Nucl. Med.* 40, 209 (1999).
136. R. C. O. Ladner, *J. Nucl. Med.* 43, 119 (1999).
137. F. Trejtnar, M. Laznicek, A. Laznickova, and S. J. Mather, *J. Nucl. Med.* 41, 177 (2000).
138. H. Kobayashi, T. M. Yoo, I. S. Kin, M. K. Kim, N. Lee, K. O. Webber, I. Pastan, C. H. Paik, W. C. Eckelman, and J. A. Carrasquillo, *Cancer Res.* 56, 3788 (1996).
139. T. C. Behr, X. M. Goldenberg, and W. Becker, *Eur. J. Nucl. Med.* 25, 201 (1998).
140. B. Wick and B. Gruner, *Cancer Lett.* 118, 161 (1997).
141. G. M. Dubowchik and M. A. Walker, *Pharmacol. Ther.* 83, 67 (1999).
142. U. Nielsen, D. Kirpotin, E. Pickering, K. Hong, J. Park, M. Refaat Shalaby, Y. Shao, C. Benz, and J. Marks, *Biochim. Biophys. Acta* 1591, 109 (2002).
143. M. Dorvillius, V. Garambois, D. Pourquie, M. Gutowski, P. Rouanet, J. C. Mani, M. Pugniere, N. E. Hynes, and A. Pelegrin, *Tumour Biol.* 23, 337 (2002).
144. M. Theobald, J. Biggs, D. Dittmer, A. Levine, and L. Sherman, *Proc. Natl. Acad. Sci.* 92, 11993 (1995).
145. T. Fojo, *Drug Resistance Updates* 5, 209 (2002).
146. S. Kessner, A. Krause, U. Rothe, and G. Bendas, *Biochim. Biophys. Acta* 1514, 177 (2001).

147. S. A. Mousa, *Curr. Opin. Chem. Biol.* 6, 534 (2002).
148. D. Hallahan, L. Geng, S. Qu, C. Searfone, T. Giorgio, E. Donnelly, X. Gao, and J. Clanton, *Cancer Cell* 3, 63 (2003).
149. E. H. Moase, W. Qi, T. Ishida, Z. Gabos, B. M. Longenecker, G. L. Zimmermann, L. Ding, K. Krantz, and T. M. Allen, *Biochim. Biophys. Acta* 1510, 43 (2001).
150. E. B. Kullberg, M. Nestor, and L. Gedda, *Pharm. Res.* 20, 229 (2003).
151. S. E. Duff, C. Li, J. M. Garland, and S. Kumar, *FASEB J.* 17, 984 (2003).
152. J. P. Codde, A. J. Lumsden, S. Napoli, M. A. Burton, and B. N. Gray, *Anticancer Res.* 13, 539 (1993).
153. A. Hagiwara, T. Takahashi, K. Sawai, C. Sakakura, I. Shirasu, M. Ohgaki, T. Imanashi, J. Yamasaki, Y. Takemoto, and N. Kageyama, *Anti-Cancer Drugs* 8, 666 (1997).
154. E. P. Sipos, T. F. Witham, R. Ratan, P. C. Burger, J. Barahan, K. W. Li, S. Piantadosi, and H. Brem, *Neurosurgery* 48, 392 (2001).
155. J. Fricker and F. Writer, *Drug Discov. Today* 6, 387 (2001).
156. M. Bolla, D. Gonzalez, P. Warde, J. B. Dubois, R. O. Mirimanoff, G. Storme, J. Bernier, A. Kuten, C. Sternberg, T. Gil, L. Collette, and M. Pierart, *N. Engl. J. Med.* 337, 295 (1997).
157. F. J. Lejeune, *J. Clin. Invest.* 110, 433 (2002).
158. C. Halin, V. Gafner, M. E. Villani, L. Borsi, A. Berndt, H. Kosmehl, L. Zardi, and D. Ner, *Cancer Res.* 63, 3202 (2003).
159. A. G. Cochran, *Curr. Opin. Chem. Biol.* 5, 654 (2001).
160. R. Krishna and L. D. Mayer, *Eur. J. Cancer Sci.* 11, 265 (2000).
161. S. Bennis, C. Chapey, P. Couvreur, and J. Robert, *Eur. J. Cancer* 30, 89 (1994).
162. A. Astier, B. Doat, M. J. Ferrer, G. Benoit, J. Fleury, A. Rolland, and R. Leverage, *Cancer Res.* 48, 1835 (1988).
163. H. Rahman, *J. Natl. Cancer Inst.* 84, 1909 (1992).
164. P. Gill, J. Wernz, D. Scadden, P. Cohen, G. Mukwaya, J. von Roenn, M. Jacobs, S. Kempin, I. Silverberg, G. Gonzales, M. Rarick, A. Myers, F. Shepherd, C. Sawka, M. Pike, and M. Ross, *J. Clin. Oncol.* 14, 2353 (1996).
165. S. A. Longman, *Cancer Chem. Pharm.* 36, 91 (1995).
166. I. Brigger, C. Dubernet, and P. Couvreur, *Adv. Drug Deliv. Rev.* 54, 631 (2002).
167. V. P. Torchilin and V. S. Trubetskoy, *Adv. Drug Deliv. Rev.* 16, 141 (1995).
168. A. E. Gulyaev, S. E. Gelperina, I. N. Skidan, A. S. Antropov, G. Y. Kivman, and J. Kreuter, *Pharm. Res.* 16, 1564 (1999).
169. S. Mitra, U. Gaur, P. C. Gosh, and A. N. Mitra, *J. Controlled Release* 74, 317 (2001).
170. R. Reszka, P. Beck, I. Fichtner, M. Hentschel, L. Richter, and J. Kreuter, *J. Pharm. Exp. Ther.* 280, 232 (1997).
171. A. Quintana, E. Raczka, L. Piehler, I. Lee, A. Myc, I. J. Majoros, A. Patri, T. P. Thomas, J. Mulé, and J. Baker, Jr., *Pharm. Res.* 19, 1310 (2002).
172. M. Y. Khalil, J. R. Grandis, and D. M. Shin, *Expert Review of Anticancer Therapy* 3, 367 (2003).
173. S. Shukla, G. Wu, M. Chatterjee, W. Yang, M. Sekido, L. A. Diop, R. Muller, J. J. Sudimack, R. J. Lee, R. F. Barth, and W. Tjarks, *Bioconjug. Chem.* 14, 158 (2003).
174. H. Hug, M. Los, W. Hirt, and K. M. Debatin, *Biochemistry* 38, 13906 (1999).
175. J. Hernando, M. Van Der Schaaf, E. M. Van Dijk, M. Sauer, M. F. Garcia-Parajo, and N. F. Van Hulst, *J. Phys. Chem. A.* 107, 43 (2003).
176. J. F. R. Kerr, A. H. Wyllie, and A. R. Currie, *Br. J. Cancer* 26, 239 (1972).
177. E. Finkel, *Science* 292, 624 (2001).
178. N. Zurgil, Y. Shafran, D. Fixler, and M. Deutsch, *Biochem. Biophys. Res. Commun.* 290, 1573 (2002).
179. D. R. Green and J. C. Reed, *Science* 281, 5381 (1998).
180. M. Demoy, T. Minko, P. Kopeckova, and J. Kopecek, *J. Controlled Release* 69, 185 (2000).
181. K. M. Heiskanen, M. B. Bhat, H. W. Wang, J. Ma, and A. L. Nieminen, *J. Biol. Chem.* 274, 5654 (1999).
182. Assays For Apoptosis, Section 15.5, Molecular Probes, Inc. Available at: www.probes.com (2002).
183. H. Madden, *Biomedical Products* 1, 42 (2001).
184. G. Majno and J. Joris, *Am. J. Pathol.* 146, 3 (1995).
185. A. Strebel, T. Harr, F. Bachmann, M. Wernli, and P. Erb, *Cytometry* 43, 126 (2001).
186. J. Liu, M. Bhalgat, C. Zhang, Z. Diwu, B. Hoyland, and D. H. Klaubert, *Bioorg. Med. Chem. Lett.* 9, 3231 (1999).
187. H. Hug, M. Los, W. Hirt, and K. M. Debatin, *Biochemistry* 38, 13906 (1999).
188. H. Z. Zhang, S. Kasibhatla, J. Guastella, B. Tseng, J. Drewe, and S. X. Cai, *Bioconjug. Chem.* 14, 458 (2003).
189. M. Melvin, J. Tomlinson, G. Park, C. Day, G. Saluta, G. Kucera, and R. Manderville, *Chem. Res. Toxicol.* 15, 734 (2002).
190. S. Yasuhara, Y. Zhu, T. Matsui, N. Tipirneni, Y. Yasuhara, M. Kaneki, A. Rosenzweig, and J. A. J. Martyn, *J. Histochem. Cytochem.* 51, 873 (2003).
191. R. Sgonc and J. Gruber, *Exp. Gerontol.* 33, 525 (1998).
192. L. F. R. Span, A. H. M. Pennings, G. Vierwinden, J. B. M. Boezeman, R. A. P. Raymakers, and T. De Witte, *Cytometry* 47, 24 (2002).
193. A. M. Steff, M. Fortin, C. Arguin, and P. Hugo, *Cytometry* 45, 237 (2001).
194. D. M. Tinucane, E. Bossy-Wetzel, N. J. Waterhouse, T. G. Cotter, and D. H. Green, *J. Biol. Chem.* 274, 2225 (1999).

195. K. J. Kelly, R. M. Sandoval, K. W. Dunn, B. A. Molitoris, and P. C. Dagher, *Am. J. Cell Physiol.* 284, C1309 (2003).
196. J. Liu and Y. Lu, *J. Am. Chem. Soc.* 124, 15208 (2002).
197. C. Berney and G. Danuser, *J. Biophys.* 84, 3992 (2003).
198. S. R. Stauffer and J. E. Hartwig, *J. Am. Chem. Soc.* 125, 6977 (2003).
199. K. Q. Luo, V. C. Yu, Y. Pu, and D. C. Chang, *Biochem. Biophys. Res. Commun.* 283, 1054 (2001).
200. I. V. Gopich and A. Szabo, *J. Phys. Chem. B.* 107, 5058 (2003).

CHAPTER 15

Design of Protein and Enzyme Mimetics

Garland R. Marshall,[1] Dennis P. Riley[2]

[1]*Center for Computational Biology, Departments of Biochemistry and Molecular Biophysics and of Biomedical Engineering, Washington University, St. Louis, Missouri, USA*
[2]*Kereos, Inc., St. Louis, Missouri, USA*

CONTENTS

ISBN: 1-58883-048-9

Handbook of Theoretical and Computational Nanotechnology
Edited by Michael Rieth and Wolfram Schommers
Volume 6: Pages (719–755)

1. INTRODUCTION

This chapter focuses on practical solutions to difficult problems, namely, de novo design and synthesis of novel proteins with specified properties including unique catalytic activities. Since these capabilities are just materializing on the horizon [1–3], one can only outline the different approaches, limited successes and failures, and speculate on practical solutions to the problems that limit current applications and constrain success.

Proteins provide structural integrity and molecular recognition as well as catalytic processing of most chemical conversions within biological systems. It has become clear with the advent of genomics and proteomics just how complex the interactions of proteins are within living systems. The number of expressed proteins in humans has recently been estimated to be approximately 30,000. In *E. coli*, the fruit fly [4], and yeast [5], the number of expressed proteins is less, around 6000–15,000, but each one, on average, has been estimated to interact with up to six other proteins [6]. In addition, protein expression and interactions were dynamic when the changes in expression and interaction during the cell cycle in yeast were determined [7].

1.1. Protein Engineering

As a working definition of protein engineering, we will use is design of novel three-dimensional protein scaffolds for orientation of amino acid side-chain functionality for molecular recognition and catalysis. Applications could include therapeutics, biomaterials, biosensors, and industrial catalysts.

1.2. Current Status of Protein Engineering

One major difficulty with protein engineering is the lack of ability to predict the three-dimensional structure of a protein from amino acid sequence alone. This is due to the small energy differences between several different protein folds, each of which can bury hydrophobic surface in a compact structure and satisfy most internal hydrogen-bonding groups. To distinguish between such alternative folds requires an accurate potential and adequate sampling of configurational space for each candidate fold to approximate the entropy of the hydrated system [8–10].

1.3. Fold Space

The primary goal of protein engineering as defined is to find a sequence of amino acids that will stabilize a particular three-dimensional structure with desirable attributes. Examples exist in the literature where particular folds have been stabilized by optimization of the amino acid sequences. Nevertheless, optimization utilizing a particular scaffold does not guarantee that the sequence may not, in fact, stabilize another fold, perhaps novel and not included in the database of known folds used in threading, even better.

The basic problem lies with the combinatorial nature of combination of torsional variables describing any particular fold. One description of this problem is referred to as Levinthal's paradox [11]. For example, an average protein of 100 residues has 198 backbone degrees of freedom plus approximately 250 side-chain degrees of freedom assuming $2\frac{1}{2}$ rotatable bonds as a reasonable average for side chains. Only exploring the three trans isomers for each rotatable bond $= 3^{448}$ combinations to be explored. This number is so large that there is no way any protein would have folded in the lifetime of the universe (estimated 10^{17} seconds since Big Bang) assuming reasonable times (picoseconds) for conformational transitions.

Nevertheless, proteins fold. It should be obvious that a brute-force approach to evaluating each candidate fold is implausible regardless of the computational power available. The problem is simplified, somewhat, by the fact that atoms do not overlap and the resulting excluded-volume effect helps prune the combinatorial explosion [11–13].

Because of this conceptual problem, there is considerable interest in protein folding, the mechanisms by which proteins manage to fold and avoid the combinatorial conundrum [14–16]. Nevertheless, for the purposes of this discussion, the folded state is assumed to be at equilibrium, a thermodynamic state function, and therefore, independent of folding pathway.

1.4. Cut-and-Paste

Many examples of protein engineering use prefolded subunits as scaffolds and simple transfer/evolve binding sites, or scramble protein domains [17, 18] to generate new functionality. This cut-and-paste approach assumes that the scaffolds and domains are sufficiently stable that they will retain the same three-dimensional structures despite any perturbation in amino acid sequences involved. The Hellinga group has focused on bacterial periplasmic binding proteins as functional scaffolds for grafting desired binding sites [19] with considerable success. Metal-binding sites [20] such as those for zinc [21] have been appended [22, 23] and eliminated [24] from proteins. Others have used covalent modification of existing proteins to generate enzyme activity [25, 26], or modified existing enzymes to cause a predicted change in function [27, 28].

Several recent examples of successful cut-and-paste applications are given below to illustrate the power of this approach:

1. Conversion of protein scaffold to enzyme—Dwyer et al. [29] were able to design a novel triose phosphate isomerase active site within the bacterial ribose-binding protein. By knowing the geometrical for the critical amino acids residues requires for catalytic activity in triose phosphate isomerases, an appropriate geometry was found in the ribose-binding protein that could accommodate the active site. Kaplan and DeGrado [30] have reported phenol oxidase activity in a de novo-designed four-helix bundle containing two iron-binding sites (Fig. 1).
2. Change of specificity of bacterial receptors as biosensors—Looger et al. [31] used the ribose-binding protein to which a fluorophore was attached in order to signal the binding of ribose. The binding site was then mutated to recognize TNT selectively and other small ligands instead of similar analogs of the ligands as shown in Fig. 2. Similar results were reported for the nerve gas agent soman [32].
3. Ligand activation of protein splicing—some bacteria have the capability of splicing two protein segment together through excision of an intervening small cysteine protease, or intein [33]. By fusing a domain of the receptor responsible for thyroid-hormone recognition to an intein, Skretas and Wood [34] generated an expression system that allows an inactive protein to be expressed within a cell, and then spliced together to generate an activate protein by the addition of thyroid hormone.

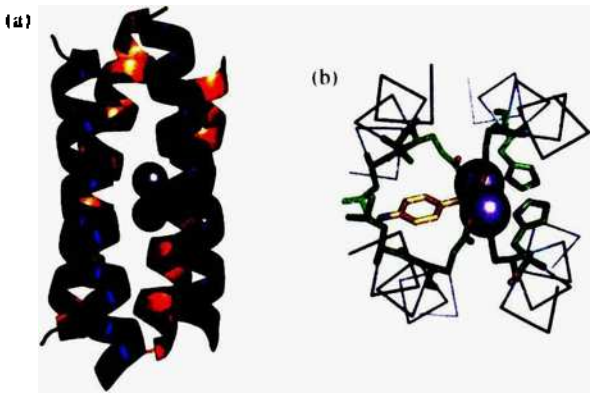


Figure 1. Schematic diagram of 4-helix bundle with two iron-binding sites (left); Model of active site with substrate interaction with diiron catalytic site C. M. Taylor, adapted from original.

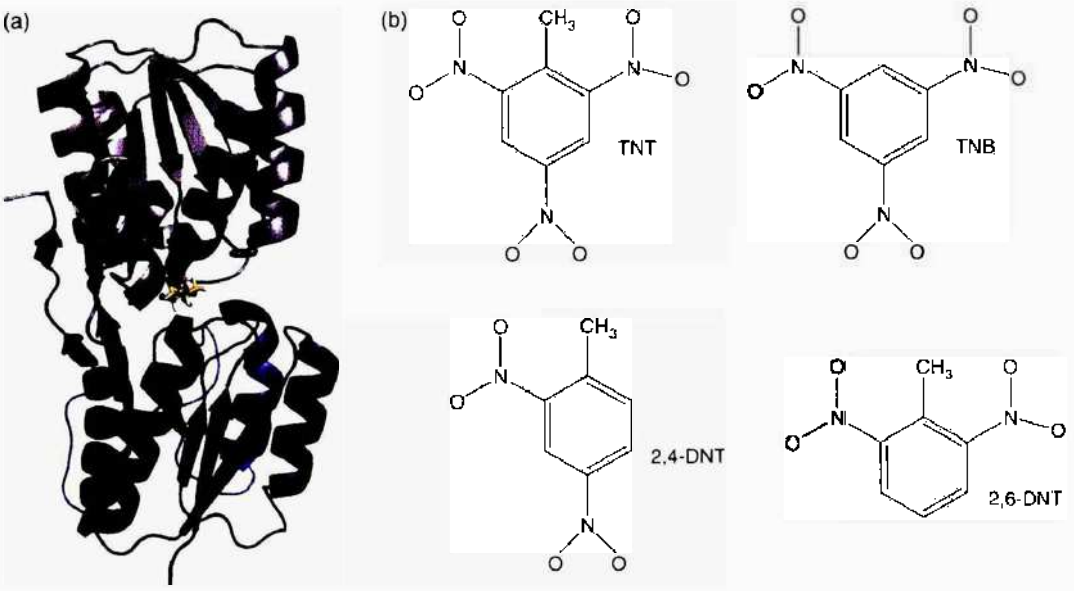


Figure 2. Conversion of ribose-binding protein (a) into a biosensor for TNT (b) that can discriminate TNB, 2,4-DNT, and 2,6-DNT.

These and many other examples from the literature illustrate the creative way in which novel proteins with designed properties can be generated by using of preexisting protein building blocks. Our final goal, however, is the ability to design proteins de novo from first principles.

1.5. De Novo Protein Design

Considerable effort has been expended to design proteins from first principles to test the functional level of our understanding of protein motifs and stability [35, 36]. The first serious effort that we recall is that of the Richardsons [37] to design a novel beta-barrel protein, betabellin. They enlisted the synthetic expertise of Bruce Erickson to prepare the designed protein and iterative cycles of design, synthesis, and disappointment ensued until the 15th design [38] gave the desired structure. A major difficulty with designing proteins with beta-sheets are their tendency to aggregate [39].

Examples of several successful miniproteins that have been engineered are shown in Fig. 3. Included are helices and helix bundles, sheet, beta-hairpins, and so on. These examples are mean to be illustrative, and other examples are becoming common in the literature [40–47].

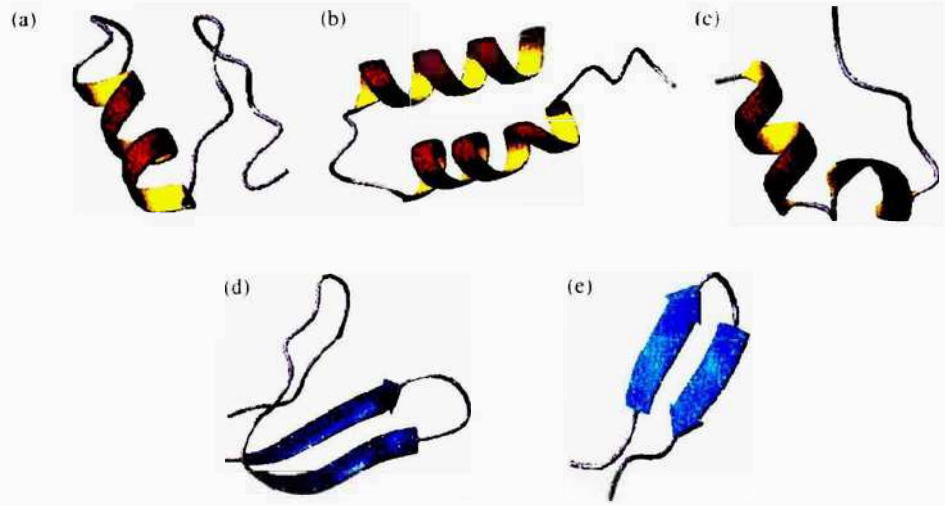


Figure 3. Examples of miniproteins that have been designed (optimized) to have the folds shown. (Top) predominately helical (a) Mayo et al. [42], (b) Wells et al. [43], (c) Andersen et al. [44]. (Bottom) Predominately sheet (d) Imperiali et al. [24, 45] and (e) Serrano et al. [46, 47].

Figure 4. Coiled-coil helix motif interacting with DNA.

Helix bundles have been the goal of many efforts in de novo protein design [40]. Hecht and co-workers have used binary patterning of polar and nonpolar amino acids [41] to guide folding in accord with the lattice-model studies of Dill and co-workers [14]. Recently, the DeGrado group has generated catalytic activity by adding a diiron catalytic center [30].

Oligomerization of designed units to explore the basis of protein/protein interactions has also been an active field. Hodges and co-workers have thoroughly explored the physical basis of coiled-coil interactions [48–51] (Fig. 4) and DeGrado and colleagues [52, 53] have successfully designed coiled-coil systems.

Helix bundles have been designed by DeGrado and co-workers [54–60]. The Imperiali group has focused on transformations of a zinc-finger, a betabetaalpha structure [24]. Recently, homotetramerization of a four-helix bundle has been engineered [61], and further design on the same system has led to a heterodimer (Fig. 5) [62, 63].

Molecular Interactions. To design proteins, one must understand the basis of molecular recognition; that is, how specificity in biological systems arises. Peptides, or small domains of proteins, continue to dominate molecular interactions in biological systems. Interactions of peptides with proteins [64] and, specifically, with G-coupled protein receptors, have been reviewed. What forces stabilize a protein-protein interaction? What is the basis of stability of one conformation of a protein domain versus alternatives?

Whether we are discussing intermolecular event, such as binding or aggregation, or intramolecular events, such as folding or conformational changes, the change in free energy of the system between the two states dominates what is observed.

$$\Delta G = \Delta H - T\Delta S$$

ΔG, change in free energy (folding or binding) = change in enthalpy ΔH minus the change in entropy ΔS times temperature T.

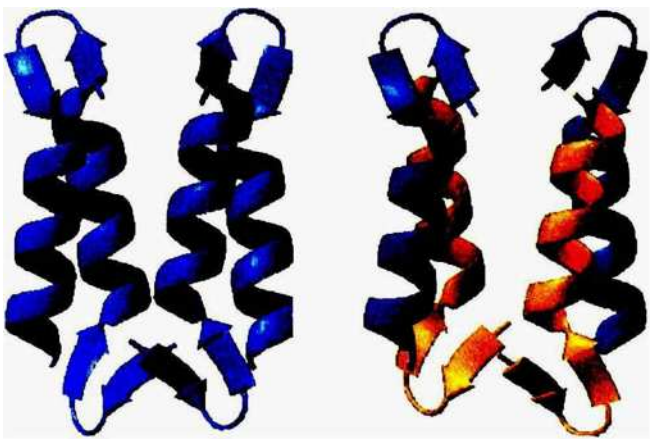


Figure 5. Designed homodimer (left) of two-helix bundles and conversion by design to heterodimer [63].

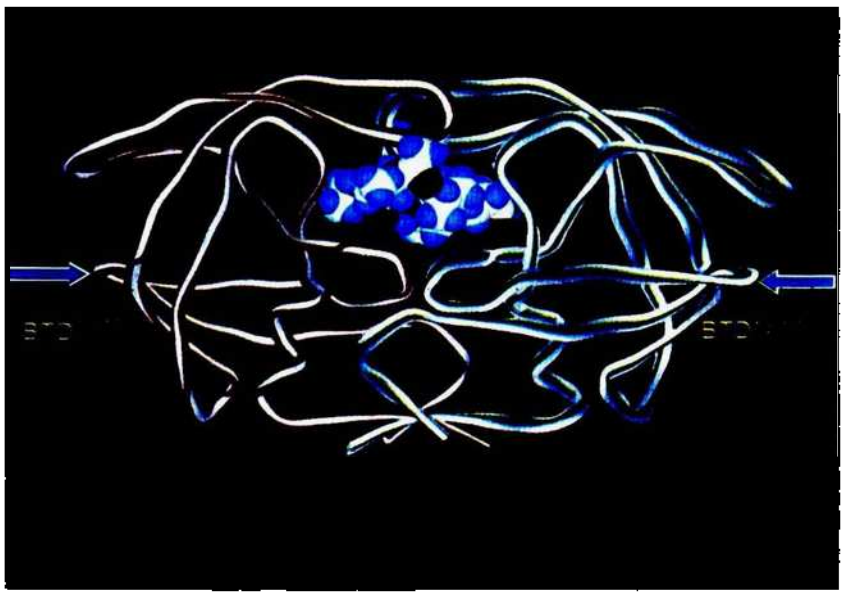


Figure 6. Schematic ribbon diagram of HIV protease homodimer complexed with inhibitor MVT-101. Two sites of incorporation of BTD to stabilize a reverse turn in the unfolded state are shown with arrows.

One major hurdle is the current focus on optimizing enthalpy that is largely dominated by packing. Multiple folds are compact with similar sums of van der Waals positive interactions and electrostatic forces are often similar between folds, distinguishing the correct fold by design is problematic. This has led us to a focus on using preorganization to steer a sequence into the designed fold by manipulating entropy.

One example from the literature illustrates the approach (Fig. 6). In the HIV-1 protease homodimer, a bicyclic dipeptide reverse-turn mimetic BTD (Fig. 7) was introduced in both monomers to replace the dipeptide unit Gly-Gly[16,17] that forms a reverse turn in the crystal structure [65] of HIV protease complexed with an inhibitor MVT-101. This forced a reverse-turn at those two positions in the unfolded state, reducing the entropic penalty for folding and resulted in an approximate 1.5-kcal/mol increase in protein stability [66].

To quote Baca et al. [66], "The free energy change associated with the folding of a protein into the native state can be expressed as:

$$\Delta G_{folding} = \Delta H_{chain} - T\Delta S_{chain} + \Delta G_{solvent}$$

where ΔH_{chain}, ΔS_{chain}, and $\Delta G_{solvent}$ are the changes in intrachain binding energy, chain entropy, and solvation free energy upon protein folding. The change in chain entropy on folding is unfavorable, so any conformational restriction of the polypeptide chain toward the folded structure should reduce the entropic cost of folding and result in a more stable three-dimensional fold."

The difficulty with this approach was the previous necessity to incorporate such semirigid segments of the protein chain by chemical synthesis of the entire protein since they could not normally be incorporated by protein expression. The recent development of expressed protein ligation [67, 68] offers a potential solution to this logistical bottleneck.

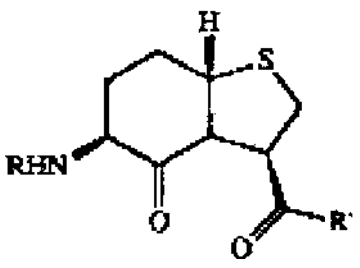


Figure 7. Bicyclic reverse-turn mimetic BTD used by Baca et al. [66] to replace Gly-Gly[16,17] in synthetic HIV protease.

2. MOLECULAR MECHANICS AND MOLECULAR SIMULATIONS

The use of force fields, which characterize atomic interactions by classical physics, provides a basis for molecular modeling and simulations [69]. Improvements in this approach through use of multipole electrostatics and polarizability have finally provided a tool that can calculate thermodynamic properties of systems to within experimental error [70, 71]. Thus, it is now feasible to calculate the affinity of one protein for a ligand or another protein with confidence assuming that adequate degrees of freedom within the interacting molecules are explored. Combinations of molecular mechanics with quantum mechanics now allow exploration of chemical reactions in which bond lengths are variables [72].

2.1. Empirical Observations

The increasing availability of crystal structures of protein complexes [73] has allowed the characterization of the interfaces between the proteins in the complex with the goal of understanding the interactions which stabilize such complexes and determine specificity [74–79]. Force field potentials have been developed based on statistical arguments and an inverse Boltzman approximation to derive energies of atomic interactions [9, 80]. Another approach is characterization of the contact-surface area in protein-protein complexes that generally varies from 500 to 5000 square angstroms with some even larger [78, 81]. It was initially claimed that recognition sites on proteins are often flat and indistinguishable in shape from other parts of the protein surface [76, 82]. As a significant part of the interaction at the interface of complexes appears to be hydrophobic, it is not surprising that hydrophobic amino acid residues are enriched in the interface in hydrophobic patches of 200–400 square angstroms [76, 77, 83]. The relationship between buried surface area and affinity has been reviewed [84].

2.2. Recognition "Hot Spots"

Some side chains within the interface play a more significant role than others in the energetics of binding and in the determination of the relative orientation of the two proteins. Experimentally, this information on "hot spots" has been obtained by systematic mutagenesis of side chains within the interface to alanine and determination of the changes in binding affinity. Bogan and Thorn [85] collected a database of 2,325 alanine mutants for which the change in free energy of binding upon mutation to alanine had been measured; an updated database ASEdb is available at http://www.asedb.org. Analysis of the database by Bogan and Thorn [85] generated several observations; amino acid side chains in hot spots are located near the center of protein/protein interfaces, are generally solvent inaccessible, and are self-complementary across protein/protein interfaces; that is, they align and pack against one another. Out of 31 contact residues involved in the interaction of growth hormone with its receptor, for example, two tryptophan residues of the receptor accounted for more than 75% of the free energy of binding as determined by mutation to alanine [86]. In the case of the growth hormone itself, eight of the 31 side chains involved in the interface accounted for approximately 85% of the binding energy; thus the genesis of the "hot-spot" theory [87] as a basis for inhibitor design and drug discovery. To emphasize the importance of side chains in peptide recognition, in the case of recognition of the hormone somatostatin by its G-protein coupled receptors (GPCRs), many of the amide bonds can be reduced [88], the direction of the peptide backbone reversed, or even the whole peptide backbone replaced by a saccharide with recognition retained at the receptor [89, 90]. Similar studies on the tripeptide hormone, thyrotropin TRH, Glp-His-Pro-NH_2, led to a CNS-active analog with the backbone amides replaced by a cyclohexyl scaffold [91] that did not release TSH. Our own studies [92, 93] determining the bioactive conformation of TRH have led to polycyclic analogs that retain full activity at the endocrine receptor responsible for TSH release (Fig. 8) [94].

2.3. Side-Chain Recognition

Side-chain recognition is dominant in peptide-receptor complexes. "Aromatic residues are found consistently to play a special role in the recognition and activation of receptors...

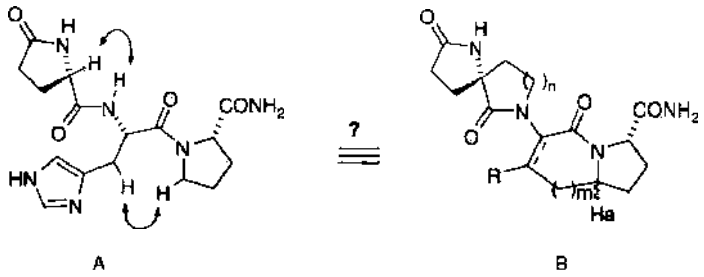


Figure 8. Analysis of TRH (A) analogs by the active analog approach [94] led to a proposal for the receptor-bound conformation compatible with internal cyclization to generate polycyclic analogs (B) [92, 93].

The rigid arrangement of atoms and the resulting fixed large-surface area of aromatic side chains, such as tyrosine, tryptophan, histidine and phenylalanine, combine to maximize the potential free energy of interaction as the entropic cost of assuming a particular geometry has already been paid" [95]. Charged groups, particularly the planar guanidium of Arg and the carboxyl groups of Glu and Asp are also often essential recognition hot spots in peptide messages providing a geometrical array of hydrogen bond donors and/or acceptors. Asn and Gln also have planar groups with distinctive hydrogen-bond geometries. These observations are also valid within the interfaces of protein/protein interactions [79] as formation of salt bridges across the intermolecular interface is highly favorable [96, 97]. From the analysis of Bogan and Thorn of amino acid preference in hot spots, similar conclusions support those of Marshall derived from structure-activity studies on peptide ligands: tryptophan is most highly enriched almost four-fold, (Trp having the largest planar surface), followed by Arg, Tyr, Ile, Asp, and His, respectively, with 50% enrichment for His. One significant difference between peptide/ligand binding and protein/protein interfaces, of course, is that interfaces between proteins are most often composed of amino acid sequences that are not contiguous, while the small size of peptides usually means that the peptide member has an interacting surface composed of adjacent amino acids, such as occur in a reverse turn, or on the surface of an α-helix. One additional difference is that the peptide has little if any intrinsic structure in solution, and the entropic cost of formation of the complex is, therefore, much greater.

The problems associated with inhibiting protein/protein interactions has been recently reviewed [81, 98, 99]. In this overview, the current status of efforts to mimic the structural motifs seen in protein/protein interactions with druglike organic molecules will be discussed with regard to reverse-turn, α-helix, β-strand and other noncontiguous hot spots (or epitopes). While the noncontiguous motifs are the most challenging, recent systematic efforts to identify and classify such motifs as well as virtual screening for such mimetics have begun to appear [100, 101].

2.4. Logical inference

Given the fact that side-chain recognition dominates biological interactions, one can use peptidomimetics in protein design to preorganize segments of the protein fold without worrying about recognition of the peptide backbone. In addition incorporation of peptidomimetics will tend to prevent degradation of the designed protein, prolonging its half-life and increasing the probability of oral bioavailability. In essence, incorporation of peptidomimetics in the place of secondary structures will force the resulting hybrid (chimeric) protein to adopt the desired fold by eliminating other potential folds that do not have the same pattern of secondary structural elements aligned in identical locations. The repertoire of secondary structural peptidomimetics to be incorporated into designed proteins is quite comprehensive.

2.5. β-Stand Motifs

A significant motif of peptide recognition involves interactions with the peptide backbone as commonly seen in proteolytic enzymes. The linear recognition motif in proteolytic enzymes allows alignment of the peptide backbone within the active site to give precise orientation to

the enzymatic functional groups responsible for amide-bond hydrolysis. In protein-protein recognition, β-strand recognition usually manifests itself as β-sheet assembly in complex formation. As an example, HIV protease is a homodimer with each of two active-site aspartyl residues contributed from each monomer. A four-stranded β-sheet structurally stabilizes the dimer and is responsible for more than 80% of the stabilization energy [102]. Several attempts have been made to inhibit dimer formation with β-sheet mimetics [103–105].

2.6. α-Helix Motif

The most common secondary structure element in molecular recognition is the α-helix. Besides the obvious significant role of α-helices in regulation of expression by binding to nucleic acids, α-helices have been shown to play a major role in protein/protein recognition. The prototype for experimental study was ribonuclease S (Fig. 9) in which the amide bond between residues 20 and 21 of ribonuclease A had been cleaved by subtilisin [106]. The cleaved enzyme remained enzymatically active and the 20-residue S-peptide could be reversibly dissociated, and the energetics of recognition studied [107] by calorimetry, where a mutation of Met-13 to glycine eliminates half the binding energy (some of which must be due to increased entropy of the glycine mutant in the free peptide). Another system of more current biological interest in which the α-helix plays a dominant role in complex formation is the binding of a helix of p53 to a hydrophobic cleft on the surface of Hdm2. Recently, Vassilev et al. [108] have described a small molecular weight drug candidate for the treatment of cancer that inhibits this interaction by mimicking the interaction of the three amino acid residues Leu, Phe, Trp) crucial for binding of the helix [109]. A short helical octapeptide had previously been described [1, 110] as a nanomolar inhibitor that contained two α,α-dialkylamino acids to stabilize the helical conformation [111] and orient the three side chains of Leu, Phe, and Trp. Developing inhibitors of protein/protein complexes where the interacting surfaces arise from discontinuities in the peptide backbone is naturally more problematic.

3. RECEPTOR-BOUND CONFORMATION OF PEPTIDES

Past approaches to determine the conformation of biologically active peptides when bound to receptors of unknown three-dimensional structure have achieved limited success. In part, this is due to the serial manner in which the research has been done in the past, as well as the synthetic difficulty in removing the rotational degrees of freedom from a peptide without concomitant loss of receptor recognition. Little evidence supports direct interaction of the amide bonds of the peptide backbone in receptor recognition, in contrast to recognition by proteolytic enzymes. To emphasize the importance of side chains, in the case of somatostatin,

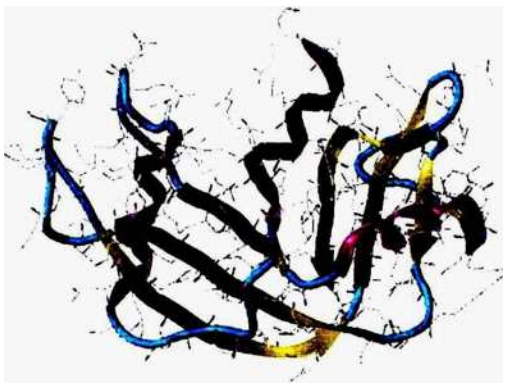


Figure 9. Ribonuclease S derived from ribonuclease A by treatment with subtilisin to cleave the amide bond between residues 20 and 21 to generate a reversible helix-binding system whose readout is enzymatic activity [106]. Structure of ribonuclease A (PDB accession code 1KF5.pdb), with secondary structural elements shown. Helices are magenta, sheets are yellow, and loop regions are cyan. The red arrow indicates the beginning of the α-helix of S-peptide, at Ala4. The green arrow indicates the subtilisin cleavage site.

many of the amide bonds can be reduced [112], the direction of the peptide backbone can be reversed, or even the whole peptide backbone can be replaced by a saccharide with recognition retained at the receptor [113, 114]. Similar studies on the tripeptide hormone, thyrotropin TRH, Glp-His-Pro-NH_2, led to a CNS-active analog with the backbone amides replaced by a cyclohexyl scaffold [91] that did not release TSH. Our own studies [115–117] determining the bioactive conformation of TRH have led to polycyclic analogs that retain activity at the endocrine receptor responsible for TSH release, but do not show the anticipated increase in binding affinity to be expected if the receptor-bound conformation were mimicked exactly.

Moreover, the chemical structure of peptides leads to degradation in biological systems limiting oral bioavailability and duration of action. Peptidomimetics, more drug-like structures that interact with peptide receptors, are sought through combinatorial chemistry and high-throughput screening. Combinatorial chemistry [118] has taught us that many different chemical structures can interact with a given G-protein coupled receptor (GPCR), and yield therapeutic candidates with nanomolar affinities upon optimization. Frequently, nonpeptide lead compounds have been isolated using high-throughput screening against peptide receptors [119]; thus, the concept of privileged organic scaffolds has emerged. The essence of the *privileged organic scaffold* concept [120, 121], is that a chemical scaffold, proved successful in one receptor system, such as a GPCR can be used to generate a library that can then be successfully screened against other GPCRs. Combinatorial chemists often use this concept to develop leads that interact with GPCRs, integrins or other receptors [122]. Certainly, the wide variety of organic scaffolds that has resulted from screening against particular receptors has provided multiple opportunities for lead optimization and circumvention of patent protection.

3.1. Peptidomimetic Scaffolds

Examples of privileged organic scaffolds include, for instance, the benzodiazepine scaffolds (Fig. 10) used by Evans et al. [120] that are thought to mimic a reverse turn (Fig. 11) [124]; benzodiazepine analogs continue to generate leads against multiple peptide receptors

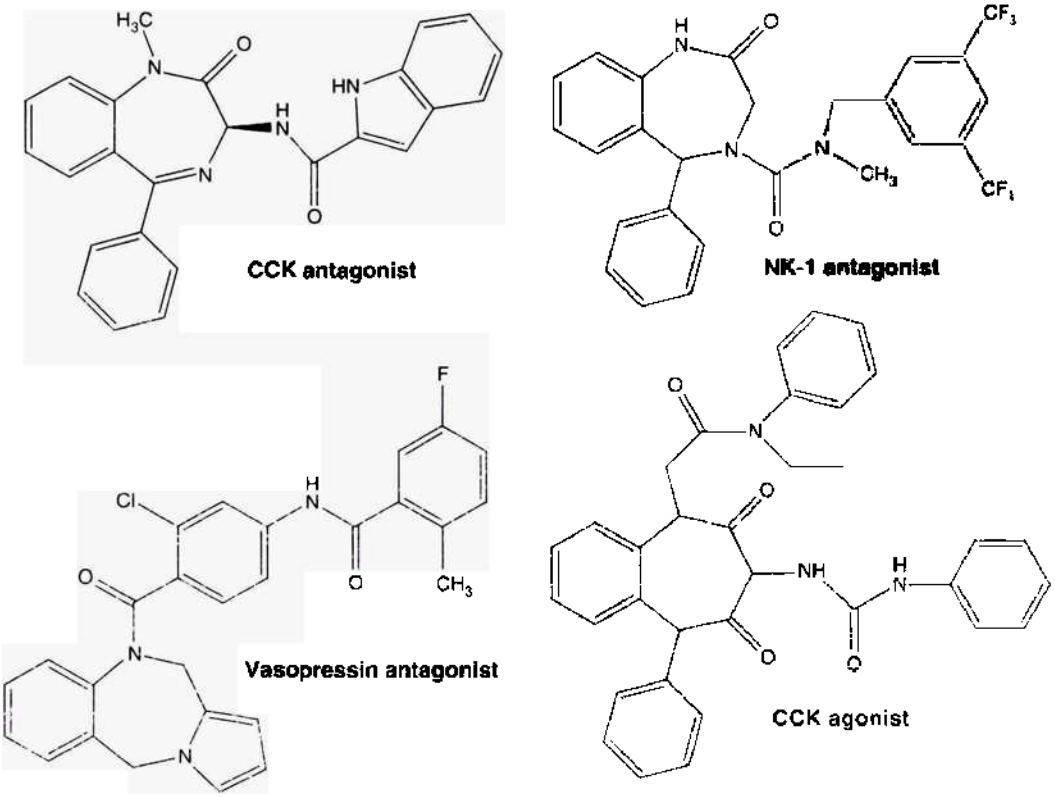


Figure 10. The benzodiazepine scaffolds suggested as privileged in deriving drugs that interact with GPCRs.

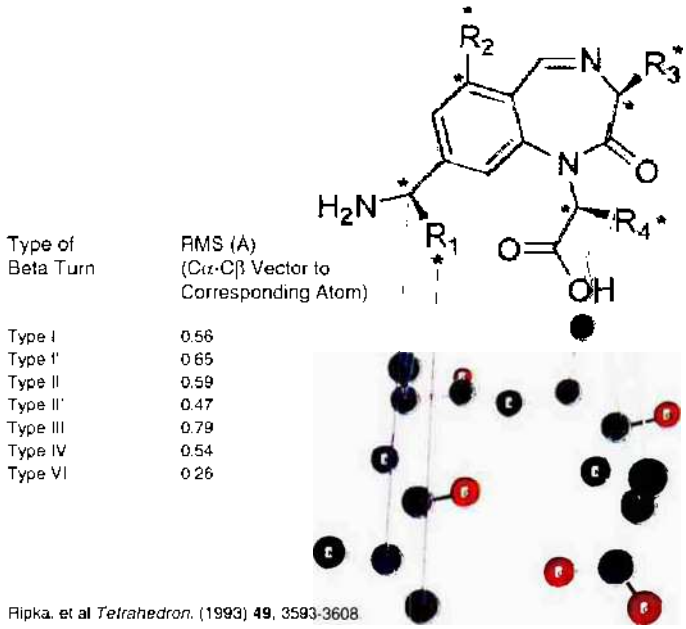


Figure 11. Geometrical overlap of Cα-Cβ vectors of peptide reverse-turn with substituents of benzodiazepines scaffold [124].

[125–128]. Haskell-Luevano et al. [129] screened a library of 951 compounds based upon the β-turn motif and identified the first two nonpeptidic heterocyclic micromolar agonists associated with the melanocortin-1 receptor. Another example is the rapid identification of selective agonists of the five somatostatin-receptor subtypes through combinatorial chemistry [130]. Rigid arrangement of atoms and the resulting large, fixed surface area of the important "side chains" in peptide terms in the privileged scaffolds maximize the potential free energy of interaction, as the entropic cost of assuming a specific geometry recognizable by the receptor has already been paid in these rigidified compounds.

Introduction of conformational constraints, however, is a "double-edged sword." The more one rigidifies a ligand, the less likely it will orient the functionality responsible for complementary interaction with the receptor. Unless, of course, the structure of the complex of the receptor and a ligand to be mimicked is known. Then, one can attempt to design a preorganized ligand that can present a similar binding surface to the receptor. However, many of the known privileged scaffolds, such as those for reverse-turn mimetics [131], including our own [132, 133] present synthetic difficulties in orienting the peptide-mimetic side chains on the scaffold [131, 134]. Cyclic peptide scaffolds, however themselves can serve as privileged scaffolds, because data obtained by structure-activity studies of the parent peptide and efforts to determine the receptor-bound conformation are often decisive in library design and lead optimization. Excellent examples of the evolution of peptide to peptidomimetic drugs are HIV protease inhibitors; one can easily trace the evolution from transition-state analogs of peptide substrates to the nonpeptide analogs used in the clinics. For example, selective agonists [130, 135] for all five somatostatin receptor subtypes were derived from work at Merck on peptide analogs of somatostatin [130]. Analogs of the somatostatin agonist L-363,301 c[Pro-Phe-D-Trp-Lys-Thr-Phe] yielded an antagonist of the neurokinin-1 receptor as well as μ- and δ-opioid receptor antagonists [136]. Similar overall philosophy was employed by the Bartlett group on α-amylase inhibitors [137] who designed cyclic peptides to mimic a turn motif from an inhibitory protein, tendamistat. Kessler has advocated cyclic heterochiral penta- and hexapeptides as conformational scaffolds for probing receptor recognition, where a recognition motif such as RGD or LDT for integrin receptors is systematically shifted around cyclic peptide backbone structures containing different chiralities to spatially sample various conformations [138–149], ultimately yielding compounds with nanomolar affinities and high selectivity. Solution conformational studies of the active cyclic peptides were performed to identify the conformational preferences associated with the most active scaffold. The Kessler group, in collaboration with Merck KgaA, has used

the results from their "spatial screening" with constrained cyclic peptides to guide the development of selective nanomolar nonpeptide molecule inhibitors for $\alpha v\beta 6$, $\alpha v\beta 5$ and $\alpha v\beta 3$ integrins [149]. One peptidic $\alpha v\beta 3$ inhibitor cRazaGDfNMe (V) was reported in Phase II clinical studies and formed the basis for design of nanomolar nonpeptidic clinical candidates [150]. Here and throughout the text the D-amino acids are shown in one-letter code with a lower case. Porcelli et al. [151] used this approach to discover a novel substance P antagonist. While the evidence is clear that such an approach works, expediency has led industry largely to abandon this approach in favor of combinatorial chemistry and high-throughput screening. The other reason preventing more wide use of cyclic peptides as conformational templates was the fact that even cyclopentapeptides do not represent a conformationally rigid scaffold [152] so that translation from an active cyclopentapeptide to a nonpeptide drug was problematic.

A diverse library of highly constrained peptide mimetics could be used as parallel probes to interrogate the receptor regarding three-dimensional (3D) recognition. The diversity of unambiguous 3D orientations of side chains in such a library is a measure of its conformational information content. It is our belief that development of such libraries would provide an efficient way to determine if side-chain recognition is present in a therapeutic target.

3.2. Secondary-Structure Mimetics

Several groups have been prominent in the design/development of reverse-turn mimetics [153, 154] particularly heterochiral chimeric dipeptides analogs of Pro-D-pro and D-pro-Pro as β-strand nucleation sites [155, 156], α,α-dialkyl amino acids as helical nucleation sites [111], and metal complexes of chiral azacrowns as reverse-turn mimetics [153]. Others have developed semirigid β-strand analogs [157] and more recently organic semirigid α-helical segments [2, 158, 159]. The basic approach focuses on preorganization to decrease entropy loss on binding through introduction of nucleation sites to limit conformational flexibility. As an example of the impact of preorganization on energetics, an unusual bicyclic dipeptide reverse-turn analog BTD was incorporated into HIV protease by chemical synthesis and showed the anticipated increase in stability [66].

3.3. Helical Peptidomimetics

Helices are found in proteins predominately as α-helices (3.6 residues/turn), but occasionally as 3_{10}-helices (3 residues/turn) as shown in Fig. 12. Transitions between the two in isolated helices are extremely facile; the barrier has been estimated [160] between the two conformation of the helices shown in Fig. 12 to be on the order of 3–4 kcal/mol. Helix geometry of the peptide backbone allows the single hydrogen bonds of either helical conformer to make two weaker bifurcated hydrogen bonds in the transition state. The lowness of this barrier

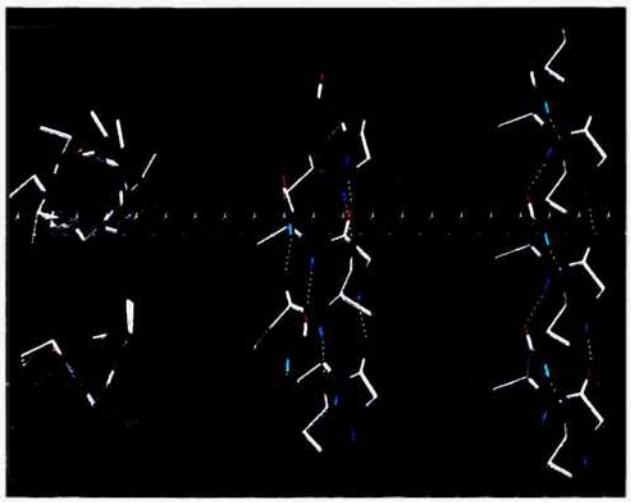


Figure 12. Ideal helices: α-helix (top left and middle); 3_{10}-helix (bottom left and right).

suggests that small peptide helices can easily be induced to bind in either conformer or a mixture of the two helical states by their receptor. We hope to design helical peptidomimetics that can distinguish between the two helical states by fixing the position and relative orientation of side chains.

The problem of organic helix mimetics is simply to design an appropriate organic scaffold with limited conformations that can orient attached R-groups to correctly mimic the orientation seen in α- and 3_{10}-helices (Fig. 13). There are 3.6 residues/turn of the α-helix, with a rise of 1.5 Å per residue. Since an α-helix often interacts by a common surface of the i, $i+3$, $i+4$, and $i+7$ residues, the characteristic axial rise between these residues is 4.5 Å or 6.0 Å, respectively. Looking down the helical axis, residues are projected at $-60°$ and $40°$ for $i \rightarrow i+3$ and $i \rightarrow i+4$ interactions, respectively. Just as reverse turns are common recognition motifs in biological systems, helix recognition is common in DNA/RNA-protein [161–163] and protein/protein interactions as exemplified below. We have chosen to simplify the potential surface of a potential peptide therapeutic by using preorganization to direct the conformation to that required for 3D complex formation with the target protein of the pathogen. It is estimated that elimination of a single rotational degree of freedom by preorganization stabilizes the receptor-bound conformation should enhance affinity by approximately 1.2–1.6 kcal/mol assuming complete loss of rotational freedom [164]. Thus, preorganization of a seven-residue helical segment with multiple 12–14 rotational degrees of freedom in the peptide backbone, for example, fixed into its bound conformation should enhance the binding affinity by several orders of magnitude. For this to be a viable option, we must demonstrate and develop appropriate methodology for the routine synthesis of helical peptidomimetics. Jacoby suggested 2,6,3′5′-substituted biphenyls as better than allenes, alkylidene cycloalkanes and spiranes as α-helix mimetics of side-chain positions i, $i+1$, $i+3$, and $i+4$ [158]. The Hamilton group suggested the terphenyl scaffold as α-helix mimetics of side-chain positions i, $i+1$, $i+3$, and $i+4$ [159]. More recently, the Schepartz group has suggested helical β-peptides as helical peptidomimetics and demonstrated their utility in the p53/hdm2 system [165]. Kelso et al. have used a motif of HXXXH complexed with $Pd(en)^{2+}$ to preorganize peptides into α-helices [166]. Taylor has advocated the use of side-chain lactams for a similar purpose [167]. Templates that mimic the surface of 3_{10}-helices have virtually been ignored and represent a new direction for our research.

Marshall and Bosshard [168] predicted by computation in 1972 that the achiral α, α-dialkyl amino acid, α-methylalanine, or aminoisobutyric acid Aib would severely restrict the Φ, Ψ torsion angles of that residue to those associated with right- or left-handed helices (both α- and 3_{10}-helices). Subsequent experimental validation of that prediction is abundant, from our own lab [169–171] as well as others [172–176]. An example where α,α-dialkyl amino acids were used to induce a helix of the peptide in water that enhanced binding involves the p53/hdm2 helix recognition (Fig. 14) [1].

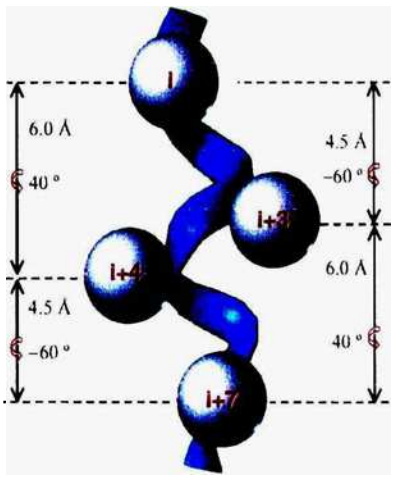


Figure 13. Idealized α-helix geometry with 4.5 or 6 Å rise between residues i, $i+3$, $i+4$, and $i+7$ of an α-helix, all located on same face of helix.

Figure 14. Enhanced binding activity of p53(16–27) analogs to hdm2 resulted when two Aib-like residues were included to induce helicity. NMR conformation of the most potent analog in water [1].

3.4. Bak/Bcl-xL System

Surface mimetics of the Bak BH3 domain are attractive therapeutic targets in oncology to bind to Bcl-xL and preclude it from complexing with proapoptotic Bak. The NMR structure [177] shows the 16-residue BH3 domain peptide (random coil in solution) from Bak residues 72 to 87, $K_d \cong 300$ nm binds in a helical conformation to a hydrophobic cleft on the surface of Bcl-xL (Fig. 15). From alanine scans of the Bak-peptide, it was clear that four hydrophobic residues Val74, Leu78, Ile81, Ile85 along one edge of the helix are critically involved in binding. Furthermore, it has been shown that the α-helix propensity of these peptides is decisive for strong binding [178]. The Hamilton group (Yale) has used this system to demonstrate the elegant concept of helical peptidomimetics and its application to therapeutic targets. The terphenyl analog with two carboxylic acids and the isobutyl, 1-naphthalenemethylene, isobutyl sequence (Fig. 15) shows the strongest binding to Bcl-xL with a *KD* value of 114 nM [2]. This work was extended with another class of helical scaffolds, the trispyridylamide (Fig. 16). In this case, those analogs had maximal binding affinities only in the low μ molar range [3].

Two questions are raised. What is an optimal helical scaffold with regard to conformational rigidity (decrease entropy loss during binding) and helical-surface mimicry to retain favorable enthalpic interactions as well as displacement of solvent? Can one design and predict the binding affinity of a helical peptidomimetic prior to synthesis?

Inhibition of HIV entry by complexation with gp41. A terphenyl helical peptidomimetic has been shown [179] to mimic the most recently introduced therapeutic Fuzeon®, T20, Enfuvirtide®, a 36-residue peptide targeting gp41 and preventing HIV/cell-membrane fusion (Fig. 17). Mimetic 1a effectively disrupts N36/C34 complexation with an IC_{50} value of 13.18 ± 2.54 mg/mL. The effects of antagonist 1a on HIV-1-mediated fusion were studied using a dye-transfer, cell-fusion assay; 1a shows inhibition of HIV-1-mediated cell-to-cell fusion with an IC_{50} value of 15.70 ± 1.30 mg/mL (Fig. 18).

3.5. Synthetic Feasibility and Side-Chain Orientation

Many of the known privileged scaffolds, however, such as those for reverse-turn mimetics [180], including our own [132, 133] present synthetic difficulties in orienting the peptidemimetic side chains on the scaffold [134, 180]. Cyclic peptide scaffolds, on the other hand,

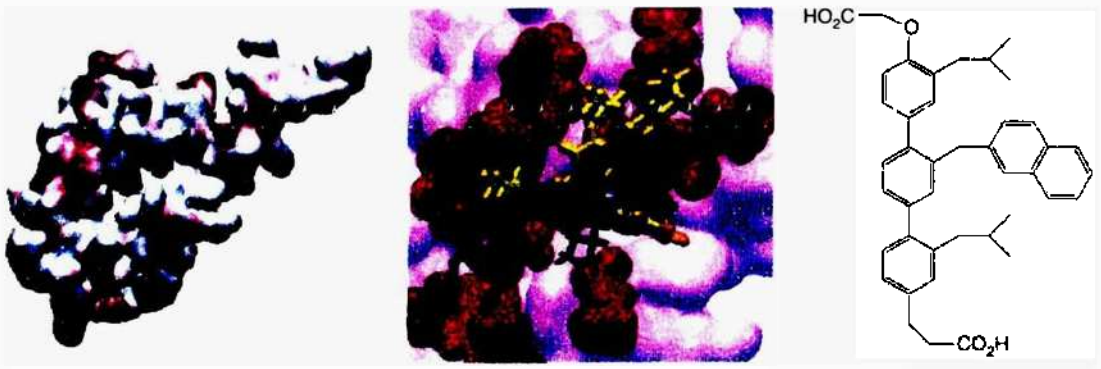


Figure 15. Helical peptidomimetic based on terphenyl scaffold for inhibition of Bcl-xL by the Hamilton group [2]. A diagram from Kutzki et al. [2] shows complex of Bak peptide with Bcl-xL and presumed overlap with peptide.

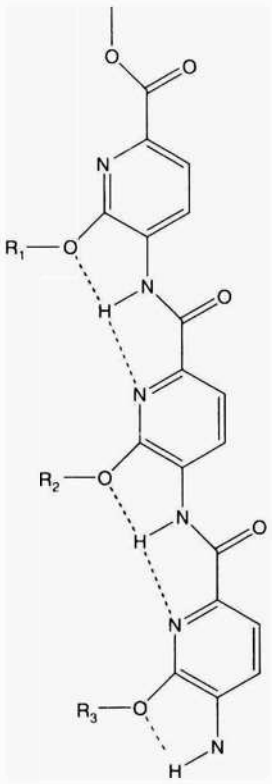


Figure 16. Trispyridylamides as potential helical peptidomimetics [3].

themselves can serve as privileged scaffolds, because data obtained by SAR studies of the parent peptide and efforts to determine the receptor-bound conformation are often decisive in library design and lead optimization. Excellent examples are HIV protease inhibitors where one can easily trace the evolution from transition-state analogs of peptide substrates to the nonpeptide analogs used in the clinics. For example, selective agonists for all five somatostatin receptor subtypes were derived from work at Merck on peptide analogs of somatostatin [130, 135]. Analogs of the somatostatin agonist L-363,301 c-Pro-Phe-D-Trp-Lys-Thr-Phe- yielded an antagonist of the neurokinin-1 receptor as well as μ- and δ-opioid receptor antagonists [136]. A similar overall philosophy was employed by the Bartlett group on α-amylase inhibitors [137] who designed cyclic peptides to mimic a reverse-turn motif from an inhibitory protein, tendamistat. Kessler has advocated cyclic heterochiral penta- and hexapeptides as conformational scaffolds for probing receptor recognition, where a

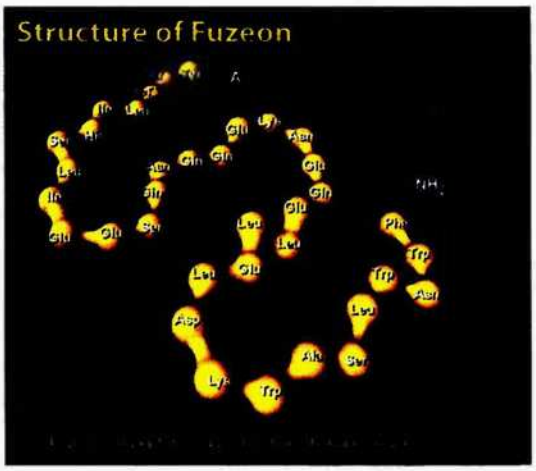


Figure 17. Structure of Fuzeon®, the latest approved drug for the treatment of HIV/AIDS (from Trimeris Web page).

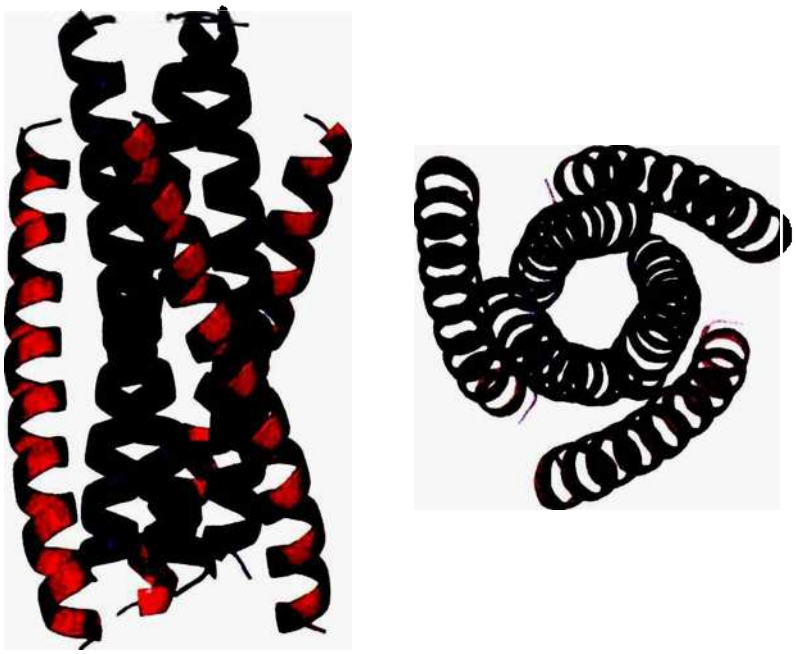


Figure 18. Sequence and orthogonal views of hexahelical complex of N-terminal and C-terminal segments of gp41 whose formation is thought to trigger fusion. The outside three helices are the ones being mimicked by Fuzeon™.

recognition motif such as RGD or LDT for integrin receptors is systematically shifted around cyclic peptide backbone structures containing different chiralities to spatially sample various conformations [138–149], ultimately yielding compounds with nanomolar affinities and high selectivity.

3.6. Enzyme Design

It has been a goal of synthetic chemistry [181] for decades to design small molecular-weight enzyme mimetics in order to test their understanding of any subtleties of their mechanisms of catalysis [256–259]. Most of these efforts have had limited success [182–186], even when combined with selection [187] as in the case of abzymes [260–263]. Nevertheless, an enzyme mimetic of one of nature's most efficient catalysts, superoxide dismutase, has recently been designed that duplicates its catalytic efficiency. What follows is a discussion of the insights into catalysis and subsequent research that allowed this achievement as a case study.

4. SOD MIMETICS AS PROTOTYPES

Low molecular-weight catalysts that mimic a natural enzymatic function have potential utility for the treatment of diseases characterized by the overproduction of a deleterious metabolic by-product or foreign gene product. The discovery and development of pentaazamacrocyclic ligand complexes of manganese(II) as functional mimics of superoxide-dismutase (SOD) enzymes serves as an example of rational design of a catalytic drug, a synzyme or enzyme mimetic. Two very important design factors were critical for design of a drug. Since the metal center in a complex was to serve as the active site for superoxide-dismutase activity, it was necessary to not only be able to design a catalyst that affords the proper shape, atomic type of ligand donors, and donor number to afford fast catalytic rates, but it was essential that the metal be held/bound tightly so that no redox-active metal is allowed to dissociate from the complex *in vivo* to prevent toxic side effects.

We have been investigating for many years the development of synthetic enzymes (*synzymes*) as an approach to managing the types of diseases outlined above. As an example of this approach, we describe the discovery and subsequent development of highly *stable* Mn(II) complexes that possess high SOD activity—the two key attributes for a drug targeting superoxide. Key to success in designing and synthesizing highly stable and highly active SOD mimics was the development of a detailed understanding of the mechanism of action of the

native enzyme itself as well as of this particular class of SOD mimetics, and the subsequent development of a computer-aided design (CAD) paradigm based on molecular mechanics (MM) that made it possible to study and screen a large number of possible structures prior to embarking on complicated syntheses of highly substituted and constrained metal ligands.

4.1. Background

Superoxide dismutases are a class of oxido-reductase enzymes that catalyze the dismutation of superoxide and contain either Cu, Fe, or Mn at the active site, recently reported is a unique nickel-containing superoxide dismutase [188, 189]. The enzyme catalyzes the one-electron reduction product of molecular oxygen (Eqs. [1] and [2], where M^n is the metalloenzyme in the reduced state and M^{n+1} is the enzyme in the oxidized state) to oxygen and hydrogen peroxide [190, 191]. In effect the enzyme simply shuttles an electron between superoxide and its protonated form with its specificity due to the electrochemical potential of the two. The relative simplicity of this mechanism has led to a variety of other approaches to SOD mimetics [192–196]. SOD enzymes have been shown to have efficacy in animal models of disease states proposed to be, in part, mediated by superoxide, such as myocardial ischemia-reperfusion injury [197–199], inflammation [200–202], and cerebral ischemia-reperfusion injury [203–205]. Evidence for superoxide as a mediator in multiple disease states continues to accrue, for example, Parkinson's disease and neuronal apoptosis [206–210], cancer [211–215], and AIDS [216].

$$O_2 + M^{n+1} \longrightarrow O_2 + M^n \tag{1}$$

$$HO_2 + M^n \xrightarrow{H^+} H_2O_2 + M^{n+1} \tag{2}$$

Because of the limited therapeutic applications of natural SOD enzymes arising from a number of negatives associated with their protein nature, such as their lack of oral activity; inability to gain access to the intracellular space of cells, such as mitochondria, one site where superoxide is produced; immunogenicity when derived from nonhuman sources; bell-shaped, dose-response curves; short half-lives; and cost. We pursued the concept of designing synthetic, low molecular-weight mimetics of SOD enzymes that could overcome such limitations, and hence serve as pharmaceutical candidates affording a new and promising approach to the treatment of a variety of diseases in which superoxide plays a deleterious role.

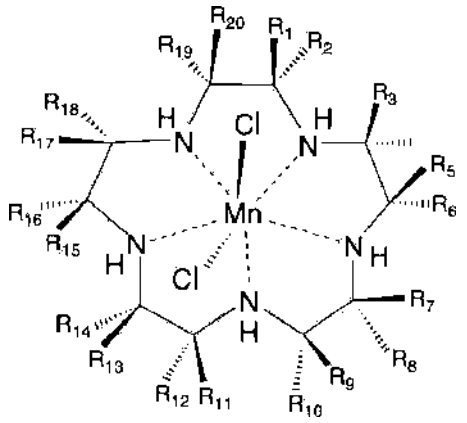
It was not our intent to transfer the active site of an superoxide-dismutase enzyme onto a small-molecule scaffold but rather to understand the mechanism of catalysis and duplicate the mechanism in a small-molecular weight, druglike molecule. A comparison of the structure of the active site of manganese SOD and M40403, the clinical candidate for an SOD mimetic is shown in Fig. 19.

4.2. Initial Structure-Activity Studies—Catalyst/Drug Design

Initial efforts were focused on the synthesis of manganese-based complexes as low molecular-weight SOD mimics; this was based largely on considerations of potential toxicity. Of the

Figure 19. M40403 SOD mimetic (left) compared with active site of human manganese superoxide dismutase (right). Note that inner shell ligands of manganese (red) are five nitrogens (yellow) of the pentaazacrown ring with two axial chloride ligands (green) in M40403 while the manganese ligands are three imidazole nitrogens and one carboxylate oxygen in the enzyme.

three metals known (Fe, Mn, Cu) at the time to catalyze SO to hydrogen peroxide and oxygen, manganese is the least toxic to mammalian systems as the free aquated metal ion, and also the least likely of the three M^{2+} ions to react with hydrogen peroxide to generate hydroxyl radicals (Fenton chemistry). Early design efforts focused on the synthesis and screening of complexes of Mn(II) that would possess high chemical and thermodynamic stability. Much of the early synthetic efforts focused on complexes of Mn(II) with macrocyclic ligands that provide enhanced kinetic stability as compared to open or noncyclic ligands.



From our initial structure-activity relationship (SAR) studies, we discovered that a class of manganese II-based complexes, I, incorporating the macrocyclic ligand 1,4,7,10,13-pentaazacyclopentadecane (where R_1–R_{20} can be any number of defined substituents with known stereochemistry) effectively catalyzed the dismutation of superoxide [217]. The discovery of this class of Mn(II) complex as *functional* mimics of the SOD enzymes resulted from synthesis and screening studies utilizing a large number of different macrocyclic complexes of Mn(II) [218]. Even though more than 40 complexes with different macrocylic ligands containing Mn(II) were synthesized, only the Mn(II) complex of the unsubstituted ligand derived from **I** (**Ia**—the 1,4,7,10,13-pentaazacyclopentadecane ligand, R_1-R_{20} = H), [Mn([15]aneN$_5$)Cl$_2$], afforded a catalytic activity as monitored by stopped-flow, kinetic analysis of superoxide decay [219, 220]. This specificity of superoxide-dismutase activity was somewhat surprising; for example, the Mn(II) complex with the analogous 16-membered ([16]aneN$_5$)-ligand was at least two orders of magnitude less active. Simple chemical changes to the ligand, such as replacing one NH with an oxygen to give an ether linkage, eliminated all SOD activity. This complex, [Mn([15]aneN$_5$)Cl$_2$], is an excellent catalyst for the dismutation of superoxide to oxygen and hydrogen peroxide; possessing a k_{cat} of $4.1 \times 10^{+7}$ M^{-1} s^{-1} at pH = 7.4; that is, about 20% of the activity of the native human MnSOD enzyme [217]. The complex crystallizes in either a seven-coordinate geometry, as exemplified by the *trans*-dichloro derivative with a planar macrocyclic ligand conformation, or as a six-coordinate complex, as in the case of a nitrato nitrate complex, which possesses a folded ligand geometry and a six-coordinate pseudo-octahedral geometry [221, 222]. In each case the complexes are high-spin d^5 white materials possessing an average Mn(II) to N bond distance of 2.28 Å.

The [Mn([15]aneN$_5$)Cl$_2$] complex possesses a reasonable thermodynamic stability at pH = 7.4 ($\log K$ = 10.7) and excellent kinetic stability with the complex completely intact at pH = 7.4 with no metal dissociation for many hours even in the presence of EDTA [217], which possesses a higher binding affinity for Mn(II) than does the [15]aneN$_5$ macrocycle. Additionally, this complex possesses excellent oxidative stability with the Mn(II) to Mn(III) oxidation occurring at +0.77v (SHE) [217]. The observed stability of this complex was sufficient to assess this SOD mimic in a variety of *in vitro* and *in vivo* models of superoxide-mediated injury. Most notably, this complex exhibited efficacy in *in vitro* [218] and *in vivo* models of inflammation [219], myocardial ischemia-reperfusion injury [223–226], vascular relaxation, and restenosis [227, 228]. In addition, this and other complexes of this class of SOD mimetics derived from **I** have superior properties to the SOD enzymes in regard to their normal dose-response curve (no deleterious effects observed at high doses in animal models), cellular

permeability (dependent on the nature of the R groups), extended *in vivo* stability, lack of immunogenicity, and projected lower cost of goods.

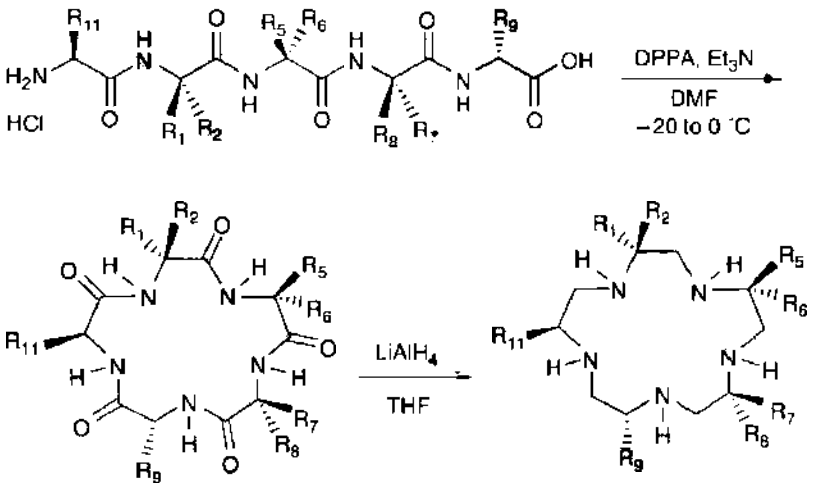
With positive biological data in hand, it was concluded that such synzymes might prove useful as therapeutic agents for the treatment of inflammation and myocardial ischemia-reperfusion injury, and other diseases mediated, in part, by overproduction of superoxide [229, 230]. As an added potential benefit, these SOD mimics have the unique property of potentiating nitric oxide, a vasorelaxant and antithrombotic [228, 229], by reducing or eliminating the diffusion-controlled reaction between superoxide anion and nitric oxide that yields peroxynitrite, another toxic metabolite derived from oxygen [231]. This nitric oxide reaction with superoxide has been measured to proceed with a nearly diffusion-controlled rate, but we have observed that the SOD mimics, **I**, limit the formation of peroxynitrite *in vivo* , and hence potentiate the lifetime of nitric oxide *in vivo* . This observation provided a basis for testing the compound in an animal model of platelet-mediated (NO blocks platelet aggregation) thrombosis in injured and stenotic arteries. The complex was shown to be protective as restenosis of the damaged vessel was inhibited by the mimetic when administered over a two-week period [228]. Subsequent studies reveal that the complex is effective in blocking reperfusion injury in an isolated rabbit-heart model [224], and *in vivo* in a canine model of cardiac reperfusion injury [225].

4.3. Development of Improved SOD Mimics Derived from Ia

The fact that the synthetic low molecular-weight functional SOD mimic, $[Mn([15]aneN_5)Cl_2]$, **Ia**, possesses biological activity of pharmacological relevance to human disease states, provided the basis for pursuing this class of molecules for the development of a pharmaceutical agent. There were two key features necessary for improving the $[Mn([15]aneN_5)Cl_2]$ complex for use as a human pharmaceutical agent: (1) increase the chemical stability via increasing the preorganizational rigidity of the macrocyclic ligand, thereby increasing the inherent kinetic stability of the complex [232] and (2) increase the SOD activity, so that increased catalytic activity would translate into a lower dose of mechanism-based drug, and hence, a diminished exposure to the metal-based drug reducing potential side effects.

Attempts to develop more active and more stable complexes probed the role that substituents (both on the N and C atoms of the parent macrocyclic ring) would exert on both the catalytic SOD activity and the overall chemical stability of the resultant complexes. How structural factors would affect these two key parameters were not immediately obvious; since it was not known at the outset how derivatized ligand systems would affect the catalytic activity. Thus, the number of substitutents, their placement, and their stereochemistry could all be critical design elements for maximizing catalytic activity and chemical stability. At the outset, we found that the nitrogen-substituted complexes examined possessed no measurable catalytic activity. Thus, we focused on the synthesis of C-substituted derivatives, **I**. Unfortunately, standard procedures for the synthesis of azacrown macrocyclic ligands employing Atkins-Richman [233] chemistry were inadequate for stereo-defined centers, as both the cyclization yields were poor, owing to higher oligomer formation, and poor yields from the acid hydrolysis/detosylation, owing to stabilized carbonium ion-induced ring cleavage/elimination. While this chemistry is sufficient for preparing unsubstituted and even monosubstituted macrocycles, yields for polysubstituted ligands were unworkably low. Consequently, better synthetic methods to such C-substituted ligands were successfully developed which used amide moieties as protecting groups for the amines during the cyclization [234]. Such methods included acid chloride cyclization, bis-chloroacetamide protection schemes, and cyclic pentapeptide techniques (Scheme 1) which made it possible to run high-concentration reactions and achieve high yields of the desired pentaazamacrocycle (>50%, in most cases) [233–235]. Additionally, combinations of these methods could be successfully incorporated into a single ligand preparation [235].

Employing such synthetic methodologies made it possible to synthesize highly substituted and stereochemically defined ligands that probed the effect that stereochemically defined substituents exert on the stability and catalytic activity. For these studies, we employed both methyl and fused-cycloalkyl substituents on the macrocyclic ring carbons to elicit the effect



Scheme 1.

of stereochemistry on stability and rate. Examples of ligands employed are shown in Fig. 20. All of the resultant Mn(II) complexes were characterized and shown to be similar to the parent unsubstituted complex; for example, high-spin d^5 Mn(II) *trans*-dichloro complexes with $E_{1/2}$SHE in the range of +0.74–0.78 v [235]. Both the thermodynamic and kinetic stabilities of these complexes were assessed [235], and several very significant features were discovered. First, increasing the number of C-substituents always increased the thermodynamic and kinetic stabilities of the complexes in a nonlinear manner; that is, stability increases geometrically with the number of substituents, so that, for example, the Mn(II) complex

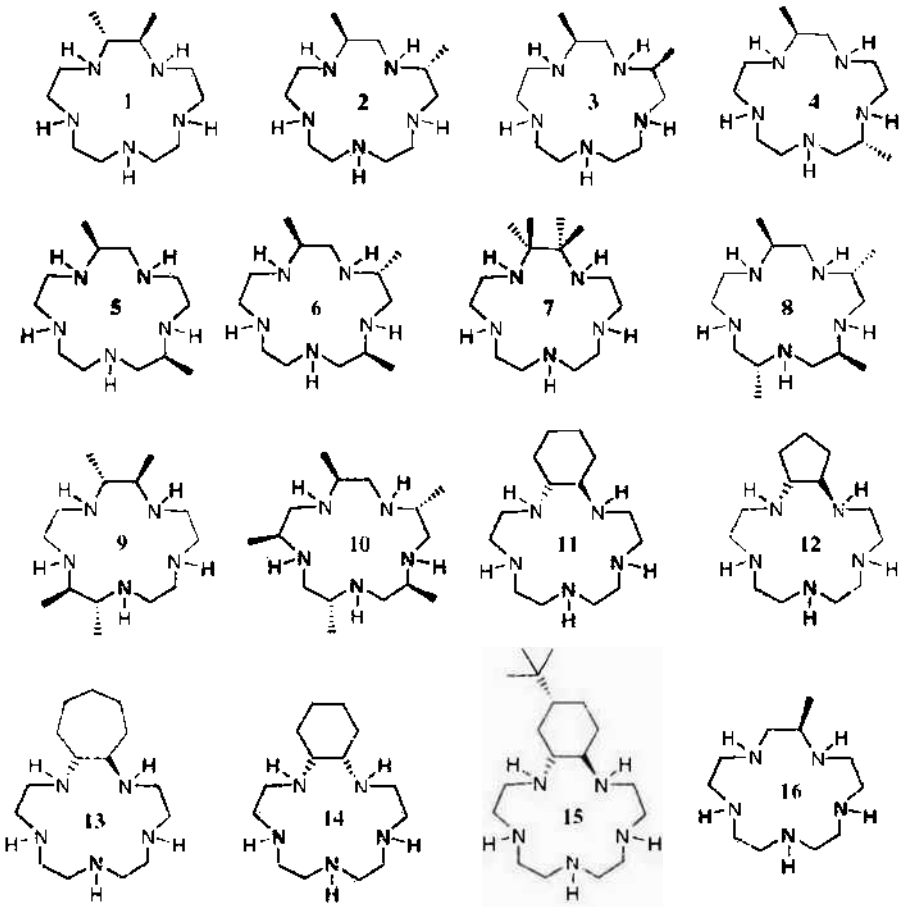
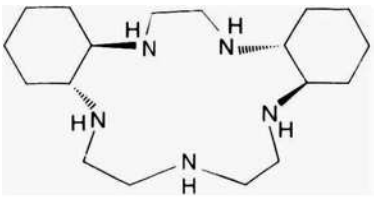
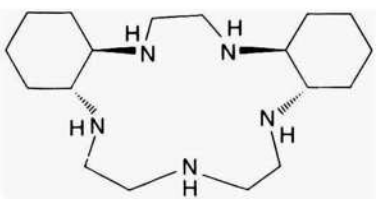


Figure 20. Ligands utilized as their Mn(II) complexes to probe the effects of the number and stereochemistry and placement of substituents on the stability and catalytic rate of SOD dismutation.

of the monomethyl ligand, 16, is about twice as stable as the unsubstituted complex, Ia, the Mn(II) complex of the pentamethyl ligand, 10, was more than 160 times more stable than the unsubstituted complex [235]. We also observed that the *trans*-cyclohexano complex, [Mn(*trans*-cyclohexano(15)aneN$_5$)Cl$_2$], derived from ligand 11 (see Table 1 for representative examples of complexes with k_{cat} and k_{diss} values), is more than twice as active as a SOD catalyst $k_{cat} = 9.1 \times 10^{7}$ M^{-1} s^{-1} at pH = 7.4 [236] and possessed both an improved thermodynamic stability log K of 11.6 and kinetic stability two times over the unsubstituted complex.



17



18

[Mn(*trans*-cyclohexano(15)aneN$_5$)Cl$_2$]. Additionally, we observed that the stereochemistry of the methyl substituents exerts a large effect on the catalytic rate, but had little effect on the stability of the complex as long as the number of substituents remains constant. A most remarkable aspect of the effect that substituents exert on the catalytic rate was revealed with the complexes of the macrocycles containing two *trans*-fused cyclohexano groups differing in the stereochemistry of their substitution. Such a change can dramatically increase stability and have profound effects on catalytic activity. For example, the k_{cat} for the Mn(II) complex of the all *R*-ligand, 17, or its all *S*-enantiomer, [Mn2*R*,3*R*,8*R*,9*R*-bis-cyclohexano[(15)aneN$_5$]Cl$_2$], at pH 7.4 is 1.2×10^{8} M^{-1} s^{-1} and possesses an increased thermodynamic stability log K = 13.3 and an enhanced kinetic stability at any pH over a 100-fold slower rate of dissociation results when compared to the unsubstituted complex, Ia. In contrast, the isomeric complex containing the 2*R*,3*R*,8*S*,9*S*-bis-cyclohexano-[(15)aneN$_5$] ligand, 18, possesses a similar stability profile, but has no catalytic SOD activity [236].

Clearly, the stereochemical orientation of the carbon substituents exerts a major role in determining the ability of a Mn(II) complex of this class to function as an SOD catalyst. Thus, our synthesis and characterization efforts, which were focused on the role that C-substituents have on both the catalytic SOD activity and the overall chemical stability of the resultant complexes, had revealed that the interplay between position, number, and stereochemistry of substituents in dictating catalytic activity was subtle and unobvious.

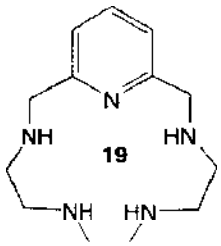
It should be noted that the complex derived from ligand 18 was characterized by X-ray crystallography and shown to have the Mn(II) arranged in a pentagonal bipyramidal geometry with *trans*-dichloro ligands and a planar macrocyclic orientation [222]. This complex

Table 1. List of representative complexes and their catalytic rate constants for the dismutation of superoxide at pH = 7.4 and their kinetic stabilities as a function of [H+].

Complex	k_{cat} (pH = 7.4) $\times 10^{-7}$ M^{-1} s^{-1}	k_{diss} (M^{-1} s^{-1})
Mn([15]aneN$_5$)Cl$_2$	4.1	2814
Mn(**10**)Cl$_2$	3.9	18
Mn(**11**)Cl$_2$	9.1	1375
Mn(**17**)Cl$_2$	12.1	28
Mn(**18**)Cl$_2$	<0.1	26
Mn(**19**)Cl$_2$	3.7	16
Mn(**20**)Cl$_2$	2.9	23
Mn(**28**)Cl$_2$	6.1	4.8
Mn(**29**)Cl$_2$	15.0	4.0
Mn(**30**)Cl$_2$	<0.1	5.9
Mn(**31**)Cl$_2$	4.56	1.42
Mn(**32**)Cl$_2$	2.0	<0.1
Mn(**33**)Cl$_2$	200	~0.7

and that of the parent unsubstituted complex, as well as the Mn(II) complex of ligand 7 both also characterized by X-ray crystallography show the same geometry and orientation of NH's. In each case the NH pattern is such that they alternate about the ring in an up : down : up : down : up orientation, so that the two sides of the macrocyclic ring in the complex are chemically distinct. We also observed this same NH pattern with the Cd(II) complex of the unsubstituted complex [237].

Our initial design goal was to maximize the number of substituents on the macrocyclic ring, thereby increasing the preorganizational rigidity of the macrocyclic ligand, and thus increasing the stability of the MnL complex to dissociation [232]. But we had no initial way of knowing in advance whether increasing the number of substituents on the parent macrocyclic ring would affect the catalytic activity in a beneficial or deleterious fashion. This necessitated that we develop a detailed mechanistic understanding of how these Mn(II) macrocyclic ligand complexes function as catalysts. This was predicated on the strong belief that once such information was in hand, the compounds could be subjected to computer-aided design (CAD) techniques, such as molecular mechanics (MM) calculations, to gain a quantitative and hence predictive understanding of how position, number, and stereochemistry of the substituents would effect the catalytic rate.



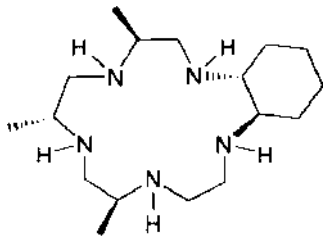
19

Since the sole pathway for the loss of Mn(II) from these pentaazacrown ligands is a dissociative one for these macrocyclic ligand complexes, they possess a high inherent stability [235]. In water, though, there exists a proton-dependent (first-order in $[H^+]$) pathway for dissociation, and at or near physiological pH, this is the major pathway for dissociation of the ligand from Mn(II) [235]. Thus, in addition to the rigidity of the ligand, the basicity of the ligand itself, is also critical to determining the kinetics rate of dissociation. An excellent example of this effect is observed with the Mn(II) complex of the pyridino ligand 19. This complex is nearly 175 times more kinetically stable at any pH than the parent unsubstituted ligand complex with Mn(II), Ia. This is undoubtedly due to some contribution from the rigidity conferred on the ligand by the pyridine substitution, but its complex with Mn(II) is even more stable than those of the biscyclohexyl ligand (17 and 18) complexes with Mn(II). This enhanced stability is most likely attributable to the much lower basicity of the ligand itself; namely, a measure of the affinity of the ligand for proton. For the aliphatic substituted ligands, three pK_a values are observed, with the values generally in the range: 10.5–11.0, 9.2–9.5, and 5.0–5.9. With the ligand, 19, these values are lower reflecting the weaker basicity of the ligand: 9.4, 8.8, and 5.3, and its lower affinity for a proton [235].

The structure of the Mn(II) complex of ligand 19, $[Mn(pyridin[15]aneN_5)Cl_2]$, is similar to other complexes of this class in that it crystallizes as a *trans*-dichloro pentagonal bipyramidal Mn(II) seven-coordinate complex with a planar array of the five nitrogens of the macrocyclic ligand, while possessing an alternating up : down : up : down N-H orientation in relation to the plane defined by the metal and the five nitrogens [238]. While it possesses an improved stability profile as compared to the parent unsubstituted complex Ia, it retains a SOD catalytic rate constant equivalent to it. Also, there are an even number of NH's, thus, the two sides of the macrocyclic plane, defined by the metal and the five nitrogens, are chemically equivalent.

The X-ray structure determination of the Mn(II) complex with ligand 20 has been published [239], revealing that this complex, as all the others that have been structurally characterized in this family of C-substituted [15]aneN_5 Mn(II) complexes, possesses the alternating NH pattern described above. It is of interest to note that this complex crystallizes with a water at one coordination site *trans* to a chloro ligand with an ionic chloride in the

crystal lattice. Further, the water is bound on the side of the macrocycle which possesses two *cis* non-adjacent NH's and the chloro ligand on the side with three *cis* NH's. At the time this complex was synthesized, its design was based on the empirical premise that more C-substitution is desirable for stability, but with no understanding of how to arrange the substituents to achieve a fast rate for the SOD reaction. In fact, the catalytic activity of this complex is only about 70% of the parent unsubstituted complex at pH = 7.4 and almost six times less active than the Mn(II) complex of ligand 17; thus, this complex represents an excellent example of the dilemma that was faced for the design of highly activity catalysts; namely, while it is very stable, it has lower catalytic activity. Clearly, a sound mechanistic footing was necessary to solve the problem of how to design a maximally effective SOD mimic with its catalytic activity being at least as good as that of the Mn(II) complex with ligand 17, but with significantly enhanced stability.



20

4.4. Mechanistic Studies

The need to understand the mechanistic details for the catalytic superoxide dismutation reaction was a critical element for the design of highly substituted complexes possessing both high stability and high activity. From kinetic and mechanistic studies, we have determined that the Mn(II) complexes of these penta-azacrown ligands function via a catalytic cycle in which the rate-determining step is oxidation of Mn(II) to Mn(III) [221, 222]. Two independent pathways operate for most of these complexes: (1) an outer-sphere, proton-coupled electron transfer from a bound water to an incoming hydroperoxy radical (Eq. [3]):

$$Mn^{II}(L)(OH_2) + HO_2^{\cdot} \xrightarrow[k_1]{} Mn^{III}(L)(OH) + HOOH \qquad (3)$$

and (2) an inner-sphere substitution involving coordination of superoxide anion to Mn(II) in a vacant axial site (Eq. [4]), followed by fast protonation of the bound superoxo creating a pseudo-octahedral Mn(III)-hydroperoxo complex (Eq. [5]):

$$Mn^{II}(L) + O_2^{\cdot-} \xrightarrow[k_2]{} Mn^{III}(L)(OO^{2-}) \text{ (inner-sphere substitution)} \qquad (4)$$

$$Mn^{III}(L)(OO^{2-}) + H^{+} \xrightarrow[fast]{} Mn^{III}(L)(O_2H) \text{(oxidation)} \qquad (5)$$

Isotope studies using D_2O as the reaction solvent were particularly helpful in providing further insight into these pathways. Using the fast catalyst derived from ligand 18 in both H_2O and D_2O, rate constants for both the proton-independent and dependent pathways for oxidation of Mn(II) were measured. In H_2O the pH-independent exchange rate-limited path (Eq. [2]) possesses a rate constant of 1.58×10^{7} s^{-1}, consistent with known aquo ligand exchange rates on Mn(II) [240]. In D_2O this rate is increased by ~10% consistent with water-exchange rates on the Mn(II) ion having a role in dictating this rate. Other complexes showed similar small effects on the magnitude of this rate constant when measured in D_2O. This pathway is constrained to never be much faster than $\sim 1 \times 10^{7}$ M s^{-1} (the water-exchange rate for Mn(II)) owing to the limit imposed by the need to generate a vacant coordination site by loss of water from the metal.

The measured rates for the outer-sphere pH dependent pathway of Mn(II) oxidation with this complex and that of Mn(II) complex with ligand 11 were both dramatically lowered in D_2O showing an isotope effect of nearly six consistent with H-atom transfer for the

proton-dependent pathway of oxidation of Mn(II) [222]. This pathway is most significant and contributes all of the rate enhancement observed for those complexes with faster rates than $1 \times 10^{+7}$ M^{-1} s^{-1}.

For either pathway to be maximally efficient requires that the barrier to electron transfer must be minimal; that is, the ligand-reorganizational barrier to electron transfer must be small conforming to Marcus theory for electron transfer [241, 242]. Thus, the precursor Mn(II) complex should adopt a six-coordinate pseudo-octahedral geometry similar to that required by the corresponding Mn(III) complex. Loss of an axial aquo ligand followed by folding of one of the secondary amine NH's of the macrocyclic ring so that it occupies an axial site would generate a pseudo-octahedral complex, $[Mn^{II}(L)X]^{n+}$, accommodating the requirement for an octahedral six-coordinate Mn(III) complex. If the macrocyclic [15]aneN$_5$ ligand possesses C-substituents which, due to intramolecular steric repulsions and angle strains, force the ligand to adopt this folded pseudo-octahedral geometry about the spherically symmetrical Mn(II) ion, then the Mn(II) complex would be poised to undergo facile electron-transfer as the ligand-reorganizational barrier would be minimized; that is, the shape of the reduced complex and its oxidized product will be similar. This view conforms to Pauling's hypothesis concerning the basis of enzyme catalysis, that is, stabilizaton of the transition state. It is in this manner that the stereochemistry of the substituents would be expected to exert a major effect on the rate of electron transfer via either pathway. This need to rearrange the ligand from a planar geometry into a folded conformation, stablizng a pseudo-octahedral geometry on Mn(II), may be the reason why some complexes of this family show no measurable rate. It was intriguing to speculate that the steric constraints imposed on the folding by the presence of a substituent positioned on a carbon of the ring may favor or inhibit a particular fold, and thereby promote or block the pathway for electron-transfer.

Clearly, there were some very intriguing results that any theory must be able to explan. For, example the two ligands 17 and 18 afford complexes that were either extremely actveor inactive. Thus, any theory that attempts to rationalize the observed reaction-rate constants for catalytic dismutation of superoxide must predict why the Mn(II) complex of ligand 18 has no catalytic rate.

4.5. Computer-Aided Design (CAD)

In an attempt to develop a theory of the details of this SOD catalysis that could corecly predict the effects that substitutents exert on the catalytic rate, molecular modeling was used. The premise for the modeling paradigm was that the macrocyclic pentaaza ligand possesses C-substituents, which due to intramolecular steric repulsions and angle strains, force the ligand to adopt various degrees of folded pseudo-octahedral geometry about the spherecaly symmetrical Mn(II) ion. If the Mn(II) complex is constrained in a geometry which resenbles a six-coordinate pseudo octahedral geometry, then the Mn(II) complex would be poised o undergo facile electron-transfer as the ligand-reorganizational barrier from electron transfer would be minimized. It was in this manner that the stereochemistry of the substituents oud be expected to exert a major effect on the rate of electron transfer via either pathway.

To better understand the effects that substituents exert, we used a combination of moecular mechanics calculations and synthesis with the goal of rational design of highly substituted, chemically stable, synthetic enzymes. It needs to be stressed that while the prediction of properties and reactivity of small carbocyclic ring systems based on conformational control is widely practiced and relatively straightforward [243], the conformational analyss of large, macrocyclic-ring, metal complexes undergoing redox chemistry offered a more dificut challenge.

To develop effectively a rational modeling paradigm not only required a detailed understanding of the mechanism of the rate-determining step in the catalytic process, but also a comprehensive database of chemical structures with their rate data so that the theoreical model could be subjected to stringent validation. Additional ligands (shown in Fig. 21) were also synthesized. The Mn(II) complexes of these ligands were then utilized as our source of rate and stability data for this CAD study. The Mn(II) complex of each ligand was completely characterized in terms of physical properties (e.g., the kinetic stability to dissociaion

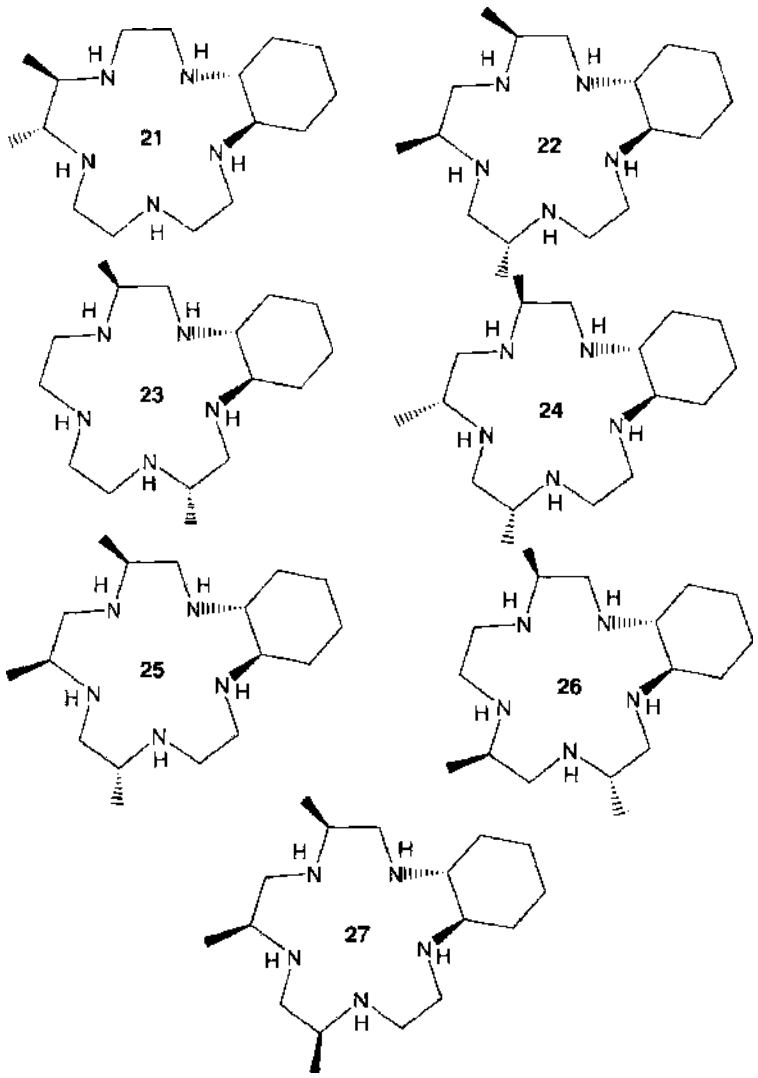


Figure 21. Ligands utilized to form the database for substituted $[Mn([15]aneN_5)X_2]$ complexes for probing the effects of substituents for molecular modeling (MM) studies.

in aqueous media) and the kinetics of their superoxide-dismutase catalytic activity [239]. In all cases, as we noted previously [235], increasing the number of the substituents on the macrocyclic ligand increases the kinetic stability of the complexes to dissociation in water.

The catalytic activities of these database complexes follow no apparent trend regarding the number of substituents or their stereochemisty, other than that they both clearly have effects on the overall rate and on the two competing pathways for oxidation of Mn(II) during the catalytic cycle. Most striking was that there was either no measurable k_{is} (inner-sphere rate) or no k_{os} (outer-sphere rate) for some of these complexes. Complexes with the ligands 10, 18, and 21 had no measurable k_{os}. While these complexes were utilized as the original database to construct the modeling paradigm, complexes from ligands 28–31 (Fig. 22) were synthesized based on predictions of the model and provided tests of the predictive power of the MM calculations. For the complexes derived from ligands 28–31, the kinetic stabilities and catalytic rate constants, k_{cat} pH=7.4, were also measured and are listed in Table 1.

While there are a number of computational methods available that could be used to predict structures and calculate the thermodynamic properties of a given structure for a coordination complex, the simplest level of calculation would be the use of molecular mechanics MM assuming appropriate parameters are available [244]. We have completed the X-ray structure determinations and refinements of many complexes in this class and reported five crystal structures of this class of Mn(II) penta-azacrown complexes [238]. This extensive X-ray structural database allowed us to refine the commercial modeling parameters and utilize an improved R_{avg} for the Mn-Nitrogen distance, 2.283 Å, and also an R_{avg} for the Mn-Cl

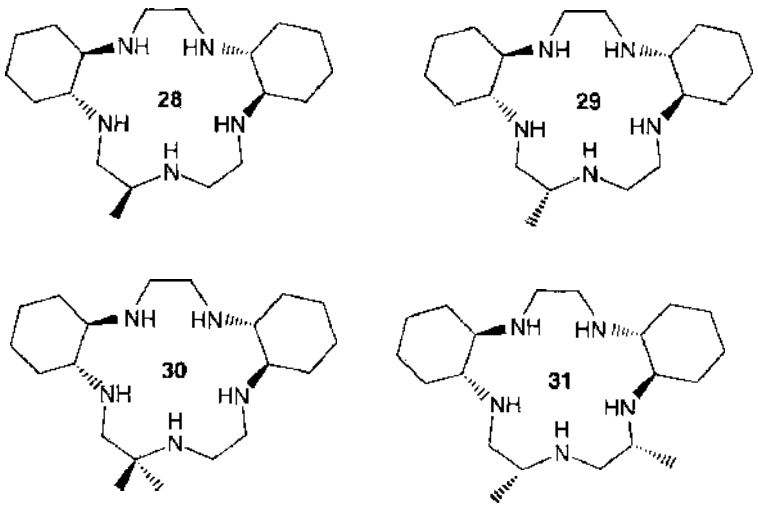


Figure 22. Ligands utilized for testing the predictive ability of the folding paradigm for MM calculations.

distance, 2.616 Å. In all cases where the NH stereochemical pattern is alternating, each side of the plane of the macrocyclic ring is chemically unique, since one side will have two nonadjacent NHs, while the other side possesses three (two in the case of pyridino ligands) NHs in a *cis* orientation. This motif provides a basis for probing the structural effects using MM calculations.

The design premise required that the ligand dictates catalytic activity by promoting or preventing a particular folded structure. A good catalyst will be one in which the Mn(II) center is constrained in a geometry which promotes rapid electron transfer; that is, pseudo-octahedral geometry adopted by Mn(III). Consequently, the folding of an NH out of the plane defined by the metal and the five nitrogens of the ligand into an axial site governs the ability of the corresponding complex to function as a catalyst. The MM calculations can be used to determine the relative energies of all the possible folds for each complex (five possible, since each nitrogen can fold into an axial site, but only on the side of the macrocycle which the NH is located) and their relative energies compared. Although it should be emphasized that with some ligands due to their symmetry, there may be fewer unique binding modes. For example, the pyridin[15]aneN5 ligand, 19, posses a higher symmetry than other ligands of this class displayed here. It possesses four NHs and the two axial *trans* coordination sites are equivalent. As a consequence, there exists only two unique folding motifs with this Mn(II) complex.

Our initial foray into modeling using MM calculations involved trying to probe the relative energetics of two types of six-coordinate structures generated via loss of a ligand from a *trans*-seven-coordinate Mn(II) as depicted in Fig. 23. In solution and solid state, both the six- and seven-coordinate geometry exist for these complexes; thus, in solution, a dynamic equilibrium of such structures is expected [222]. Two initial six-coordinate structures could be generated via the loss of an axial ligand from the seven-coordinate structure; that is, the complex labeled A in Fig. 23, which has the vacant site on the side in which three NHs are

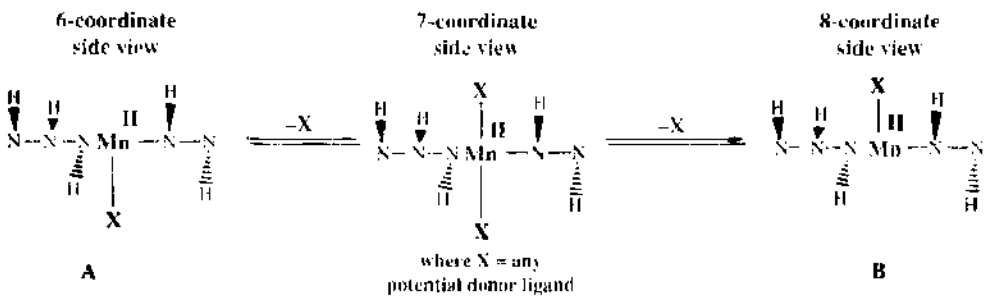


Figure 23. Depiction of the equilibrium between two possible six-coordinate intermediate structures generated via ligand dissociation from a seven-coordinate intermediate Mn(II) complex.

cis, and the structure labeled B, which has the vacant site on the side with two nonadjacent NHs *cis*. Clearly, the two intermediates are chemically quite distinct, and for each, an energy could be calculated using the MM approach. Invariably, structure A was found to be at a much lower energy than the structure B (in the range of 4–8 kcal/mol for chloro as axial ligand and 8–12 kcal/mol for the aquo complex). The magnitude of the difference suggests that structure A should be a logical starting point as a common intermediate leading to catalysis for all complexes.

Using this approach and assuming that the six-coordinate aquo complex is the Mn(II) complex in water which leads to productive chemistry, the simplest entry to the outersphere pathway presents itself as the aquo complex of structure A (where $X = H_2O$). This suggests that for this path to be the viable fast rate for a given complex, the substituents would force one of the three *cis* NHs on the side of the vacant coordination site out of the plane defined by the Mn(II) and the five nitrogen atoms in effect, folded into the axial pseudo O_h site. When this was investigated for these complexes generated from the ligands specified above, we find, in general, that there is a most preferred NH lowest-energy conformer folding for one of the three *cis* NH nitrogens into an axial pseudo O_h site. This was the modeling paradigm that evolved for the outer-sphere correlation.

For those ligands in which there are five secondary NH donors, the innersphere, superoxide-binding pathway, apparently uses a folding motif generating a six-coordinate intermediate in which one of the NHs on the side with two *cis* NHs folds into the pseudo-axial site. For all the complexes we observed, there exists a unique lower-energy folded conformer. This is consistent with the possibility that the inner-sphere pathway employs a folded structure that is unique and different from the outer-sphere pathway. It is also consistent with the inner-sphere path being accessed by an intermediate *common* to the outer-sphere path.

A scenario consistent with these hypotheses follows. The outersphere path correlates with folding one of the three *cis* NH's fold into an axial site, and the inner-sphere path could utilize that same vacant site to bind superoxide. This negatively charged ligand will labilize the aquo ligand *trans*- to it and generate a vacant site on the side of the plane of the MnL complex with the two non-adjacent *cis* NHs. Thus, loss of the bound *trans*-water and subsequent folding of either of the NHs on the side of the two *cis* NHs would generate a folded six-coordinate pseudo-octahedral complex. This correlation was validated by the modeling efforts, and it suggests that one common intermediate leads (geometry A in Fig. 23) to catalysis [245].

At the outset, we could make no claim as to the validity of this explanation, other than it is consistent with the general mechanistic understanding and characterization of the chemistry of these complexes. For all the complexes utilized in the database (as previously described), an evaluation of the energy difference (ΔE) of the folded six-coordinate structure for the Mn(II) and the corresponding Mn(III) complex were performed for each NH folded into an axial site. From this evaluation of different folded structures, each complex will possess a lowest-energy folded structure for the six-coordinate complex derived from an NH on the side of the three NHs occupying a pseudo-axial site, and also a lower-energy structure for the six-coordinate complex derived from an NH on the side of two *cis* nonadjacent NHs occupying a pseudo-axial site. Thus, a correlation of the energy difference, ΔE ($\Sigma_{Mn(III)} - \Sigma_{Mn(II)}$), for a series of complexes for each of the two types of folds potentially correlating with unique inner-sphere and outer-sphere folded structures can be performed, where the ground state energy of both Mn(III) and Mn(II) are approximated by the summation of the various contributing energy terms [246, 247]:

$$\Sigma_{Mn(III)} = U_{electrostatics} + U_{H\text{-}bonding} + U_{bond\ stretch} + U_{angle\ strain} + U_{torsional\ strain} + U_{dihedral\ strain} + U_{Van\ der\ Waals} + U_{angle\ deformation} \quad (6)$$

and

$$\Sigma_{Mn(II)} = U'_{electrostatics} + U'_{H\text{-}bonding} + U'_{bond\ stretch} + U'_{angle\ strain} + U'_{torsional\ strain} + U'_{dihedral\ strain} + U'_{Van\ der\ Waals} + U'_{angle\ deformation} \quad (7)$$

For a series of complexes, a correlation of ΔE for the folded geometries both the inner-sphere and the outer-sphere was simplified by comparing a modified energy $\Delta E^{\#}$ in which the contribution of $(U_{electrostatics} - U'_{electrostatics}) + (U_{H\text{-}bonding} - U'_{H\text{-}bonding})$ are removed since they are maintained invariant amongst the series of complexes. Plots of $\Delta E^{\#}$ $(\Sigma_{Mn(III)} - \Sigma_{Mn(II)})_{inner\text{-}sphere}$ versus k_{is} values (inner-sphere rate constant), and $\Delta E^{\#}$ $(\Sigma_{Mn(III)} - \Sigma_{Mn(II)})_{outer\text{-}sphere}$ versus k_{os} values (outer-sphere rate constant) were generated. Linear plots of $\Delta E^{\#}$ versus k_{is} or k_{os} would indicate that the correlation is valid. In the cases described here with the ligands used for the database, excellent correlations were obtained relating the folding of one of the three *cis*-NHs as the outer-sphere H-atom transfer path, and the folding of one of the two *cis* NHs as the inner-sphere SO binding pathway.

In general, we find that there is a clear energy preference for a single, unique NH to fold on each side of the macrocycle for each pathway; that is, the energy differences between possible NH folding modes were large (sometimes as large as 70–100 kcal), thus, ΔE $(E_{Mn(III)0} - E_{Mn(II)})$ for the three outer-sphere folds and for the two inner-sphere folds for each complex always affords a clear lowest-energy choice for each type of fold with each complex. The orientation of the NHs is an important aspect of the strain energies required to fold the ligands. Implicit in this exercise is the assumption that the orientation of the NHs relative to the substituents is known. For those complexes in which crystallographic data exists, the orientations are known. Using this data, certain structural relationships, for example, a *trans*-cyclohexano substituent dictates that the NHs α to the substituted carbons of the macrocyclic ring must also be *trans*, and MM calculations of the free ligands and for the Mn(II) planar seven-coordinate complexes, the NH pattern can be established with a high level of confidence for any complex of this family.

The goal of such correlations is to determine whether: (1) the folding paradigm is indeed consistent, (2) there are unique folding patterns for the inner-sphere and separate unique folding patterns for the outer-sphere pathways, and (3) if the MM calculation/modeling paradigm can be successfully used to predict structures to be synthesized so as to optimize the number of substituents for maximum stability while retaining high catalytic activity. This modeling paradigm based on MM calculations successfully addressed all three points. This meant that progress in developing highly substituted complexes with high catalytic activity could be made without the need to rely on trial-and-error synthesis, that can be exceedingly time consuming for the synthesis of such highly substituted and stereo-defined ligands. Indeed, the excellent correlations allowed us to test the model in a predictive manner utilizing the four ligands of Figure 22.

From the standpoint of ease of synthesis and the desire to maximize the stability with the minimum number of substituents, we chose to further elaborate the bis-cyclohexano structure of the highly activity catalyst derived from ligand 17. Note that this complex of this ligand with four of the macrocyclic ring carbons bearing a substituent achieves equivalent stability to those complexes bearing five methyl substituents. Thus, synthetically it was attractive to add an additional methyl to this ligand. Intuitively, to increase stability most efficiently, we chose to put the additional substituent on one of the chelate rings that did not possess a substituent. Thus, structures 28 and 29, each bearing an additional methyl substituent, were modeled. The two complexes only differ in the stereochemistry of the added methyl substituent: 28 (*S*-Me) and 29 [100]. The predictions based on the use of these correlations are that the all *R* complex 29 should be a very good catalyst, especially via the outer-sphere pH dependent path, while complex 28 with the *S*-Me would be much less active via this mode, and they should possess comparable activities via the inner-sphere path. As predicted, complex 29 is in fact very active with a large pH dependence; about three times that of complex 28. In fact, complex 29 possesses a catalytic rate at pH = 7.4 in excess of $1.5 \times 10^{+5}$ $M^{-1} s^{-1}$ and was the most active catalyst which we had synthesized up to that point.

On the basis of this level of success, it was intriguing to consider more highly substituted structures and whether they, in fact, would have catalytic activity. One type of substituent that we have observed to enhance stability is the *gem*-dimethyl; thus the *gem*-dimethyl complex 30 was subjected to the modeling paradigm. The results were quite clear; namely, there did not exist any folds that were of a low energy; that is, neither the inner- or outer-sphere

pathway would be predicted to operate with this structure. The ligand and complex were synthesized, and, indeed, only a trace of catalytic activity was observed.

The desire to increase stability by increasing the number of substituents on the macrocycle while retaining good catalytic activity led us to consider whether there were carbons of ligand 29 on which one could add a substituent and retain good activity. This exercise utilizing the modeling led to the prediction that, of the five carbon atoms of the macrocyclic ring of 29 which were devoid of substituents, the 14-position could bear a methyl (or other substituent) and retain reasonable activity, but only if the stereochemistry of that additional substituent is *R*, generating the all *R* hexasubstituted-ligand structure 31. The modeling, in fact, predicted that the catalytic activity for such a complex 31 should be mid-range within this family via both pathways. The actual activity was not quite as high as predicted, but was nevertheless very good considering that the complex bore six substituents. In fact, this complex had a better catalytic activity and possessed a kinetic stability of more than 2000 times that of the original unsubstituted parent complex, $[Mn([15]aneN_5)Cl_2]$, and a binding constant, $\log K \sim 15.5$. An important structural insight into the driving force for the folding can be made with this complex. The 14*R*-Me substituent of the ligand 31 is in an axial position in the planar seven-coordinate Mn(II) complex structure. This high-energy strained conformation is dramatically relieved when the outer-sphere fold occurs generating the six-coordinate structure. This folding places the 14*R*-Me into an equatorial orientation; thus, relieving strain and providing a driving force for the fold (Fig. 24).

Two complexes require comment; the Mn(II) complexes of ligand 10 and ligand 18, since they exhibited unique inactivity. For the complex with ligand 10, no outer-sphere H-atom pathway was observed in its SOD catalysis. From the modeling, it was quite clear that such a path could not occur because the fold corresponding to the outer-sphere fold with this ligand was extremely unfavorable; that is, the fold of one of the NHs on the side of the three *cis* NHs. The complex with the bis-cyclohexyl ligand 18 was also quite distinctive in that little or no catalytic activity was observed. Again the modeling was predictive of this; that is, all NH folds were very high energy indicating that this ligand preferred a planar geometry, and hence would be a poor catalyst (as observed).

In general, the linear correlations developed for this system of Mn(II) complexes lends support to the theory that folding is a critical aspect of catalytic activity for these synthetic enzymes, and that the substituents (relative position and stereochemistry) influence the favorability of the folding in a predictable manner. Thus, we are able to model any given complex for the inner-sphere or the outer-sphere folding energetics and calculate the expected rate constant for each path for any ligand pattern! This paradigm has worked extremely well at correlating and predicting activity, and has made it possible to design highly complicated molecules and test their potential utility as a catalyst without the extraordinary effort required to first synthesize such a structure. We have been able to literally test hundreds of combinations of structures in this fashion, and were able to gain a fair degree of certainty regarding a structure prior to embarking on the complicated synthesis. In no case examined

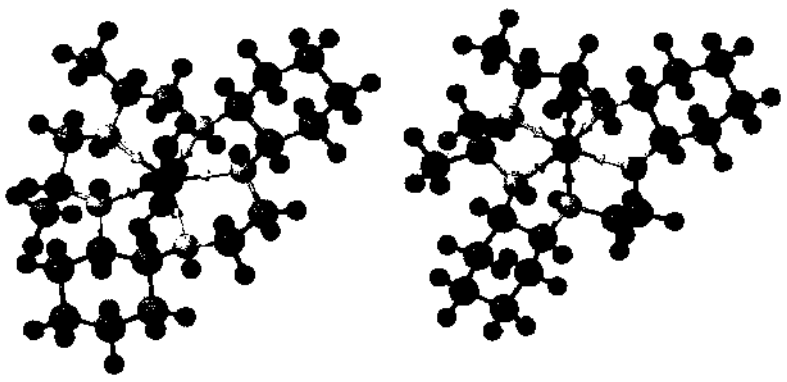
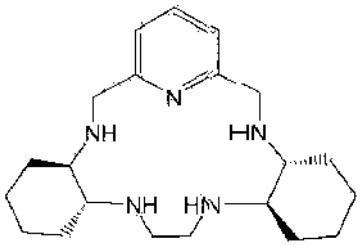


Figure 24. Molecular structure of the complex [Mn(Ligand **31**)] from MM calculations for the seven-coordinate bis(aquo) complex and the six-coordinate mono(aquo) complex showing the folded lowest-energy structure for this complex.

to date did this MM modeling paradigm fail to predict the approximate activity of a complex. Biological results generated with complexes that were designed using the MM technique and methods described here have provided additional confidence that these methods hold great promise for the design of improved human pharmaceutical agents. Continuing efforts aimed at developing catalysts with activities approaching that of the Cu/Zn SOD enzymes ($k_{cat} > 1 \times 10^{+9}$ M^{-1} s^{-1}) and with enhanced stabilities continue. By building on these modeling successes and our detailed knowledge of the factors affecting stability, we have designed ligands that not only possess dramatically enhanced stabilities and competent catalytic rates, but can be readily synthesized in few steps with commercially available building blocks. This last point is very important and it cannot be overemphasized that the ease/cost of synthesis and development are major elements in drug design. One notable success is M40403 (Mn(II) complex of ligand 32), which was successfully developed at MetaPhore Pharmaceuticals and introduced into human clinical testing in 2001 [248–250].



32

The synthesis of the Mn(II) complex of ligand 32 is achieved by a very atom efficient and chemically high-yield, template-ring closure in which the linear tetraamine, N,N′-bis-[(1*R*,2*R*)-2-aminocyclohexyl]-1,2-diaminoethane, is condensed in a mole ratio of 1 : 1 : 1 with $MnCl_2$ and 2,6-pyrdinediacarboxaldehyde affording in a quantitative chemical yield and with complete stereochemical retention of the all *R* stereochemistry [251]. The relative ease of synthesis of such a synthetic enzyme in high yield accomplishes/solves one of the major design goals of any drug substance; namely, the cost/difficulty of manufacture. In addition to solving this design goal, this ligand also addresses the desire to have a ligand that can bind the Mn(II) very tightly and still afford a reasonable catalytic rate for superoxide dismutation at physiological pH. The marriage of the very high stability imparting characteristics (lower basicity and rigidity) of the monopyridyl substituted ligand 19 combined with the propensity of the *R,R,R,R*-bis-cyclohexano ligand 17 to impart folding of one of the secondary NH-donors into an axial binding site into one ligand was intellectually quite appealing. Modeling suggested that such a complex should be a comparable catalyst as compared to that of the complex based on ligand 19, but not as good as that derived from ligand 17 since it is more constrained to a planar geometry. In fact, at pH=7.4, it was found that this was the case with most of the catalytic activity due to the inner-sphere superoxide binding pathway and only a small contribution to the overall rate being due to outer-sphere pH dependent (highly-folded six-coordinate intermediate) pathway [244]. Nevertheless, the incredibly high kinetic and thermodynamic stability of this complex, coupled with the ease of synthesis and that the molecule was very active in a variety of pharmacologically relevant models of human disease [249, 250, 252–254], all contributed to the decision to pursue this compound as a human drug candidate.

Assuming that increasing the catalytic activity of the SOD mimetic will translate into a decreased human dose, it would be desirable to further increase the catalytic activity (Fig. 25). Since the native SOD enzymes possess catalytic rate constants on the order of 10^{+9} $M^{-1}s^{-1}$, we attempted to design a low molecular-weight mimic utilizing the M40403 platform as a starting point for further structural modification/optimization studies using CAD, with the proviso/constraint that any structure that is selected for development following CAD must possess a reasonable synthesis route affording drug product with a pharmaceutically acceptable manufacturing cost.

One molecule that met the threefold design criteria of enhanced catalytic activity, ease of synthesis, and high stability is a dimethyl-substituted derivative of the ligand 32. This new

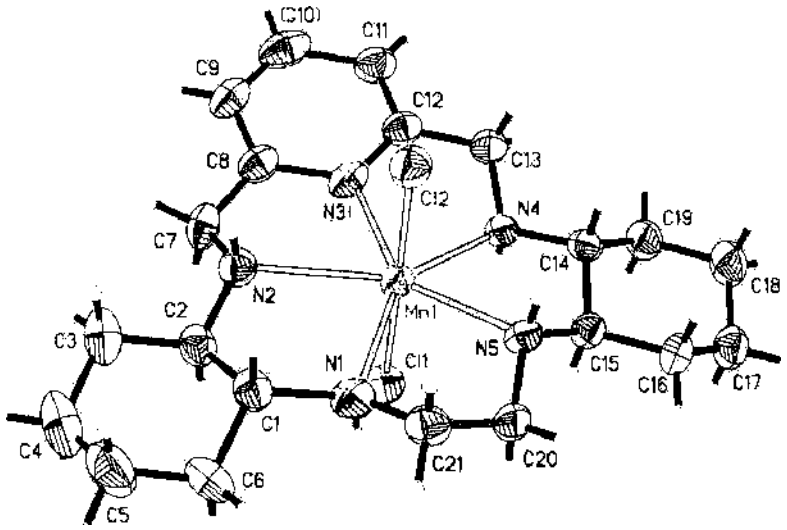
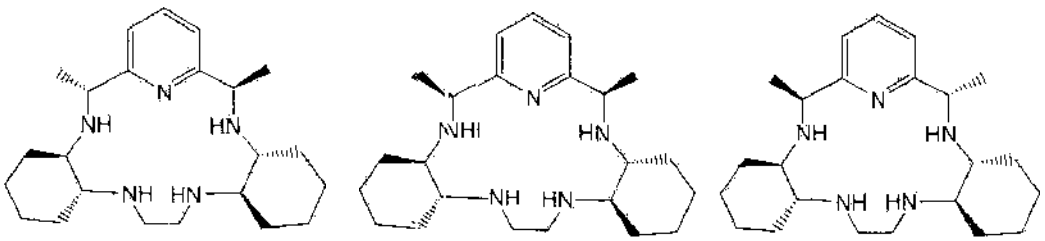


Figure 25. Ortep drawing for the M40403 complex [Mn(ligand **32**)Cl_2] showing the 50% probability ellipsoids for the non-hydrogen atoms and the hydrogen atoms off the secondary amines. Reprinted with permission from [244], K. Aston et al., *Inorg. Chem.* 40, 1779 (2001). © 2001, American Chemical Society.

ligand possesses the *S*, *S*-dimethyl stereochemistry and is readily synthesized in high yield via the template condensation utilizing 2,6-diacetypyridine in place of 2,6-dicarboxaldehyde [244]. Two other isomers isomers are possible when the *R*,*R*-diamminocyclohexane tetraamine is used and they have also been synthesized and characterized.



33 34 35

Modeling clearly predicts that the Mn(II) complex of the ligand 35 should be highly favored to fold into an octahedral geometry about the Mn(II) ion, while the complex with the ligand, 33, of all *R*-stereochemistry, is constrained to be planar and is unable to fold, even as the Mn(III) ion. The *R*,*S*-ligand, 34, when complexed to Mn(II) has an energetics for folding that is comparable to that with ligand 32, the unsubstituted analog. Measurement of the catalytic rates for these complexes validates these results and reveals that the complex with ligand 33 possesses no catalytic rate for the superoxide-dismutation reaction and that the complex with ligand 34 possesses a catalytic rate profile virtually identical to that of M40403 (complex with ligand 32). The most significant result is that the complex with ligand 35 possesses a catalytic rate constant for SOD activity greater than 10^{-9} M^{-1} s^{-1} at pH = 7.4—essentially as fast as the native Cu/Zn SOD enzyme!

The crystal structure of the dichloro complex with this ligand 35 was determined and found to possess a very unique solid-state structure (Fig. 26). This complex crystallizes in a monoclinic space group ($P2_1$) with two distinctly different molecules in the unit cell. One form is the trans-dichloro seven-coordinate geometry observed with M40403, but the second molecule possesses a six-coordinate geometry in which one of the chlorides has moved away from the metal by nearly half an angstrom becoming a non-coordinated (ionic) chloride and one of the nitrogens has folded toward the vacated axial site generating a six-coordinate complex with one of the secondary NHs (on the unsubstituted chelate ring) folded into the axial site. The observed <ClMnN angle for this folded nitrogen is 120.3 degrees while the calculated angle is 124.6 degrees [244].

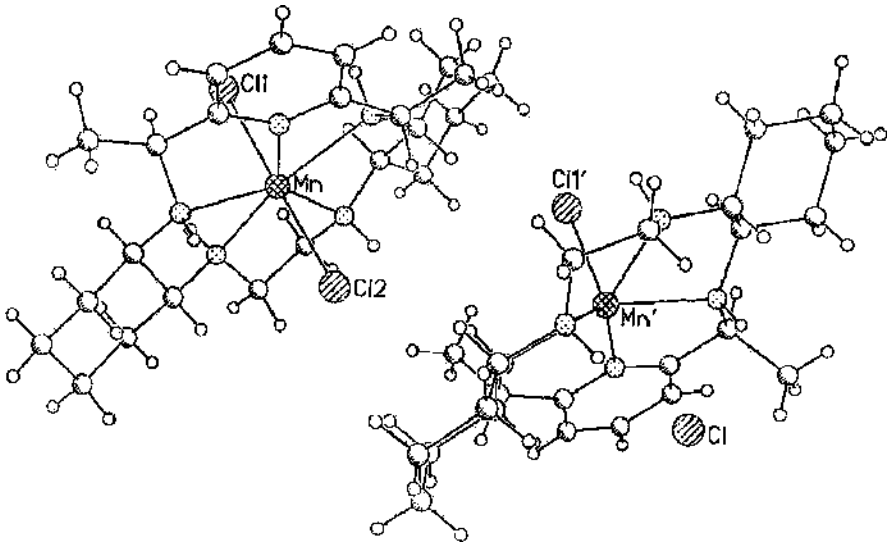


Figure 26. Ortep drawing for the complex [Mn(ligand 35)Cl_2] showing the 50% probability ellipsoids for the non-hydrogen atoms and the hydrogen atoms of the secondary amines. Reprinted with permission from [244]. K. Aston et al., *Inorg. Chem.* 40, 1779 (2001). © 2001, American Chemical Society.

An understanding of the details of the mechanism of action of these synthestic superoxide dismutase catalysts has made it possible to devise a computer modeling paradigm that allows one to design highly substituted (and hence highly stable) complexes that possess high catalytic activity. Molecular mechanics calculations have made it possible to correctly predict how substituents and their stereochemistry affect the energetics of ligand folding and thus affect catalytic activity.

It should be noted though that the complexes discussed and the methodologies described here are aimed solely at the design of a core structure for an optimized catalyst. The compounds presented here only present a hydrocarbon surface; where, in fact, the optimum drug for a synthetic enzyme for any particular disease may well require a different log P, or different type of functional substituent, to optimize the localization of the compound to optimize drug efficacy, minimize toxicity, and facilitate clearance of the drug. The chemistry presented here for designing and synthesizing such core catalyst structures, optimized for stability and catalytic activity, are amenable to the construction of complexes that have pendant functionality, for example, pendant alchols, amines, amides, esters, and acids. Such chemistry is ready available from standard peptide methods using the available chiral pool and unnatural chiral synthetic amino acids. Thus, by adapting a promising core catalyst structure, such as that achieved with the Mn(II) complex with ligand 35, it is possible to develop optimized structures with varying functional substituents tailored to meet the need of the disease state. It is in such a manner that human pharmaceutical agents with minimal toxicity and maximal efficacy could be realized.

4.6. Targeting Receptors with Azacrown-Metal Complexes

A remarkable geometrical resemblance between the orientation of hydrogens on the metal complexes of penta-azacrowns and side-chain orientations in reverse-turn conformations of peptides has been described by Reaka et al. [153]. As cyclic pentapeptides can serve as precursors to pentaazacrowns, the use of amino acids in the preparation of chiral azacrowns is obvious. Marshall and colleagues [255] have designed metal-complexes of pentaazacrowns to serve as RGD and α-amylase inhibitors.

5. CONCLUSIONS

Remarkable progress has been made in the ability to design and stabilize protein folds and, in some cases, to attribute useful designed properties to the designed proteins. In most cases, however, success has depended on a combination of design, genetation of chimeras,

and some form of selection/optimization. In most cases, this is due to current inability to search fold space and the absence of scoring functions that can unambiguously predict the experimentally observed structure of a given amino acid sequence.

In this review, we advocate the use of synthetic organic chemistry combined with protein expression to generate "protein mimetics" that are limited to the protein folds that can produce the desired functionality. Examples of synthetic mimetics of protein secondary-structure elements are given that can readily be adapted as constraints to drive a chimeric protein to assume the designed fold. In addition, a detailed discussion of a synthetic mimetic of a catalytic active site (superoxide dismutase) provides a guide to development of modular catalytic sites that can be engineered into designed protein scaffolds to provide the basis for substrate specificity.

ACKNOWLEDGMENTS

As are readily apparent in the references, the authors are indebted to many talented colleagues for their contributions to their research described as well as to the National Institutes of Health for support. As should also be obviously, the topic chosen for review is vast, the literature overwhelming and growing exponentially, and the time available limited. References were chosen for illustration, and no attempt to be comprehensive is claimed.

REFERENCES

1. C. Garcia-Echeverria et al., *J. Med. Chem.* 43, 3205 (2000).
2. O. Kutzki et al., *J. Am. Chem. Soc.* 124, 11838 (2002).
3. J. T. Ernst et al., *Angew. Chem. Int. Ed. Engl.* 42, 535 (2003).
4. L. Giot et al., *Science* 302, 1727 (2003).
5. A. H. Y. Tong et al., *Science* 303, 808 (2004).
6. A. H. Tong et al., *Science* 295, 321 (2002).
7. U. de Lichtenberg et al., *Science* 307, 724 (2005).
8. A. Berglund et al., *Proteins: Struct. Funct. Bioinform.* 54, 289 (2004).
9. C. Zhang et al., *Protein Sci.* 13, 400 (2004).
10. K. M. Misura and D. Baker, *Proteins* 59, 15 (2005).
11. R. Zwanzig, A. Szabo, and B. Bagchi, *Proc. Natl. Acad. Sci. USA* 89, 20 (1992).
12. L. L. Looger and H. W. Hellinga, *J. Mol. Biol.* 307, 429 (2001).
13. F. Offredi et al., *J. Mol. Biol.* 325, 163 (2003).
14. K. A. Dill, *Protein Sci.* 8, 1166 (1999).
15. J. D. Bryngelson et al., *Proteins* 21, 167 (1995).
16. M. S. Evans, T. F. T. Clarke, and P. L. Clark, *Protein Pept. Lett.* 12, 189 (2005).
17. C. Vogel et al., *Curr. Opin. Struct. Biol.* 14, 208 (2004).
18. L. Holm and C. Sanders, *Proteins* 33, 88 (1998).
19. M. A. Dwyer and H. W. Hellinga, *Curr. Opin. Struct. Biol.* 14, 495 (2004).
20. D. P. Barondeau and E. D. Getzoff, *Curr. Opin. Struct. Biol.* 14, 765 (2004).
21. I. L. Alberts, K. Nadassy, and S. J. Wodak, *Protein Sci.* 7, 1700 (1998).
22. Y. Lu, S. M. Berry, and T. D. Pfister, *Chem. Rev.* 101, 3047 (2001).
23. O. Maglio et al., *Proc. Natl. Acad. Sci. USA* 100, 3772 (2003).
24. M. D. Struthers, R. P. Cheng, and B. Imperiali, *Science* 271, 342 (1996).
25. C. M. Tann, D. Qi, and M. D. Distefano, *Curr. Opin. Chem. Biol.* 5, 696 (2001).
26. D. Qi et al., *Chem. Rev.* 101, 3081 (2001).
27. T. M. Penning and J. M. Jez, *Chem. Rev.* 101, 3027 (2001).
28. L. Baltzer and J. Nilsson, *Curr. Opin. Biotechnol.* 12, 355 (2001).
29. M. A. Dwyer, L. L. Looger, and H. W. Hellinga, *Science* 304, 1967 (2004).
30. J. Kaplan and W. F. DeGrado, *Proc. Natl. Acad. Sci. USA* 101, 11566 (2004).
31. L. L. Looger et al., *Nature (London)* 423, 185 (2003).
32. M. Allert et al., *Proc. Natl. Acad. Sci. USA* 101, 7907 (2004).
33. H. Paulus, *Annu. Rev. Biochem.* 69, 447 (2000).
34. G. Skretas and D. W. Wood, *Protein Sci.* 14, 523 (2005).
35. L. Baltzer, H. Nilsson, and J. Nilsson, *Chem. Rev.* 101, 3153 (2001).
36. D. A. Moffet and M. H. Hecht, *Chem. Rev.* 101, 3191 (2001).
37. J. S. Richardson and D. C. Richardson, *Trends Biochem. Sci.* 14, 304 (1989).
38. A. Lim et al., *J. Struct. Biol.* 130, 363 (2000).
39. J. S. Richardson and D. C. Richardson, *Proc. Natl. Acad. Sci. USA* 99, 2754 (2002).
40. Y. Wei et al., *Protein Sci.* 12, 92 (2003).
41. M. W. West and M. H. Hecht, *Protein Sci.* 4, 2032 (1995).

42. B. I. Dahiyat and S. L. Mayo, *Science* 278, 82 (1997).
43. M. A. Starovasnik, A. C. Braisted, and J. A. Wells, *Proc. Natl. Acad. Sci.* 94, 10080 (1997).
44. J. W. Neidigh, R. M. Fesinmeyer, and N. H. Andersen, *Nat. Struct. Biol.* 9, 425 (2002).
45. J. J. Ottesen and B. Imperiali, *Nat. Struct. Biol.* 8, 535 (2001).
46. M. T. Pastor et al., *Proc. Natl. Acad. Sci.* 99, 614 (2002).
47. M. Lopez de la Paz et al., *Proc. Natl. Acad. Sci.* 99, 16052 (2002).
48. M. S. Chana et al., *J. Pept. Res.* 65, 209 (2005).
49. J. R. Litowski and R. S. Hodges, *J. Biol. Chem.* 277, 37272 (2002).
50. J. G. Adamson, N. E. Zhou, and R. S. Hodges, *Curr. Opin. Biotechnol.* 4, 428 (1993).
51. O. D. Monera et al., *J. Biol. Chem.* 268, 19218 (1993).
52. N. L. Ogihara et al., *Protein Sci.* 6, 80 (1997).
53. A. Lombardi, J. W. Bryson, and W. F. DeGrado, *Biopolymers* 40, 495 (1996).
54. W. F. DeGrado, L. Regan, and S. P. Ho, *Cold Spring Harbor Symp. Quant. Biol.* 52, 521 (1987).
55. B. Lovejoy et al., *Science* 259, 1288 (1993).
56. S. T. Walsh et al., *Biochemistry* 40, 9560 (2001).
57. J. W. Bryson et al., *Protein Sci.* 7, 1404 (1998).
58. G. Ghirlanda et al., *J. Mol. Biol.* 319, 243 (2002).
59. Y. Zhu et al., *Proc. Natl. Acad. Sci. USA* 100, 15486 (2003).
60. S. J. Lahr et al., *J. Mol. Biol.* 346, 1441 (2005).
61. M. H. Ali et al., *Proc. Natl. Acad. Sci. USA* 101, 12183 (2004).
62. M. H. Ali et al., *Structure (Camb)* 13, 225 (2005).
63. S. T. Walsh et al., *Protein Sci.* 12, 520 (2003).
64. G. R. Marshall, *Curr. Opin. Struct. Biol.* 2, 904 (1992).
65. M. Miller et al., *Science* 246, 1149 (1989).
66. M. Baca, P. F. Alewood, and S. B. Kent, *Protein Sci.* 2, 1085 (1993).
67. B. Ayers et al., *Biopolymers* 51, 343 (1999).
68. T. W. Muir, D. Sondhi, and P. A. Cole, *Proc. Natl. Acad. Sci. USA* 95, 6705 (1998).
69. J. W. Ponder and D. A. Case, *Adv. Protein Chem.* 66, 27 (2003).
70. P. Ren and J. W. Ponder, *J. Phys. Chem. B* 107, 5933 (2002).
71. A. Grossfield, P. Ren, and J. W. Ponder, *J. Am. Chem. Soc.* 125, 15671 (2003).
72. V. Guallar and R. A. Friesner, *J. Am. Chem. Soc.* 126, 8501 (2004).
73. I. A. Vakser and A. Sali, *COMBASE*, http://guitar.rockefeller.edu.
74. J. Janin, *Biochimie* 77, 497 (1995).
75. S. J. Hubbard and P. Argos, *Protein Sci.* 3, 2194 (1994).
76. S. Jones and J. M. Thornton, *Proc. Natl. Acad. Sci. USA* 93, 13 (1996).
77. P. Lijnzaad, and P. Argos, *Proteins* 28, 333 (1997).
78. L. Lo Conte, C. Chothia, and J. Janin, *J. Mol. Biol.* 285, 2177 (1999).
79. Z. Hu et al., *Proteins* 39, 331 (2000).
80. S. Miyazawa and R. L. Jernigan, *Proteins* 34, 49 (1999).
81. A. V. Veselovsky et al., *J. Mol. Recognit.* 15, 405 (2002).
82. B. Ma et al., *Proc. Natl. Acad. Sci. USA* 100, 5772 (2003).
83. C.-J. Tsai et al., *J. Mol. Biol.* 260, 604 (1996).
84. K. N. Houk et al., *Angew. Chem. Int. Ed. Engl.* 42, 4872 (2003).
85. A. A. Bogan and K. S. Thorn, *J. Mol. Biol.* 280, 1 (1998).
86. B. C. Cunningham and J. A. Wells, *J. Mol. Biol.* 234, 554 (1993).
87. T. Clackson and J. A. Wells, *Science* 267, 383 (1995).
88. Y. Sasaki et al., *J. Med. Chem.* 30, 1162 (1987).
89. J. Boer et al., *J. Med. Chem.* 44, 2586 (2001).
90. R. Hirschmann et al., *J. Med. Chem.* 41, 1382 (1998).
91. G. L. Olson et al., *J. Med. Chem.* 38, 2866 (1995).
92. J. C. Simpson et al., *Bioorg. Med. Chem.* 10, 291 (2002).
93. L. D. Rutledge et al., *J. Med. Chem.* 39, 1571 (1996).
94. G. R. Marshall et al., in "Computer-Assisted Drug Design" (E. C. Olson and R. E. Christoffersen, Eds.), p. 205, American Chemical Society, Washington, DC, 1979.
95. G. R. Marshall, *Biopolymers* 60, 246 (2001).
96. L. N. Drozdov-Tikhomirov et al., *J. Biomol. Struct. Dyn.* 19, 279 (2001).
97. D. Xu, C. J. Tsai, and R. Nussinov, *Protein Eng.* 10, 999 (1997).
98. B. A. Chrunyk et al., *Biochemistry* 39, 7092 (2000).
99. T. Berg, *Angew. Chem. Int. Ed. Engl.* 42, 2462 (2003).
100. U. Reineke et al., *Nat. Biotechnol.* 17, 271 (1999).
101. D. P. Fairlie, M. L. West, and A. K. Wong, *Curr. Med. Chem.* 5, 29 (1998).
102. M. J. Todd, N. Semo, and E. Freire, *J. Mol. Biol.* 283, 475 (1998).
103. M. D. Sholtz and J. Chmielewski, *Bioorg. Med. Chem. Lett.* 9, 2431 (1999).
104. R. Zutshi, M. Brickner, and J. Chmielewski, *Curr. Opin. Chem. Biol.* 2, 62 (1998).
105. M. J. Bowman and J. Chmielewski, *Biopolymers* 66, 126 (2002).
106. F. M. Finn and K. Hofmann, *Acc. Chem. Res.* 6, 169 (1973).
107. R. Varadarajan et al., *Biochemistry* 31, 1421 (1992).

108. L. T. Vassilev et al., *Science* 303, 844 (2004).
109. P. H. Kussie et al., *Science* 274, 948 (1996).
110. P. Chene et al., *J. Mol. Biol.* 299, 245 (2000).
111. E. E. Hodgkin et al., *Biopolymers* 30, 533 (1990).
112. Y. Sasaki et al., *J. Med. Chem.* 30, 1162 (1987).
113. R. Hirschmann, *Angew. Chem. Int. Ed. Engl.* 30, 1278 (1991).
114. J. Boer et al., *Angew. Chem. Int. Ed. Engl.* 40, 3870 (2001).
115. L. D. Rutledge et al., *J. Med. Chem.* 39, 1571 (1996).
116. Y. S. Tong et al., *Tetrahedron* 56, 9791 (2000).
117. J. C. Simpson et al., *Bioorg. Med. Chem.* 10, 291 (2002).
118. A. W. Czarnik and S. H. DeWitt, (Eds.), "A Practical Guide To Combinatorial Chemistry," 1st Edn., American Chemical Society, Washington, DC, 1997, p. 1.
119. R. A. Wiley and D. H. Rich, *Med. Res. Rev.* 13, 327 (1993).
120. B. E. Evans et al., *J. Med. Chem.* 31, 2235 (1988).
121. A. A. Patchett and R. P. Nargund, *Annu. Rep. Med. Chem.* 35, 289 (2000).
122. D. A. Horton, G. T. Bourne, and M. L. Smythe, *Chem. Rev.* 103, 893 (2003).
123. J. P. Schneider and J. W. Kelly, *Chem. Rev.* 95, 2169 (1995).
124. W. C. Ripka et al., *Tetrahedron* 49, 3593 (1993).
125. B. K. Blackburn et al., *J. Med. Chem.* 40, 717 (1997).
126. Y. Shigeri et al., *Life Sci.* 63, L. 151 (1998).
127. E. K. Dziadulewicz et al., *Bioorg. Med. Chem. Lett.* 9, 463 (1999).
128. W. H. Miller et al., *J. Med. Chem.* 43, 22 (2000).
129. C. Haskell-Luevano et al., *J. Med. Chem.* 42, 4380 (1999).
130. S. P. Rohrer et al., *Science* 282, 737 (1998).
131. S. Hanessian et al., *Tetrahedron* 53, 12789 (1997).
132. F. Cornille et al., *J. Am. Chem. Soc.* 117, 909 (1995).
133. U. Slomczynska et al., *J. Org. Chem.* 61, 1198 (1996).
134. L. Halab, F. Gosselin, and W. D. Lubell, *Biopolymers* 55, 101 (2000).
135. L. H. Yang et al., *Proc. Natl. Acad. Sci. USA* 95, 10836 (1998).
136. P. W. Schiller et al., *J. Med. Chem.* 43, 551 (2000).
137. F. A. Etzkorn et al., *J. Am. Chem. Soc.* 116, 10412 (1994).
138. H. Kessler and B. Kutscher, *Tetrahedron* 26, 177 (1985).
139. M. Anmailley et al., *FEBS Letters* 291, 50 (1991).
140. M. Gurrath et al., *Eur. J. Biochem.* 210, 911 (1992).
141. G. Muller, M. Gurrath, and H. Kessler, *J. Comp.-Aided Mol. Design* 8, 709 (1994).
142. M. Pfaff et al., *J. Biol. Chem.* 269, 20233 (1994).
143. R. Haubner et al., *J. Am. Chem. Soc.* 118, 7461 (1996).
144. R. Haubner et al., *J. Am. Chem. Soc.* 118, 7881 (1996).
145. R. Haubner, D. Finsinger, and H. Kessler, *Angew. Chem. Int. Ed. Engl.* 36, 1374 (1997).
146. J. Wermuth et al., *J. Am. Chem. Soc.* 119, 1328 (1997).
147. M. A. Dechantsreiter et al., *J. Med. Chem.* 42, 3033 (1999).
148. J. Boer et al., *J. Med. Chem.* 44, 2586 (2001).
149. S. L. Goodman et al., *J. Med. Chem.* 45, 1045 (2002).
150. G. A. Sulyok et al., *J. Med. Chem.* 44, 1938 (2001).
151. M. Porcelli et al., *Biopolymers* 50, 211 (1999).
152. G. V. Nikiforovich et al., *J. Am. Chem. Soc.* 122, 3262 (2000).
153. A. J. Reaka, C. M. Ho, and G. R. Marshall, *J. Comput. Aided. Mol. Des.* 16, 585 (2002).
154. H. B. Lee et al., *J. Org. Chem.* 69, 701 (2004).
155. D. K. Chalmers and G. R. Marshall, *J. Amer. Chem. Soc.* 117, 5927 (1995).
156. Y. Takeuchi and G. R. Marshall, *J. Amer. Chem. Soc.* 120, 5363 (1998).
157. A. B. Smith III et al., *Bioorg. Med. Chem.* 7, 9 (1999).
158. E. Jacoby, *Bioorg. Med. Chem. Lett.* 12, 891 (2002).
159. B. P. Orner, J. T. Ernst, and A. D. Hamilton, *J. Am. Chem. Soc.* 123, 5382 (2001).
160. M. L. Smythe, S. Huston, and G. R. Marshall, *J. Am. Chem. Soc.* 115, 11594 (1993).
161. C. O. Pabo and R. T. Sauer, *Annu. Rev. Biochem.* 53, 293 (1984).
162. C. W. Muller, *Curr. Opin. Struct. Biol.* 11, 26 (2001).
163. D. E. Draper, *J. Mol. Biol.* 293, 255 (1999).
164. G. R. Marshall, R. H. Head, and R. Ragno, in "Thermodynamics in Biology" (E. Di Cera, Ed.), p. 87, Oxford University Press, Oxford, 2000.
165. J. A. Kritzer et al., *J. Am. Chem. Soc.* 126, 9468 (2004).
166. M. J. Kelso et al., *J. Am. Chem. Soc.* 126, 4828 (2004).
167. J. W. Taylor, *Biopolymers* 66, 49 (2002).
168. G. R. Marshall and H. E. Bosshard, *Circ. Res.* 31, Suppl. 2,143 (1972).
169. P. Van Roey et al., *Int. J. Pept. Protein. Res.* 22, 404 (1983).
170. P. Van Roey et al., *Int. J. Pept. Protein. Res.* 19, 499 (1982).
171. G. R. Marshall et al., *Proc. Natl. Acad. Sci. USA* 87, 487 (1990).
172. B. V. Prasad and P. Balaram, *CRC Crit. Rev. Biochem.* 16, 307 (1984).

173. C. Toniolo et al., *Biopolymers* 33, 1061 (1993).
174. N. H. Andersen, Z. Liu, and K. S. Prickett, *FEBS Lett.* 399, 47 (1996).
175. Y. Higashimoto et al., *J. Biochem. (Tokyo)* 125, 705 (1999).
176. Y. Inai, N. Ousaka, and T. Okabe, *J. Am. Chem. Soc.* 125, 8151 (2003).
177. M. Sattler et al., *Science* 275, 983 (1997).
178. J. W. Chin and A. Schepartz, *Angew. Chem. Int. Ed. Engl.* 40, 3806, 3922 (2001).
179. J. T. Ernst et al., *Angew. Chem. Int. Ed. Engl.* 41, 278 (2002).
180. S. Hanessian et al., *Tetrahedron* 53, 12789 (1997).
181. R. Breslow, *Chem. Rec.* 1, 3 (2001).
182. Y. K. Wu and P. Ahlberg, *Acta. Chem. Scand.* 46, 60 (1992).
183. M. Z. Atassi and T. Manshouri, *Proc. Natl. Acad. Sci. USA* 90, 8282 (1993).
184. M. J. Corey and E. Corey, *Proc. Natl. Acad. Sci. USA* 93, 11428 (1996).
185. E. Perez-Paya, R. A. Houghten, and S. E. Blondelle, *Pept. Res.* 7, 286 (1994).
186. J. Svenson, N. Zheng, and I. A. Nicholls, *J. Am. Chem. Soc.* 126, 8554 (2004).
187. A. Fernandez-Gacio, M. Uguen, and J. Fastrez, *Trends Biotechnol.* 21, 408 (2003).
188. D. P. Barondeau et al., *Biochemistry* 43, 8038 (2004).
189. J. Wuerges et al., *Proc. Natl. Acad. Sci. USA* 101, 8569 (2004).
190. J. M. McCord and I. Fridovich, *J. Biol. Chem.* 244, 6049 (1969).
191. I. Fridovich, *J. Biol. Chem.* 264, 7761 (1989).
192. S. S. Ali et al., *Free Radic. Biol. Med.* 37, 1191 (2004).
193. D. Decraene et al., *J. Invest. Dermatol.* 122, 484 (2004).
194. A. Nishiyama et al., *J. Am. Soc. Nephrol.* 15, 306 (2004).
195. Y. G. Abashkin and S. K. Burt, *Inorg. Chem.* 44, 1425 (2005).
196. I. Batinic-Haberle et al., *Arch. Biochem. Biophys.* 343, 225 (1997).
197. S. W. Werns and B. R. Lucchesi, *Free Radic. Biol. Med.* 4, 31 (1988).
198. J. M. McCord, *J. Free Radic. Biol. Med.* 2, 307 (1986).
199. B. A. Omar et al., *Free Radic. Biol. Med.* 11, 517 (1991).
200. Y. Oyanagui, *Biochem. Pharmacol.* 25, 1465 (1976).
201. M. T. Droy-Lefaix et al., *Free Radic. Res. Commun.* 12–13, Pt 2, 725 (1991).
202. M. Shingu et al., *Rheumatol. Int.* 14, 77 (1994).
203. P. H. Chan et al., *Ann. Neurol.* 29, 482 (1991).
204. Y. Ando et al., *Brain Res.* 477, 286 (1989).
205. G. Yang et al., *Stroke* 25, 165 (1994).
206. H. X. Deng et al., *Science* 261, 1047 (1993).
207. D. R. Rosen et al., *Nature* 362, 59 (1993).
208. R. H. Brown, Jr., *Cell* 80, 687 (1995).
209. C. M. Troy and M. L. Shelanski, *Proc. Natl. Acad. Sci. USA* 91, 6384 (1994).
210. L. J. Greenlund, T. L. Deckwerth, and E. M. Johnson, Jr., *Neuron* 14, 303 (1995).
211. A. Bravard et al., *Int. J. Cancer* 51, 476 (1992).
212. S. L. Church et al., *Proc. Natl. Acad. Sci. USA* 90, 3113 (1993).
213. D. K. St. Clair et al., *Free Radic. Biol. Med.* 16, 275 (1994).
214. S. E. Safford et al., *Cancer Res.* 54, 4261 (1994).
215. N. Yoshizaki et al., *Int. J. Cancer* 57, 287 (1994).
216. S. C. Flores et al., *Proc. Natl. Acad. Sci. USA* 90, 7632 (1993).
217. D. P. Riley and R. H. Weiss, *J. Am. Chem. Soc.* 116, 387 (1994).
218. M. M. Hardy et al., *J. Biol. Chem.* 269, 18535 (1994).
219. R. H. Weiss et al., *J. Biol. Chem.* 271, 26149 (1996).
220. D. P. Riley, W. J. Rivers, and R. H. Weiss, *Anal. Biochem.* 196, 344 (1991).
221. D. P. Riley and R. H. Weiss, *J. Am. Chem. Soc.* 116, 387 (1994).
222. D. P. Riley et al., *J. Am. Chem. Soc.* 119, 6522 (1997).
223. J. L. Zweier, *J. Biol. Chem.* 263, 1353 (1988).
224. K. S. Kilgore et al., *J. Mol. Cell. Cardiol.* 26, 995 (1994).
225. S. C. Black et al., *J. Pharmacol. Exp. Ther.* 270, 1208 (1994).
226. C. M. Venturini et al., in "The Biology of Nitric Oxide 3: Physiological and Clinical Aspects" (S. Moncada et al., Eds.), p. 65, Portland Press, London, 1994.
227. T. P. Kasten et al., *Proc. Soc. Exp. Biol. Med.* 208, 170 (1995).
228. Y. Y. Meng et al., *J. Am. Coll. Cardiol.* 25, 269 (1995).
229. R. H. Weiss and D. P. Riley, in "Inorganic Chemistry in Medicine" (N. Farrell, Ed.), Royal Society of Chemistry, London, in press.
230. R. H. Weiss et al., in "The Oxygen Paradox" (K. J. A. Davies and F. Ursini, Eds.), p. 641, CLEUP University Press, Padova, Italy, 1995.
231. J. S. Beckman et al., *Nature (London)* 364, 584 (1993).
232. L. F. Lindoy, "The Chemistry of Macrocyclic Ligand Complexes," Cambridge University Press, Cambridge, 1989.
233. P. J. Lennon et al., *Tetrahedron Lett.* 35, 853 (1994).
234. K. W. Aston et al., *Tetrahedron Lett.* 35, 3687 (1994).
235. D. P. Riley et al., *Inorg. Chem.* 35, 5213 (1996).

236. R. Weiss et al., "Proceedings of the Society of Magnetic Resonance and the European Society for Magnetic Resonance in Medicine and Biology," Abstract 43.
237. G. W. Franklin, D. P. Riley, and W. L. Newmann, *Coord. Chem. Rev.* 174, 133 (1998).
238. D. P. Riley, in "Advances in Supramolecular Chemistry" (G. W. Gokel, Ed.), p. 217, Cerberus Press, Miami, 2000.
239. D. P. Riley, *Chem. Rev.* 99, 2573 (1999).
240. M. Eigen, *Pure. Appl. Chem.* 6, 105 (1963).
241. R. A. Marcus, *Annu. Rev. Phys. Chem.* 15, 155 (1964).
242. D. H. Macartney and D. W. Thompson, *Inorg. Chem.* 26, 2195 (1989).
243. U. Burkert and N. L. Allinger, "Monograph" (M. C. Caserio, Ed.), Vol. 177, p. 339, American Chemical Society, Washington, DC, 1982.
244. K. Aston et al., *Inorg. Chem.* 40, 1779 (2001).
245. D. P. Riley, *Adv. Supramolecular Chem.* 6, 217 (2000).
246. R. D. Hancock, *Acc. Chem. Res.* 23, 253 (1990).
247. R. D. Hancock and A. E. Martell, *Chem. Rev.* 89, 1875 (1989).
248. D. Salvemini et al., *Science* 286, 304 (1999).
249. H. Macarthur et al., *Proc. Natl. Acad. Sci. USA* 97, 9753 (2000).
250. D. Salvemini, D. P. Riley, and S. Cuzzocrea, *Nat. Rev. Drug. Discov.* 1, 367 (2002).
251. D. Salvemini and D. P. Riley, *Drugs Future* 25, 1027 (2000).
252. D. Salvemini et al., *Science* 286, 304 (1999).
253. S. Cuzzocrea et al., *Br. J. Pharmacol.* 132, 19 (2001).
254. S. Cuzzocrea, C. Thiemermann, and D. Salvemini, *Curr. Med. Chem.* 11, 1147 (2004).
255. W.-J. Zhang, M. E. Weber, and G. R. Marshall, in "Peptide Revolution: Genomics, Proteomics & Therapeutics," American Peptide Society, Boston, MA, 2004.
256. G. Ashkenasy et al., *Chemistry* 8, 4017 (2002).
257. R. Boulatov et al., *J. Am. Chem. Soc.* 124, 11923 (2002).
258. D. Lee et al., *J. Am. Chem. Soc.* 124, 3993 (2002).
259. S. J. Chiou et al., *Inorg. Chem.* 39, 4347 (2000).
260. P. G. Schultz and R. A. Lerner, *Science* 269, 1835 (1995).
261. D. Hilvert, *Annu. Rev. Biochem.* 69, 751 (2000).
262. R. Ricoux et al., *J. Immunol. Methods* 269, 39 (2002).
263. F. Tanaka, *Chem. Rev.* 102, 4885 (2002).
264. M. Q. Xu and T. C. Evans, Jr., *Methods* 24, 257 (2001).

CHAPTER 16

Quantum Information Processing in Nanostructures

Alexandra Olaya-Castro, Neil F. Johnson

Centre for Quantum Computation, Clarendon Laboratory, Department of Physics, University of Oxford, Oxford, United Kingdom

CONTENTS

ISBN: 1-58883-048-9

Handbook of Theoretical and Computational Nanotechnology
Edited by Michael Rieth and Wolfram Schommers
Volume 6: Pages (757–777)

1. INTRODUCTION

What can be accomplished in a world governed by quantum correlations? This question has brought together researchers from a wide range of backgrounds—from fundamental physics through to information technology—in the common pursuit of understanding, designing, and eventually building quantum information processing (QIP) systems.

The fundamental unit of quantum information is the quantum bit, or *qubit*, which is a quantum version of the classical binary bit (i.e., 0 and 1). Thanks to the enormous amount of work in the QIP field over the past decade, we now know that it should be possible to use qubits to realize fundamentally new and more powerful methods for computation and communication. In other words, one ought to be able to harness the "spookiness" of quantum mechanics, in particular quantum correlations, to achieve a revolutionary form of information processing. It has been shown, for example, that algorithms for factorizing large integers can be realized faster in a suitably chosen QIP system than in a conventional computer [1]. The teleportation of a quantum state between spatially distinct locations has been experimentally demonstrated [2, 3]. An application of increasing importance in this field concerns the use of QIP to simulate other quantum systems [4–6]. Interesting by-products of this research include a new understanding of physical behavior at a quantum phase transition [7–9]. The promise of further exciting applications and a deeper fundamental understanding of quantum mechanics has sparked off a very active QIP research community spanning physicists, mathematicians, materials scientists, chemists and computer scientists.

1.1. The Challenge of Scalability

Any practical implementation of a QIP system needs to meet several stringent criteria in order to operate successfully [10]. First of all, the qubits (which can also be thought of as "quantum memories") must be sufficiently isolated so that they can then be directly and conditionally manipulated in a controlled environment. They need to be initialized precisely and then efficiently measured. The effective interactions among qubits should also be carefully tuned, and a set of universal quantum operations should be made possible in order to perform any other required quantum gate. Most importantly, the system must be scalable to more than a few qubits. Consequently, a large-scale QIP system is expected to include of the order tens to hundreds of qubits arranged in a configurable way, depending on the quantum routine to be achieved. In order to maintain the quantum correlations, parallel addressing of spatially separated units is also required, with operation times smaller than both local and nonlocal decoherence times. Scalability and robustness are hence arguably two of the most demanding requirements facing the practical implementation of quantum information processing.

1.2. Nanostructures for Scalable Quantum Information Processing

The clearest demonstrations of QIP involving massive particles have come from the field of interacting atoms/ions and photons [11]. Individual atoms or ions, which can be strongly or weakly coupled to the photon field, are manipulated in controlled environments with well understood couplings and well defined decay channels, thereby offering an ideal experimental setup to study systematically the basic principles of quantum computation and communication. In fact, trapped ions in optical lattices have become an ideal test ground for the study of more complex QIP architectures [12, 13]. Despite these achievements, there is a growing feeling within the QIP community that the potential integration of basic science and technology, which will be required for building large-scale quantum processors, will be best achieved in solid-state systems. After all, the transistor gave us classical computing, so why shouldn't a quantum version give us quantum computing?

In the past decade, there have been a number of proposals of solid-state systems in which quantum information processing might be achieved experimentally [14–19]. All these proposals are conceptually similar in the sense that they require manipulation of the coherent dynamics of a quantum system that is embedded in a complex environment. Hence, they all share the same challenge of having to resist highly nontrivial decoherence sources.

Given that decoherence times in solid-state systems are short, the structures themselves must be sufficiently small that the quantum information can be manipulated and/or transferred in a time that is shorter than the decoherence time. Remarkably, such small structures, or *nanostructures*, are now beginning to be built. In fact, the incredible advances in solid-state nanofabrication mean that artificial and natural nanostructures, such as quantum dots, are becoming prime candidates for building a robust large-scale QIP system.

Semiconductor quantum dots (QDs) [20–22] are structures with the dimensions of the order a few nanometers. Their electronic and optical properties can be tailored through the quantum confinement, which results from their finite size. In particular, they exhibit physical features similar to individual atoms but are typically a few orders of magnitude larger and have an energy scale that is a few orders of magnitude smaller. Because of their atom-like properties [21] but their key difference in size, they can be regarded as localized and addressable units for storing and manipulating quantum information. Quantum information can be encoded in QDs through a variety of effective two-level systems—for example, electron spin or charge excitation. Even more interesting is the possibility of building an artificial "molecule" by coupling together two QDs. As we will discuss later, the nature of the interaction between QDs not only offers different perspectives for QIP but also establishes a direct link with biological and organic systems.

Although the earliest proposals for QIP with QDs envisioned spin interactions controlled by electrical gates [14, 23], it is now thought that schemes based on all-optical control of the quantum system and its interactions are possibly more desirable, as they offer two principal advantages. First, the use of ultrafast laser technology [24] means that quantum operations can be carried out within the coherence time. Second, the interaction between a QD (acting as a stationary quantum memory) and photons (acting as a channel for information transmission) make this setup attractive for both the manipulation and transfer of quantum information over relatively large distances. In this latter context, QDs can also be integrated with optical nanocavities offering a means for exploiting the techniques of cavity quantum electrodynamics [25, 26].

With this backdrop, the aim of this chapter is to survey the current state of theoretical proposals of QIP in semiconductor QDs, focusing on schemes that exploit optical control of qubit interactions. Readers interested in other solid-state proposal are invited to consult Ref. [27]. The chapter is organized as follows. General theoretical aspects of the field of QIP are reviewed in Section 2. Theoretical proposals for all-optical QIP using QDs are reviewed in Section 3. Finally, future trends and developments are discussed in Section 4.

2. THEORETICAL BACKGROUND

The coherent superposition and entanglement of quantum states are the features lying at the heart of any QIP protocol. They account for quantum interference effects within a single system (superposition) or between different quantum systems (entanglement). In any experimental situation, however, the unavoidable interaction of a quantum system with its environment leads to decoherence processes that eventually destroy superpositions and entanglement and yield statistical mixtures of states. Understanding such complex decoherence mechanisms and finding approaches to make quantum interference effects robust in the face of such decoherence are still the major challenges facing many QIP proposals.

2.1. Basic Quantum Information Processing Toolbox: Superposition and Entanglement

Quantum coherence represents the ability of a quantum system to be in a superposition of different states and is the crucial element that leads to the definition of the fundamental unit of quantum information: the quantum bit (qubit). Qubits are quantum systems that can be represented to a good approximation as a two-state system. As a consequence, the state of an isolated qubit can be expressed as a *coherent superposition* of these two states:

$$|\Phi\rangle = \alpha|0\rangle + \beta|1\rangle \tag{1}$$

where $|\alpha|^2$ and $|\beta|^2$ are the probabilities that the system be found in state $|0\rangle$ or $|1\rangle$, respectively. Such a coherent superposition implies that there is always a basis in which the value of the qubit is well-defined, as opposed to an incoherent mixture in which the qubit's state becomes a statistical mixture irrespective of the basis used [28].

A single two-state quantum system can be represented as an effective spin-$\frac{1}{2}$ particle in a local magnetic field [29], whose Hamiltonian can be written as

$$\mathcal{H}_0(t) = B(t) \cdot \sigma \tag{2}$$

Here, $\sigma_{x,y,z}$ are the Pauli matrices describing the qubit to be manipulated. Full control over the coherent dynamics of the system is possible if at least two components of the effective field $B(t)$ can be switched arbitrarily. This Hamiltonian corresponds to an ideal physical scheme for realizing single-qubit unitary operations described by the operator $U(t, 0) = \mathcal{T} \exp(\int_0^t \mathcal{H}_0(t')dt')$ ($\mathcal{T}$ denotes time-ordering). For instance, under the action of $U(t, 0)$, it is possible to perform a Hadamard gate [30], which transforms the states $|0\rangle$ and $|1\rangle$ into coherent superpositions of the type described in Eq. (1) as $H|0\rangle \mapsto (1/\sqrt{2})(|0\rangle + |1\rangle)$ and $H|1\rangle \mapsto (1/\sqrt{2})(|0\rangle - |1\rangle)$.

However, any physical realization of a qubit—for example, an electron's spin up or down, an optically excited QD with one exciton or no exciton, or the flux in a SQUID—is embedded in a weakly interacting environment with a large number of degrees of freedom. This yields transient behavior with only partial coherence. Thus, a density matrix formalism is more convenient to describe the qubit dynamics, where the reduced density operator in the basis $\{|0\rangle, |1\rangle\}$ can be written as

$$\hat{\rho}_q = \begin{pmatrix} \rho_{00} & \rho_{01} \\ \rho_{10} & \rho_{11} \end{pmatrix} \tag{3}$$

The diagonal elements ρ_{00} and ρ_{11} denote the populations of each level and the off-diagonal elements ρ_{10} and ρ_{01} define the coherences between states $|0\rangle$ and $|1\rangle$. The population of the excited level ρ_{11} is typically described as having a characteristic decay time known as T_1, whereas the typical decay time of the coherences ρ_{10} and ρ_{01} is known as the dephasing or coherence time T_2. If, however, the qubit dynamics are explored over timescales significantly smaller than T_2 using ultrafast spectroscopy, the dephasing processes may not have sufficient time to be completed and hence the effects of memory in these processes must be accounted for. The Markov approximation in which memory effects are neglected yielding a characteristic exponential decay may no longer be sufficiently accurate. Consequently, T_2 may depend on the temporal correlations between the qubit system and the environment and within the environment itself. In semiconductor nanostructures, such non-Markovian effects are under active experimental and theoretical investigation [31–33].

The processing of quantum information requires qubits to be coherently manipulated on timescales shorter than T_2. From the experimental point of view, a fundamental probe of coherent manipulation of a qubit is the identification of Rabi oscillations, which are produced in resonantly driven two-state systems [34] and have no classical analog. They correspond to a sinusoidal time-evolution of the population difference in a two-level system for timescales short as compared to the coherence time.

Entanglement occurs as a result of quantum interference among states of composite systems, giving rise to nonclassical correlations between spatially separated quantum systems. An essential manifestation of these nonclassical correlations is that a measurement performed in one of the subsystems determines the state of the other even if they are significantly far apart. For a long time, this was considered a bizarre if not "spooky" feature of quantum mechanics. However, the modern view of entanglement is to see it as a fundamental nonlocal resource that can be used in any QIP protocol. The power of this resource has been demonstrated in quantum communication schemes performed with entangled photons, including state teleportation [2, 3], cryptographic key distribution [35], and quantum dense coding [36]. Entanglement of massive particles has been experimentally achieved in cavity QED [37], ion traps [38], and superconducting qubits [39, 40].

The most general entangled state of N qubits can be expressed as a superposition of 2^N states with complex and time-dependent amplitudes a_i with $i = 0, 1, \ldots, 2^N - 1$:

$$|\Psi\rangle = a_0(t)|0_1 0_2 0_3 \ldots 0_N\rangle + a_1(t)|1_1 0_2 0_3 \ldots 0_N\rangle + \cdots + a_{2^N-1}(t)|1_1 1_2 1_3 \ldots 1_N\rangle \quad (4)$$

As long as this state cannot be factorized into a product of single qubit states, then there will be nonclassical correlations that can be exploited to store more information than a comparable set of classical counterparts, and that can be exploited to perform parallel operations on different qubits. Therein lies the extraordinary potential for quantum computation.

The quantification of entanglement has become a central problem in quantum information theory. In the case of bipartite quantum systems, several measures have been proposed (see Ref. [41]) including the entanglement of formation [42], which has the clear meaning of being the asymptotic number of Bell pairs required to prepare the state using only local unitary transformations and classical communication. Intimately related to this measure is the notion of concurrence [43]. Although the bipartite case is well understood, a general formulation of multipartite entanglement remains an outstanding open problem.

2.2. Universal Resources for Quantum Information Processing

One of the key issues that has been addressed in QIP science is to determine what subsets of physical resources are capable of achieving universal quantum computation and, more generally, how to quantify such resources for different QIP tasks [44]. The idea of universality is to identify a set of gates that are sufficient to perform any other quantum operation on arbitrary N qubits—for example, in order to transform any initial separable state of N qubits into an arbitrary entangled state of the type described in Eq. (4).

It is well-known that certain two-qubit gates are universal for quantum computation when assisted with arbitrary one-qubit gates [45]. An example is the controlled-not (CNOT) operation which transforms as

$$|m, n\rangle \rightarrow |m, m \oplus n\rangle \quad m \in \{0, 1\} \quad (5)$$

where $\oplus$ denotes addition modulo 2. Recently, it has become clearer what general class of two qubit gates are universal. It turns out that any interaction creating entanglement between any pair of qubits is universal for quantum computation [46–48]. In particular, it has been demonstrated that a fixed *entangling* two-qubit interaction can be used to perform universal quantum computation, when assisted by single-qubit transformations between applications of the entangling gate V [48]. This can be expressed as follows

$$U = (A_1 \otimes B_1)e^{iV}(A_2 \otimes B_2) \quad (6)$$

where A_i, B_j are one qubit gates and $V = \sum_{X,Y,Z} \theta_\alpha \sigma_\alpha \otimes \sigma_\alpha$ with $\frac{-\pi}{4} \leq \theta_\alpha \leq \frac{-\pi}{4}$. The key point in the demonstration is that by using single-qubit gates, any gate U can be transformed into $W = e^{i(\alpha\sigma_Z \otimes \sigma_Z}$ with $0 < |\phi| < \pi/2$, which is a natural gate for implementing the CNOT operation.

2.2.1. Fundamental Two-Qubit Interactions

The physical implementation of a two-qubit gate depends on both the system's network structure and the nature of the *effective* interaction coupling together two qubits. The network structure refers to whether each qubit can be coupled to any other qubit, or whether each qubit can only be coupled to a few other qubits (e.g., nearest neighbors). For a given quantum task, this underlying structure is related to the number of operations required to achieve that task. This, however, also depends on the microscopic nature of the interaction between qubits. For instance, it has been discussed that a CNOT operation can be more efficiently realized if the available interaction is given by an Ising Hamiltonian rather than when it is described by an isotropic Heisenberg interaction [49]. The nature of the interaction also

determines whether it is possible to use recoupling to encode qubits and hence reduce the number of single-qubit operations [50].

In a large number of proposals for QIP in solid-state systems, the fundamental two-qubit gate is generated by an exchange interaction between qubits i and j, which has the general form

$$\mathcal{H}_{\text{int}}(t) = \sum_{\alpha=X,Y,Z} J^{\alpha}(t)\sigma_i^{\alpha} \cdot \sigma_j^{\alpha} \tag{7}$$

This interaction may arise naturally in the system, or it might be effectively created by manipulating local variables like individual energy transitions or individual couplings to a common collective mode.

The isotropic Heisenberg model, which corresponds to $J^{\alpha}(t) \equiv J(t)$, describes the fundamental two-qubit interaction in proposals for the implementation of spin-based quantum computation [14, 15, 23, 51]. It has been shown that the Heisenberg interaction is sufficient to perform universal quantum gates when a logical qubit is encoded in two or more qubits [52]. These schemes require tunability of the interaction, which may prove very difficult to achieve experimentally. As an alternative, schems for quantum computation with an always-on Heisenberg interaction have been proposed [53]. By tuning individual energy transitions via local or global addressing, the effective interaction between qubits in a linear array becomes an effective Ising interaction: $\mathcal{H}_{\text{int}} \simeq J\sum \sigma_i^Z \sigma_{i+1}^Z$. The interesting point here is that up to single-qubit rotations, the CNOT operation is a natural two-qubit gate for the Ising Hamiltonian as discussed in Refs. [48, 49], and hence it is a resource for universal quantum computation.

The anisotropic XXZ model, which corresponds to the case where $J^X(t) = \pm J^Y(t) \neq J^Z(t)$, is the natural interaction of electrons on a helium surface [18]. The scaling properties of entanglement close to the phase transition described by this model have been discussed [9], as well as schemes for encoded QIP [50].

The XY model, corresponding to the case $J^X(t) = J^Y(t)$, $J^Z = 0$, has been intensively studied in the context of QIP [7, 9, 49, 54]. These studies include the generation of so-called W states [55] in a one-dimensional XY model [54] as well as scaling properties of entanglement in the vicinity of a quantum phase transition described by the XY interaction in a transverse magnetic field [7, 9]. Moreover, it has been shown that the effective dipole–dipole interaction dominates the generation of entangled states in a wide range of systems, from excitons in quantum dots [16] through to both atomic systems [56] and quantum-dot spins in cavities [25]. In addition, it has been shown that a natural two-qubit gate for this interaction is a simultaneous CNOT and SWAP operation [49]. The XY interaction therefore has an important potential role to play in QIP implementations.

2.2.2. Cavity QED as a Resource for QIP

The nature of the photon-matter interaction in cavity quantum electrodynamics (QED) has very important implications for the processing of QIP [57] for several reasons. In the strong coupling regime, it is possible to use the cavity field as an intermediary or "bus" to create entanglement between qubits. Several two-qubit entangling protocols using this scheme have been proposed [56, 58, 59] with some of them being experimentally realized [37]. Cavity QED thus provides a scenario for exploring more complex QIP protocols in "clean" environments. In addition, qubit-cavity systems are arguably the most promising candidates for the distribution of quantum information in a quantum network [60]. For example, qubits driven by lasers and strongly coupled to spatially separated cavities can be entangled via photons travelling from one cavity to another [60]. Finally, cavity QED has well-defined decoherence channels. Subject to high-efficiency measurements, this provides an experimental setup for studying the conditional evolution of open quantum systems [57].

In cavity-assisted entanglement, noninteracting qubits can be prepared in a entangled state through either dispersive [56] or resonant [59] interactions with a common quantized field. Entanglement is achieved in the transitory regime where the strong interaction between qubits and a vacuum cavity field is switched on and off on a timescale shorter than the cavity-field decay time and the dipole decay time. Such an interaction can be effectively

created using a chain of single qubits, for example via a stream of flying qubits that enter and leave the cavity, as well as via a set of qubits that simultaneously interact with the field.

The most general situation is described in the interaction picture by the Hamiltonian ($\hbar = 1$) [61]

$$\mathcal{H}_c = \sum_{i=1,2} f_i(t_i, t, \tau_i)\{e^{-i\delta t}a^\dagger \sigma_i^- + e^{i\delta t}\sigma_i^+ a\} \tag{8}$$

where $\sigma_i^+ = |1_i\rangle\langle 0_i|$, $\sigma_i^- = |0_i\rangle\langle 1_i|$ with $|1_i\rangle$ and $|0_i\rangle$ ($i = 1, 2$) being the excited and ground states of the ith qubit. Here, $a^\dagger$ and a are respectively the creation and annihilation operators for the cavity photons, and δ is the detuning between the qubit transition frequency and the cavity field frequency. The time-dependent coupling of the cavity field with the ith qubit, which is injected at t_i and interacts during a time τ_i, is given by a time-window function,

$$f_i(t_i, t, \tau_i) = [\Theta(t - \tau_i) - \Theta(t - \tau_i - t_i)]\gamma_i(t) \tag{9}$$

where $\gamma_i(t)$ is the time-dependent qubit-field coupling strength. For the situation in which the second qubit interacts with the cavity just after the first did (i.e., $t_2 = t_1 + \tau_1$), the Hamiltonian corresponds to the Jaynes–Cummings interaction for each qubit separately. Under resonant conditions $\delta = 0$, the initial maximum entanglement between a first qubit and the field is converted into qubit–qubit entanglement during the interaction of the second qubit with the cavity mode [59]. Einstein–Podolsky–Rosen (EPR) atomic pairs and three particle entangled states have experimentally been achieved using this sequential scheme [37].

For the case of two qubits that simultaneously interact with the same cavity field, we have $t_1 = t_2$ and $\tau_1 = \tau_2$. A far richer control space can be explored if we generalize the setup to different and time-dependent qubit-cavity couplings. This latter situation is described by a generalized two-qubit *dynamical Dicke* (*DD*) model. Under nonresonant conditions where $\delta \gg \gamma_1\gamma_2$, the common quantized field is only virtually excited. This gives rise to an effective interaction corresponding to the Heisenberg XY model [56], in which a coherent energy transfer takes place between pairs of qubits. This scheme has been experimentally realized with two Rydberg atoms crossing a nonresonant cavity [62].

When both qubits are on resonance with the cavity mode and are simultaneously interacting, the generation of entanglement can be controlled using asymmetric qubit-cavity couplings [59]. This scheme exploits a trapping vacuum condition in which the cavity-qubit state becomes separable, leaving the cavity photon number unchanged but the two-qubit subsystem in an entangled state. Such a trapping condition does not arise for identical couplings.

2.3. General Aspects of Decoherence

Decoherence is generally a fast process whose timescale depends primarily on the size and temperature of the system but may also depend on other factors as well (e.g., sample preparation, errors in the prescribed unitary evolution, different sources of thermal and quantum noise). The original meaning of decoherence was specifically designated to describe the loss of coherence in the off-diagonal elements of the density operator in the energy eigenbasis [63]. More recently, with the additional goal of better understanding the effects of environmental interactions on QIP and other forms of quantum control, decoherence has been defined as nonunitary dynamics induced by system–environment couplings [64, 65].

This nonunitary dynamics is not limited to the known relaxation processes of dissipation and dephasing, traditionally associated with T_1 and T_2 processes. Nonunitary or "irreversible" dynamics also includes processes in which the system is conditioned to follow a specific quantum path by a measurement at certain times, that is, conditional quantum evolution or processes in which correlations in the bath properties may play an important role in the system's dynamics (non-Markovian effects).

Irreversible dynamics is generally seen as an undesirable consequence of QIP hardware. However, recent studies have shown that decoherence channels can actually be exploited to favor quantum coherence (e.g., in order to prepare two-qubit entangled states [66, 67] or to perform high-fidelity quantum gate operations [68]). Here, we briefly summarize three approaches that give insight into how to avoid decoherence sources, or how to use them in

order to promote the persistence of quantum coherence. The first two approaches can be derived within the general framework of processes for which all memory effects in the environment can be discarded, that is, Markovian dynamics that can be described by Lindblad-like equations [63]. The third type of approach, in contrast, is aimed at understanding and exploiting the bath's memory effects.

2.3.1. Conditional Quantum Evolution

Conditional quantum evolution, also known as the quantum jump approach, has mainly been studied in connection with optical quantum systems [69]. In this approach, the environment or bath functions are subject to a continuous series of measurements that have the potential to interrupt the unitary evolution of the system at infinitesimal time intervals. Conditioned on a particular observation record for the decay channels, the system follows non-Hermitian dynamics in which the effective time-evolution does not preserve the norm of the state. This approach has been reviewed in Ref. [69].

The possibility of using decay channels to prepare entangled states have been discussed in the case of noninteracting qubits [66] as well as for interacting qubits [67] that are identically coupled to the quantized mode inside a leaky cavity. Subject to the condition that no photon is detected, they both demonstrate that a successful measurement performed on a sufficiently large timescale generates an uncorrelated state for the qubits and cavity, leaving the two-qubit subsystem in its antisymmetric maximally entangled state. In the framework of this conditional evolution, it has also been shown that dissipation can be exploited to implement fast two-qubits gates with ground state qubits coupled to a common vibrational mode [68].

A common criticism of the above schemes is that the gate operation is probabilistic, as there is always a nonzero probability of obtaining an undesirable outcome (i.e., a photon is detected and hence the gate operation has failed). Therefore, in a broader perspective, a primary motivation for research on conditional quantum evolution is the prospect of developing general methods for real-time feedback control of open quantum systems [57].

2.3.2. Decoherence-Free Subspaces and Subsystems

The central idea of decoherence-free subspaces is the identification of the dynamical symmetries in the system–environment interaction, in order to construct a basis of collective states that are robust to dissipation and dephasing [70, 71]. In this approach, all qubits are identically coupled to the same environment such that the qubit-permutation symmetry gives rise to collective decoherence that can be effectively overcome if the qubits are encoded in decoherence-free states. These states have been identified as singlets or one-dimensional irreducible representations of the algebra generating the dynamical symmetry [71].

Following these ideas, there have been proposals for inducing subspaces that are effectively decoupled via external time-dependent Hamiltonians. This is known as dynamical decoupling [72], and it appears to be a promising alternative for suppressing decoherence in solid-state systems that are subject to strong low-frequency noise [73]. A further generalization and unification of various schemes to avoid decoherence is provided by the concept of noiseless subsystems [74], in which the information can be encoded in higher dimensional representations. This generalization has become the basis of a full theory of universal and fault tolerant QIP on decoherence-free states [75]. A recent review of the theory of decoherence-free subspaces and it extension to subsystems is given in Ref. [65].

2.3.3. Non-Markovian Approaches

A particular motivation for understanding non-Markovian effects in the dynamics of an open quantum system comes from experimental and theoretical studies in semiconductor nanostructures that have shown that the relaxation-time approximation may be questionable in the ultrafast optical regime within which these systems are being explored for QIP. The reason is that these approximations overestimate decay effects on the short timescales within which the system is dynamically evolving [31–33].

These studies tie in to the more fundamental goal of advancing the theory of open quantum systems in order to develop analytical methods in which correlations in the environment are incorporated [63]. A specific goal is to construct evolution equations for the

reduced density matrix that generalize the Markovian Lindblad master equation in order to include bath memory effects, while remaining both numerically and analytically tractable [76–78]. Such non-Markovian extensions are required in order to preserve complete positivity, thereby ensuring that the system dynamics is compatible with the joint system-environment unitary evolution.

The importance of considering correlation effects within the bath lies in the fact that such interference effects might actually reduce the effective decay-rate for coherence. For example, it has recently been shown that decoherence control is possible through a Parrondo-like effect [79] in which two correlated decoherence sources are made to effectively cancel.

3. ALL-OPTICAL QUANTUM INFORMATION PROCESSING IN SEMICONDUCTOR NANOSTRUCTURES

The possibility of controlling the coherent dynamics of an open quantum system is a fascinating topic from both technological and fundamental physics perspectives. In semiconductor nanostructures, the availability of ultrafast lasers with their wide ranges of pulse widths, wavelengths, pulse energies, and pulse repetition rates, has made possible the coherent control of matter dynamics in the *transient* regime, that is, before relaxation processes destroy the coherence created by an ultrafast optical excitation [24]. For optically induced polarization of a quantum dot, the timescale of the transient regime has been found to range from tens of picoseconds [80] at low temperatures, down to several hundreds of femtoseconds at room temperature [81]. At low temperatures, the mechanism of decoherence is mainly determined by radiative decay [80], whereas at higher temperatures, it is dominated by pure dephasing processes that are themselves driven by interactions of the charge carriers with acoustic phonons [82].

Unprecedented levels of coherent control have been demonstrated on superpositions of exciton and biexciton states in a single quantum dot, using ultrafast optics [83–86] (see Fig. 1). Earlier experiments showed coherent manipulation of the exciton wave function [80], coherent evolution between two different excitonic states with orthogonal polarizations [83], as well as Rabi oscillations between the vacuum state of excitons and a linearly polarized excitonic state [84]. Furthermore quantum superpositions have been reported between the ground state and the biexciton state [85] as well as exciton-to-biexciton Rabi oscillations [86]. These experimental achievements have been integrated in a setup to perform an all-optical quantum gate in a single quantum dot [86]. Although this system is not scalable by simply adding arbitrary numbers of excitons into a dot, it demonstrates the potential for ultrafast optically driven manipulation in scalable architectures based on multidot arrays.

Proposals for large-scale QIP in semiconductor QDs (Fig. 2) have been made using two different realizations of the qubit: the excess electron spin [14, 23, 87, 88] and an electron–hole pair excitation or exciton [16, 17, 89–92]. The former qubit benefits from a weak coupling to the environment, yielding longer coherence times that are of the order of

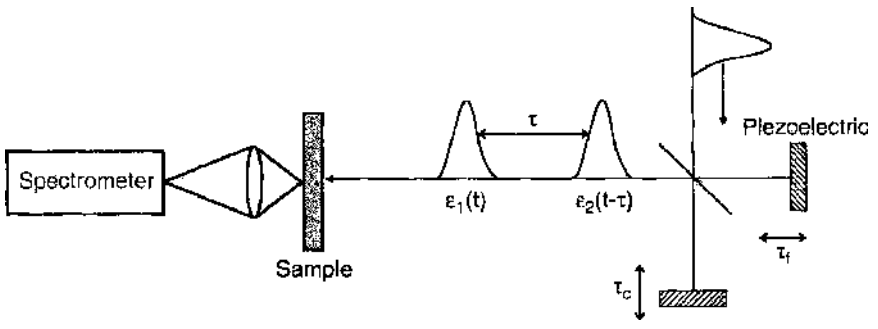


Figure 1. Schematic diagram of the experimental setup used for ultrafast coherent control of an exciton in a dot in Ref. [80]. Wave interferometry is achieved by using two phase-locked optical pulses ε_1 and ε_2 delayed a time $\tau = \tau_f + \tau_c$. The pulses interact with the QD at times $t = 0$ and $t = \tau$, creating a coherent superposition of the dipole excited state.

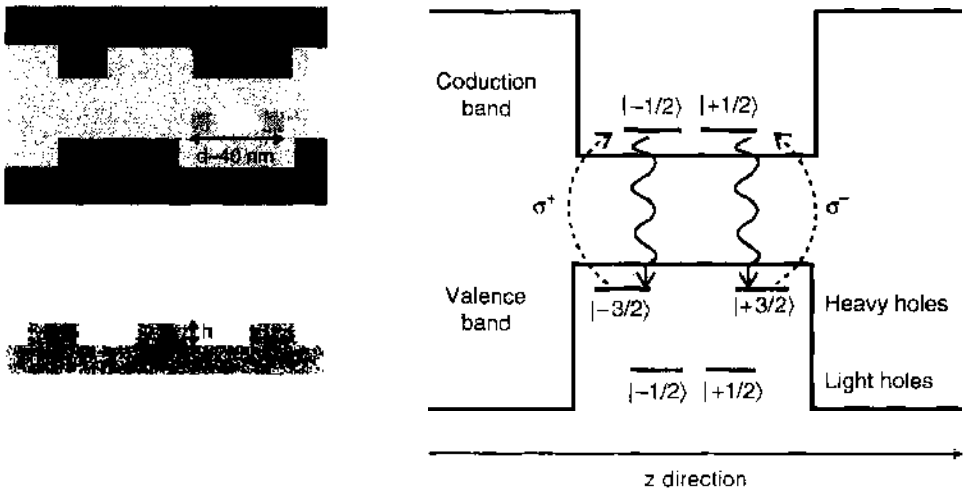


Figure 2. Left panel: Schematic diagrams of quantum dots. An island formed due to fluctuations in a quantum well width (left top) and self-assembled quantum dots grown by the Stranski–Krastanow process (left bottom). The base diameter d and the height h are in the range 20–40 nm and 3–6 nm, respectively, depending on the growing conditions. These later dots offer high potential for scalable architectures for QIP. Right panel: Relevant energy levels of a *III–V* semiconductor quantum dot. Strong confinement is assumed to be in the z direction. Here, $\sigma^{\pm}$ indicates right (left)-hand circularly polarized light.

microseconds [93]. However, it lacks a fast and easy tunability of coupling between qubits. The latter candidate provides an extremely fast interaction between qubits, of the order picoseconds for several proposals, but it suffers from short dephasing times. In order to achieve a high ratio between coherence and gate operation times, hybrid schemes that merge these various advantages have recently been proposed [94–97]. In particular, these latter approaches use optical excitations to control the coupling between QD spins. Schemes based on spin-flip Raman transitions mediated by a high-Q semiconductor microcavity have also been proposed [25, 87, 88].

The common goal of these approaches is to use ultrafast technology to achieve all-optical control of the qubit state and inter-qubit interactions in semiconductor QDs, and to then scale up this procedure to produce large-scale QIP. In this section, we review proposals aimed at this end.

3.1. Ultrafast Schemes with Excitons as Qubits

Optically driven semiconductor QDs lie at the heart of many QIP proposals [16, 17, 89–92]. These systems offer discrete energy levels that can be quickly and accurately addressed in space and can be built using well established semiconductor fabrication technology. The natural qubit in the QD system is given by the absence ($|0\rangle$) and presence ($|1\rangle$) of a ground-state electron–hole pair, the so-called exciton as illustrated in Fig. 3. Experimental evidence suggests that the decoherence time of this qubit is mainly limited by the radiative lifetime [80].

Several QIP schemes have proposed exploiting exchange and/or direct Coulomb interactions between spatially separated excitonic qubits in coupled QD structures [16, 17, 89, 91]. The Coulomb exchange interaction in QD molecules gives rise to a nonradiative resonant energy transfer (i.e. *Förster* process), which corresponds to the exchange of a virtual photon, thereby destroying an exciton in a dot and then re-creating it in a close by dot (see left-hand side of Fig. 4). Quiroga and Johnson [16] and Reina et al. [89] have discussed how to exploit the Förster interaction to prepare both Bell and Greenberger–Horne–Zeilinger (GHZ) entangled states of excitons, using far-field light excitation to globally address two and three quantum dots in a spatially symmetric arrangement. The Hamiltonian describing the formation of single excitons within the individual QDs and its interdot Förster hopping is given by [16]

$$\mathcal{H}(t) = H_0 + H_I + H_{xx}(t) \tag{10}$$

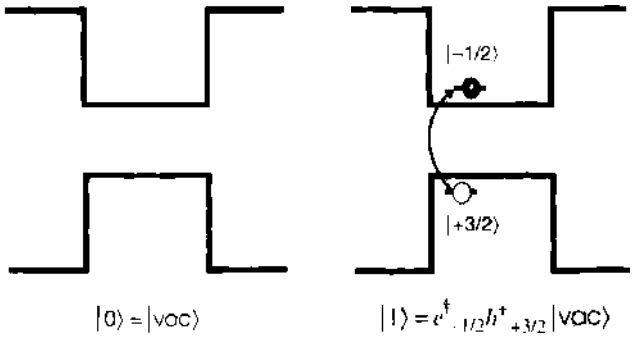


Figure 3. Logical states of an excitonic qubit. $|0\rangle$ denotes vacuum of excitations and $|1\rangle$ referes to the presence of a groundstate exciton made up of a hole with angular momentum $-3/2$ and an electron with angular momentum $1/2$.

Here the single-exciton Hamiltonian is given by

$$H_0 = \frac{\varepsilon}{2}\sum_{l=1}^{N}(e_l^\dagger e_l - h_l h_l^\dagger) \tag{11}$$

the interdot Förster interaction can be written as

$$H_F = -\frac{V_F}{2}\sum_{l,l'}^{N}(e_l^\dagger h_{l'} e_{l'} h_l^\dagger + h_l e_{l'}^\dagger h_{l'}^\dagger e_l) \tag{12}$$

the coupling of the carrier system with a classical laser of amplitude $E(t)$ is described by

$$H_{ext}(t) = E(t)\sum_{l=1}^{N} e_l^\dagger h_l^\dagger + E^*(t)\sum_{l=1}^{N} h_l e_l \tag{13}$$

and all constant energy terms have been ignored. The electron (hole) creation operator in the lth QD is designated by $e^\dagger(h^\dagger)$, ε is the QD band gap, and V_F denotes the strength of the Förster coupling between dots. The equidistant configuration of the multidot system favors the definition of global quasispin operators, which in turn enable an understanding of the dynamics of the multiexciton system in terms of the dynamics of its associated global quasispin [16]. Starting with the initial condition of a vacuum of excitons, the preparation of a Bell state such as $|\Psi\rangle = \alpha|00\rangle + \beta|11\rangle$, where $|11\rangle$ denotes the simultaneous presence of two excitons in a double dot structure, follows from the application of a finite rectangular pulse of

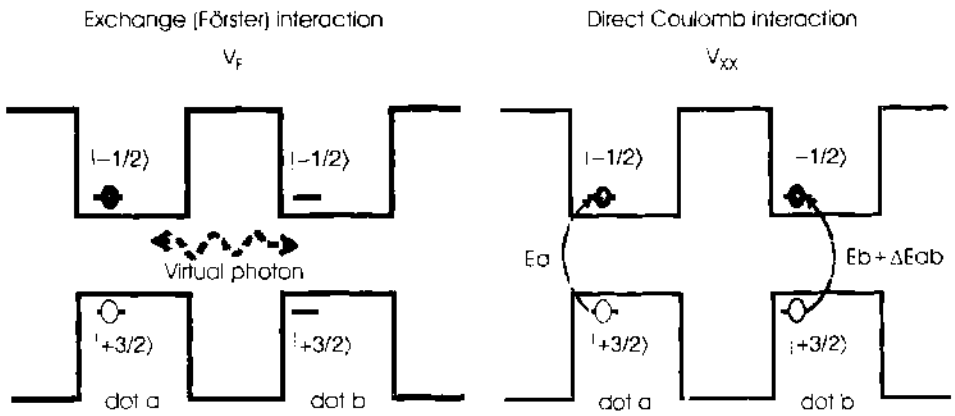


Figure 4. Schematic illustration of inter-dot interactions. Resonant energy transfer (Förster) process thereby an exciton is destroyed in one quantum dot and re-created on the other quantum dot via exchange of a virtual photon (left side). When each dot contains an exciton, direct electrostatic Coulomb interaction between quantum dots takes place. This leads to individual energy shifts that in presence of a static electric field in the xy plane become significant and can be exploited to generate an energy selective two-qubit gate (right side).

the form $E(t) = A\cos(\omega t)$ with central frequency ω and sub-picosecond duration. When the multidot system is arranged in a linear configuration where exciton hopping takes place only between nearest neighbors, the Hamiltonian $\mathcal{H}(t)$ takes the form of the one-dimensional XY Heisenberg model [98]. The derivation of the effective Hamiltonian relies on introducing the following local 1/2−pseudospin operators for electron-hole pairs in each dot:

$$\begin{aligned}\sigma_l^x &= \frac{1}{2}(e_l^\dagger h_l^\dagger + h_l e_l)\\ \sigma_l^y &= \frac{-i}{2}(e_l^\dagger h_l^\dagger - h_l e_l)\\ \sigma_l^z &= \frac{1}{2}(e_l^\dagger e_l - h_l h_l^\dagger)\end{aligned} \tag{14}$$

Hence, without the optical field term, the effective XY Hamiltonian reads

$$\mathcal{H}_{\text{eff}}(t) = \varepsilon \sum_{l=1}^{N} \Delta\sigma_l^z - V_F \sum_{l=1}^{N-1} (\sigma_l^x \sigma_{l+1}^x + \sigma_l^y \sigma_{l+1}^y) \tag{15}$$

where Δ is the detuning from the resonant excitation and $\sigma_l^{\pm} = \sigma_l^x \pm i\sigma_l^y$. An interesting point that emerges from this analysis is the difference of timescales required to generate entanglement in the different configurations. In particular, it turns out that GHZ states require longer times in a symmetric multidot arrangement than in a linear configuration [98]. It would be interesting to check whether such a statement also applies to a more general QIP protocol.

The direct electrostatic interaction between two excitons (see right-hand side of Fig. 4) changes the energy of the excitonic transition and is known as a "bi-excitonic" shift [20]. In the presence of an in-plane electric field, the excitons acquire a permanent electric dipole, and the bi-excitonic shift becomes significant [99]. This electrostatic dipole–dipole interaction has been proposed as a physical mechanism for implementing a controlled-NOT operation in a double dot structure, as discussed by Biolatti et al. [17]. Assuming that the distance between the dots is large enough to prevent single-particle tunneling but at the same time that they are sufficiently close to assure a strong interdot Coulomb interaction, the effective Hamiltonian governing the dynamical evolution of the system can be written as [17]

$$\mathcal{H}_{\text{eff}}(t) = H_0 + H_{xx} = \sum_{l=a,b} E_l \hat{n}_l + \frac{1}{2}\sum_{l\neq l'} \Delta E_{ll'} \hat{n}_l \hat{n}_{l'} \tag{16}$$

where $\hat{n}_l$ denotes the excitonic occupation number operator with eigenvalues 0 and 1, with 0 and 1 denoting the absence and presence of an exciton in the lth QD respectively. Here, E_l denotes the groundstate exciton energy; $\Delta E_{ll'}$ is the bi-excitonic shift in the presence of an electric field, which only arises if the qubits in dots l and l' are in state $|1\rangle$ as illustrated in the right-hand side of Fig. 4. This state-dependent interaction can be exploited to design conditional operations with properly adjusted two-color laser pulses. For example, in the case of two coupled dots a and b, the transition $|1_a 0_b\rangle \mapsto |1_a 1_b\rangle$ could be achieved on the sub-picosecond timescale by the following sequence: First apply a π rotation of the state of qubit a ($|0\rangle \mapsto |1\rangle$) and then apply a second pulse with frequency $E_b + \Delta E_{ab}$ to perform a π rotation of qubit b. We note that this analysis neglects the presence of the Förster interaction. In addition, one of the possible difficulties with this scheme is the need for an external electric field, as this would require the presence of electrical contacts that not only increase the complexity of the setup but would also imply an additional source of decoherence due to electromagnetic fluctuations. In order to circumvent this requirement, the same group [90] has proposed the use of QD structures with built-in electric fields, as observed in GaN dots.

The interplay between the resonant energy transfer (V_F) and the interdot bi-exciton binding energy ($\Delta E_{ll'} = V_{xx}$) has been studied numerically by Lovett et al. [91, 92]. These authors have shown that by taking into account both the Förster interaction and the bi-exciton binding energy, it is possible to develop an energy-selective approach to prepare entangled states

of excitons and to perform the CNOT operation in QD molecules. They consider two QDs a and b having different sizes, which implies that in the absence of interdot interactions, there is an energy difference between the excitonic transitions in the dots denoted by Δ_0. Hence two regimes, determined by the ratio V_F/Δ_0, can be explored. When the Förster coupling is dominant ($V_F/\Delta_0 \gg 1$) and the initial state is $|0_a 1_b\rangle$, which denotes a single exciton in dot a and no-exciton in dot b, the system can naturally evolve into a singlet that is maximally entangled state. When $V_F/\Delta_0 \ll 1$, the bi-exciton binding energy V_{xx} becomes dominant and a CNOT operation can be implemented, which is driven by pulses of different frequencies. In this case, entangled states can also be prepared by a properly designed laser-pulse sequence starting from the vacuum of excitons.

Inter-dot interactions in the presence of interband excitations have recently been described in the context of a multipolar quantum electrodynamics (QED) Hamiltonian [100]. This treatment allows one to understand the previously discussed interactions in terms of the exchange of transverse photons and electrostatic contributions, but also points out the physical mechanism to induce dipole-forbidden transitions that are mediated by an optical near-field [100].

All these schemes share the advantage of providing fast two-qubit operations within the sub-picosecond timescale, as a result of resonant energy transfer or the interdot-biexciton binding energy. However, this timescale is still comparable with the dephasing time of the exciton dipole. A good question at this point is then: *Is it possible to take advantage of this ultrafast interaction, and hence ultrafast technology itself, to develop a combined scheme that integrates a long coherence qubit with fast two-qubit operations?* This is precisely the question that recent proposals have been exploring [94–97]. In what follows, we summarize proposals that share the idea of exploiting electron–hole excitations to control the coupling between QD spin qubits.

3.2. Exciton-Assisted Spin-Based Quantum Computation

Single electron spins confined in quantum dots have been proposed as qubits [14, 23] because of their long relaxation times, which are well into the microsecond timescale [93]. It has been suggested that the interaction between spins could be controlled by electronic gates [14]. However an all-optical approach is clearly more desirable, as this would avoid the additional fabrication of gates and the unavoidable fluctuations in electromagnetic fields arising from these gates. As for the other all-optical approaches that have formed the focus of this article, it instead makes sense to try to exploit the major advances in ultrafast laser technology that continue to be developed [24]. Indeed, it has already been demonstrated experimentally that a single electron spin in a QD can be probed optically [101].

The idea behind spin-based QIP assisted by excitons [94–97] is to exploit virtual or resonant interband excitations to induce and control optically interactions among spin-qubits that are localized in different QDs. In these schemes, the logical qubit is defined by the spin-states of a single conduction-band electron confined in a QD: $|0\rangle = |m_z = -1/2\rangle$ and $|1\rangle = |m_z = 1/2\rangle$, as shown in Fig. 5.

The first scheme that used exciton states to induce indirect exchange interactions between electron spins localized on different QDs, was discussed by Piermarocchi et al. [94].

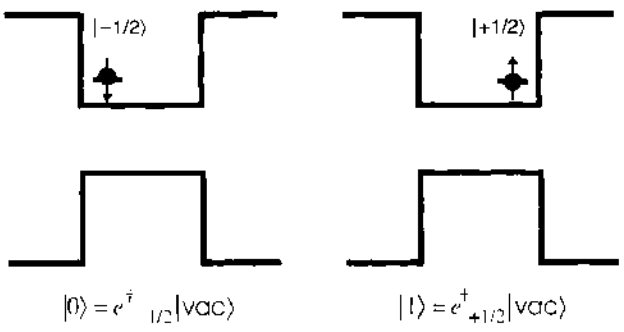


Figure 5. Logical states associated to the spin of a single electron confined in a QD.

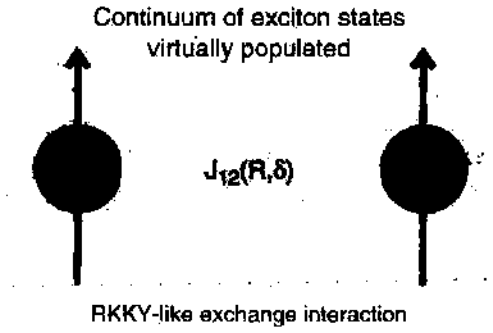


Figure 6. Schematic illustration of the effective optical RKKY exchange interaction between electrons in separated quantum dots via an intermediate continuum of exciton states produced by an off-resonance optical excitation. Here, δ is the detuning between the laser frequency and the band gap frequency in the host material; R is the distance between confined electrons.

The excess electrons confined in spatially separated QDs interact with a delocalized electron–hole pair that has been excited by an optical pulse that is itself detuned with respect to the continuum of exciton states in the host material. Keeping only the lowest order contribution, exchange Coulomb interactions between the localized and the optically excited conduction electrons lead to an effective spin–spin exchange coupling between QD spins of the type discussed in the previous section (i.e., $H_{\text{eff}} = -2J_{12}S_1S_2$ with $S_{i=a,b}$ being the electron spin in dot i). In this scheme, the coupling parameter J_{12} is always positive (ferromagnetic interaction) and depends on the detuning frequency between the laser and the excitonic transition in the continuum and the inter-dot separation, as illustrated in Fig. 6. This coupling mechanism is analogous to a Ruderman–Kittel–Kasuya–Yosida (RKKY) interaction between two well separated magnetic impurities mediated by either conduction electrons or excitons [102], except that in the current case, the intermediate electron–hole pair is produced by an external optical field.

The second scheme we discuss is based on conditional spin and exciton dynamics and employs a Pauli-blocking mechanism [95, 96]. As illustrated in Fig. 7, when the QD is excited with left-handed circularly polarized light, the presence of an excess electron in its spin-down state ($|0\rangle$) inhibits the creation of an exciton with electron angular momentum $m_z^e = -1/2$. By contrast, if the qubit is in its spin-up state ($|1\rangle$), then nothing prevents the creation of the electron–hole pair. This yields a state-selective coupling of the logical states $|0\rangle$ and $|1\rangle$ to the auxiliary state, defined as $|x^-\rangle$, which describes a QD with two electrons with opposite spins occuping the same energy level together with a single hole. This scheme also

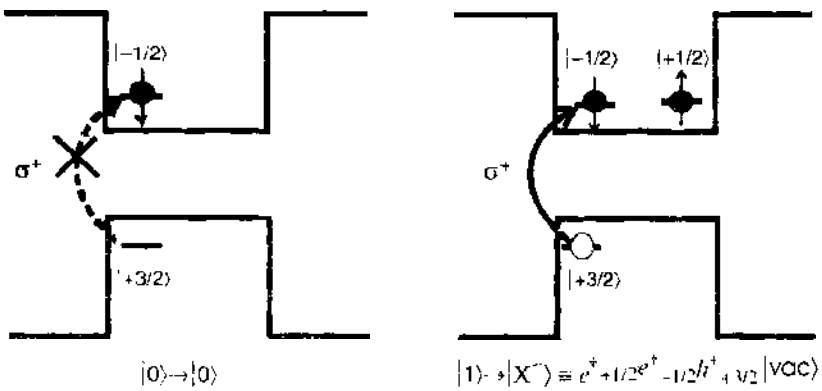


Figure 7. Optically controlled Pauli-blocking mechanism. If the spin qubit is in $|0\rangle$, a creation of an exciton with electron angular momentum $-1/2$ is inhibited (left side). If the qubit is in $|1\rangle$, then nothing forbids the creation of such exciton. If the qubit is in a coherent superposition of its logical states, then it is transformed into a charge superposition: $|0\rangle + |1\rangle \leftrightarrow |0\rangle + |X^-\rangle$.

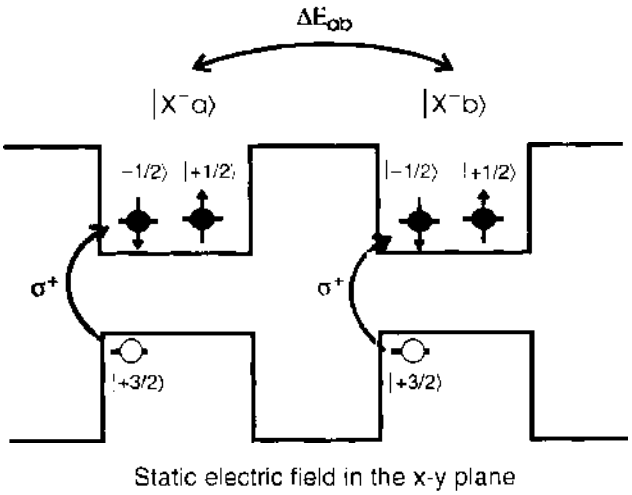


Figure 8. Bi-excitonic shift. Interaction between QD spins is controlled by the electrostatic Coulomb interaction between optically created trion states in dots $|x_a^-\rangle$ and $|x_b^-\rangle$. In presence of a static electric field, the trion states acquire permanent dipoles while the Coulomb interaction induces a significant energy shift ΔE_{ab} in exciton states. In combination with the Pauli-exclusion mechanism, this energy shift can be exploited to perform a quantum phase gate.

exploits the Pauli exclusion principle and electrostatic exciton–exciton interactions in order to implement a two-qubit phase gate. A rotation, given by an accumulated phase θ, only occurs when both qubits are in their logical state $|1\rangle$:

$$|m, n\rangle \mapsto e^{i\theta mn}|m, n\rangle \quad m \in \{0, 1\} \tag{17}$$

As discussed by Calarco et al. [96], the key mechanism here is the energy shift due to the electrostatic dipole–dipole interaction between excitons when an external static electric field is applied (Fig. 8). As mentioned earlier, this bi-excitonic shift occurs only in the case when both qubits are in their spin-up state. The effective Hamiltonian describing the situation for two adjacent QDs (a and b) is given by [96]

$$\mathcal{H}_{\text{eff}} = \sum_{\nu=a,b}\left(\frac{\Omega_\nu(t)}{2}|x^-\rangle_\nu\langle 1| + \textit{H.c.}\right) - \Delta E_{ab}|x^-\rangle_a\langle x^-| \otimes |x^-\rangle_b\langle x^-| \tag{18}$$

Here, $\Omega(t)$ is the effective Rabi frequency between the single-electron state $|1\rangle$ and the trion state $|x^-\rangle$. In this proposal, as in most of the spin-based schemes, realizing the optical rotation of a single spin represents a significant challenge. To overcome this difficulty, the authors suggest the implementation of Raman transitions via light-hole levels $m_z^h = \pm 1/2$ in situations where these states are the hole groundstates, a situation that does occur in II–VI semiconductors nanostructures.

A modification of the above scheme to achieve the spin-couplings via inter-dot resonant or Förster energy interaction (V_F) has been discussed in Ref. [97]. This modified scheme requires the excitation of a single groundstate exciton instead of inter-dot biexciton states.

3.3. Quantum Information Processing Schemes with Microcavities and Quantum Dots

As discussed in the previous section, prior work on conditional quantum dynamics has demonstrated that cavity QED could play a key role in future quantum communication or computation schemes. More specifically, it has been shown that cavity-mediated interactions provide a means for preparing entangled states, in addition to providing a means for transmitting quantum information between distant nodes in a quantum network. Following this line of thought, Imamoglu et al. [25, 87] proposed a system that relies on the use of a quantized cavity mode and applied laser fields in order to mediate the interaction between

spins of distant, doped QDs (Fig. 9). The central idea in this scheme is to implement cavity assisted spin-flip Raman transitions and to couple pairs of qubits via virtual photons in the common vacuum cavity mode.

Assuming that a uniform magnetic field is applied along the x direction, the QD qubit is defined by the spin states $|m_x = -1/2\rangle = |0\rangle$ and $|m_x = 1/2\rangle = |1\rangle$ of the single conduction-band electron. The scheme's proposers consider QD spins interacting with a x-polarized vacuum cavity mode and a y-polarized laser field in order to implement cavity-assisted spin-flip Raman transitions between the two spin states, in close analogy with atomic cavity-QED schemes [103]. By adjusting the frequencies of the individual lasers in order to establish a near two-photon resonance condition, such that the cavity is virtually excited, the following effective Hamiltonian can be obtained:

$$H_{\text{int}} = \sum_{i \neq j} \frac{\tilde{g}_{ij}(t)}{2} \left[\sigma_{10}^{i} \sigma_{01}^{j} + \sigma_{10}^{j} \sigma_{01}^{i} \right] \tag{19}$$

Here, $\sigma_{10}^{i} = |1\rangle\langle 0|$ is the spin projection operator for the ith QD, and $\tilde{g}_{ij}(t)$ corresponds to the product of the two-photon coupling coefficients for the spins in QDs i and j. It has been shown that the two-qubit coupling between any pair of QDs can be carried out in parallel and on sub-nanosecond timescales.

A modification of the above scheme so as to combine Pauli-blockade effects with the microcavity scheme has been discussed in Ref. [88]. In this modified scheme, the quantum information is defined by the states $|m_z = \pm 1/2\rangle$ of the single conduction electron. The cavity mode, which is assumed to be right-hand polarized, induces an electron–hole pair excitation in a dot only if the excess spin electron is in the state $|m_z = -1/2\rangle$. Individual lasers are assumed to be linearly polarized and have their frequencies adjusted such that the Raman transition between the spin states can occur.

3.4. Decoherence Control Through Optical Pumping

Despite its atom-like properties, there are fundamental features that distinguish a quantum dot from an atom. One of these is the completely different role that hyperfine interactions play in the two quantum systems. In contrast to valence electrons of an atom, a single electron in a dot is confined on a length scale that extends over many lattice sites. Consequently, the electron spin interacts with a few thousand randomly oriented nuclear spins. Interactions with this unpolarized bath present a decoherence mechanism for an electron spin in

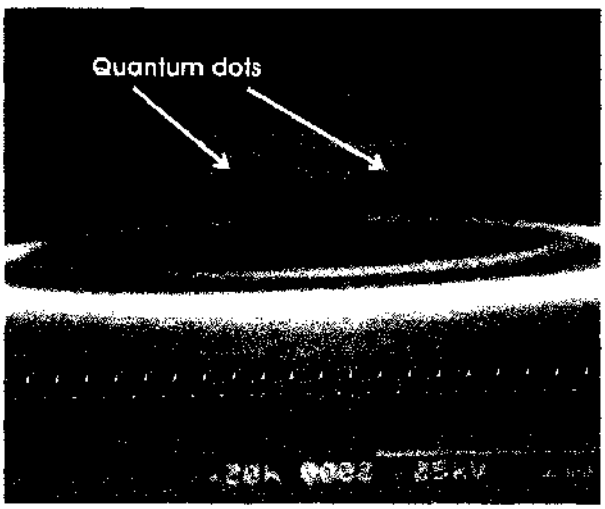


Figure 9. Scanning electron micrograph of the GaAs microdisk nanostructure. The diameter of the cavity is 4.5 μm with a corresponding cavity-mode volume of $V_{\text{eff}} \approx 20(\lambda/2n)^3$. The highest value of Q measured in this structure exceeded 1800. The cavity contains InAs QDs at locations fixed during the growth. The density of QDs is 2×10^{10} cm^{-2}. Reprinted with permission from [26], A. Kiraz, et al., *J. Opt. B: Quantum Semiclass. Opt.* 5, 129 (2003). © 2003, Institute of Physics Publishing.

a quantum dot [104]. Imamoglu et al. [105] have discussed an optical scheme to suppress this decoherence mechanism. The main idea is to use hyperfine interactions to polarize the nuclear field. This is achieved by shifting the energy of the initial spin-up electronic state via the ac-Stark effect in order to create a resonant condition for the electron-nuclear spin-flip transition.

Interestingly, Shabaev et al. [106] have shown that even in the presence of random hyperfine interactions with nuclear spins, a strong resonant optical excitation of the electron spin to an intermediate trion state provides readout capability of the spin state during the relaxation process of the trion state [106]. Combined with a permanent transverse magnetic field, such an optical excitation also provides a way to initialize the spin into a state with a well-defined phase [106].

3.5. Concluding Remarks

Optical properties of semiconductor QDs can be tailored by varying the size, shape, and composition material of the QD, thereby offering a suitable scenario to implement all-optical approaches for the coherent control of qubits and their interactions. Semiconductor nanostructures integrated with ultrafast optics technology are therefore an attractive solid-state alternative for constructing scalable and fault-tolerant architectures in order to implement quantum computation and communication, as well as quantum simulation protocols.

There is still a very long way to go before large-scale quantum processors can be made out of QD arrays. Indeed, the state of the field is such that it would be a major scientific breakthrough if someone were to demonstrate quantum entanglement involving just a few QDs, let alone control or manipulate this entanglement. However, there are many experimental groups trying to do exactly that—and eventually someone will manage. Along the way, there are many open problems that will need to be solved, and several new research themes will emerge. These open problems include a deeper understanding of decoherence mechanisms and readout for a single electron-spin, and experimental signatures of superposition and entanglement for excitonic and spin qubits and for the entanglement between a QD system and photons. In the next section, we discuss some of the trends we foresee for future developments.

4. TRENDS FOR FUTURE DEVELOPMENTS

The study of QIP in nanostructures has opened up many new questions. As a result of the presence of similar physical mechanisms of interactions in organic and semiconductor systems, one such question concerns the extent to which organic and biomolecular systems offer a viable alternative for QIP, for example, "bio-QDs" such as the LHI and LHII complexes used by nature to harvest the energy of a photon in photosynthesis. A second set of questions relates to how one can best exploit the interaction between nonclassically correlated photons and nanostructures. We address each of these briefly.

4.1. Quantum Information Processing in Organic and Biomolecular Nanostructures

Quantum dots embedded in organic structures offer novel physical properties. In particular, they allow the formation of exciton states that exhibit large oscillator strengths and strong coupling to the light. These properties yield a large coherence length and high optical nonlinealities [107] that may be exploited for QIP.

It has also been recognized that the coupling mechanisms occurring in QD molecules have the same physical origin as those present in existing natural systems, such as the light-harvesting complexes (LHI and LHII) for which excitonic interactions and energy transfer processes play a central role [108]. At nanosecond timescales, the fluorescence resonance energy transfer (FRET), which is associated with the Förster process in light-harvesting complexes and in FRET-coupled dye pairs, exhibits an incoherent dynamics. However, coherent

FRET signals might possibly be found at pico- and femtosecond timescales, as discussed earlier for coupled QDs. Due to these basic similarities between inorganic QDs and biomolecular nanostructures, novel ideas for using these latter natural structures to process quantum information are beginning to be explored [91, 109].

From our own perspective, we believe that *hybrid* bio-nano QIP systems will emerge as an important field of study in the future. This interdisciplinary field will need to combine novel ideas and understanding from the physical, chemical, and biological sciences, as well as the biotechnology and nanofabrication industries. Whether fully quantum, mixed quantum-classical, or just classical devices can be built remains to be seen. However, all three prospects are exciting, whether it be classical information processing (IP) or full quantum information processing (QIP) systems that finally emerge.

4.2. Nanostructures and Entangled Photons

We have mentioned that one can take advantage of the Förster interaction between two quantum dots, which are globally addressed by a laser beam, in order to generate Bell states of the type of $\alpha|00\rangle + \beta|11\rangle$, where the state $|11\rangle$ indicates the simultaneous presence of two excitons in the double-dot array [16]. This indicates that the formation of such entangled states can be achieved through a two-photon excitation. In fact, it has been experimentally shown that the coherent resonant dipole interaction between molecules can be probed via a two-photon transition in which both molecules are simultaneously excited [110]. An interesting question therefore arises as to what kind of quantum interference phenomena occur when the molecular (or QD) pair is excited by two photons that are entangled. Sources of entangled pairs of photons in the visible spectrum are now available [111]. In fact, sources of three and four entangled photons have very recently been reported [112, 113], which in turn allows one to extend the question beyond two molecules (or QDs) to many-particle systems. This is also an exciting area for future study, both theoretically and experimentally.

4.3. Photon Statistics and Dynamics of Nonclassical Correlations

In quantum systems with optical outputs, it is expected that experimental signatures of entanglement take the form of nonclassical statistical correlations of the emitted light [114]. In matter–light coupled systems, photon statistics has already proved to be a valuable tool for the identification of quantum signatures such as photon-antibunching in the resonance fluorescence of a two-level atom or a single quantum dot [115]. Indeed, second-order photon correlations are expected to exhibit signatures of coherent superpositions in single and double QD nanostructures [32, 33]. In a parallel development, Hanbury–Brown–Twiss experiments have been performed to measure intensity correlations in the nonlinear response of strongly and coherently coupled molecules [110], as well as in the fluorescence resonance energy transfer (FRET) in coupled dye pairs [116]. These experimental achievements suggest that photon statistics could be used to characterize the interactions in coupled quantum dots. However, further theoretical studies are required to investigate the precise details of such a characterization.

ACKNOWLEDGMENTS

We are extremely grateful to Luis Quiroga for useful discussions and critical reading of this manuscript. We acknowledge funding from the Clarendon Fund and ORS (A. O. C) and the DTI-LINK project "Computing at the Quantum Edge" (N. F. J).

REFERENCES

1. P. Shor, *SIAM J. Comp.* 26, 1484 (1997).
2. D. Bouwmeester, J.-W. Pan, K. Mattle, M. Eibl, H. Weinfurter, and A. Zeilinger, *Nature* 390, 575 (1997).
3. D. Boschi, S. Branca, F. De Martini, L. Hardy, and S. Popescu, *Phys. Rev. Lett.* 80, 1121 (1998).
4. S. Somaroo, C. H. Tseng, T. Havel, R. Laflamme, and D. G. Cory, *Phys. Rev. Lett.* 82, 5381 (1999).

5. C. H. Tseng, S. S. Somaroo, Y. S. Sharf, E. Knill, R. Laflamme, T. F. Havel, and D. G. Cory, *Phys. Rev. A* 61, 12302 (2000).
6. G. Vidal, *Phys. Rev. Lett.* 93, 040502 (2004).
7. A. Osterloh, L. Amico, G. Falci, and R. Fazio, *Science* 416, 608 (2002).
8. T. J. Osborne and M. A. Nielsen, *Phys. Rev. A* 66, 032110 (2002).
9. G. Vidal, J. I. Latorre, E. Rico, and A. Y. Kitaev, *Phys. Rev. Lett.* 90, 227902 (2003).
10. D. P. DiVincenzo, *Fortschr. Phys.* 48, 771 (2000).
11. C. Monroe, *Nature* 416, 238 (2002).
12. F. Schmidt-Kaler, H. Häffner, M. Riebe, S. Gulde, G. P. T. Lancaster, T. Deuschle, C. Becher, C. F. Roos, J. Eschner, and R. Blatt, *Nature* 422, 408 (2003).
13. S. Gulde, M. Riebe, G. P. T. Lancaster, C. Becher, J. Eschner, H. Häffner, F. Schmidt-Kaler, I. L. Chuang, and R. Blatt, *Nature* 421, 48 (2003).
14. D. Loss and D. P. DiVincenzo, *Phys. Rev. A* 57, 120 (1998).
15. B. E. Kane, *Science* 393, 133 (1998).
16. L. Quiroga and N. F. Johnson, *Phys. Rev. Lett.* 83, 2270 (1999).
17. E. Biolatti, R. C. Iotti, P. Zanardi, and F. Rossi, *Phys. Rev. Lett.* 85, 5647 (2000).
18. P. M. Platzman and M. I. Dykman, *Science* 284, 1967 (1999).
19. Y. Makhlin, G. Schön, and A. Shnirman, *Rev. Mod. Phys.* 73, 357 (2001).
20. L. Jacak, P. Hawrylack, and A. Wojs, "Quantum Dots," Springer, Berlin, 1998.
21. D. Bimberg, M. Grundmann, and N. N. Ledentosow, "Quantum Dot Heterostructures," John Wiley & Sons, Chichester, 1999.
22. P. Harrison, "Quantum Wells, Wires and Dots," Wiley, New York, 2001.
23. G. Burkard, D. Loss, and D. P. DiVincenzo, *Phys. Rev. B* 59, 2070 (1999).
24. J. Shah "Ultrafast Spectroscopy of Semiconductors and Semiconductor Structures," Springer, Berlin, 1998.
25. A. Imamoglu, D. D. Awschalom, G. Burkard, D. P. DiVincenzo, D. Loss, M. Sherwin, and A. Small, *Phys. Rev. Lett.* 83, 4204 (1999).
26. A. Kiraz, C. Reese, B. Gayral, L. Zhang, W. V. Schoenfeld, B. D. Gerardot, P. M. Petroff, E. L. Hu, and A. Imamoglu, *J. Opt. B: Quantum Semiclass. Opt.* 5, 129 (2003).
27. See for instance the proceedings of the Discussion Meeting at The Royal Society in November 2002. *Philos. Trans. R. Soc. Lond. A* 361 (2003).
28. D. Bouwmeester and A. Zeilinger, in "The Physics of Quantum Information" (D. Bouwmeester, A. Ekert, and A. Zeilinger, Eds.), p. 1. Springer-Verlag, Berlin, 2000.
29. M. E. Rose "Elementary Theory of Angular Momentum," John Wiley & Sons, New York, 1967.
30. M. A. Nielsen and I. L. Chuang, "Quantum Computation and Quantum Information," Cambridge University Press, Cambridge, UK, 2000.
31. P. Borri, W. Langebin, S. Schenider, U. Woggon, R. L. Sellin, D. Ouyang, and D. Bimberg, *Phys. Rev. Lett.* 87, 157401 (2001).
32. F. J. Rodríguez, L. Quiroga, and N. F. Johnson, *Phys. Rev. B* 66, 161302(R) (2001).
33. L. Quiroga, F. J. Rodríguez, and N. F. Johnson, *Microelectronics Journal* 35, 95 (2004).
34. L. Allen and J. H. Eberly, "Optical Resonance and Two-Level Atoms," Dover, New York, 1987.
35. C. H. Bennett, *Phys. Rev. Lett.* 68, 3121 (1992).
36. C. H. Bennett and S. Wiesner, *Phys. Rev. Lett.* 69, 2881 (1992).
37. J. M. Raimond, M. Brune, and S. Haroche, *Rev. Mod. Phys.* 73, 565 (2001).
38. C. A. Sackett, D. Kielpinski, B. E. King, C. Langer, V. Meyer, C. J. Myatt, M. Rowe, Q. A. Turchette, W. M. Itano, D. J. Wineland, and C. Monroe, *Nature* 404, 256 (2000).
39. A. J. Berkley, H. Xu, R. C. Ramos, M. A. Gubrud, F. W. Strauch, P. R. Johnson, J. R. Anderson, A. J. Dragt, C. J. Lobb, and F. C. Wellstood, *Science* 300, 1548 (2003).
40. T. Yamamoto, Yu. A. Pashkin, O. Astafiev, Y. Nakamura, and J. S. Tsai, *Nature* 425, 941 (2003).
41. D. Bruss, *J. Math. Phys.* 43, 4237 (2002).
42. W. K Wootters, *Quant. Inf. Comp.* 1, 27 (2001).
43. W. K. Wootters, *Phys. Rev. Lett.* 80, 2245 (1998).
44. M. A. Nielsen, C. M. Dawson, J. L. Dodd, A. Gilchrist, D. Mortimer, T. J. Osborne, M. J. Bremner, A. W. Harrow, and A. Hines, *Phys. Rev. A* 67, 052301 (2003).
45. A. Barenco, C. Bennett, R. Cleve, D. DiVincenzo, N. Margolus, P. Shor, T. Sleator, J. Smolin, and H. Weinfurther, *Phys. Rev. A* 52, 346 (1995).
46. W. Dür, G. Vidal, J. I. Cirac, N. Linden, and S. Popescu, *Phys. Rev. Lett.* **87**, 137901 (2001).
47. C. H. Bennett, J. I. Cirac, M. S. Leifer, D. W. Leung, N. Linden, S. Popescu, and G. Vidal, *Phys. Rev. A* 66, 012305 (2002).
48. M. J. Bremner, C. M. Dawson, J. L. Dodd, A. Gilchrist, A. W. Harrow, D. Mortimer, M. A. Nielsen, and T. J. Osborne, *Phys. Rev. Lett.* 89, 247902 (2002).
49. N. Schuch and J. Siewert, *Phys. Rev. A* 67, 032301 (2003).
50. D. A. Lidar and L.-A. Wu, *Phys. Rev. Lett.* 88, 017905 (2002).
51. R. Vrijen, E. Yablonovitch, K. Wang, H. W. Jiang, A. Balandin, V. Roychowdhury, T. Mor, and D. DiVincenzo, *Phys. Rev. A* 62, 012306 (2000).
52. D. P. DiVincenzo, D. Bacon, J. Kempe, and K. B. Whaley, *Nature* 408, 339 (2000).
53. S. C. Benjamin and S. Bose, *Phys. Rev. Lett.* 90, 247901 (2003).
54. X. Wang, *Phys. Rev. A* 64, 012313 (2001).

55. W. Dür, G. Vidal, and J. I. Cirac, *Phys. Rev. A* 62, 017901 (2000).
56. S. B. Zheng and G.-C. Guo, *Phys. Rev. Lett.* 85, 2392 (2000).
57. H. Mabuchi and A. C. Doherty, *Science* 298, 1372 (2002).
58. X. X. Yi, X. H. Su, and L. You, *Phys. Rev. Lett.* 90, 097902 (2003).
59. A. Olaya-Castro, N. F. Johnson, and L. Quiroga, *Phys. Rev. A* 70, 020301(R) (2004).
60. J. I. Cirac, P. Zoller, H. J. Kimble, and H. Mabuchi, *Phys. Rev. Lett* 78, 3221 (1997).
61. A. Olaya-Castro, N. F. Johnson, and L. Quiroga, *J. Opt. B: Quantum Semiclass. Opt.* 6, S730 (2004).
62. S. Osnaghi, P. Bertet, A. Auffeves, P. Maioli, M. Brune, J. M. Raimond, and S. Haroche, *Phys. Rev. Lett.* 87, 037902 (2001).
63. H.-P. Breuer and F. Petruccione, "The Theory of Open Quantum Systems," Oxford University Press, Oxford, 2002.
64. W. H. Zurek *Rev. Mod. Phys.* 75, 715 (2003).
65. D. A. Lidar and K. B. Whaley, in "Irreversible Quantum Dynamics" (F. Benatti and R. Floreanini, Eds.), Vol. 62, p. 83. Spring Lecture Notes in Physics. Springer, Berlin, 2003.
66. M. B. Plenio, S. F. Huelga, A. Beige, and P. L. Knight, *Phys. Rev. A* 59, 2468 (1999).
67. S. Nicolosi, A. Napoli, A. Messina, and F. Petruccione, *Phys. Rev. A* 70, 022511 (2004).
68. A. Beige *Phys. Rev. A* 67, 020301(R) (2003).
69. M. B. Plenio and P. L. Knight, *Rev. Mod. Phys.* 70, 101 (1998).
70. L.-M Duan and G.-C. Guo, *Phys. Rev. A* 57, 737 (1998).
71. D. A. Lidar, I. L. Chuang, and K. B. Whaley, *Phys. Rev. Lett.* 81, 2594 (1998).
72. L. Viola, *Phys. Rev. A* 66, 012307 (2002).
73. K. Shiokawa and D. A. Lidar, *Phys. Rev. A* 69, 030302(R) (2004).
74. P. Zanardi, *Phys. Rev. A* 63, 012301 (2001).
75. J. Kempe, D. Bacon, D. A. Lidar, and K. B. Whaley, *Phys. Rev. A* 63, 042307 (2001).
76. S. Daffer, K. Wodekiewicz, J. D. Cresser, and J. K. Melver, *Phys. Rev. A* 70, 010304 (2004).
77. H.-P. Breuer, *Phys. Rev. A* 70, 012106 (2004).
78. A. Shabani and D. A. Lidar, eprint quat-ph/0404077.
79. C. F. Lee, N. F. Jonson, F. J. Rodriguez, and L. Quiroga, *Fluctuation and Noise Letters* 2, L305 (2002).
80. N. H. Bonadeo, J. Erland, D. Gammon, D. Park, D. S. Katzer, and D. G. Steel, *Science* 282, 1473 (1998).
81. P. Borri, W. Langbein, J. M. Mork, J. M. Hwam, F. Heinrichsdorf, M.-H. Mao, and D. Bimberg, *Phys. Rev. B* 60, 7784 (1999).
82. B. Krummheuer, V. M. Axt, and T. Khun, *Phys. Rev. B* 65, 195313 (2002).
83. G. Chen, N. H. Bonadeo, D. G. Steel, D. Gammon, D. S. Katzer, D. Park, and L. J. Sham, *Science* 289, 1906 (2000).
84. T. H. Stievater, Xiaoqin Li, D. G. Steel, D. Gammon, D. S. Katzer, D. Park, C. Piermarocchi, and L. J. Sham, *Phys. Rev. Lett.* 87, 133603 (2001).
85. G. Chen, T. H. Stievater, E. T. Batteh, Xiaoqin Li, D. G. Steel, D. Gammon, D. S. Katzer, D. Park, and L. J. Sham, *Phys. Rev. Lett.* 88, 117901 (2002).
86. Xiaoqin Li, Yanwen Wu, D. Steel, D. Gammon, T. H. Stievater, D. S. Katzer, D. Park, C. Piermarocchi, and L. J. Sham, *Science* 301, 809 (2003).
87. A. Imamoglu, *Fortschr. Phys.* 48, 987 (2000).
88. M. Feng, I. D'Amico, P. Zanardi, and F. Rossi, *Phys. Rev. A* 67, 014306 (2003).
89. J. H. Reina, L. Quiroga, and N. F. Johnson, *Phys. Rev. A* 62, 012305 (2000).
90. S. De Rinaldis, I. D'Amico, E. Biolatti, R. Rinaldi, R. Cingolani, and F. Rossi, *Phys. Rev. B* 65, 081309(R) (2002).
91. B. W. Lovett, J. H. Reina, A. Nazir, B. Kothari, and G. A. D. Briggs, *Phys. Lett. B* 315, 136 (2003).
92. B. W. Lovett, J. H. Reina, A. Nazir, and G. A. D. Briggs, *Phys. Rev. B* 68, 2053 (2003).
93. J. M. Kikkawa and D. D. Awschalom, *Phys. Rev. Lett.* 80, 4313 (1998).
94. C. Piermarocchi, P. Chen, and L. J. Sham, *Phys. Rev. Lett.* 89, 167402 (2002).
95. E. Pazy, E. Biolatti, T. Calarco, I. D'Amico, P. Zanardi, F. Rossi, and P. Zoller, *Europhys. Lett.* 62, 175 (2003).
96. T. Calarco, A. Datta, P. Fedichev, E. Pazy, and P. Zoller, *Phys. Rev. A* 68, 012310 (2003).
97. A. Nazir, B. W. Lovett, S. D. Barrett, T. P. Spiller, and G. A. D. Briggs, *Phys. Rev. Lett.* 93, 150502 (2004).
98. P. Zhang, C. K. Chan, Q.-K. Xue, and X.-G. Zhao, *Phys. Rev. A* 67, 012312 (2003).
99. E. Biolatti, I. D'Amico, P. Zanardi, and F. Rossi, *Phys. Rev. B* 65, 075306 (2002).
100. S. Sangu, K. Kobayashi, A. Shojiguchi, and M. Ohtsu, *Phys. Rev. B* 69, 115334 (2004).
101. J. A. Gupta, R. Knobel, N. Samarth, and D. D. Awschalom, *Science* 292, 2458 (2001).
102. C. A. R. Sâ de Melo, *Phys. Rev. B* 51, 8922 (1995).
103. T. Pellizari, S. A. Gardiner, J. I. Cirac, and P. Zoller, *Phys. Rev. Lett.* 75, 3788 (1995).
104. A. V. Khaetskii, D. Loss, and L. Glazman, *Phys. Rev. Lett.* 88, 186802 (2002).
105. A. Imamoglu, E. Knill, L. Tian, and P. Zoller, *Phys. Rev. Lett.* 91, 017402 (2003).
106. A. Shabaev, Al. L. Efros, D. Gammon, and I. A. Merkulov, *Phys. Rev. B* 68, 201305 (2004).
107. N. Q. Huong and J. L. Birman, *Phys. Rev. B* 61, 13131 (2000).
108. X. Hu, T. Ritz, A. Damjanovic, F. Autenrieth, and K. Shulten, *Q. Rev. Biophys.* 35, 1 (2002).
109. S. K. Sekatskii, M. Chergui, and G. Dietler, *Europhys. Lett* 63, 21 (2003).
110. C. Hettich, C. Schmitt, J. Zitzmann, S. Kühn, I. Gerhardt, and V. Sandoghdar, *Science* 298, 385 (2002).

111. J. Volz, C. Kurtsiefer, and H. Weinfurther, *Appl. Phys. Lett.* 79, 869 (2001).
112. P. Walther, J.-W. Pan, M. Aspelmeyer, R. Ursin, S. Gasparoni, and A. Zeilinger, *Nature* 429, 158 (2004).
113. M. W. Mitchell, J. S. Lundeen, and A. M. Steinberg, *Nature* 429, 161 (2004).
114. A. Olaya-Castro, F. J. Rodriguez, L. Quiroga, and C. Tejedor, *Phys. Rev. Lett.* 87, 246403 (2001).
115. P. Michler, A. Imamoglu, M. D. Mason, P. J. Carson, G. F. Strouse, and S. L. Buratto, *Nature* 406, 968 (2000).
116. A. J. Berglund, A. C. Doherty, and H. Mabuchi, *Phys. Rev. Lett.* 89, 068101 (2002).

CHAPTER 17

Theoretical and Computational Models for Neuro, Genetic, and Neuro–Genetic Information Processing

Nikola Kasabov, Lubica Benuskova

Knowledge Engineering and Discovery Research Institute, School of Information Technology, Auckland University of Technology, Auckland, New Zealand

CONTENTS

1. INTRODUCTION

Neuroscience, along with the information and mathematical sciences, has developed a variety of theoretical and computational models to model complex brain functions [1]. Along with this development, artificial neural networks—computational models that adopt principles

ISBN: 1-58883-048-9

Handbook of Theoretical and Computational Nanotechnology
Edited by Michael Rieth and Wolfram Schommers
Volume 6: Pages (779–816)

from the nervous system—have been developed into powerful tools for learning from data and generalization [2–6]. Artificial neural networks have been applied to a large amount of complex problems such as classification, prediction, diagnosis, and planning, not only for brain and molecular data modeling [7–9], but across all disciplines of science and engineering [10]. This chapter is a review of existing computational models of brain functions and of neural network models.

With the recent advancements in genetic research and with the successful sequencing of the human and other genomes, more information is becoming available about the interaction between brain functions and genes, about genes related to brain diseases (e.g., epilepsy [11], mental retardation [12], etc.), and about gene-based treatment of them [13]. It is well accepted now that brain functions are better understood and treated if information from molecular and neuronal level is integrated. For this reason, computational models that combine genetic and neuronal information are needed for modeling and prognosis. Such models are called neuro-genetic models. These models are further discussed in the section on future development with the anticipation that models of integrated molecular and neuronal information processing will be an important part of the theoretical and computational nanotechnology of the future.

2. NEURO-INFORMATION PROCESSING

2.1. Neuro-Information Processing in the Brain

Many functions are associated with neural cells (neurons) and with neural networks in the brain [14]. An ensemble of neurons operates in concert, thus defining the functions of a neural network that lead to such outcomes as perception of sound or brain disease (e.g., epilepsy [11] or mental retardation [12], etc.). At the level of the whole brain, complex dynamic interactions are observed, and cognitive functions are performed (e.g., learning, visual pattern recognition, speech and language processing, etc.).

It is estimated that there are from 10^{11} to 10^{12} of neurons in the human brain [14]. Three quarters of these neurons form a 4–6-mm-thick cerebral cortex that constitutes a heavily wrinkled brain surface. The cerebral cortex is thought to be a seat of cognitive functions, like perception, imagery, thinking, and so on. The cortex cooperates with evolutionary, older subcortical nuclei that are located in the middle of the brain, in and around the so-called brainstem (Fig. 1). Subcortical structures and nuclei comprise, for instance, thalamus, basal ganglia, hypothalamus, and dozens of other groups of neurons with more or less specific functions in modulating the state of the whole brain. For example, the input from all sensory organs comes to the cortex preprocessed in the thalamus. All the brain parts, either cortical or subcortical, are directly or indirectly heavily interconnected, thus forming a huge recurrent neural network.

Figure 1 shows a schematic functional division of the human cerebral cortex. One-third of the cortex is devoted to the processing of visual information in the primary visual cortex and higher-order visual areas. Association cortices take about one-half of the whole cortical surface. In the parietal-temporal-occipital association cortex, sensory, and language data are being associated. Memory and emotional data are associated in the limbic association cortex (internal and bottom portion of hemispheres). The prefrontal association cortex takes care of all associations, evaluation, planning ahead, and attention.

At the border between the frontal and parietal lobes, there is a somatic sensory cortex, which processes touch and other tactile signals (temperature, pain, etc.) from the body surface and interior. In front of this cortex, there is a primary motor cortex, which issues signals for voluntary muscle movements including speech. These signals are preceded by the preparation and anticipation of movements that takes place in the premotor cortex. The plan of actions and their consequences, as well as the inclusion and exclusion of motor actions into and from the overall goal of an organism, are performed within the prefrontal association cortex. Subcortical basal ganglia and cerebellum also participate in preparation and tuning of motor outputs in the sense of particular movements. For instance, the cerebellum executes routine automatized movements like walking, biking, driving, and so on. Language processing

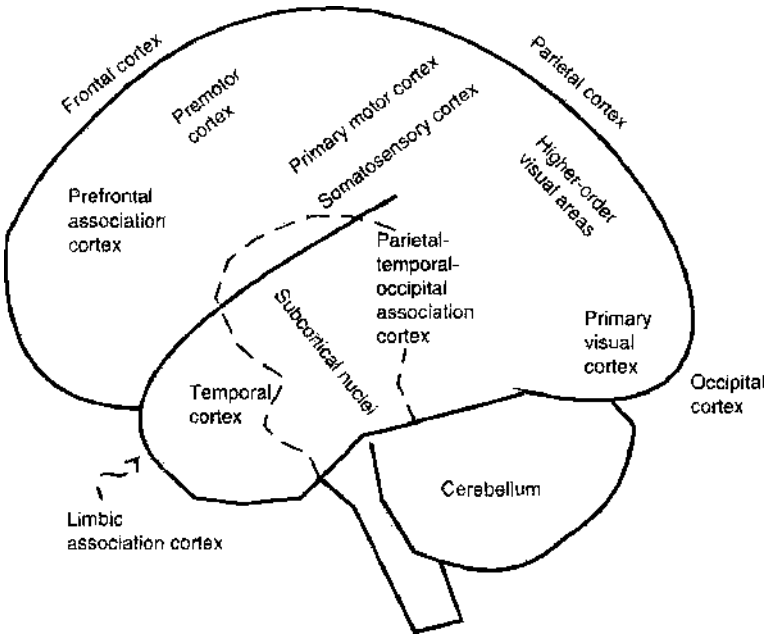


Figure 1. Gross anatomical and functional division of the human cerebral cortex. The same division applies for the other; in this case, the right hemisphere. A dashed curve marks the position of evolutionary older subcortical nuclei in the brainstem of the brain. Each of the depicted areas has far more subdivisions.

takes place within the temporal cortex, parietal-temporal-occipital association cortex, and frontal cortex. We want to point out that there are far more anatomical and functional subdivisions within each of the mentioned areas. Invaluable detailed information comes from the study of patients with mental deficits caused by injuries of particular brain areas. At present, noninvasive imaging techniques such as functional magnetic resonance (fMRI), positron emission tomography (PET), electroencephalogram (EEG), and others provide a rich source of information about the dynamics and organization of work within the healthy and diseased brain.

2.1.1. Processing of Signals by Neurons

Neuro-information processing in the brain depends not only on the properties of neural networks but also on the properties of processing units—neurons. A neuron (Fig. 2) receives and sends out electric and chemical signals. The place of signal transmission is a synapse. In the synapse, the signal can be nonlinearly strengthened or weakened. The strength or efficacy of synaptic transmission is also called a synaptic weight. One neuron receives and sends out signals through from 10^3 to 10^5 synapses. Dendrites (i.e., numerous bushy cell extensions) and soma constitute the input surface. Electrical signals transmitted by synapses can have a positive or negative electric sign. In the former case, we speak about excitatory synapses, and in the latter case about inhibitory synapses. Most of excitatory synapses are

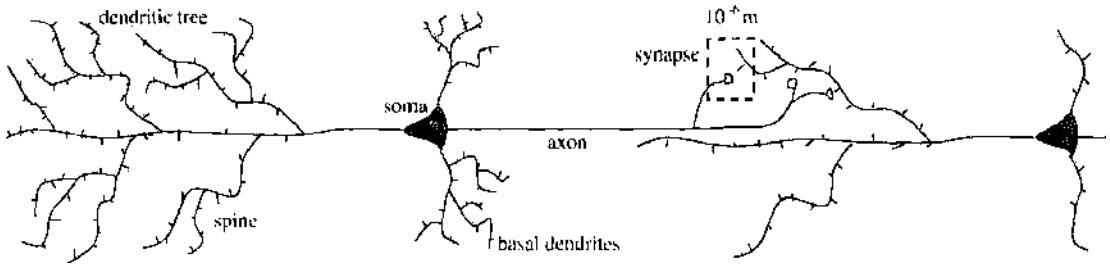


Figure 2. Schematic illustration of a neuron and its parts. There is a synapse at every dendritic spine. Synapses are also formed on the dendritic shafts and on the soma.

formed on spines, tiny extensions on dendrites. Spines are very important devices in relation to learning and memory.

When the sum of positive and negative contributions (signals) weighted by synaptic weights gets bigger than a particular value, called the excitatory threshold, a neuron fires; that is, it emits an output signal called a spike (Fig. 3). A spike is also called an action potential or a nerve impulse. The output frequency of a spike train (from 1 to 10^2 Hz) is proportional to the overall sum of positive and negative synaptic contributions. Spikes are produced at the initial segment of an axon (the only neuronal output extension). Then they quickly propagate along the axon toward other neurons within a network (propagation speed is 5–100 m/s). At its distant end, an axon makes thousands of branches, each of which is ended by a synaptic terminal (bouton).

2.1.2. Process of Synaptic Transmission

A synapse consists of a presynaptic terminal (bouton), synaptic cleft and postsynaptic membrane (Fig. 4). In the presynaptic terminal, there are dozens of vesicles filled with molecules of neurotransmitter (NT) ready to be released. When a presynaptic spike arrives into a terminal, calcium ions rush in and cause the fusion of vesicles with the presynaptic membrane. This process is also called exocytosis. Molecules of NT are released into the synaptic cleft (Fig. 4b) and diffuse toward the receptors within a postsynaptic membrane. Molecules of NT form a transient bond with the molecules of receptors. This causes the opening of ion channels associated with postsynaptic receptors. In the excitatory synapse, receptors are associated with sodium (Na^+) ion channels, and a positive excitatory postsynaptic potential (EPSP) is generated. In the inhibitory synapse, receptors are associated with chlorine (Cl^-) ion channels, and a negative inhibitory PSP is generated. Eventually, NT releases its bond with receptors and diffuses back to the presynaptic membrane and out of the synaptic cleft. Special molecular transporters within a presynaptic membrane take molecules of NT back inside the terminal, where they are recycled into new vesicles. This process is called a reuptake of NT. The whole synaptic transmission lasts for about 1 MS. Such a synapse is called a chemical synapse, because the transmission of an electric signal is performed in a chemical way.

The PSP, either excitatory or inhibitory, has some amplitude and duration. The amplitude and duration of the PSP depend on the number of activated receptor-ion channels and on how long they stay open, which may be miliseconds, tens of miliseconds, or hundreds of miliseconds. The duration of channel opening depends on the number of released NT molecules and on the type of receptors that are associated with ion channels. The amplitude of PSP also depends on the electric input resistance for ions, which in turn depends on the size and shape of a postsynaptic spine and dendrites and on the distance of synapse from soma. For instance, a short and stubby dendritic spine has a much smaller electric resistance than a long and thin spine. All these pre- and postsynaptic factors determine the weight (strength, efficacy) of a particular synapse.

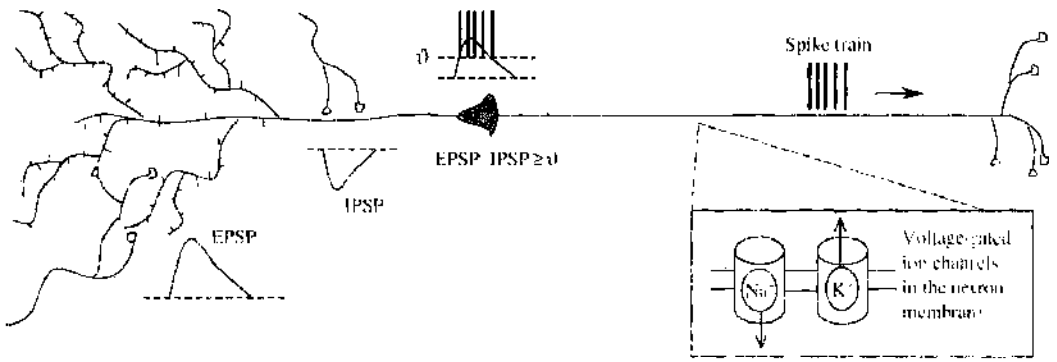


Figure 3. Electric synaptic potentials and axonal ion channels responsible for spike generation and propagation. EPSP = excitatory postsynaptic potential, IPSP = inhibitory postsynaptic potential, ϑ = excitatory threshold for an output spike generation.

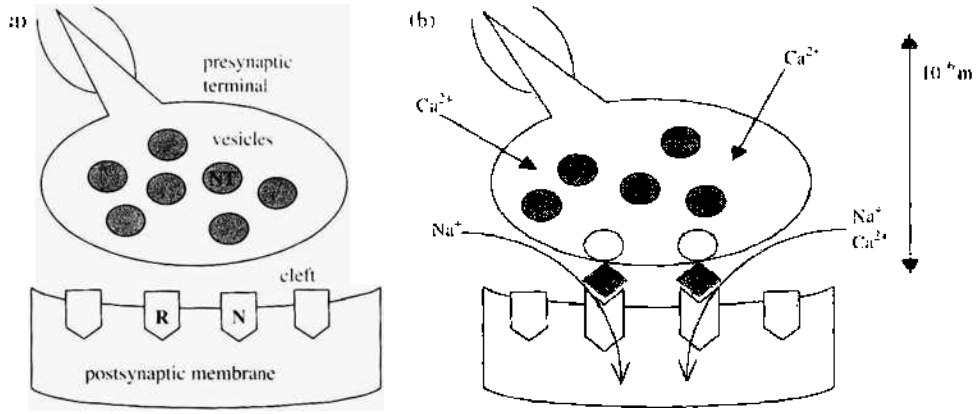


Figure 4. Scheme of synaptic transmission. (a) A synapse is ready to transmit a signal. (b) Transmission of electric signal in a chemical synapse. NT = neurotransmitter. R = AMPA-receptor-gated ion channel for sodium. N = NMDA-receptor-gated ion channel for sodium and calcium.

Within a postsynaptic membrane, there are also kinds of receptors that are not associated with an ion channel, but instead with an enzyme. When the overall amount of released NT reaches some critical concentration, these receptor–enzyme complexes activate particular cytoplasmatic enzymes, the so-called second messengers. Second messengers trigger chains of various biochemical reactions that may lead to a change in synaptic weight, or even to transient changes in gene expression leading to alteration in biomolecular synthesis of receptors, neurotransmitters, and enzymes. Thus, second messengers may act locally within a synapse itself, or they may activate further (third, and so on) messengers that carry the message to the genome of a neuron, thus causing a change in its biochemical machinery related to signal processing. Therefore, it is now widely accepted that the activity of a neuron itself influences its processing of information, and even its life itself, whether it survives or not.

2.1.3. Learning Takes Place in Synapses

For major discoveries in the field of synaptic mechanisms of learning, the 2000 Nobel prize for medicine went to the neuroscientists Eric R. Kandel and Paul Greengard. The third laureate, Arvid Carlsson, got the prize for discoveries of actions of the neurotransmitter dopamine. At present, it is widely accepted that learning is accompanied by changes in the synaptic weights in cortical neural networks [14]. Changes of synaptic weights are also called synaptic plasticity. In 1949, the Canadian psychologist Donald Hebb formulated a universal rule for these changes: "When an axon of cell A excites cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells so that A's efficiency as one of the cells firing B is increased" [15].

In cerebral cortex and in hippocampus of humans and animals, learning takes place in excitatory synapses formed on dendritic spines that use glutamate as their neurotransmitter. In the regime of learning, glutamate acts on specific postsynaptic receptors, the so-called NMDA receptors (*N*-methyl-D-aspartate). NMDA receptors are associated with ion channels for sodium and calcium (see Fig. 5). The influx of these ions into spines is proportional to the frequency of incoming presynaptic spikes. Calcium acts as a second messenger, thus triggering a cascade of biochemical reactions that lead either to the long-term potentiation of synaptic weights (LTP) or to the long-term depression (weakening) of synaptic weights (LTD). In experimental animals, it has been recorded that these changes in synaptic weights can last for hours, days, or even weeks and months, up to a year. Induction of such long-term synaptic changes involves transient changes in gene expression [16].

A subcellular switch between LTD and LTP is the concentration of calcium within spines [17]—an LTD/LTP threshold. In turn, the intraspine calcium concentration depends on the intensity of synaptic stimulation; that is, on the frequency of presynaptic spikes: more presynaptic spikes means more glutamate within synaptic cleft. Release of glutamate must coincide with a sufficient depolarization of the postsynaptic membrane to remove the magnesium

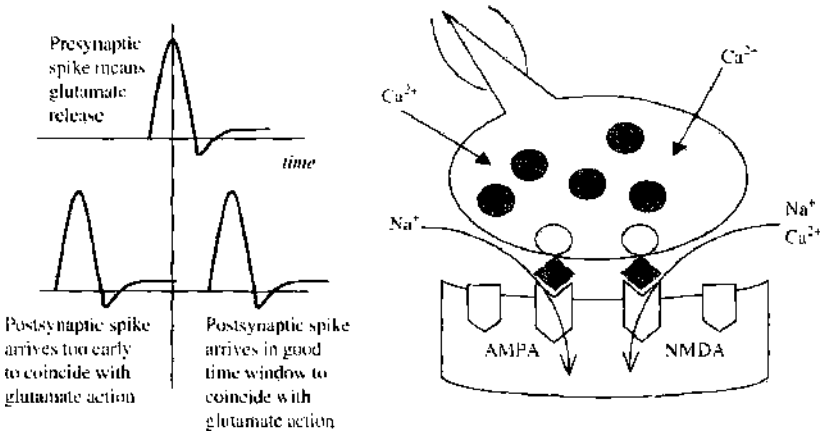


Figure 5. Spike timing dependent plasticity (STDP). Ions of sodium and calcium enter the postsynaptic spine through the NMDA-receptor gated ion channels. When postsynaptic neuron fires, the postsynaptic spikes back-propagate to spines. When their timing coincides with glutamate action, more calcium enters spines, and thus there is a bigger chance to achieve LTP.

block of the NMDA receptor. The greater the depolarization, the more ions of calcium enter the spine. Postsynaptic depolarization is primarily achieved via AMPA receptors; however, recently, a significant role of backpropagating postsynaptic spikes has been pointed out (see, e.g., Fig. 5) .[18]. Calcium concentrations below or above the LTD/LTP threshold switch on different enzymatic pathways that lead either to LTD or LTP, respectively. However, the current value of the LTD/LTP threshold (i.e., the properties of these two enzymatic pathways) can be influenced by levels of other neurotransmitters, by an average previous activity of a neuron, and possibly by other biochemical factors as well. This phenomenon is called metaplasticity, a plasticity of synaptic plasticity [19]. Dependence of the LTD/LTP threshold on different postsynaptic factors is the subject of the Bienenstock, Cooper, and Munro (BCM) theory of synaptic plasticity [20] (for a nice overview see, e.g., Ref. [21]). The BCM theory of synaptic plasticity has been successfully applied in computer simulations to explain experience-dependent changes in the normal and ultrastructrally altered brain cortex of experimental animals [22, 23].

The ease with which LTD and LTP can be evoked in the developing and the adult brain are not the same. One of the factors responsible for this difference may be the genetically programmed difference in NMDA receptor composition [24]. The NMDA receptor is made up of an NR1 subunit, which is obligatory for channel function, and a selection of developmentally and regionally regulated NR2 subunits (A–D). For example, the glutamate-evoked positive current has a longer duration in receptors containing NR2B subunits than in those containing NR2A subunits. The proportion of NR2B subunits is higher in young animals than in adults, which may account for the greater degree of synaptic plasticity seen in young animals.

2.1.4. Summary of Cortical Plasticity

Experimental developmental neuroscience brings abundant evidence that in the developing brain strong genetic programs determine the overall pattern of hierarchical organization and connections between brain areas. Nature provides an anatomical and physiological "scaffold" in the sense of a framework that outlines the structures to be formed it later. A newborn brain is not a "tabula rasa;" however, nurture and experience can shape it dramatically [14]. For instance, congenitally deaf kittens can develop almost normal auditory cortical maps and normal hearing after experiencing sounds through implanted artificial chochlea [25].

Information needed to specify precise subtle differentiation of neurons and subtle patterns of interneuronal connectivity far surpasses that which is contained in genetic programs. Genetic programs provide for a vast overproduction of abundant and redundant synaptic connections in the developing brain. Individual differences in experience cause selective

pruning of the majority of these synapses. Only synapses that mediate the genuine individual experience of an individual remain. Process of experience-dependent synaptic pruning during early stages of brain development may constitute the basis of brain and mind individuality. Early developmental overproduction of redundant synapses lasts only for some time after birth. Time windows, that is, beginnings and durations, are different for different brain systems. They also differ for different animal species. In general, what lasts weeks and months in rats, cats, and monkeys usually lasts for years in humans.

Later in adolescence and adulthood, a new experience is "burned" into the brain not by a selective pruning of existing redundant synapses but, instead, by a selective creation of new connections and by changing the efficacies of the synaptic transmission of existing connections. This does not mean that synapses cannot be removed as a result of experience later in life, or that new connections cannot be created because of experience early in development. They can, but the prevailing process of experience-dependent cortical plasticity at different ages is not the same. Some cortical areas retain the capacity of synaptic plasticity for their whole course of life. These are association cortical areas, highest-order sensory areas, and premotor and emotional cortical areas of the brain. The capacity of the brain to be plastic and change its microstructure as a result of experience is of profound importance for discovering the rules of the mind/brain relation.

2.2. Coding and Representation of Information in the Brain

2.2.1. Principles of Information Representation in the Brain

The first principle of representation (and coding) of information in the brain is redundancy. Redundancy means that every piece of information (meant in any sense) is stored, transmitted, and processed by a redundant number of neurons and synapses so that it does not become lost when neural networks undergo damage; for instance, as a result of aging. When neural networks get damaged, their performance does not drop down to zero abruptly, like in a sequential computer, but instead it degrades gracefully. Computer models of neural networks also confirm the idea that a degradation of performance with the loss of neurons and synapses is neither total nor linear, but instead, neural networks can withstand quite substantial damage and still perform well.

Next, the contemporary view on the nature of neural representation is such that information (in the sense of content or meaning) is represented by place in the cortex (or, in general, in the brain). However, this placement is a result of both genetically determined anatomical framework and shaping by input through the process of experience-dependent plasticity. For instance, a sound pattern for the word "apple" is represented (coded) in the auditory areas of the temporal cortex. It is represented as a spatial pattern of active versus inactive neurons. This neural representation is associated (connected) through synaptic weights, with the neural representation of a visual image of apple in the parietal cortex, with the neural representation of an apple odour in the olfactory cortex, with memories on the grandma garden and facts about apples being represented in some other areas of the brain, and so forth. Neural representations (i.e., distributions or patterns of active neurons) within particular areas and their associations between areas appear as a result of learning (synaptic plasticity). Different objects are represented by means of different patterns or by distributions of active neurons within cortical areas. Thus, each object is represented by means of distributed networks of neurons, whereas the activity of neurons in these networks is also distributed.

Perception is an active process. Instead of a passive processing of all electrical signals that may or may not arrive from hierarchically lower processing levels, cortical neural networks should be able to use fragments of activity patterns to fill in the gaps, and thus quickly re-create the whole neural representation. The filling-in process can be nicely modeled by means of model neural networks; for instance, those of the Hopfield type [26] (Fig. 6).

Explicit neural representations (patterns of activities) are implicitly stored in the matrix of synaptic weights through which neurons in the network are interconnected. The weight distribution of storing a particular object representation is created as a result of an experience-dependent synaptic plasticity (learning). When a sufficiently large portion of this

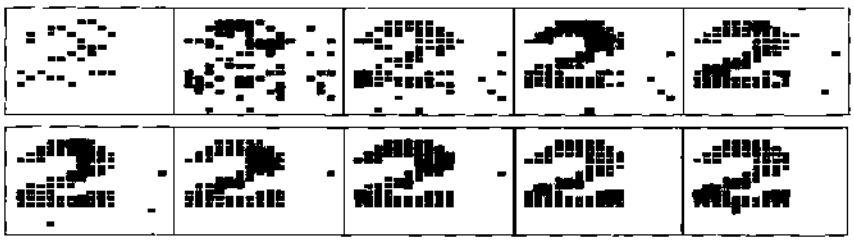


Figure 6. Illustration of spontaneous re-creation of neural representation after only few input impulses (figure in the uppermost left corner). Black pixel represents a firing neuron, whereas blank pixel represents a silent neuron. Between each pattern of activity from left to right (1 ms time frame), neurons in the network exchange only one impulse. Thus, basically, after exchanging only two to three spikes, the memory pattern is re-created. Network can reverberate the restored memory pattern until a different external input arrives.

neural representation is activated from outside the network, few electric signals along all the synapses in the network quickly switch on the correct remaining neurons in the representation. Neural representations in the sense of patterns of activity have a holistic character. Patterns of activity are being recalled (restored) as a whole.

2.2.2. Problem of Neural Coding: The Brain is Fast, Neurons are Slow

Neurons within and between different brain areas send messages to each other by means of output spikes. Transmission time at one synapse takes about 1 ms. Neural representations of objects "communicate with each other;" that is, neurons within and between these representations send messages to each other. At present, the nature of these messages, that is, the nature of a neural code, is a mystery. It is because the recognition of sensory objects takes only about 200 ms [27], although it is performed by tens of different cortical areas in which the information processing involves billions of neurons. Even when taking into account the parallel processing of information, one hierarchical area of the cortex is left with only about 10–30 ms to perform its share of processing, the results of which are sent to hierarchically higher areas. Therefore, recently, a number of hypotheses have emerged based both on theoretical and experimental investigations, dealing with this problem.

These hypotheses can be divided into two categories: spike-timing hypotheses and rate code hypotheses [28].

2.2.2.1. Coding Based on Spike Timing

2.2.2.1.1. Reverse Correlation The first option is that the information about the salience of the object feature is encoded in the exact temporal structure of the output spike train. Let us say that two neurons fire three spikes within 20 ms. The first neuron fires a spike train with this temporal structure | ||, and the second neuron with this temporal structure |||. By means of the techniques of reverse correlation, it is possible to calculate which stimulus exclusively causes which temporal pattern of which neuron. The main proponents of this theory are Bialek and his coworkers, who have made its successful verification in the fly visual system [29].

2.2.2.1.2. Time to the First Spike Let, at time instant t_0, a stimulus arrive to the neural network. Neurons that fire the first (let us say in a window of 10 ms) carry the information about the stimulus features. The rest of neurons and the rest of impulses are ignored. This theory is favored by S. Thorpe [30].

2.2.2.1.3. Phase Information about the presence of the feature is encoded in the phase of neuron's impulses with respect to the reference background oscillation. Either they are in a phase lead or in a phase lag. The information can also depend on the magnitude of this phase lead (lag). This coding is preferred by people investigating hippocampus [31].

2.2.2.1.4. Synchronization Populations of neurons that represent features belonging to one object can be bound together by synchronous firing. Such a binding of features was discovered in the laboratory of W. Singer in the cat visual cortex [32]. It was also detected in the human cortex during perception [27].

2.2.2.2. The Rate Code

2.2.2.2.1. (Temporal) Average Rate This coding is still being considered for stationary stimuli that last up to around 500 ms or longer, so neurons have enough time to count (integrate) impulses over long time. Neurons that have the highest frequency signalize the presence of the relevant feature.

2.2.2.2.2. Rate as a Population Average An average frequency is not calculated as a temporal average but, rather, as a population average. One feature is represented by a population of many (10,000) neurons, for instance, in one cortical column. On the presence of a feature, most of the neurons are activated. When we calculate the number of spikes in a 10-ms window of all these neurons and divide this number by the number of neurons activated, we will get approximately the same average frequency as when calculating a temporal average rate of any of these neurons (provided they all fire with the same average rate). This idea has been thoroughly investigated in Ref. [33]. It has been also shown, that by means of population averaging we can get a reliable calculation of the neurons' average rates even in the case when they have a Poisson-like distribution of output spikes. Populations that relay the highest number of spikes signal the presence of the relevant feature.

On seeing all these options, one gets inevitably confused. Every hypothesis has some experimental support; thus, which one is correct, and which one is not? Is it possible that different brain areas use different codes? In one area it is the spike-timing code and in another area it is a rate code. Or does the brain switch between different codes according to the task to be performed? These possibilities remain to be explored.

Let us reason now, However, that all these hypotheses do not have to be, in fact, mutually exclusive. For instance, let us imagine that a population of neurons representing one feature of an object has a synchronized firing. Those neurons that have the highest output rates have also the shortest time to the first spike. Their synchronization guarantees that all of them have the same phase difference with respect to the background oscillation. Actually, the background oscillation might have helped to synchronize them. The information about the intensity of a feature in the stimulus (feature salience) is encoded in the average frequency of the whole population of neurons (within 10 ms) and is relayed on to the next population of neurons in the processing hierarchy. An actual temporal pattern of spikes relayed by one population of neurons indeed corresponds only to one stimulus. Thus, maybe the mentioned options are only different angles under which we can see the same process.

3. ARTIFICIAL NEURAL NETWORKS

Artificial neural networks (ANNs) are massively parallel computational systems inspired by biological neural networks. They can have different architectures and properties of their processing elements. Illustration of a general architecture of an ANN is shown in Fig. 7.

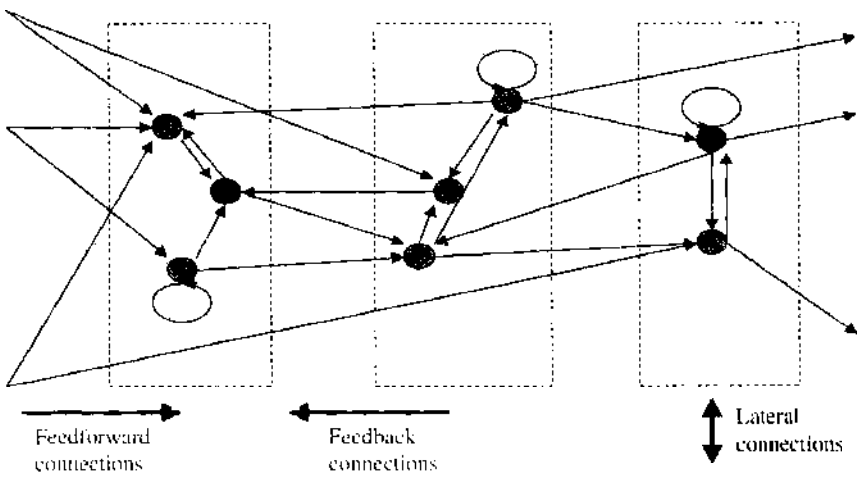


Figure 7. Formal neurons in ANNs are organized into layers (dashed rectangles). Connections between neurons within an ANN can be of the feed-forward, feedback, or lateral type.

There are different variations derived from this general architecture; for instance, there can be only feed forward connections between layers, or there can be only feedback connections in the one and only layer with the feed forward input, or there can be feedback connections allowed only between some layers, and so on.

In general, ANNs perform mappings of vectors from m-dimensional input space on vectors from n-dimensional output space. We can also say that ANNs learn how to associate vectors from one space to vectors from another space. Firts of all, there are many models of ANNs, each having different particular properties. However, they all are adaptible systems that can reorganize their internal structure based on their experience in the process called training or learning. ANNs are very often referred to as connectionist systems [34].

Basic processing elements (processing units) of ANNs are formal neurons based on the rate code hypothesis of a neural code according to which the information being sent from one neuron to another is encoded in the output rate of its spike train. Thus, the input–output function of an ith element of an ANN is

$$o_i(t) = f\left[\sum_{\forall k} w_{ik}(t)\, x_k(t)\right] \tag{1}$$

where $o_i, x_k, w_{ik} \in \Re$ are the output rate of the ith neuron, input rate of the kth input, and the synaptic weight between the kth and ith element, respectively. The function f is the so-called activation or transfer function of a neuron. It can be linear or nonlinear, continuous and differentiable, or binary, depending on which ANN model we are working with.

Different models of ANNs differ with respect to their architecture, transfer functions of their elements, and the rules used to modify the weights between neurons in the process of learning [5].

3.1. General Classification Scheme

Let us introduce a general classification scheme of ANNs that will lead to explanation of their properties, capabillities, and drawbacks [34].

3.1.1. General Notions

Most of the known ANN training algorithms are influenced by a concept introduced by Donald Hebb [15]. He proposed a model for unsupervised learning in which the synaptic strength (weight) is increased if both the source and the destination neurons become simultaneously activated. It is expressed as

$$w_{ij}(t+1) = w_{ij}(t) + c\, o_i(t)\, o_j(t) \tag{2}$$

where $w_{ij}(t)$ is the weight of the connection between the ith and jth neuron in the network at the moment t, and o_i and o_j are the output signals of the neurons i and j at the same moment t. The weight $w_{ij}(t+1)$ is the adjusted weight at the next time time moment $(t+1)$. Usually some kind of weights normalization is applied after each adjustment to prevent growth to infinity.

In general terms, a connectionist system $\{S, W, P, F, J, L\}$ that is defined by its structure S, its connection weights W, its parameter set P, its function F, its goal function J, and a learning procedure L, learns if the system optimizes at least part of its structure S and its function F when observing events $z1, z2, z3, \ldots$ from the problem space Z.

Through a learning process, the system improves its reaction to the observed events and captures useful information that may be later represented as knowledge.

The goal of a learning system is defined as to find the minimum of an objective function $J(S)$ named "the expected risk function" [2, 6]. The function $J(S)$ can be represented by a loss function $Q(Z, S)$ and an unknown probability distribution $\Pi(Z)$.

Most of the learning systems optimize a global goal function over a fixed part of the structure of the system. In NN, this part is a set of predefined and fixed number of connection weights (i.e., the set number of elements in the set W). As an optimization procedure, some known statistical methods for global optimization are applied [6]: for example, the gradient

descent method. Final structure S is expected to be globally optimal (i.e., optimal for data drawn from the whole-problem space Z).

In the case of a changing structure S and changing (e.g., growing) part of its connections W, where the input stream of data is continuous and its distribution is unknown, the goal function could be expressed as a sum of local goal functions J', each one optimized in a small subspace of $Z' \subset Z$ as data are drawn from this subspace. While the learning process is taking place, the number of dimensions of the problem space Z may also change over time. These scenarios would be reflected in different models of learning, as is explained next.

3.1.2. Different Models of Learning in Connectionist Systems

There are many methods for learning that have been developed for connectionist architectures (for a review, see Ref. [1]). It is difficult and quite risky to try to put all the existing methods into a clear classification structure (which should also assume "slots" for new methods), but it is necessary to define the scope of the evolving connectionist systems paradigm.

A connectionist classification scheme is graphically represented in Table 1, and is explained below. On the one hand, this scheme is a general one, as it is valid not only for connectionist learning models but also for other learning paradigms; for example, evolutionary learning, case-based learning, analogy-based learning, and reasoning. On the other hand, the scheme is not a comprehensive one, as it does not present all existing connectionist learning models. It is only a working classification scheme needed for the purpose of this work.

A (connectionist) system that learns from observations $z1, z2, z3, \ldots$ from a problem space Z can be designed to perform learning in different ways. The classification scheme in Table 1 outlines the main questions and issues and their alternative solutions when constructing a connectionist learning system. Now, let us offer an explanation of individual issues and alternatives.

1. *What space has the learning system developed in?*
   a. *The learning system has developed in the original problem space Z*: The structural elements (nodes) of the connectionist learning system are points in the d-dimensional original data space Z. This is the case in some clustering and prototype learning systems. One of the problems here is that if the original space is high dimensional (e.g., a 30,000 gene expression space) it is difficult to visualize the structure of the system and observe some important patterns. For this purpose, special visualization techniques, such as principle component analysis (PCA), or Sammon mapping, are used to project the system structure S into a visualization space V.
   b. *The learning system has developed in its own machine space M*: The structural elements (nodes) of the connectionist learning system are created in a system (machine) space M, different from the d-dimensional original data space Z. An example is the self-organizing map (SOM) neural network [35]. SOMs develop in a two-, three-, or more dimensional topological spaces (maps) from the original data.
2. *Is the space open?*
   a. An open problem space is characterized by unknown probability distribution $P(Z)$ of the incoming data and a possible change in its dimensionality. Sometimes the dimensionality of the data space may change over time, involving more, or fewer, dimensions; for example, adding new modalities to a person identification system.
   b. A closed problem space has a fixed dimensionality and either a known distribution of the data or a distribution that can be approximated in advance through statistical procedures.
3. *Is learning online?*
   a. *Batch-mode, offline learning*: In this case, a predefined learning (training) set of data $P = \{z1, z2, \ldots, zp\}$ is learned by the system through propagating this data set several times through the system. Each time, the system optimizes its structure W, based on the average value of the goal function over the whole data set P. Many traditional algorithms, such as the back-propagation algorithm, use this type of learning [36, 37].

Table 1. A general classification scheme of ANN models.

Issue	Alternatives
What space has the learning system developed in?	(a) The learning system has developed in the original data space Z (b) The learning system has developed in its own machine space M
Is the space open?	(a) An open data space (b) A closed space that has a fixed dimensionality
Is learning online?	(a) Batch-mode, offline learning (b) Online, pattern mode, incremental learning (c) Combined online and offline learning
Is learning lifelong?	(a) Single-session learning (b) Lifelong learning
Are there desired output data and in what form?	(a) Unsupervised learning (b) Supervised learning (c) Reinforcement learning (d) Combined learning
Is evolution of populations of individuals over generations involved in the learning process?	(a) Individual, development-based learning (b) Evolutionary learning, (population-based learning over generations) (c) Combined
Is the structure of the learning system of a fixed size, or it is evolving?	(a) Fixed-size structure (b) Dynamically changing structure
How structural modifications in the learning system partition the problem space?	(a) Global partitioning (global learning) (b) Local partitioning (local learning)
What knowledge representation is facilitated in the learning system?	(a) No explicit knowledge representation is facilitated in the system (b) Memory-based knowledge (c) Statistical knowledge (d) Analytical knowledge (e) Symbolic knowledge (f) Combined knowledge (g) Meta-knowledge (h) "Consciousness" (the system knows about itself) (i) "Creativity" (e.g., generating new knowledge)
If symbolic knowledge is represented in the system, of what type is this knowledge?	(a) Propositional rules (b) First-order logic rules (c) Fuzzy rules (d) Semantic maps (e) Schemata (f) Meta-rules (g) Finite automata (h) Higher-order logic
If the systems' knowledge can be represented as fuzzy rules, what type of fuzzy rules are they?	(a) Zadeh-Mamdani fuzzy rules (b) Takagi-Sugeno fuzzy rules (c) Other types of fuzzy rules (e.g., type 2 fuzzy rules)
Is learning active?	(a) Active learning in terms of data selection, filtering and searching for relevant data (b) Passive learning—the system accepts all incoming data

b. *Online, pattern mode, incremental learning*: Online learning is concerned with learning each data example separately as the system operates (usually in a real time), and the data might exist only for a short time. After observing each data example, the system makes changes in its structure (the W parameters) to optimize the goal function J. A typical scenario for online learning is when data examples are drawn randomly from a problem space and fed into the system one by one for training. Although there are chances of drawing the same examples twice or several times, this

is considered as a special case, in contrast to the offline learning, when one example is presented to the system many times as part of the training procedure. Methods for online learning in NN are studied in Refs. [38–40]. In Ref. [41], a review of some statistical methods for online learning, mainly gradient descent methods applied on fixed-size connectionist structures, is presented.

Some other types of learning, such as incremental learning and lifelong learning, are closely related to online learning.

Incremental learning is the ability of a NN to learn new data without fully destroying the patterns learned from old data and without the need to be trained on either old or new data. According to Schaal and Atkeson [42], incremental learning is characterized by the following features:

- Input and output distributions of data are not known, and these distributions may change over time
- The structure of the learning system W is updated incrementally
- Only limited memory is available, so that data have to be discarded after they have been used.

Online learning, incremental learning, and lifelong learning are typical adaptive learning methods. Adaptive learning is aiming at solving the well-known stability/plasticity dilemma, which means that the system is stable enough to retain patterns learned from previously observed data while being flexible enough to learn new patterns from new incoming data.

Adaptive learning is typical for many biological systems and is also useful in engineering applications, such as robotic systems and process control. Significant progress in adaptive learning has been achieved as a result of the adaptive resonance theory (ART) [4] and its various models, which include unsupervised models (ART1–3, FuzzyART) and supervised versions (ARTMAP, FuzzyARTMAP-FAM) [43].

c. *Combined online and offline learning*: In this mode the system may work for some of the time in an online mode, after which it switches to offline mode, and so forth. This is often used for optimization purposes, when a small "window" of data from the continuous input stream can be kept aside, and the learning system that works in an online mode can be locally or globally optimized through offline learning on this window of data through "window-based" optimization of the goal function $J(W)$.

4. *Is the learning process lifelong?*
   a. *Single-session learning*: The learning process happens only once over the whole set P of available data (and it may even take many iterations during training). After that, the system is set in operation and never trained again. This is the most common learning mode in many existing connectionist methods.
   b. *Lifelong learning*: Lifelong learning is concerned with the ability of a system to learn from continuously incoming data in a changing environment during its entire existence. Growing as well as pruning may be involved in the lifelong learning process, as the system needs to restrict its growth while always maintaining a good learning and generalization ability.

5. *Are there desired output data, and in what form are they available?* The availability of examples with desired output data (labels) that can be used for comparison with what the learning system produces on its outputs defines four types of learning.
   a. *Unsupervised learning*: There are no desired output data attached to the examples $z1, z2, z3, \ldots$. The data are considered as coming from an input space Z only.
   b. *Supervised learning*: There are desired output data attached to the examples $z1, z2, z3, \ldots$. The data are considered as coming in (x, y) pairs from both an input space X and an output space Y that collectively define the problem space Z (Fig. 8). The connectionist learning system associates data from the input space X with data from the output space Y.

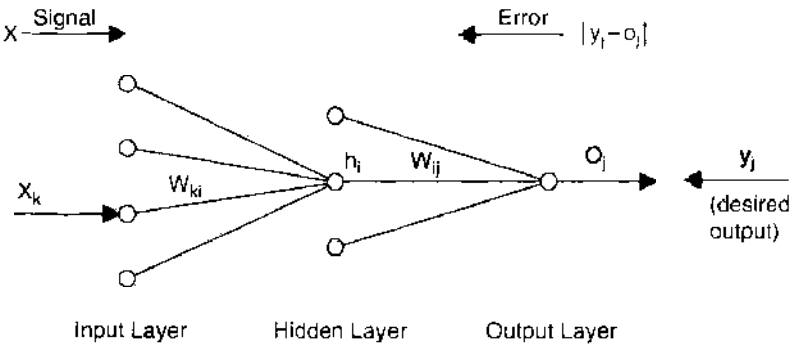


Figure 8. Illustration of a multilayer perceptron and supervised learning by means of the error back-propagation algorithm. Back-propagated error is proportional to the difference between the desired and actual output.

c. *Reinforcement learning*: In this case, there are no exact desired output data, but some hints about the "goodness" of the system reaction are available. The system learns and adjusts its structural parameters from these hints. In many robotic systems, a robot learns from the feedback from the environment that may be used as a qualitative indication of the correct movement of the robot.
d. *Combined learning*: This is the case in which a connectionist system can operate in more than one of the above learning modes.

6. *Is evolution of populations of individuals over generations involved in the learning process?*
   a. *Individual, development-based learning*: A system is developed independently and is not part of a development process of a population of individual systems.
   b. *Evolutionary learning population-based learning over generations*: Here, learning is concerned with the performance of not only an individual system but also, a population of systems that improve their performance through generations. The best individual system is expected to emerge—to evolve from such populations.

      Evolutionary computation (EC) methods, such as genetic algorithms (GA), have been widely used for optimizing ANN structures [44–46]. Such ANNs are called evolutionary neural networks. They use ideas from Darwinism. Most of the evolutionary computation methods developed so far assume that the problem space is fixed; that is, the evolution takes place within a predefined problem space, and this space does not change dynamically. Therefore, these methods do not allow for modeling real, online adaptation. In addition, they are very time consuming, which also prevents them from being used in real-world applications.

7. *Is the structure of the learning system of a fixed size, or it is evolving?*

   Here we will refer again to the bias/variance dilemma (see, e.g., Refs. [4, 47, 48]). With respect to an ANN structure, the dilemma states that if the structure is too small, the ANN is biased to certain patterns, and if the NN structure is too large, there are too many variances that may result in overtraining, poor generalization, and so forth. To avoid this problem, an ANN structure should change dynamically during the learning process, thus better representing the patterns in the data and the changes in the environment.
   a. *Fixed-size structure*: This type of learning assumes that the size of the structure S is fixed (e.g., number of neurons, number of connections) and that through learning, the system changes some structural parameters (e.g., W—the values of connection weights). This is the case in many multilayer perceptron ANNs trained with the back-propagation algorithm [2, 6, 36, 37, 49–51].
   b. *Dynamically changing structures*: According to Heskes and Kappen [52], there are three different approaches to dynamically changing structures: constructivism, selectivism, and a hybrid approach.

      Connectionist constructivism is about developing ANNs that have a simple initial structure and that grow during its operation through inserting new nodes. This theory is supported by biological facts (see Ref. [40]). The insertion can be controlled

either by a similarity measure of input vectors or by the output error measure, or by both, depending on whether the system performs an unsupervised or supervised mode of learning. A measure of difference between an input pattern and already stored patterns is used for deciding whether to insert new nodes in the adaptive resonance theory models ART1 and ART2 [4] for unsupervised learning. There are other methods that insert nodes based on the evaluation of the local error. Such methods are growing cell structure and growing neural gas [39]. Other methods insert nodes based on a global error to evaluate the performance of the whole ANN. One such method is the Cascade–Correlation Method [53]. Methods that use both similarity and output error for node insertion are used in Fuzzy ARTMAP [43] and also in EfuNN (Evolving Fuzzy NN) [34].

Connectionist selectivism is concerned with pruning unnecessary connections in a NN, which starts its learning with many, in most cases redundant, connections [54, 55]. Pruning connections that do not contribute to the performance of the system can be done by using several methods: Optimal-Brain Damage [56], Optimal Brain Surgeon [41], and Structural Learning with Forgetting [57].

8. *How do structural modifications affect the partitioning of the problem space?* When a connectionist model is created, either in a supervised or in an unsupervised mode, the nodes and the connections partition the problem space Z into segments. Each segment of the input sub-space is mapped onto a segment from the output sub-space in case of a supervised learning. The partitioning in the input sub-space imposed by the model can be one of the following two types:
   a. *Global partitioning (global learning)*: Learning causes global partitioning of the space. Partitioning hyperplanes can be modified either after every example is presented (in the case of online learning) or after all examples are presented for one iteration (in the case of a batch mode learning).

      Through the gradient descent learning algorithm, the problem space is partitioned globally. This is one of the reasons why global learning in multilayer perceptrons suffers from the catastrophic forgetting phenomenon [58, 59]. Catastrophic forgetting (also called unlearning) is the inability of the system to learn new patterns without forgetting previously learned patterns. Methods to deal with this problem include rehearsing of the NN on a selection of past data or on generated new data points from the problem space [58].

      Other techniques that use global partitioning are support vector machines (SVMs) (see Ref. [60]) for a comparative study of ANN, fuzzy systems, and SVM). Through learning, the SVMs optimize the positioning of the hyperplanes to achieve maximum distance from all data items on both sides of the plane.
   b. *Local partitioning (local learning)*: In the case of a local learning, the structural modifications of the system affect the partitioning of only a small part of the space in which the current data example is drawn from. Examples are given in Fig. 9a and 9b, where the space is partitioned by circles and squares into a two-dimensional space. Each circle or square is the subspace defined by a neuron. The activation of each neuron is defined by local functions imposed on its subspace. An example of such local functions is kernels, as shown in Fig. 9a. Kernels K are defined by formulas as given below

$$K(x) = \frac{\exp(-x^2/2)}{\sqrt{2\pi}} \quad \text{while } \sum K(x) = 1 \quad \text{for all } x \in Z \tag{3}$$

Other examples of local partitioning are when the space is partitioned by hypercubes and fractals in a three-dimensional space.

Before creating a model, it is important to choose which type of partitioning will be more suitable for the task in hand. In the evolving connectionist systems presented later, the partitioning is local. Local partitioning is easier to adapt in an online mode, faster to calculate, and does not cause catastrophic forgetting.

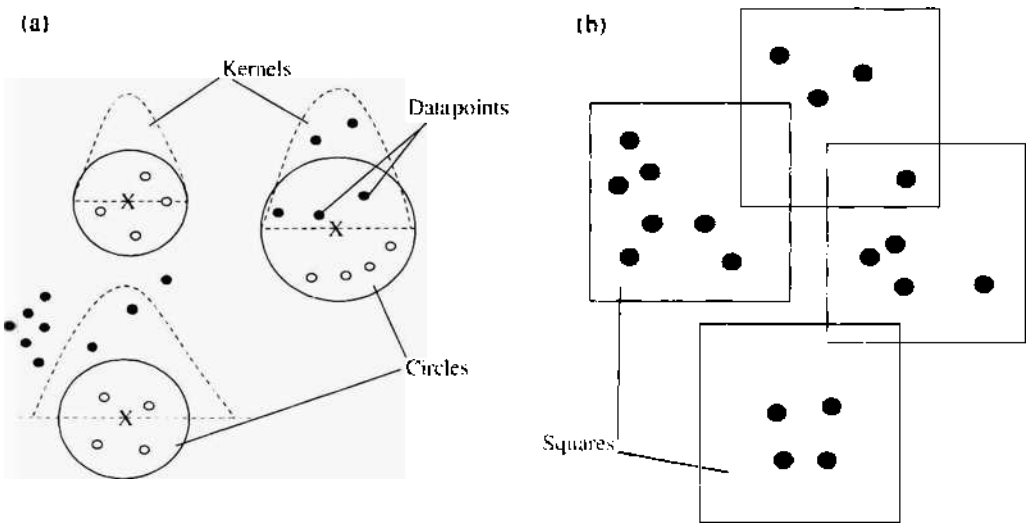


Figure 9. Partitioning of the problem space Z by (a) circles and (b) squares into a two-dimensional space. Reprinted with permission from [34], N. Kasabov, "Evolving Connectionist Systems. Methods and Applications in Bioinformatics, Brain Study and Intelligent Machines." Springer, London, 2003. © 2003, Springer-Verlag.

9. *What knowledge representation is facilitated in the learning system?* It is a well-known fact that one of the most important characteristics of the brain is that it can retain and build knowledge. However, it is not known yet how exactly the activities of the neurons in the brain are transferred into knowledge.

   For the purpose of the discussion in this chapter, knowledge can be defined as the information learned by a system that the system can interpret in different ways and can use in inference procedures to obtain new facts and new knowledge.

   Traditional ANNs and connectionist systems have been known to be poor facilitators of representing and processing knowledge, despite some early investigations [61, 62]. However, some of the issues of knowledge representation in connectionist systems have already been addressed in the so-called knowledge-based neural networks (KBNNs) [63–65]. KBNNs are ANNs that are prestructured in a way that allows for data and knowledge manipulation, which includes learning, knowledge insertion, knowledge extraction, adaptation, and reasoning. KBNNs have been developed either as a combination of symbolic artificial intelligence (AI) systems and NN [66], or as a combination of fuzzy logic systems and NN [67–71]. Rule insertion and rule extraction operations are examples of how a KBNN can accommodate existing knowledge along with data, and how it can "explain" what it has learned. There are different methods for rule extraction that are applied to practical problems [72–77].

   Generally speaking, learning systems can be distinguished on the basis of the type of knowledge they represent.

   a. *No explicit knowledge representation is facilitated in the system*: An example for such connectionist system is the traditional multilayer perceptron network trained with the backpropagation algorithm [36, 37, 49, 51, 78].
   b. *Memory-based knowledge*: The system retains examples, patterns, prototypes, and cases; for example instance-based learning [79], case-based reasoning systems [80], and exemplar-based reasoning systems [81].
   c. *Statistical knowledge*: The system captures conditional probabilities, probability distribution, clusters, correlation, principal components, and other statistical parameters [5].
   d. *Analytical knowledge*: The system learns an analytical function f: X? Y, that represents the mapping of the input space X into the output space Y. Regression techniques and kernel regressions in particular are well established [5, 82].
   e. *Symbolic knowledge*: Through learning, the system associates information with predefined symbols. Different types of symbolic knowledge can be facilitated in a learning system, as discussed further below.

f. *Combined knowledge*: The system facilitates learning of several types of knowledge.
g. *Metaknowledge*: The system learns hierarchical level of knowledge representation in which metaknowledge is also learned; for example, which piece of knowledge is applicable at what time.
h. *"Consciousness" of a system*: The system becomes "aware" of what it is, what it can do, and where its position among the rest of the systems in the problem space is.
i. *"Creativity" of a system*: An ultimate type of knowledge would be such knowledge that allows the system to act creatively, to create scenarios, and possibly to reproduce itself; for example, a system that generates other systems (programs) and improves in time based on its performance in the past.

10. *What type of symbolic knowledge is facilitated by the system?* If we can represent the knowledge learned in a learning system as symbols, different types of symbolic knowledge can be distinguished.

    - Propositional rules
    - First-order logic rules
    - Fuzzy rules
    - Semantic maps
    - Schemata
    - Meta-rules
    - Finite automata
    - Higher-order logic

11. *If the systems' knowledge can be represented as fuzzy rules, what types of fuzzy rules are facilitated by the system?* Different types of fuzzy rules can be used; for example, Zadeh-Mamdani fuzzy rules [83, 84], Takagi-Sugeno fuzzy rules [85], and other types of fuzzy rules (e.g., type-2 fuzzy rules; for a comprehensive reading, see Ref. [86]).

    Generally speaking, different types of knowledge can be learned from a process or from an object in different ways, all involving the human participation. These ways include direct learning by humans, simple problem representation as graphs, analytical formulas, using ANN for learning and rule extraction, and so forth these forms can be viewed as alternative and possibly equivalent forms in terms of final results obtained after a reasoning mechanism is applied on them. Elaborating analytical knowledge in a changing environment is a very difficult process involving changing parameters and formulas with the changing data. If evolving processes are to be learned in a system and also understood by humans, neural networks that are trained in an online mode and their structure are interpreted as knowledge are now the most promising models.

12. *Is the learning process active?* Humans and animals are selective in terms of processing only important information. They are searching actively for new information [87, 88]. Similarly, we can have two types of learning in an intelligent system: active learning in terms of data selection and filtering, and searching for relevant data, and passive learning, when the system accepts all incoming data.

3.1.3. Brief Overview of Major Problems With Existing Learning Systems

Despite the successfully developed and used ANNs, fuzzy systems, GAs, hybrid systems, and other methods and techniques for adaptive training, there are a number of problems with them. The main problems are listed below.

a. *Difficulty in preselecting the system architecture*: Usually an ANN model has a fixed architecture (e.g., a fixed number and organization of neurons and connections). This makes it difficult for the system to adapt to new data of unknown distribution. A fixed ANN architecture would definitely prevent the ANN from learning in a lifelong learning mode.
b. *Catastrophic forgetting*: When an ANN learns a new item, it forgets the old ones. This phenomenon was explained in the global learning paradigm above.

c. *Excessive training time required*: Training an ANN in a batch mode usually requires many iterations of propagating data through the ANN structure. This may not be acceptable for an adaptive online system that would require a fast adaptation.
d. *Lack of knowledge representation facilities*: Many of the existing ANN architectures capture statistical parameters during training but do not have linguistic meaning. This problem is called the "black box" problem. It occurs when only limited information is learned from the data, and essential aspects, which may be more appropriate and more useful for the future work of the system, are missed forever.

To overcome these problems, improved and new connectionist and hybrid methods and techniques are required both in terms of learning algorithms and system development.

3.2. Evolving Connectionist Systems

3.2.1. Evolutionary Computation

Evolutionary computation (EC) is concerned with population-based search and optimization of individual systems through generations of populations [89–92]. In other words, some property (or properties) of an individual will be improved not only through an individual development but also through natural selection. Methods of EC in principle include two stages: a stage of creating a new population of individuals, and a stage of development of the individual systems, so that a system develops and evolves through interaction with the environment that is also based on the genetic material embodied in the system.

The most popular among the EC techniques are the GAs. They are computational models for the optimization of complex combinatorial and organizational problems with many variants, by employing analogy with nature's evolution. GAs were introduced for the first time in the work of John Holland [91]. They were further developed by him and other researchers [89–92].

The terms used in the GA are analogous to the terms used to explain the evolution processes. They are:

- *gene*: a basic unit, which defines a certain characteristic (property) of an individual;
- *chromosome*: a string of genes, used to represent an individual, or a possible solution to a problem in the solution space;
- *population*: a collection of individuals;
- *crossover (mating) operation*: substrings of different individuals are taken and new strings (offsprings) are produced;
- *mutation*: random change of a gene in a chromosome;
- *fitness (goodness) function*: a criterion that evaluates how good each individual is;
- *selection*: a procedure of choosing a part of the population that will continue the process of searching for the best solution while the other parts of the individuals "die."

The main steps of a GA are outlined here:

1. Initialize a population of n individuals P
2. REPEAT
   2a {apply a crossover operation between the individuals from P to create an offspring set of individuals R}
   2b {apply a fitness function to evaluate the fitness of the individuals in R}
   2c {apply a selection criteria to select the fittest individuals from R in a new set P}
   2d {apply a mutation operator on the individuals from P}

UNTIL {an individual from P has reached a desired fitness or end of the procedure is reached}

When using the GA method for a complex multioptional optimization problem, there is no need for in-depth problem knowledge, or a need for many data examples stored beforehand. What is needed here is merely a "fitness" or "goodness" criterion for the selection of the most promising individuals (they are partial solutions to the problem). This criterion may require a mutation as well, which is a heuristic approach of a "trial-error" type. This implies keeping (recording) the best solutions at each of the stages.

The simple genetic algorithms introduced by John Holland are characterized by: simple, binary genes (i.e., the genes take values of 0 and 1 only); simple, fixed single-point crossover operation, in which the crossover operation is done by choosing a point where a chromosome is divided into two parts swapped with the two parts taken from another individual (see Fig. 10); and fixed-length encoding (i.e., the chromosomes had a fixed length of six genes).

The main issues in using genetic algorithms relate to the choice of genetic operations (crossover, selection, mutation).

Genetic algorithms comprise a great deal of parallelism. Thus, each of the branches of the search tree for best individuals can be used in parallel with the others. This allows for an easy realization of the genetic algorithms on parallel architectures.

Genetic algorithms are search heuristics for the "best" instance in the space of all possible instances. The following issues are important for any genetic algorithm.

- *The encoding scheme*, that is, how to encode the problem in terms of genetic algorithms (what variables to choose as genes, how to construct the chromosomes, etc).
- *The population size*—how many possible solutions should be kept for further development;
- *The crossover operations*—how to combine old individuals and produce new, more prospective ones;
- *The selection criteria;*
- *The mutation operator*—when and how to apply mutation.

In short, the major characteristics of the genetic algorithms are the following.

First, they are heuristic methods for search and optimization. As opposed to the exhaustive search algorithms, the GAs do not produce all variants to select the best one. Therefore, they may not lead to the perfect solution but, rather, to one, which is closest to it taking into account the time limits. But nature itself is imperfect too (partly because the criteria for perfection keep changing), and what seems to be close to perfection according to one "goodness" criterion may be far from it according to another.

Second, they are adaptable, which means that they have the ability to learn, to accumulate facts and knowledge without having any previous knowledge. They begin only with a "fitness" criterion for selecting and storing individuals (partial solutions) that are "good" and dismissing those which are "not good."

GAs can be incorporated in learning modules as part of an expert system or of other information processing systems.

Other EC techniques include evolutionary strategies (these techniques use only one chromosome and a mutation operation, along with a fitness criterion, to navigate in the solution (chromosomal) space); evolutionary programming—These are EC techniques applied to the automated creation and optimization of sequence of commands (operators) that constitute a program (or an algorithm) to solve a given problem [90].

The theory of GA and the other EC techniques includes different methods for selection of individuals from a population, different cross-over techniques, and different mutation techniques.

Selection is based on fitness. A common approach is proportional fitness (i.e., "if I am twice as fit as you, I have twice the probability of being selected.") Roulette wheel selection gives chances to individuals according to their fitness evaluation. Other selection techniques include tournament selection (every time of selection, the roulette wheel is turned twice,

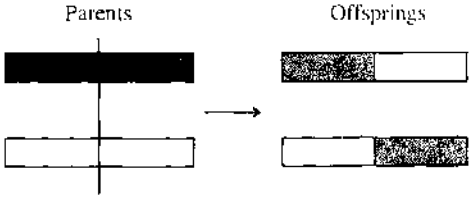


Figure 10. Schematic illustration of the operation of single-point crossover in GA. Parenting individuals exchange parts (in this case one half) of their chromosomes to create exactly two offsprings.

and the individual with the highest fitness is selected), rank ordering, and so on [93]. An important feature of the selection procedure is that fitter individuals are more likely to be selected. The selection procedure can also involve keeping the best individuals from the previous generation (if this principle was used by the nature, Michelangelo would have been alive today, he is one of the greatest artists ever, with the best genes in this respect). This operation is called elitism.

After the best individuals are selected from a population, a cross-over operation is applied between these individuals. Different single or multiple-point crossover operations can be used. Then the selected individuals undergo mutation.

Mutation can be performed in the following ways: For a binary string, just randomly "flip" a bit; for a more complex structure, randomly select a site, delete the structure associated with this site, and randomly create a new substructure

Some EC methods just use mutation (no crossover; e.g., evolutionary strategies). Normally, however, mutation is used to search in a "local search space," by allowing small changes in the genotype (and, therefore, hopefully in the phenotype). In the field of ANNs, optimal values of parameters (weights, architecture, etc.) can be sought not only through learning but also through evolution; that is, through the process of selection and crossover of the best individual neural networks. This process can be combined with individual learning to lead to the Baldwin effect [94]. The genotypes after Baldwinian learning remain unchanged; however, learning can influence indirectly the selection process by altering the fitness of individuals, and thus eventually evolution is affected [95].

3.2.2. Artificial Intelligence Versus Emerging Intelligence

Many authors see intelligence as a set of features or fixed properties of mind that are stable and static. According to this approach, intelligence is genetically defined, given, rather than developed.

Intelligence is also seen as a constant and continuous adaptation. Darwin's contemporary H. Spencer proposed in 1855 the law of intelligence, stating that "the fundamental condition of vitality is that the internal state shall be continually adjusted to the external order" (see Ref. [96], p. 14). Intelligence is "the faculty of adapting oneself to circumstances" according to Henri Simon and Francis Binet, the authors of the first IQ test (see Ref. [97]).

In Ref. [98], intelligence is defined as "the human capacity to acquire knowledge, to acquire a set of adaptations and to achieve adaptation."

Knowledge representation, concept formation, reasoning, and adaptation are obviously the main characteristics of intelligence, on which all authors agree [99, 100]. How these features can be implemented in a computer model is the main objective of the area of AI.

AI develops methods, tools, techniques, and systems that make possible the implementation of intelligence in our computer models. This is a "soft" definition of AI, which is in contrast with the first definition of AI (the "hard" one) given by Alan Turing in 1950. According to the Alan Turing's test for AI, if a person communicates in natural language with another person or an artificial system behind a bar without being able to distinguish between the two, and even more, without being able to identify whether it is a male or a female, as the system should be able to fool the human in this respect, then the system behind the bar can be considered an AI system. The described test points to an ultimate goal of AI that is understanding concepts and a language on the one hand, and to a silly task of fooling people on the other hand, rather than toward showing how one can achieve the goal.

Many techniques of AI prove to be useful to a point but do not cope very well with the dynamic nature and the combinatorial complexity of many problems from life sciences and engineering. Table 2 shows graphically different levels of evolving processes in the human brain that cannot be modeled by the use of existing AI methods.

In a general sense, information systems should help trace and understand the dynamics of the processes, automatically evolve "rules" that would change over time, "take a shortcut" in the complex problem spaces, and improve all the time. These requirements define a subset of AI, which can be called emerging intelligence (EI). The emphasis here is not on the achievement of the ultimate goal of AI as defined by Turing but, rather, on creating systems

Table 2. Different levels of evolving processes in the human brain that cannot be modeled by the use of existing AI methods.

1. Evolutionary development Function examples: genome evolution, creation of new individuals and species
2. Brain level Function examples: cognition, speech and language, consciousness
3. Neural network level Function examples: sound perception, visual image processing
4. Whole cell, neuronal level Function examples: neuronal processes of activation and growth
5. Molecular level Function examples: DNA translation into RNA, RNA transcription into proteins

Reprinted with permission from [34], N. Kasabov, "Evolving Connectionist Systems. Methods and Applications in Bioinformatics, Brain Study and Intelligent Machines," Springer, London, 2003. © 2003, Springer-Verlag.

that learn all the time, improve their performance, and become more and more intelligent. A constructivist working definition of EI is given below. It emphasizes the dynamic and knowledge-based structural and functional self-development of a system.

EI is a feature of an information system that develops its structure and functionality in a continuous, self-organized, adaptive, and interactive way from incoming information, possibly from many sources, and performs intelligent tasks typical for humans (e.g., adaptive pattern recognition, concept formation, languages learning, and intelligent control).

David Fogel [101], in his highly entertaining and highly sophisticated book *Blondie 24—Playing at the Edge of AI, describes a case of EI as a system* that learns to play checkers online without using any instructions and improves after every game. The system uses connectionist structure and evolutionary algorithms, along with statistical analysis methods.

3.2.3. Framework for Evolving Connectionist Systems

Evolving connectionist systems (ECOS) are multimodular connectionist architectures that facilitate modeling of evolving processes and knowledge discovery [34].

An evolving connectionist system is a neural network or a collection of such networks that operate continuously in time and adapt their structure and functionality through a continuous interaction with the environment and with other systems.

The adaptation is defined through a set of parameters that is subject to change during the system operation, an incoming continuous flow of information with unknown distribution, and a goal (rationale) criteria (also subject to modification) that is applied to optimize the performance of the system over time.

If we refer to the general framework of a learning system $\{S, W, P, F, L, J\}$, as discussed previously, in ECOS, the set of parameters P can be regarded as a chromosome of "genes," and both developmental learning and evolutionary computation can be applied for the system's optimization.

ECOS comprises a small subset of all the connectionist models and systems that follow the working classification scheme discussed above. They evolve either in the problem data space or in their own system space. They learn in any of the learning modes: unsupervised, supervised, reinforced or combined with the following specific characteristics:

1. They evolve in an open space.
2. They learn in an online, pattern mode, with incremental learning, and possibly through one pass of the incoming data through the system.
3. They learn in a lifelong learning mode.
4. They learn both as individual systems and as evolutionary population systems.
5. They use constructive learning and have evolving structures.
6. They learn and partition the problem space locally, thus allowing for a fast adaptation and tracing the evolving processes over time.
7. They facilitate different types of knowledge; mostly a combination of memory-based, statistical and symbolic knowledge. The evolving connectionist models presented in

the first part of this book facilitate Zadeh–Mamdani fuzzy rules (EFuNN, Chapter 3; HyFIS, Chapter 5), Takagi–Sugeno fuzzy rules (DENFIS, Chapter 5), and type 2 fuzzy rules (Chapter 5) in Ref. [34].

Each evolving connectionist system consists of four main parts: data acquisition, preprocessing and feature evaluation, connectionist modeling part, and knowledge acquisition.

Table 3 is a generalized algorithm of the functioning of an ECOS [34]. An online processing of all this information makes it possible for the ECOS to interact with users in an intelligent way. If man–system interaction can be achieved in this way, this processing can be used to extend system–system interactions as well.

Modeling evolving processes is a difficult task, as in many cases it is not well defined in terms of global optimization and goal function. However, it is vital that a variety of methods be developed that can be applied to the number of challenging real-life applications.

4. GENE INFORMATION PROCESSING

In living systems, many dynamic, adaptive, evolving processes are observed at different levels and different stages of the development that are involved in a complex interaction. At a molecular level and a cell level, the DNA, the RNA, and the protein molecules evolve and interact in a continuous way. The genes form dynamic gene networks (GNs) that define the complexity of the living organism [7]. It is not just the number of the genes in a genome, but the interaction between them that makes one organism more complex than another. The confirmation that there might be only about 30,000 protein-coding genes in the human genome is one of the key results of the monumental work of the human genome project [102]. There is a mere one-third increase in gene numbers from a rather unsophisticated nematode (*Caenorhabditis elegans*, with about 20,000 genes) [103] to humans (and other mammals). In fact, the genomes of all mammals are so similar that it is hard to understand how they can produce such different animals. If their genes are alike, it is probably changes in when, where, and how active they are that drives the differences between species.

Table 3. Generalized ECOS algorithm, introduced in [34].

Set some preliminary parameter values for the ECOS parameters (chromosomes)
REPEAT {in a lifelong learning mode}
  IF input, or input-output data is available DO
   Read input data (or input- output data pairs if such are available)
   Evaluate the input-output features:
        (a) add new ones if necessary;
        (b) select the current most appropriate ones for the task
   Propagate input data through the NN modules and evaluate the similarity of the input data to the modules
   If there is not sufficient similarity—create new modules, or create new connections in an existing module
   Calculate the output of the system
   Calculate a feedback from the output to the system through:
        A supervised mode of learning, if output error values are calculated, or
        A reinforcement mode of learning—if just hints about the correctness of the output values are available, or
        Report the output values if the system is in a recall mode
   Modify the structure of ECOS based on the feedback
   Extract and report the current knowledge learned by the ECOS, e.g. through rule extraction techniques
   Optimize the ECOS structure based on some accumulated statistical and ECOS parameters (possibly EC methods)
  ELSE
   Apply inner structural and functional learning for structure improvement (e.g., sleep learning)
  UNTIL {the system is stopped, or interrupted}

4.1. Genes and Cellular Processes

DNA is a chemical chain that is present in the nucleus of each cell of an eukaryotic organism, and it consists of ordered double-helix pairs of small chemical molecules (bases): adenine (A), cytosine (C), guanidine (G), and thymidine (T), linked together by a deoxyribose sugar phosphate nucleic acid backbone.

The central dogma of molecular biology (see Fig. 11) states that the DNA is transcribed into RNA, which is translated into proteins [104].

DNA contains millions of base pairs, but only 5% or so is used for the production of proteins, and these are the segments from the DNA that contain genes. Each gene is a sequence of base pairs that is used in the cell to produce proteins. Genes have length of hundreds to thousands of bases.

RNA has a similar structure to DNA, but here thymidine is substituted by uridine (U). In pre-RNA, only segments that contain genes are extracted from the DNA. Each gene consists of two types of segments: exons, which are segments translated into proteins, and introns, which are segments that are considered redundant and do not take part in the protein production. Removing the introns and ordering only the exon parts of the genes in a sequence is called splicing, and this process results in the production of messenger RNA (or mRNA) sequences.

mRNAs are directly translated into proteins. Each protein consists of a sequence of amino acids, each of them defined as a base triplet, called a codon. From one DNA sequence many copies of mRNA are produced; the presence of certain gene in all of them defines the level of the gene expression in the cell and can indicate what and how much of the corresponding protein will be produced in the cell.

The above description of the central dogma of the molecular biology is very much a simplified one, but it will in understanding the rationale behind using connectionist and other information models in bioinformatics.

Genes are complex chemical structures, and they cause dynamic transformation of one substance into another during the whole life of an individual, as well as throughout the life of the human population over many generations. When genes are "in action," the dynamics of the processes in which a single gene is involved are very complex, as this gene interacts with many other genes and proteins and is influenced by many environmental and developmental factors.

Modeling these interactions, learning about them, and extracting knowledge is a major goal for bioinformatics. Bioinformatics is concerned with the application of the methods of information sciences for the analysis, modeling, and knowledge discovery of biological processes in living organisms.

The whole process of DNA transcription, gene translation, and protein production is continuous, and it evolves over time. Proteins have 3D structures that unfold over time, governed by physical and chemical laws. Proteins make some genes to express and may

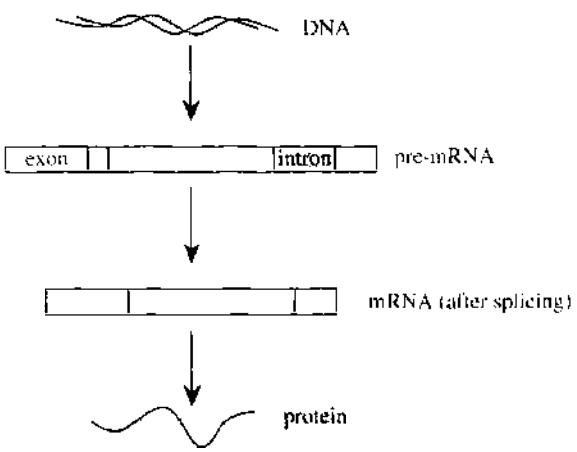


Figure 11. DNA is transcribed into RNA, which is translated into proteins—the central dogma of molecular biology.

suppress the expression of other genes. The genes in an individual may mutate, change slightly their code, and therefore express differently at a next time. Thus, genes may change, mutate, and evolve in a life time of a living organism.

4.2. Computational Models of Gene Information Processing

In a single cell, the DNA, the RNA and the protein molecules interact in a continuous way during the process of the RNA transcription from DNA (genotype), and the subsequent RNA to protein (phenotype) translation [105–109]. A single gene interacts with many other genes in this process, inhibiting, directly or indirectly, the expression of some of them, and promoting others at the same time. This interaction can be represented as a gene regulatory network (GRN) [7, 8, 110–115]. Our challenge is to create computational models of GRN from both dynamic data (e.g., gene expression data of thousands of genes over time, and also from protein data) and from static data (e.g., DNA), under different external inputs (diseases, drugs, etc.). A large amount of data on gene interactions for specific genomes, as well as on partial models, is available from public domain databases such as GenBank and PubMed (http://www.ncbi.nlm.nih.gov/), KEGG (http://www.genome.ad.jp/kegg/), Stanford Microarray Database, and many more [116]. Collecting both static and time course gene expression data from up to 30,000 genes is now a common practice in many biological, medical, and pharmaceutical laboratories in the world through the introduction of microarray technologies (see, e.g. www.ebi.ac.uk/microarray). Sophisticated information and mathematical methods are needed for the analysis, modeling, and discovery of GRN from this data.

Several generic information methods for modeling and for the discovery of variable interaction networks from time course data have been proposed and used in the domain of GRN modeling. Among them are statistical methods that include correlation techniques, linear regression, Bayesian networks, and hidden Markov models [9, 115, 117–119]; neural networks [120, 121]; evolutionary computation, and genetic algorithms in particular [116, 122, 123]; directed graphs [105, 116, 117]; Petri nets [116]; and ordinary and partial differential equations [8]. There have also been specific methods developed for the purpose of cell modeling [8, 124, 125]. A detailed survey of the elements and the pathways in the control of gene expression and the principles of their computational modeling can be found, for instance, in the books [8, 110, 113, 116] and in the review papers [126, 127]. With respect to the taxonomy of GRN models, their principles and descriptions, exhaustive reviews can be found for instance in Refs. [127, 128].

In Ref. [129], an evolving connectionist system (ECOS) is incrementally evolved from incoming data $\mathbf{X}(t_0)$, $\mathbf{X}(t_1)$, $\mathbf{X}(t_2)$, ..., representing the values of all, or some, of the gene expression variables or their clusters. Consecutive vectors $\mathbf{X}(t)$ and $\mathbf{X}(t+k)$ are used as input and output vectors, respectively, in an ECOS model, as shown in Fig. 12a. After training an ECOS on the data, rules are extracted through IF–THEN representation of the rule nodes [e.g., IF x1(t) is High (0.87) and x2(t) is Low (0.3) THEN x3(t + k) is High (0.6) and x5(t + k) is Low (0.2)]. Each rule represents a transition between the current and next state of the variables, as shown in Fig. 12b, where each rule is shown as an arrow. All rules together form a representation of the GRN. Figure 12b shows two trajectories, N_1 and N_2, that represent two GRNs, derived under different conditions, in the 2D PCA (principal component analysis) coordinate space of all n variables. By modifying a threshold for rule extraction, one can extract in an incremental way stronger or weaker patterns of relationships between the variables [77].

Using another ECOS model [130], other types of variable relationship rules can be extracted [e.g.: IF x1(t) is (0.63 0.70 0.76) and x2(t) is (0.71 0.77 0.84) and x3(t) is (0.71 0.77 0.84) and x4(t) is (0.59 0.66 0.72) THEN x5(t + k) = 1.84 − 1.26x1(t) − 1.22x2(t) + 0.58x3(t) − 0.3x4(t)], where the cluster for which the value of x_5, as defined in the rule above, is a fuzzy cluster represented through triangular membership functions defined as triplets of values for the left, center, and right points of the triangle on a normalization range of Ref. [131]. The fuzzy representation allows for models to deal with imprecise data [10]. The rules extracted from the ECOS form a representation of the GRN (see Fig. 12c).

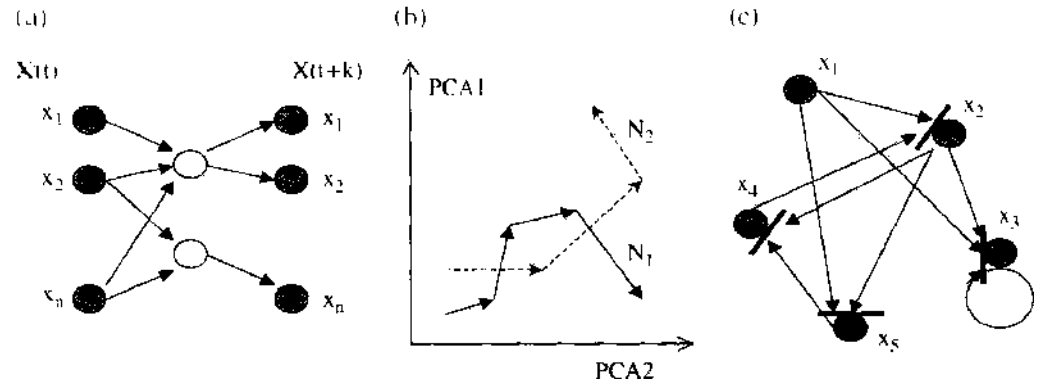


Figure 12. (a) A hypothetical ECOS for GRN modeling; (b) state transitions (rules, represented as arrows) in the two PCA dimensional space of the n variables; (c) part of a GRN extracted from an ECOS model.

Rules may change with the addition of new data, thus making it possible to identify stable versus dynamic parts of the GRN. The relationship between variables from Fig. 12c can be represented by a simple linear, or a complex nonlinear, relationship function.

5. NEURO-GENETIC INFORMATION PROCESSING

Let us consider the differences between humans and chimpanzees. After sequencing ~3 million letters of the chimp genome and comparing them with the human draft, Svante Pääbo of the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany, and his group reasoned that DNA sequence cannot be the cause of such species variation only 1.3% of letters are different [132]. These letters belong mainly to the genes expressed in the brain. Then the researchers and the team of U.S. scientists measured the levels of gene activity in the brains of humans and chimps. Chimp and human brain transcription patterns are very different. The human brain has accelerated usage of genes, with ~90% of the studied genes being more highly expressed than in the chimp brain [133], and mainly the up-regulation of genes involved in synaptic transmission and plasticity (learning and memory), energy metabolism, and growth. Higher levels of such a neuronal genetic activity are likely to have important consequences in cognitive and behavioral capacities of humans.

However, to understand these genetic differences, we also need to know what is going on in a chimp's mind; after all, we know that chimpanzees and bonobo chimpanzees have quite impressive language and learning skills [134] (see also http://www.gsu.edu/~wwwlrc/biographies/kanzi.html). Therefore, it is important to study the neuro-genetic interactions in all species to compare them with humans.

In September 2003, the news spread that the Microsoft cofounder Paul G. Allen has donated $100 million to launch a private research organization in Seattle devoted to deciphering the links between our genes and our brain (Allen Brain Atlas Project, http://www.nba.com/blazers/news/Allen_Institute_for_Brain_Scie-84505-41.html). It is rather a heroic task, considering that approximately 6000 genes are thought to be expressed only in the brain, with many more that are expressed in the brain and also in other parts of the body as well.

Therefore, it will be crucial to study, both theoretically and experimentally, the consequences of mutated genes on the activity of neural networks and on the neurological and mental deficits that can follow. It is how the links can be established, determining which gene(s) and which interactions between genes are responsible for which neuronal, and eventually which mental, function.

5.1. Neuro-Genetic Processes in the Brain

Complex faculties of brains, such as, for example, intelligence, are under the influence of both genetic and environmental factors, but what does it actually mean, for instance, for intelligence to be under the influence of genes? What are the nature and rules of this

dependence? There is vast scientific evidence that intelligence (like other mental faculties, processes, states, etc.) depends on the normal functioning of brain neural networks, and it is the set-up of brain neural networks that is under the influence of genes. Whole-brain development from conception is guided by the complex sequence of switching on and off many different genes operating in their own complex and intricate networks (GRNs). Through protein synthesis, genes determine the structure and connectivity of the brain, including all the biochemical processes involved in information processing and the mere survival of brain cells.

For the correct brain function, there is interplay between genetic and epigenetic factors, like signals from outside of neurons, either from other neurons or other cells in the brain (i.e., glia) or from somewhere else in the body (e.g., hormones). Proteins, molecules, and ions acting on neurons from outside can and do act on the genome to influence its activity. It is how the environment exerts its influence on the structure of brain neural networks.

What is the nature of the questions we can ask?

- Can we perform a reverse engineering of brain genetic networks? That is, can we identify causal regulatory interactions between genes from time-dependent multigene expression measurements?
- What are the ways that genes can determine the function of brain neural networks?

Thus, in the first research area on genes and the CNS (central nervous system), researchers have applied GRN inference techniques to an extensive survey of gene expression in CNS development. Detailed cluster analysis has uncovered waves of expression for about 112 genes that characterize distinct phases of the development of spinal cord and hippocampus [114, 135]. Thus, the first area of research covers the monitoring of the activity of many genes in parallel and the use of the models of GRN to help and guide the inference of regulatory connections between genes, resulting in a gene-interaction diagram of gene-interaction pathways. There is still a long way to go before all 6000 genes are processed in a similar way but with better techniques and better theoretical and computational models.

The second area of study is to reveal the consequences of genes mutation on neural and mental functions (see Table 4) [136]. It is crucial to study, both theoretically and experimentally, the consequences of mutated genes on the activity of neural networks and on the neurological and mental deficits that can follow. It is how the links can be established; that is, which genes and which interactions between genes are responsible for which neuronal, and eventually which mental, function. We can see that the detailed pathogenesis of these diseases is at this time unknown. This is the area into which computational models can bring new insights.

5.2. Computational Modeling of Neuro-Genetic Processes

The genes, encoded in the DNA, that are transcribed into RNA and then translated into proteins in each cell contain important information related to the brain activity. A specific gene from the genome relates to the activity of a neuronal cell in terms of a specific function. However, the functioning of the brain is much more complex than that. The interaction between the genes is what defines the functioning of a neuron. Even in the presence of a mutated gene in the genome that is known to cause a brain disease, the neurons can still function normally provided a certain pattern of interaction between the genes is maintained—a certain state of the GN [137]. In contrast, if there is no mutated gene in the genome, certain abnormalities in brain functioning can be observed as defined by a certain state of the interaction between the genes [11]. The above-cited and many other observations point to the significance of modeling a neuron and a neuronal ensemble at the gene level to predict the state of the ensemble. The process of modeling the gene interaction with the goal of brain understanding is a significant challenge to biologists, mathematicians, information and computer scientists, brain scientists, and researchers from many other areas.

5.2.1. Principles of a General Neuro-Genetic Model

In Ref. [127], a computational model of GRN of early neurogenesis in *Drosophila* is developed with the use of artificial neural networks of a Hopfield type, where each gene (gene product) is a node in a recurrent network and the values of connections express the

Table 4. Single and multiple genes related to some brain functions and abnormalities [136].

Disease	Mutations of genes identified so far	Location of genes on chromosomes	Brain abnormality	Symptoms
Alzheimer disease (AD)	PS2 (AD4) PS1 (AD3) Unknown Unknown	1 14 19 21	plaques made of fragmented brain cells surrounded by amyloid-family proteins, tangles of cytoskeleton filaments	progressive inability to remember facts and events and later to recognize friends and family
Amyotrophic lateral sclerosis (ALS)	SOD1	21	progressive degeneration of motor neuron cells in the spinal cord and brain	loss of motor control which ultimately results in paralysis and death
Angelman syndrome (AS)	UBE3A	maternally derived chromosome 15 (segment 15q11–13)	mutations in UBE3A disrupt protein break down during brain development	mental retardation, abnormal gait, speech impairment, seizures, frequent laughing, smiling, and excitability
Epilepsy (many forms)	Multiple	Multiple	abnormal cell firing in the brain	recurring seizures
Fragile X syndrome	FMR1	X	impaired synaptic function of glutamatergic synapses	the most common inherited form of mental retardation
Huntington disease (HD)	HD gene	4	dilatation of ventricles and atrophy of caudate nucleus	degenerative neurological disease that leads to dementia
Williams syndrome	LIM kinase and elastin coding sequences	7	Unknown	high competence in language, music and interpersonal relations, with low IQ

sign and strength of their interactions. Here a model is presented [138] that is based on two sets of essential genes—a set of genes $\mathbf{G}_{gen}$ that defines generic neuronal functions, and a set of genes $\mathbf{G}_{spec}$ that defines specific neuronal functions (e.g., epileptic behavior). Consider a function that describes measurement of GABA (e.g., $GABA_A$) in the synaptic cleft. The measurement of cross-membrane potential using a patch-clamp procedure provides a function of time during an event that probes spiking properties of the neuron when a certain voltage is delivered to the neuron.

The two sets of genes together form a set $\mathbf{G} = \{g_1, g_2, \ldots, g_n\}$ that will be used, and a GN of this set will be defined in the model. An example of a GN of four genes is given in Fig. 13. The expression level of each gene $g_j(t + \Delta t)$ is a function of the gene expression levels of the rest of the genes $\mathbf{G}(t)$. As a simple model, we will assume that this function is a linear function

$$g_j(t + \Delta t) = w_{j,0} + w_{j,1}g_1(t) + \cdots + w_{j,n}g_n(t) \tag{4}$$

The square matrix of gene connection weights $\mathbf{W}$ represents the GN. This model can be run for consecutive time moments.

Here we use the gene instead of proteins, using a standard assumption that if a gene is up-regulated, more proteins defined by this gene will be produced in the neuronal cell.

A set of neuronal functions (parameters) $\mathbf{P} = \{p_1, p_2, \ldots, p_m\}$ from a neural network model is related to particular genes, so that each parameter p_j is a function of the expression of several (or in the partial case, one) genes. For simplicity in our model, we will assume that one parameter p_j depends on one gene g_k through a liner function

$$p_j(t + \Delta t) = z_{j,0} + z_{j,k}g_k(t) \qquad j = k \tag{5}$$

The parameter vector $\mathbf{Z}$ defines the relationship between the selected genes $\mathbf{G}$ and the spiking characteristics of a neuron.

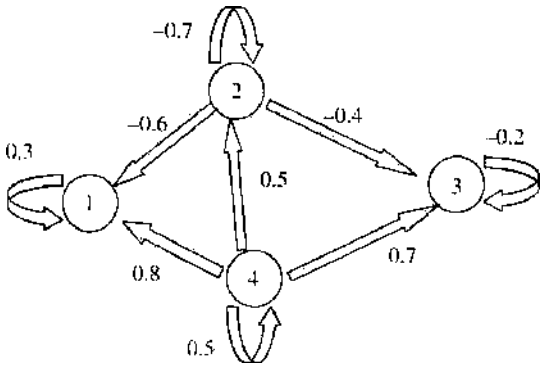


Figure 13. A hypothetical example of a GN comprising of four genes related to both generic and specific functions of a neuron. The genes are connected with ares that represent the relationship (both sign and strength) between the level of expression of this gene at time moment (t) and the next time moment ($t + \Delta t$).

This generic neuro-genetic (NGM) model can be run continuously over time in the following way:

1. Define the initial expression values of the genes **G**, $\mathbf{G}(t = 0)$ in the neuron and the matrix **W** of the GN if that is possible.
2. Run the GN and define the next moment state of the gene set $\mathbf{G}(t + \Delta t)$ using Eq. (4).
3. Define the values of the parameters **P** from the gene state **G** using Eq. (5).
4. Define the spiking activity of neuron(s) (taking into account all external inputs to the neural network).
5. Go to step 2.

We assume that the same matrix **W** defines the GN of each neuron in a SNN. The spiking activity of all neurons are defined using the algorithm in the section below for a time interval T, thus allowing us to calculate the probability distribution function (PDF) of the spiking activity of neurons in the spiking neural network (SNN).

5.2.2. Using a Genetic Algorithm to Find the Values of Parameters of a Neuro-Genetic Model

We assume that the functioning of a neural network is evaluated as its PDF of neural activity, thus making it possible to observe and model different brain functions such as epilepsy; alpha, beta, and gamma states; learning; memorizing; sleeping; and so forth. Usually, EEG data are available to test these models. In general, the functioning of a neural network can be expressed and evaluated in other terms; for instance, metabolic or other activity terms.

The task for us is to define a set of parameter values for **W** and **Z** so that the SNN has a desired PDF of neural activity, denoted here as PDF'. To solve this problem with the use of the above-described model, we can apply the methods of EC and, in particular, a GA method [116, 139, 140] with a fitness function of PDF = PDF' within a margin of tolerance.

Parameter estimation is a very difficult task in inferring GN models, mainly because of the lack of observation data relative to the number of genes involved. In this respect, EC that are robust, global optimization methods become important tools to accurately inference a GN.

EC, inspired by the Darwin theory of evolution, searches with a swarm of points based on the objective function (say, the output error) feedback of these points. It has been used for parameter estimation or optimization in many engineering applications. Unlike classical derivative-based (like Newton) optimization methods, EC is more robust against noise and multimodality in the search space. In addition, EC does not require the derivative information of the objective function and is thus applicable to complex, black box problems.

These characteristics make EC highly suitable for identifying parameters of GN models for three reasons. First, the derivative information of the underlying model (e.g., Petri Net) is usually not available. Second, data are scarce or missing (causing multimodality) and noisy, requiring a robust, global optimization algorithm that is not easily misled by suboptima

and noise. Third, qualitative inference of parameters is difficult with such small number of observations relative to the large number of genes involved.

In the GA implementation here, two chromosomes of parameters will be used—**W** and **Z**, so that for every generation (a set of values) of these chromosomes, the SNN is run for the time period of T and the PDF is evaluated. Then it is compared with the desired PDF', and if the fitness function is not satisfied, the process continuous with modified values for the parameters **W** and **Z** according to the selected GA strategy.

5.2.3. Example of a Neural Network Model with Parameters Related to Gene Net

A spiking model of a neuron—an element of the SNN—can be, for instance, inspired by the spike-response model (SRM) of a neuron [28, 141]. Neuron i receives input spikes from presynaptic neurons $i \in \Gamma_i$, where Γ_i is a pool of all neurons presynaptic to neuron i (see Fig. 14). The state of neuron i is described by the state variable $u_i(t)$ that can be interpreted as a total PSP at the membrane of soma. When $u_i(t)$ reaches the firing threshold $\vartheta_i(t)$, neuron i fires (i.e., emits a spike; see Fig. 15a). The moment of $\vartheta_i(t)$ crossing defines a firing time t_i of an output spike. The value of the state variable $u_i(t)$ is the sum of all postsynaptic potentials, i.e. (see Fig. 15b)

$$u_i(t) = \sum_{j \in \Gamma_i} \sum_{t_j \in F_j} J_{ij} \varepsilon_{ij}(t - t_j - \Delta_{ij}^{ax}) \tag{6}$$

The weight of synaptic connection from neuron j to neuron i is denoted by J_{ij}. It takes positive (negative) values for excitatory (inhibitory) connections, respectively. Depending on the sign of J_{ij}, a presynaptic spike generated at time t_j increases (or decreases) $u_i(t)$ by an amount $\varepsilon_{ij}(t - t_j - \Delta_{ij}^{ax})$. The term Δ_{ij}^{ax} is an axonal delay between neurons i and j that increases with Euclidean distance between neurons.

The positive kernel $\varepsilon_{ij}(t - t_j - \Delta_{ij}^{ax}) = \varepsilon_{ij}(s)$ expresses an individual PSP evoked by a presynaptic neuron j on neuron i. A double exponential formula can be used (see Fig. 15c)

$$\varepsilon_{ij}^{\text{synapse}}(s) = A^{\text{synapse}}\left[\exp\left(-\frac{s}{\tau_{\text{decay}}^{\text{synapse}}}\right) - \exp\left(-\frac{s}{\tau_{\text{rise}}^{\text{synapse}}}\right)\right] \tag{7}$$

where $\tau_{\text{decay/rise}}^{\text{synapse}}$ are time constants of the rise and fall of an individual PSP, A is the PSP's amplitude, and synapse = fast_excitation, fast_inhibition, slow_excitation, and slow_inhibition, respectively. These types of PSPs are based on neurobiological data [142, 143].

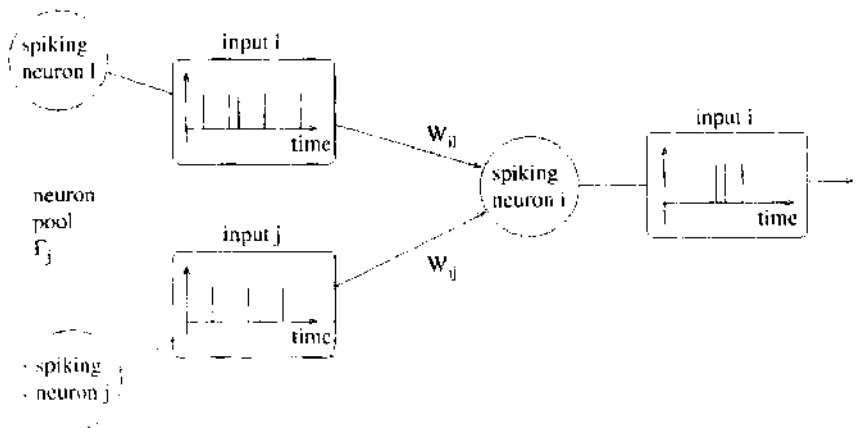


Figure 14. In response to input series of spikes from the pool of presynaptic neurons Γ_i, a neuron i generates its own series of output spikes.

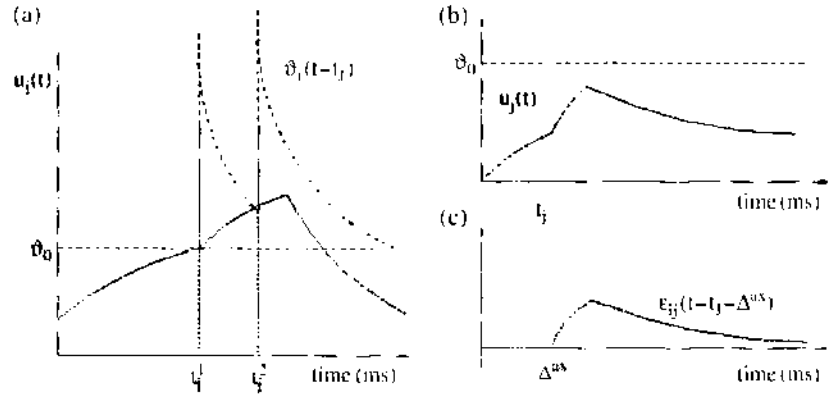


Figure 15. Spiking neuron model. (a) When the state variable $u_i(t)$ of a spiking neuron reaches the firing threshold $\vartheta_i(t)$ at time t_i, a neuron fires an output spike. However, an actual firing threshold rises after each output spike and decays back to the initial value. (b) Subthreshold temporal summation of individual postsynaptic potentials. (c) The kernel $\varepsilon_{ij}(t - t_j - \Delta_{ij}^{ax})$ describes an individual PSP evoked by the presynaptic spike fired at time t_j after some axonal delay Δ_{ij}^{ax}.

Immediately after firing an output spike at t_i, neuron's firing threshold $\vartheta_i(t)$ increases m times and then returns to its initial value ϑ_0 in an exponential fashion (see Fig. 15a).

$$\vartheta_i(t - t_i) = m \times \vartheta_0 \exp\left(-\frac{t - t_i}{\tau_{\text{decay}}^{\vartheta}}\right) \tag{8}$$

where $\tau_{\text{decay}}^{\vartheta}$ is the time constant of the threshold decay. In such a way, absolute and relative refractory periods are modeled.

External inputs from the input layer are added to the right-hand side of Eq. (5) at each time step, thus incorporating the background noise or the background oscillations. Each external input has its own weight $J_{ik}^{\text{ext\_input}}$ and $\varepsilon_k(t)$, such that

$$u_i^{\text{ext\_input}}(t) = J_{ik}^{\text{ext\_input}} \varepsilon_{ik}(t) \tag{9}$$

It is optional to add some degree of Gaussian noise to the right-hand side of Eq. (6) to obtain a stochastic neuron model instead of a deterministic one.

Figure 16 illustrates the basic architecture of a SNN. Spiking neurons within the network can be either excitatory or inhibitory. There can be as many as 10–20% of inhibitory neurons

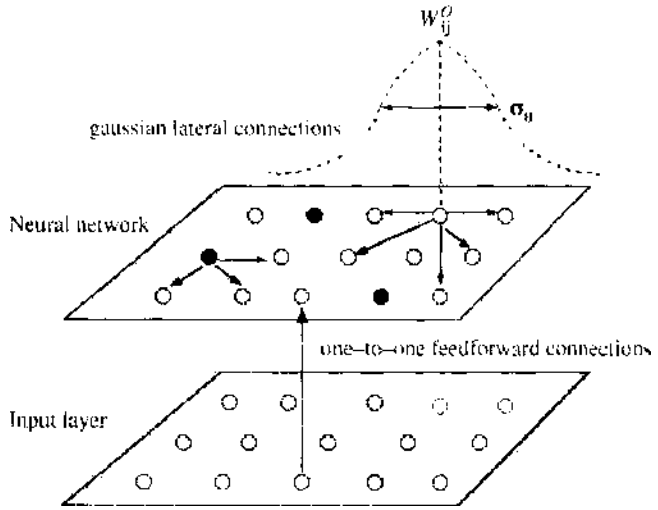


Figure 16. Architecture of the spiking neural network (SNN). About as many as 10–20% of neurons are inhibitory neurons that are randomly positioned on the grid (filled circles). Excitatory and inhibitory lateral connections decrease in strength with distance according to the Gaussian distribution. There are one-to-one feed-forward connections from the input layer.

positioned randomly on the rectangular grid of N neurons. Lateral connections between neurons have weights that decrease in value with distance from neuron i, for instance, according to a Gaussian formula, whereas the connections between neurons themselves can be established at random.

Figure 17 illustrates a record of activity of the introduced SNN. It is useful to keep a record of spiking activities of all neurons individually and in total, as well as the record of the total membrane potential, which is in fact proportional to EEG [144]. Various analytical tools are developed, for instance, for evaluation of the degree of synchrony between neurons [141] and for the evaluation of frequency spectra, like the fast Fourier transform and others [141, 145]. Presented SNN belongs among the simplest tools. There are much more detailed models of spiking neurons than SRM or the so-called integrate-and-fire (I&F) model neurons. These more detailed models include, for instance, the various ion receptor and channel kinetics [142, 146], and also include multicompartmental neuron models in which the effect of spatial—not only temporal—summation of PSPs on the neuron input surface is taken into account [147, 148].

5.2.4. Example of Relation of Genes to Particular Neural Network Parameters

Let us take as an example of the set of genes G_{spec} that define specific neuronal functions (e.g., epileptic behavior)—the set of genes that are presumably mutated in individuals suffering from the Childhood Absence Epilepsy (CAE). CAE is an idiopathic (i.e., arising from an unknown cause), generalized, nonconvulsive epilepsy. The main features are absence seizures. A typical absence is a nonconvulsive epileptic seizure, characterized by a brief (4–20 s) impairment of consciousness. This may happen up to ~200 times a day. Absence seizures occur spontaneously (i.e., they are not evoked by sensory or other stimuli [11, 149]). Absence is accompanied by a generalized, synchronous, bilateral, 2.5–4-Hz spike and slow-wave discharge (SWD) in the EEG. SWDs can start anywhere in the cortex, and from there they quickly spread to the entire cortex and thalamus [150].

Table 5 the genes that are most probably mutated in CAE, their coded proteins, the neuronal function or functions these proteins are responsible for, related parameters in the SNN, and a putative alterations in these functions, as well as changes in SNN parameters. Putative changes in the neural function can be derived from numerous studies performed on humans, rats, and mice [11, 142, 143, 149, 151–155]. It should be pointed out that other types of idiopathic epilepsies like the frontal lobe epilepsy (ADNFLE), temporal lobe epilepsy (TLE), juvenile myoclonic epilepsy (JME), adult myoclonic epilepsy (AME), and so forth

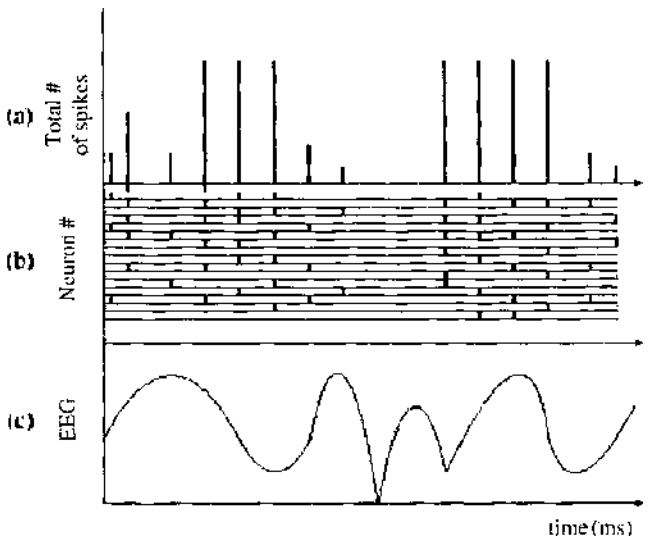


Figure 17. Temporal evolution of the network activity. (a) Total number of spikes generated by all neurons at each time step. (b) Traces of each neuron spiking activities. (c) Total sum of individual membrane potentials as a measure proportional to EEG.

Table 5. List of genes putatively mutated in CAE, their coded proteins, neuronal function these proteins are responsible for, related parameters in the SNN, and a putative alteration in this function: ↑ means increase, ↓ means decrease.

Mutated Gene	Protein	Neuronal Function and its putative Alteration	SNN Parameters
GRIK1	Ionotropic GluR5 (Kainate receptor 2 for glutamate)	Fast excitation ↑ or ↓ (depending on the place in the brain)	$A^{\text{fast\_exc}}$, $\tau^{\text{fast\_exc}}_{\text{decay, rise}}$ ↑ or ↓
GABRB3	$GABA_A$ receptor $\beta 3$ subunit	Fast inhibition ↓	$A^{\text{fast\_inh}}$, $\tau^{\text{fast\_inh}}_{\text{decay, rise}}$ ↓
GPHN	Gephyrin	Fast inhibition ↓	$A^{\text{fast\_inh}}$, $\tau^{\text{fast\_inh}}_{\text{decay, rise}}$ ↓
CHRNA4	nAChR $\alpha 4$ subunit	Both fast and slow inhibition ↓	J^{inh}_{ij} ↓
OPRM1	μ-Opioid receptor type 1	Firing threshold ↓	ϑ_0, τ^{if}_{decay} ↓

are connected to different channelopathies and receptoropathies [153, 156]; that is, to gene mutations different from CAE, so the presented table applies only to CAE.

A more detailed model of a neuron including the receptor and channel kinetics will enable us to link mutated genes to the parameters at a more detailed level. Still, it will be a hard task to determine quantitatively the values of coefficients of mutual interactions between genes themselves and between genes and their coded proteins [i.e., the values of coefficients in Eqs. (2) and (3)]. Here we have proposed GA as a means of such optimization; however, these values also can be obtained from experimental data if they are available. In any case, experimental data will serve as an ultimate test of neuro-genetic theoretical and computational models when applied to particular problems, such as, for instance, epilepsy. Once a functional neuro-genetic model of a particular neurological disease is developed, then the effects of genetic and other parameter changes can be simulated and predicted by means of the theoretical computational model. Before any neuro-genetic causal cures are going to be administered, mutated genes should be identified and their interactions with other genes and neuronal functions known to avoid any unwanted consequences of the perspective gene therapy. The gene therapy proper is another nontrivial issue. Once it is known what should be added or removed from the cell genome, reliable nanotechnology must be developed for carrying out this operation. This applies not only for possible genetic cures of epilepsy but also for any other neurological or mental disorder.

6. CONCLUSIONS AND FUTURE DEVELOPMENT

Nanotechnology has a tremendous potential for the cure of brain diseases. This approach requires a deep understanding of chemical and information processes in the brain, in single neurons and in the nuclei of these neurons, and especially in how these processes relate to each other. In this respect, there is a need for theories and computational models to model and predict the outcome of brain abnormalities and their treatment. Neuro-genetic modeling is a promissing approach that will be further developed and applied. A new field of nanomedicine can be developed in the future to deal with nanotechnology in medicine and health care [157, 158].

In between, many nanotechnical problems must be solved. For instance, the brain, similar like other tissues, responds to alien substances with a healing process. For instance, various neural probes (usually composed of silicon) become encapsulated with glial scar tissue, which can impede with normal neuron function. From a nanomaterial point of view, nanophase materials can influence interaction with proteins and other molecules that take part in cell processes in many unwanted ways [159]. Another area of future study and possible application will be the development of nanoscale logic networks and nanochips [160, 161]. They will lead not only to an unprecedented miniaturization of conventional computers but also to miniaturization of neurocomputers and neurochips. It is not only about implementation of classical ANNs in hardware but also about the development of the so-called neuromorphic systems. Neuromorphic systems are implementations in silicon of sensory and neural

systems whose architecture and design are based on neurobiology that can compete with human senses and pattern-recognition systems and run in real time [162, 163]. Researchers in this area also work on developing communication between living vertebrate neurons and electronic systems [162].

Although a lot is known about the brain, issues about its functioning, representation, and processing of information are still subjects of an intense research. Applicability of nanotechnology will depend not only on the nanotechnological progress itself but also on the progress in understanding the brain, its dynamical behavior, and how normal and disturbed neural functions affect the brain dynamics. The nature of brain dynamics is still unknown. Some researchers find evidence of chaos, whereas some are doubtfull [164]. The main proponents of a chaotic dynamic, Freeman [165] and Tsuda [166], argue in favor of chaotic itinerancy based on EEG and other neurophysiological data. According to the picture of chaotic itinerancy, a complex system such as the (human) brain evolves by steps along a trajectory in the state space. Each step corresponds to a shift from one basin of attraction to another. Attractors represent classes for abstraction and generalization. Thus, the brain states evolve aperiodically through sequences of attractors. In a closed system, the next attractor would be chosen solely by internal dynamics. In an open system, such as the brain, external inputs interfere with internal dynamics. Moreover, because of the changes induced by learning, trajectories continually change. Chaotic itinerancy occurs in sequence of cortical states marked by state transitions that appear in temporal discontinuities in neural activity patterns [165].

Experimental EEG data show that the entire cerebral cortex is constantly wandering in the fractal distributions of phase transitions that give the $1/f^{\alpha}$ form of the temporal and spatial frequency spectra (with $\alpha \in (1, 3)$, [165]). From this type of frequency spectrum, it appears that the brain maintains a state of self-organized criticality (SOC) [167]. The SOC state can form the basis of the brain's capacity to rapidly adjust to new external and internal stimuli. State changes resembling phase transitions occur continually everywhere in cortex at scales ranging from milimeters to ~0.1 m. Local neural activity can trigger a massive state change.

However, several issues of caution should be pointed out. In spite of the compelling evidence for SOC in the brain, the nature of the critical state is still unknown in neurobiological interpretation. The spatial and temporal power spectral densities (PSDs) often show the $1/f^{\alpha}$ form; however, more often this form is broken down as a result of distortions by clinically defined peaks. Therefore, the measurements of α vary widely. Aperiodic oscillations giving the $1/f^{\alpha}$ PSD are commonly referred to as chaotic. However, the brain activity is not at all consistent with low-dimensional deterministic chaos [164, 168]. It is high dimensional, noisy, non-Gaussian, and nonstationary [165]. Therefore, the conditions for the assessment of this type of dynamics are difficult to be met. Moreover, brains are open systems driven by stochastic input. Thus, it seems that the brain activity can hardly conform to the mathematical definitions of chaos. Whether the term chaotic itinerancy (or any other term from the chaotic vocabulary) is appropriate to describe state transitions in brain and cortex in particular remains open to challenge. Thus, the complex spatio-temporal activity data from the brain still await explanation.

Another issue that is unresolved at present and that will be highly relevant when it will be possible to nanotechnologically enter the brain is the issue of consciousness and other mental phenomena. Neurobiologists are trying to identify the so-called neural correlates of consciousness and to gather experimental data in support of the stream of transient semiglobal coherencies in brain electrical activity being (somehow) the basis for consciousness [169–171]. In his influential book *Shadows of the Mind*, physicist Roger Penrose brings the problem of explaining consciousness to the domain of physics [172]. His critics question the competence of physics ever having anything of importance to say about mental phenomena in general, and consciousness in particular (http://psyche.cs.monash.edu.au/psyche-index-v2.html#som). The grounds for this criticism vary, ranging from computational to neurobiological arguments. In his response to this critique [173], Penrose states that he certainly does not expect to find any answers in contemporary subatomic physics. Instead, he has been arguing for a new physics, for a radical upheaval in the very basis of physical theory. According to

him, at present, any scientific (including physical) theory does not help us to come to terms with the puzzle of mentality, including consciousness within such a physically determined universe. Even at this point, the point of mystery of mentality, there is not a general agreement among scientists. Some think that there is no mystery at all and that the consciousness and other mental phenomena emerge from a particular underlying basis, be it the specific computations [174] or specific properties of the brain processes [175]. In another influential book written on consciousness [176], philosopher and mathematician, David J. Chalmers clearly argues that consciousness and mentality are indeed genuinly puzzling and are not explainable by present theories. If one takes consciousness seriously, Chalmers says, one has to go beyond a strict materialist framework. The fundamental laws linking the physical and the experiential are yet to be discovered, although he attempts to define and search for them. Penrose also tries to link together the physical and experiential, and he sees the link in a new physical theory based on the union of Einstein's general relativity with quantum theory. However, he writes [173], we do not yet know the very form this new theory must take. It may have a character very different from that of a traditional physical theory. Penrose does not believe that any real progress will be achieved toward solving the mysteries of how mental phenomena fit in with the physical universe until there are some important changes in our picture of physical reality. However, already at this point, Chalmers asks why quantum processes (or any other specific physical processes) in microtubules (or any other brain substructures) should give rise to consciousness, any more than specific computational processes [177]. What all scientists agree on is that mentality is causally linked to the brain, whether its basis is particular computations or particular biological or physical processes. Brain nanotechnology will interfere with all these processes; therefore, a potential danger lies here for unforeseen consequences. However, nanotechnology by targeted and controlled manipulation of selected molecules might help in the search for those very crucial material phenomena in the brain that might be causally linked to consciousness and subjective experience. In any case, we expect that serious ethical issues will have to be dealt with in the future brain nanotechnology.

REFERENCES

1. M. Arbib, Ed., "The Handbook of Brain Theory and Neural Networks," 2nd edn. MIT Press, Cambridge, MA, 2003.
2. S. Amari, *Proc. IEEE* 78, 1143 (1990).
3. T. Kohonen, *Proc. IEEE* 78, 1464 (1990).
4. G. Carpenter and S. Grossberg, "Pattern Recognition by Self-Organizing Neural Networks," MIT Press, Cambridge, MA, 1991.
5. C. M. Bishop, "Neural Networks for Pattern Recognition," Oxford Univ. Press, Oxford, 1995.
6. S. Amari and N. Kasabov, Eds., "Brain-Like Computing and Intelligent Information Systems," Springer-Verlag, Singapore, 1998.
7. P. Baldi and S. Brunak, "Bioinformatics—A Machine Learning Approach," 2nd edn. MIT Press, Cambridge, MA, 2001.
8. J. Bower and H. Bolouri, Eds., "Computational Modelling of Genetic and Biochemical Networks," MIT Press, Cambridge, MA, 2001.
9. H. de Jong, *J. Comput. Biol.* 9, 67 (2002).
10. N. Kasabov, "Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering," MIT Press, Cambridge, MA, 1996.
11. V. Crunelli and N. Leresche, *Nat. Rev. Neurosci.* 3, 371 (2002).
12. H.-H. Ropers, M. Hoeltzenbein, V. Kalscheuer, H. Yntema, B. Hamel, J.-P. Fryns, J. Chelly, M. Partington, J. Gecz, and C. Moraine, *Trends Genet.* 19, 316 (2003).
13. Y. Zhang, F. Schlachetzki, and W. M. Pardridge, *Mol. Ther.* 7, 11 (2003).
14. E. R. Kandel, J. H. Schwartz, and T. M. Jessell, "Principles of Neural Science," 4th edn. McGraw-Hill, New York, 2000.
15. D. Hebb, "The Organization of Behavior," Wiley, New York, 1949.
16. W. C. Abraham, B. Logan, J. M. Greenwood, and M. Dragunow, *J. Neurosci.* 22, 9626 (2002).
17. H. Z. Shouval, M. F. Bear, and L. N Cooper, *Proc. Natl. Acad. Sci. USA* 99, 10831 (2002).
18. H. Markram, J. Lübke, M. Frotscher, and B. Sakmann, *Science* 275, 213 (1997).
19. W. C. Abraham and M. F. Bear, *Trends Neurosci.* 19, 126 (1996).
20. E. Bienenstock, L. N. Cooper, and P. Munro, *J. Neurosci.* 2, 32 (1982).
21. P. Jedlicka, *Bratislava Med. Lett.* 103, 137 (2002).
22. L. Benuskova, M. F. Diamond, and F. F. Ebner, *Proc. Natl. Acad. Sci. USA* 91, 4791 (1994).

23. L. Benuskova, V. Rema, M. Armstrong-James, and F. F. Ebner, *Proc. Natl. Acad. Sci. USA* 98, 2797 (2001).
24. T. V. P. Bliss, *Nature* 401, 25 (1999).
25. A. Kral, R. Hartmann, J. Tillein, S. Heid, and R. Klinke, *Cerebral Cortex* 12, 797 (2002).
26. J. J. Hopfield, *Proc. Natl. Acad. Sci. USA* 79, 2554 (1982).
27. E. Rodriguez, N. George, J.-P. Lachaux, J. Martinerie, B. Renault, and F. J. Varela, *Nature* 397, 434 (1999).
28. W. Maass and C. M. Bishop, Eds., "Pulsed Neural Networks," MIT Press, Cambridge, MA, 1999.
29. F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek, "Spikes—Exploring the Neural Code," MIT Press, Cambridge, MA, 1996.
30. M. Fabre-Thorpe, A. Delorme, C. Marlot, and S. Thorpe, *J. Cogn. Neurosci.* 13, 171 (2001).
31. J. Huxter, N. Burgess, and J. O'Keefe, *Nature* 425, 828 (2003).
32. P. Fries, P. R. Roelfsema, A. K. Engel, P. Koenig, and W. Singer, *Proc. Natl. Acad. Sci. USA* 94, 12699 (1997).
33. M. N. Shadlen and W. T. Newsome, *J. Neurosci.* 18, 3870 (1998).
34. N. Kasabov, "Evolving Connectionist Systems. Methods and Applications in Bioinformatics, Brain Study and Intelligent Machines," Springer-Verlag, London, 2003.
35. T. Kohonen, "Self-Organizing Maps," 2nd edn, Springer-Verlag, Berlin, 1997.
36. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Eds., Learning internal representations by error propagation, "Parallel Distributed Processing: Explorations in the Microstructure of Cognition" (D. E. Rumelhart and J. L. McClelland, Eds.), MIT Press/Bradford Books, Cambridge, MA, 1986.
37. P. Werbos, *Proc. IEEE* 87, 10 (1990).
38. J. S. Albus, *Trans. ASME J. Dynam. Syst. Meas. Control* 27, 220 (1975).
39. B. Fritzke, *Adv. Neural Information Processing Syst.* 7, 625 (1995).
40. D. Saad, Ed., "On-Line Learning In Neural Networks," Cambridge Univ. Press, Cambridge, 1999.
41. B. Hassibi and D. G. Stork, in "Advances in Neural Information Processing Systems" (D. S. Touretzky, Ed.), Vol. 4, pp. 164–171, Morgan Kaufmann, San Francisco, CA, 1992.
42. S. Schaal and C. Atkeson, *Neural Comput.* 10, 2047 (1998).
43. G. Carpenter, S. Grossberg, N. Markuzon, J. H. Reynolds, and D. B. Rosen, *IEEE Trans. Neural Networks* 3, 698 (1991).
44. X. Yao, *Intl. J. Neural Syst.* 4, 203 (1993).
45. D. B. Fogel, "Evolutionary Computation—Toward a New Philosophy of Machine Intelligence," IEEE Press, New York, 1995.
46. M. Watts and N. Kasabov, in "Proceedings of the 5th International Conference on Neural Information Processing" (S. Usui and T. Omori, Eds.), Vol. 2, pp. 793–796, IOS Press, Kitakyushu, 1998.
47. S. Grossberg, *J. Stat. Phys.* 1, 319 (1969).
48. S. Grossberg, "Studies of Mind and Brain," Reidel, Boston, 1982.
49. F. Rosenblatt, "Principles of Neurodynamics," Spartan Books, New York, 1962.
50. M. Arbib, "The Metaphorical Brain—An Introduction to Cybernetics as Artificial Intelligence and Brain Theory," Wiley Interscience, New York, 1972.
51. M. Arbib, "Brains, Machines and Mathematics," Springer, Berlin, 1987.
52. T. M. Heskes and B. Kappen, in "Mathematic Foundations of Neural Networks," pp. 199–233, Elsevier, Amsterdam, 1993.
53. C. Fahlman and C. Lebiere, "The Cascade-Corelation Learning Architecture, in Advances in Neural Information Processing Systems," (D. S. Touretzky, Ed.), Vol. 2, Morgan Kaufmann, San Francisco, CA, 1990.
54. G. A. Rummery and M. Niranjan, in Cambridge University, Engineering Department, 1994.
55. A. Sankar and R. J. Manmone, *IEEE Trans. Comput.* 42, 291 (1993).
56. Y. LeCun, J. S. Denker, and S. A. Solla, in "Advances in Neural Information Processing Systems" (D. S. Touretzky, Ed.), pp. 598–605, Morgan Kaufmann, San Francisco, CA, 1990.
57. M. Ishikawa, *Neural Networks* 9, 501 (1996).
58. A. Robins, *Connection Sci.* 8, 259 (1996).
59. D. Miller, J. Zurada, and J. H. Lilly, in "Proceedings of the IEEE International Conference on Neural Networks," Vol. 1, pp. 448–454, 1996.
60. V. Keeman, "Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models (Complex Adaptive Systems)," MIT Press, Cambridge, MA, 2001.
61. G. E. Hinton, *Artificial Intelligence* 40, 185 (1989).
62. G. E. Hinton, *Artificial Intelligence* 46, 1 (1990).
63. G. G. Towell and J. W. Shavlik, *Machine Learning* 13, 71 (1993).
64. G. G. Towell and J. W. Shawlik, *Artificial Intelligence* 70, 119 (1994).
65. I. Cloete and J. Zurada, Eds., "Knowledge-Based Neurocomputing," MIT Press, Cambridge, MA, 2000.
66. G. G. Towell, J. W. Shawlik, and M. Noordewier, "Refinement of Approximate Domain Theories by Knowledge-Based Neural Networks," in Proceedings of 8th National Conference AI, AAAI Press/MIT Press, Boston, MA, pp. 861–866, 1990.
67. T. Furuhashi, K. Nakaoka, and Y. Uchikawa, "A new Approach to Genetic Based Machine Learning and an Efficient Finding of Fuzzy Rules," Proceedings of WWW'94 Workshop, pp. 114–122, IEEE/Nagoya University, Nagoya, Japan, 1994.
68. R. Jang, *IEEE Trans. Syst. Man Cybernetics* 23, 665 (1993).
69. T. Yamakawa, H. Kusanagi, E. Uchino, and T. Miki, "A new Effective Algorithm for Neo Fuzzy neuron Model," Proceedings of the 5th IFSA World Congress, pp. 1017–1020, IFSA, Seoul, Korea, 1993.

70. W. Hauptmann and K. Heesche, "A Neural Network Topology for Bidirectional Fuzzy-Neuro Transformation," Proceedings of FUZZ-IEEE/IFES, pp. 1511–1518, IEEE/IFES, Yokohama, Japan, 1995.
71. N. Kasabov, *Neurocomputing* 13, 95 (1996).
72. Y. Hayashi, in "Advances in Neural Information Processing Systems" (R. P. Lippman, J. E. Moody, and D. S. Touretzky, Eds.), Vol. 3, pp. 578–584, Morgan Kaufmann, San Mateo, CA, 1991.
73. N. Kasabov, *Fuzzy Sets Syst.* 82, 2 (1996).
74. W. Duch, R. Adamczak, and K. Grabczewski, *Neural. Proc. Lett.* 7, 211 (1998).
75. N. Kasabov, in "Methodologies for the Conception, Design and Application of Soft Computing" (T. Yamakawa and G. Matsumoto, Eds.), pp. 271–274, World Scientific, Singapore, 1998.
76. S. Mitra and Y. Hayashi, *IEEE Trans. Neural Networks* 11, 748 (2000).
77. N. Kasabov, *IEEE Trans. Syst. Man Cybernetics B. Cybernetics* 31, 902 (2001).
78. S. Amari, *IEEE Trans. Electronic Comput.* 16, 299 (1967).
79. D. W. Aha, D. Kibler, and M. K. Albert, *Machine Learning* 6, 37 (1991).
80. M. T. Mitchell, R. Keller, and S. Kedar-Cabelli, *Machine Learning* 1, 47 (1997).
81. S. L. Salzberg, "Learning with Nested Generalized Exemplars," Kluwer, Boston, MA, 1990.
82. S. Haykin, "Neural Networks—A Comprehensive Foundation," Prentice Hall, Engelwood Cliffs, NJ, 1994.
83. L. A. Zadeh, *Information Control* 8, 338 (1965).
84. E. Mamdani, *IEEE Trans. Comput.* 26, 1182 (1997).
85. T. Takagi and M. Sugeno, *IEEE Trans. Syst. Man Cybernetics* 15, 116 (1985).
86. J. M. Mendel, "Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions," Prentice Hall, New York, 2001.
87. J. G. Taylor, "The Race for Consciousness," MIT Press, Cambridge, MA, 1999.
88. W. Freeman, "Neurodynamics," Springer-Verlag, London, 2000.
89. D. E. Goldberg, "Genetic Algorithms in Search, Optimisation and Machine Learning," Addison-Wesley, Reading, MA, 1989.
90. J. Koza, "Genetic Programming," MIT Press, Cambridge, MA, 1992.
91. J. H. Holland, "Adaptation in Natural and Artificial Systems," Univ. of Michigan Press, Ann Arbor, 1975.
92. J. H. Holland, "Emergence," Oxford Univ. Press, Oxford, 1998.
93. D. Fogel, L. Fogel, and V. Porto, *Biol. Cybernetics* 63, 487 (1990).
94. J. M. Baldwin, *Am. Nat.* 30, 441 (1896).
95. V. Kvasnicka and J. Pospichal, in "Advances in Soft Computing—Engineering Design and Manufacturing" (R. Roy, T. Furuhashi, and P. K. Chawdhry, Eds.), pp. 481–496, Springer-Verlag, London, 1999.
96. K. Richardson, "The Making of Intelligence," Phoenix, London, 1999.
97. A. Newell and H. A. Simon, "Human Problem Solving," Prentice Hall, Engelwood Cliffs, NJ, 1972.
98. H. C. Plotkyn, "The Nature of Knowledge," Penguin, London, 1994.
99. E. Rosch and B. B. Lloyd, Eds., "Cognition and Categorization," Lawrence Erlbaum, Mahwah, NJ, 1978.
100. E. E. Smith and D. L. Medin, "Categories and Concepts," Harvard Univ. Press, Cambridge, MA, 1981.
101. D. Fogel, "Blondie 24 Playing at the Edge of AI," Morgan Kaufmann, San Diego, CA, 2002.
102. J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, and G. G. Sutton, *Science* 291, 1304 (2001).
103. C. Elegans, Sequencing Consortium, *Science* 282, 2012 (1998).
104. F. Crick, *Nature* 227, 561 (1970).
105. S. Kauffman, *J. Theoret. Biol.* 44, 167 (1974).
106. J. L. deRisi, V. R. Iyer, and P. O. Brown, *Science* 275, 680 (1997).
107. B. Sobral, in "From Jay Lush to Genomics: Visions for Animal Breeding and Genetics" (J. M. Dekkers, S. J. Lamont, M. F. Rothschild, Eds.), Iowa State Univ. Press, Ames, 1999.
108. P. A. Pevzner, "Computational Molecular Biology: An Algorithmic Approach," MIT Press, Cambridge, MA, 2000.
109. J. R. Koza, W. Mydlowec, G. Lanza, J. Yu, and M. A. Keane, "Reverse Engineering of Metabolic Pathways from Observed Data using Genetic Programming," in Proceedings of the Pacific Symposium on Biocomputing, Vol. 6, pp. 434–445, ISCB, Hawaii, 2001.
110. J. Collado-Vides and R. Hofestadt, Eds., "Gene Regulation and Metabolism, Post-Genomic Computational Approaches," MIT Press, Cambridge, MA, 2002.
111. L. Hunter, *Can. Artificial Intelligence* 35, 10 (1994).
112. J. Dow, G. Lindsay, and J. Morrison, "Biochemistry Molecules, Cells and the Body," Addison-Wesley, Boston, MA, 1995.
113. J. Collado-Vides, B. Magasanik, and T. F. Smith, Eds., "Integrative Approaches to Molecular Biology," MIT Press, Cambridge, MA, 1996.
114. P. D'Haeseleer, X. Wen, S. Fuhrman, and R. Somogyi, "Linear Modeling of mRNA Expression Levels during CNS Development and Injury," Proceedings of the Pacific Symposium on Biocomputing, Vol. 4, pp. 41–52, ISCB, Hawaii, 1999.
115. P. D'Haeseleer, S. Liang, and R. Somogyi, *Bioinformatics* 16, 707 (2000).
116. G. Fogel and D. Corne, "Evolutionary Computation for Bioinformatics," Morgan Kaufmann, San Francisco, CA, 2003.
117. T. Akutsu, S. Miyano, and S. Kuhara, "Identification of Genetic Networks from a Small Number of Gene Expression Patterns under the Boolean Network Model," Proceedings of the Pacific Symposium on Biocomputing, Vol. 4, pp. 17–28, ISCB, Hawaii, 1999.
118. S. Gomez, S. Lo, and A. Rzhetsky, *Genetics* 159, 1291 (2001).

119. A. Lindlof and B. Olsson. *Information Sci.* 146, 103 (2002).
120. J. Vohradsky. *J. Biol. Chem.* 276, 36168 (2001).
121. J. Vohradsky. *FASEB J.* 15, 846 (2001).
122. S. Ando, E. Sakamoto, and H. Iba, "Evolutionary Modelling and Inference of Genetic Networks," Proceedings of the 6th Joint Conference on Information Sciences, pp. 1249–1256. Cary, NC. 2002.
123. Mimura and H. Iba, "Inference of a Gene Regulatory Network by Means of Interactive Evolutionary Computing," Proceedings of the 6th Joint Conference on Information Sciences, pp. 1243–1248. Cary, NC. 2002.
124. K. W. Kohn and D. S. Dimitrov, in "Computer Modeling and Simulation of Complex Biological Systems" (S. S. Iyengar, Ed.), pp. 101–123. CRC Press, Boca Raton, FL, 1998.
125. J. Vides, B. Magasanik, and T. Smith, "Integrated Approaches to Molecular Biology," MIT Press, Cambridge, MA, 1996.
126. M. A. Savageau, *Chaos* 11, 142 (2001).
127. G. Marnellos and E. D. Mjolsness, in "Modeling Neural Development" (A. vanOoyen, Ed.), pp. 27–48. MIT Press, Cambridge, MA, 2003.
128. L. F. A. Wessels, E. P. vanSomeren and M. J. T. Reinders, in "Proceedings of the Pacific Symposium on Biocomputing," p. 508–519. ISCB, Hawaii, 2001.
129. N. Kasabov and D. Dimitrov, in "Proceedings of the ICONIP'2002 –International Conference on Neuro-Information Processing," IEEE Press, Singapore, 2002.
130. N. Kasabov and Q. Song, DENFIS: *IEEE Trans. Fuzzy Syst.* 10, 144 (2002).
131. The MathWorks, "Neural Network Toolbox User's Guide," Vol. 4. The Math Works Inc., Natick, MA, 2001.
132. W. Enard, P. Khaitovich, J. Klose, S. Zoelner, F. Heisig, and S. Paabo, *Science* 296, 340 (2002).
133. M. Caceres, J. Lachuer, M. A. Zapala, J. C. Redmond, L. Kudo, D. H. Geschwind, D. J. Lockhart, T. M. Preuss, and C. Barlow, *Proc. Natl. Acad. Sci. USA* 100, 13030 (2003).
134. S. Savage-Rumbaugh and R. Lewin, "Kanzi: The Ape at the Brink of the Human Mind," Wiley, New York, 1994.
135. R. Somogyi, S. Fuhrman, and X. Wen, in "Computational Modeling of Genetic and Biochemical Networks" (J. M. Bower and H. Bolouri, Eds.), pp. 119–157. MIT Press, Cambridge, MA, 2001.
136. In "Genes and Disease," National Centre for Biotechnology Information (NCBI), 2003; http://www.ncbi.nlm.nih.gov/books/bv.fegi?call=bv.View..ShowSection&rid=gnd.chapter.75.
137. R. Morita, E. Miyazaki, C. G. Fong, X.-N. Chen, J. R. Korenberg, A. V. Delgado-Escueta, and K. Yamakawa, *JH8, Biochem. Biophys. Res. Comm.* 248, 307 (1998).
138. N. Kasabov, L. Bezuskova, and S. G. Wysoski, "Computational Neurogenetic Modelling of Gene Networks within Neural Networks," Proceeding of IEEE International Conference NN, Vol. 2, pp. 1203-1028. IEEE, Budapest, 2004.
139. J. H. Holland, "Adaptation in Natural and Artificial Systems," Univ. of Michigan Press, Ann Arbor, MI, 1975.
140. D. E. Goldberg, "Genetic Algorithms in Search, Optimization and Machine Learning," Addison-Wesley, Reading, MA, 1989.
141. W. Gerstner and W. M. Kistler, "Spiking Neuron Models," Cambridge Univ. Press, Cambridge, 2002.
142. A. Destexhe, *J. Neurosci.* 18, 9099 (1998).
143. A. V. Semyanov, *Neurophysiology* 34, 71 (2002).
144. J. W. Freeman, "Mass Action in the Nervous System," Academic Press, New York, 1975.
145. W. J. Freeman, M. D. Holmes, B. V. Burke, and S. Vanhatalo, *Clin. Neurophysiol.* 114, 1053 (2003).
146. P. Kudela, P. J. Franaszczuk, and G. K. Bergey, *Biol. Cybernetics* 88, 276 (2003).
147. K.-H. Yang, P. J. Franaszczuk, and G. K. Bergey, *Biol. Cybernetics* 89, 242 (2003).
148. J. M. Bower and D. Beeman, "The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System," 2nd ed. TELOS/Springer-Verlag, New York, 1998.
149. C. Marini, L. A. Harkin, R. H. Wallace, J. C. Mulley, I. E. Scheffer, and S. F. Berkovic, *Brain* 126, 230 (2003).
150. H. K. M. Meeren, J. P. M. Pijn, E. L. J. M. VanLuijtelaar, A. M. L. Coenen, and F. H. LopesdaSilva, *J. Neurosci.* 22, 1480 (2002).
151. A. Contractor, G. T. Swanson, A. Sailer, S. O'Gorman, and S. F. Heinemann, *J. Neurosci.* 20, 8269 (2000).
152. R. A. Deisz, *Neuropharmacology* 38, 1755 (1999).
153. R. M. Gardiner, *Epilepsy Res.* 36, 91 (1999).
154. G. E. Homanics, T. DeLorey, L. L. Firestone, J. J. Quinlan, and A. Handforth, *Proc. Natl. Acad. Sci. USA* 94, 4143 (1997).
155. S. Jones, S. Sudweeks, and J. L. Yakel, *Trends Neurosci.* 22, 555 (1999).
156. O. K. Steinlein and J. L. Noebeles, *Curr. Opin. Genet. Dev.* 10, 286 (2000).
157. R. A. Freitas, "Nanomedicine, Volume I: Basic Capabilities," Landes Bioscience, Georgetown, TX, 1999.
158. R. A. Freitas, "Nanomedicine, Volume IIA: Biocompatibility," Landes Bioscience, Georgetown, TX, 2003.
159. T. J. Webster, M. C. Waid, J. L. McKenzie, R. L. Price, and J. U. Ejiofor, *Nanotechnology* 15, 48 (2004).
160. A. S. Sadek, K. Nikolic, and M. Forshaw, *Nanotechnology* 15, 192 (2004).
161. G. Bauer, J. Hassmann, H. Walter, J. Haglmueller, C. Mayer, and T. Schalkhammer, *Nanotechnology* 15, 1289 (2004).
162. L. S. Smith and A. Hamilton, Eds., "Neuromorphic Systems: Engineering Silicon from Neurobiology," Progress in Neural Processing, World Scientific Publishing, London, 1998.
163. P. Tikovic, M. Voros and D. Durackova, *J. Elec. Eng.* 52, 68 (2001).
164. J. Theiler, *Phys. Lett. A* 196, 335 (1995).
165. W. J. Freeman, *Chaos* 13, 1 (2003).

166. I. Tsuda, *Behav. Brain Sci.* 24, 793 (2001).
167. P. Bak, C. Tang, and K. Wiesenfeld, *Phys. Rev. Lett.* 59, 381 (1987).
168. L. Benuskova, M. Kanich, and A. Krakovska, in "Proceedings of the World Congress on Neuroinformatics" (F. Rattay, Ed.). ARGESIM/ASIM-Verlag, Vienna, 2001.
169. C. Koch and F. Crick, in "Large-Scale Neuronal Theories of the Brain" (C. Koch and J. L. Davis, Eds.). MIT Press, Cambridge, MA, 1994.
170. W. Singer, in "Understanding Representation in the Cognitive Sciences" (A. Riegler, M. Peschl, and A. vonStein, Eds.). Kluwer Academic/Plenum Publishers, New York, 1999.
171. G. M. Edelman and G. Tononi, "Consciousness. How Matter Becomes Imagination." Penguin Books, London, 2000.
172. R. Penrose, "Shadows of the Mind: A Search for the Missing Science of Consciousness." Oxford Univ. Press, Oxford, 1994.
173. R. Penrose, Beyond the Doubting of a Shadow, *PSYCHE:* 2 (1996); http://psyche.cs.monash.edu.au/v2/psyche-2-23-penrose.html.
174. D. C. Dennett, "Consciousness Explained." Penguin Books, New York, 1991.
175. J. Searle, "Consciousness and Language." Cambridge Univ. Press, Cambridge, MA, 2002.
176. D. J. Chalmers, "The Conscious Mind: In Search of a Fundamental Theory." Oxford Univ. Press, Oxford, 1996.
177. D. J. Chalmers, Minds, Machines, and Mathematics, *PSYCHE:* 2 (1996); http://psyche.cs.monash.edu.au/v2/psyche-2-09-chalmers.html.

CHAPTER 18

Molecular Computation Using Hairpins and Secondary Structures of DNA

Masami Hagiya

Department of Computer Science, Graduate School of Information Science and Technology, University of Tokyo, Japan

CONTENTS

1. INTRODUCTION

Molecular computing analyzes the computational power of molecular reactions and applies it to engineering and medicine. Since DNA molecules with any given base sequence can be synthesized and there is selective hybridization between complementary DNA base sequences, DNA is mainly used for molecular computation.

Just as proteins fold into three-dimensional structures through interactions among their amino bases, DNA molecules can also take various forms other than the ordinary double helix through hybridization between complementary base sequences, including hairpins, bulges, and various junctions and tiles. Our group has been studying the computational

ISBN: 1-58883-048-9

Handbook of Theoretical and Computational Nanotechnology
Edited by Michael Rieth and Wolfram Schommers
Volume 6: Pages (817–839)

power of simple hairpin DNA structures and implementing various computational models based on hairpins. This has revealed that even hairpins have interesting computational applications and suggests that more complex structures will allow more powerful computation. The work is explained in detail in Section 3 in this chapter.

Before explaining computational models based on hairpins, we survey the current directions of molecular computing research in Section 2. The survey includes research on DNA self-assembly, in particular, computation by self-assembly of DNA tiles as advocated by Seeman and Winfree. Research on self-assembly has led to research on nanorobotics (i.e., research efforts to implement dynamic molecular devices). We then summarize future perspectives in the field, including self-assembly and nanorobotics. At the end of Section 2, we also emphasize the importance of establishing systematic design principles for molecules and reactions because the field still relies on ad hoc trial and error for designing and implementing molecular systems. We call these systematic design principles "molecular programming."

Molecular computing is still an emerging field, but like traditional research fields, it is based on a deep understanding of nature, and an engineering discipline to apply this deep understanding is developing. For molecular computing, a deep understanding of nature means understanding the computational power of molecular reactions, and an engineering discipline refers to the principles for designing and implementing molecular systems. We expect that progress in molecular computing will contribute greatly to both science and engineering.

2. PERSPECTIVES OF MOLECULAR COMPUTING

2.1. DNA and Molecular Computing

Molecular computing (or biomolecular computing) is a research field that proposes, analyzes, and implements computational models based on molecular reactions. A number of computational models using DNA molecules have been proposed, including the model proposed by Adleman [1] and refined by Lipton and others [2–4], and the model of computation by DNA self-assembly proposed by Seeman and Winfree [5].

After proposing computational models, their computational power is analyzed. As with conventional computational models, the analysis is made with respect to time and space; in molecular computation, this roughly corresponds to the numbers of experimental operations and molecules, respectively. Recently, efforts have been made to analyze molecular computation in ways that are more faithful to actual molecular reactions, regarding reactions as random processes and including errors and reaction yields in the analysis. The time needed to finish each experimental operation is also estimated by analyzing the time needed to reach some stable molecular state.

Needless to say, the technology for implementing proposed computational models is also one of the main research topics in molecular computing. For example, a method for designing a good set of DNA sequences to use for computation is still one of the most important requirements for implementing molecular computation [6–9].

Finally, this discipline aims to apply molecular computational models to other fields, including biotechnology, nanotechnology, and molecular evolution. Although the original motivation behind molecular computing (i.e., realizing a new kind of computer based on molecular reactions that outperforms existing electronic computers) is still a dream, information processing at the molecular level can be applied to various technologies. In particular, molecular computation, which can handle molecular inputs and outputs directly, is expected to lead to new types of biotechnology, including genetic analysis, medical diagnosis, and drug delivery. Needless to say, self-assembly of DNA is also a uniquely promising method for bottom-up nanotechnology.

To realize such applications, molecular computation should be implemented efficiently and robustly. In the early period of molecular computing, computer scientists were mostly interested in the theoretical analysis of molecular computation. Currently, they are becoming increasingly serious about applying their computational expertise to the design and implementation of molecular systems. In other words, the focus of their research has shifted from the computational models themselves, to their design and implementation. We call

this research trend "molecular programming," and its goal is to establish systematic design principles for molecular computation.

In this section, we briefly review the field of molecular computing. First, we explain two important molecular computation models: those of Adleman, and those of Seeman and Winfree. Then, we describe the efforts to implement nanoscale machines using DNA molecules, by emphasizing applications in biotechnology and nanotechnology. Next, we summarize the future perspectives of molecular computing. Finally, we touch on some methods for designing and implementing molecular systems, which we call "molecular programming."

2.2. Adleman's First DNA Computer

Since Adleman's work [1] has been cited and explained in a huge number of books and papers, we will explain his work very briefly. It is a method for solving combinatorial (or optimization) problems using a simple "generate and test" strategy. It is also called "data-parallel" computation by molecules because each DNA molecule holds a distinct piece of data and each experimental operation is applied simultaneously (in parallel) to all the molecules in the tube.

Adleman's model of molecular computation consists of two steps. First, candidate solutions of a given combinatorial problem are generated randomly as DNA molecules. Adleman was clever in the problem he chose for his seminal work. He chose the Hamiltonian path problem, which involves finding a path in a given directed graph that visits each node in the graph exactly once. This problem is a simple version of the so-called traveling salesman problem, the problem of finding the optimal flight route for a traveling salesman who wants to visit each city exactly once in his sales travel.

Theoretically, the Hamiltonian path problem is classified as nondeterministic polynomial time complete (NP-complete). In general, problems called NP-complete are computationally very difficult to solve and require an exponential amount of computation time.

To solve the Hamiltonian path problem, Adleman encoded each node (city) and each edge between nodes (single flight between cities) as a 20-base single-stranded DNA molecule. The 20-base sequence encoding an edge consists of two parts. The first 10-base sequence can hybridize with the sequence of the node from which the edge emanates, and the last 10-base sequence can hybridize with the node that the edge reaches (Fig. 1). In this chapter, thick straight lines denote single-stranded DNA molecules (or fragments thereof), and arrows denote the direction of each strand from 3′ to 5′. Thin straight lines denote hydrogen bonds between complementary bases. Note that the number of bases in figures does not always reflect actual implementations.

Encoding the problem in this manner, if all the strands encoding the nodes and edges of a given graph are put into a tube, then double-stranded DNA molecules, each encoding a path over the graph, are generated randomly. Note that this process of random generation based on DNA hybridization is completely autonomous, unlike other computations required to solve the Hamiltonian path problem.

After double-stranded DNA molecules are generated by hybridization, the tube is incubated with ligase so that adjacent sequences in each double-stranded DNA molecule are connected by covalent bonds. As a result, a random pool of paths over the given graph is

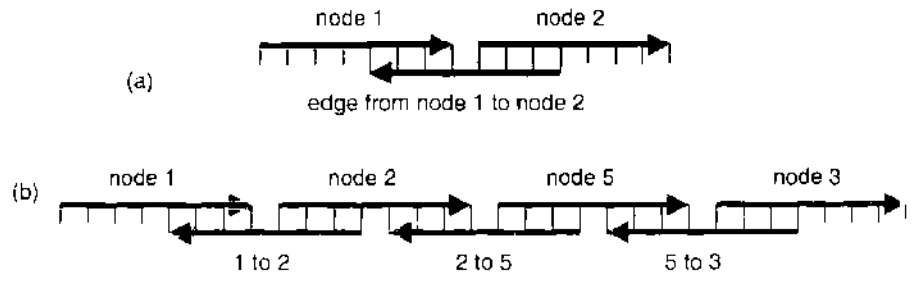


Figure 1. (a) Hybridization of an edge and two nodes. The sequence encoding the edge from node 1 to node 2 hybridizes with the sequences of node 1 and node 2. (b) A path over a graph. Successive hybridization generates a path over the graph as a double-stranded DNA molecule.

obtained, in which each molecule represents a candidate solution for the problem (i.e., a Hamiltonian path over the graph). Formally, a Hamiltonian path is a path over the given graph, beginning and ending with the specified start and goal nodes, which visits each node exactly once.

In the second step of Adleman's computational model, the pool of candidates is processed in sequence to extract real solutions to the problem (i.e., the Hamiltonian paths of the problem). We will explain this second step briefly. First, the pool is processed by polymerase chain reaction (PCR) using the appropriate primers, so that only those paths that begin with the specified start node and ending with the goal node are amplified. Second, molecules of the right length are extracted by gel electrophoresis. Note that molecules encoding a Hamiltonian path should have a fixed length because a Hamiltonian path consists of a fixed number of nodes. The resulting pool is then processed by affinity separation using magnetic beads to extract those paths that contain all the nodes of the graph. This last extraction process requires a number of operations proportional to the number of nodes in the graph because each node (other than the start and goal nodes) should be used for affinity separation. In this second step, each extraction operation is applied to all the data in the tube in parallel, and the number of operations grows proportionally with the size of the given problem. These two features and the first step of the random generation by DNA hybridization characterize Adleman's model of DNA computation.

Although Adleman's seminal work initiated the research field of molecular computing (especially DNA computing), the model still has a long way to go to outperform electronic computers at solving combinatorial (or optimization) problems, as was once expected. The main obstacle to the model can be summarized as the "scale-up problem." Nevertheless, there has been steady progress with the model. Adleman et al. solved a 20-variable SAT problem (another form of combinatorial problem) [10]. Yoshida and Suyama solved a 10-variable 43-clause SAT problem using a more sophisticated algorithm [11]. With the current implementation technology, however, it is not efficient to scale up the size of the problems that can be solved using molecular computation to the level of the most powerful algorithms for those problems on electronic computers.

Regardless of whether the original goal is achieved, Adleman's work will continue to remain important as a proof of the concept that molecules can really compute because it showed that molecules could actually solve a concrete mathematical problem.

Ironically, Adleman's model can be used as a benchmark for measuring the performance of newly developed biotechnology devices, such as automatic robots and microreactors, which perform experimental operations. Solving a mathematical problem is an appropriate benchmark because the result of the computation is known in advance (with no ambiguity) and the size of a benchmark can be adjusted easily according to the features of the new device.

A more positive future direction of Adleman's model is to use data-parallel computation for multiplexing genomic analysis, such as gene expression profiling and single-nucleotide polymorphism (SNP) typing, as Sakakibara and Suyama advocate [12]. Adleman's model can also be used to consider molecular evolution, as discussed in Section 2.5.

2.3. Computation by DNA Self-Assembly

One of the most important models for molecular computation other than Adleman's is undoubtedly the model of computation using DNA self-assembly proposed by Seeman and Winfree. In Adleman's model, the random generation of molecules, which are candidate solutions for the given problem, results from autonomous DNA hybridization. Seeman and Winfree's model gives this process the main role in computation [13]. Unlike the random generation of double-stranded DNA molecules in Adleman's model, Seeman and Winfree's model uses both linear double-stranded DNA and various structured DNA molecules for self-assembly (Fig. 2).

These structured molecules include so-called DNA tiles, which are unique in that they self-assemble further to form planar structures. Each DNA tile (DAO tile in the figure) consists of four single-stranded DNA molecules (i.e., it consists of two double-stranded DNA molecules coupled by exchanging their strands at two points). In this sense, it is also called a "double-crossover" tile. Since a DNA tile has four sticky single-stranded ends, it

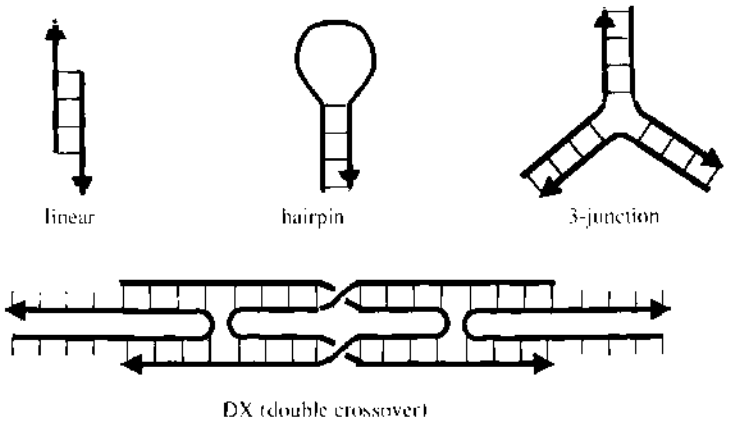


Figure 2. Various kinds of structured DNA molecules.

can hybridize with four other tiles, forming a planar structure (Fig. 3). The sticky ends hybridize selectively with their complementary sequences. Therefore, DNA tiles with different sequences as their sticky ends behave like different components during self-assembly.

Seeman studied such structured DNA molecules and their self-assembly for years before Winfree realized the computational power of self-assembly for molecular computing [14, 15]. Winfree's analysis of the computational power of DNA self-assembly can be succinctly stated as follows [13]. The self-assembly of linear double-stranded molecules has the power of regular languages in formal language theory, in the sense that the self-assembly and ligation of linear molecules produce a set of strands that can be considered a regular language over the alphabet {A, G, C, T}. Similarly, the self-assembly and ligation of linear molecules with hairpins and three junctions can produce context-free languages. Finally, the self-assembly and ligation of DNA tiles are universally computable, in the sense that they can simulate the execution of any Turing machine. This is due to the fact that two-dimensional tiling is universally computable in general, because it can simulate any one-dimensional system of cellular automata, which in turn can simulate any Turing machine.

Winfree and Seeman first succeeded in constructing planar structures of stripes consisting of two kinds of DNA tile [16]. At the time when Winfree published this result, he also mentioned the construction of Sierpinski's triangle, a typical two-dimensional fractal figure, as an interesting example of self-assembly (Fig. 4). After constructing stripes, he had to wait for several years before he recently succeeded in constructing Sierpinski's triangle from DNA [17].

Seeman and Winfree's model is now central to research in molecular computing. It now receives more attention than Adleman's model because it can contribute to the so-called

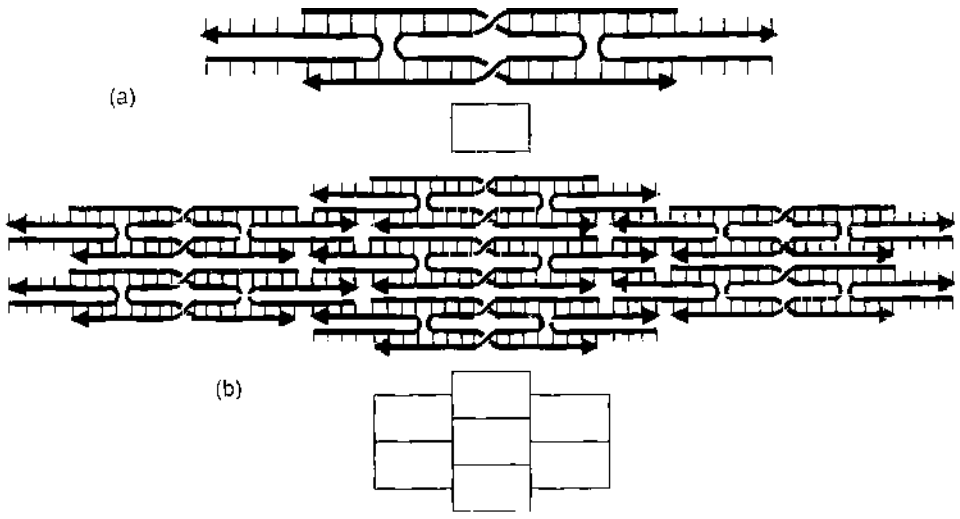


Figure 3. (a) A DNA tile (double crossover molecule) and its schematic representation. (b) Self-assembly of DNA tiles and its schematic representation.

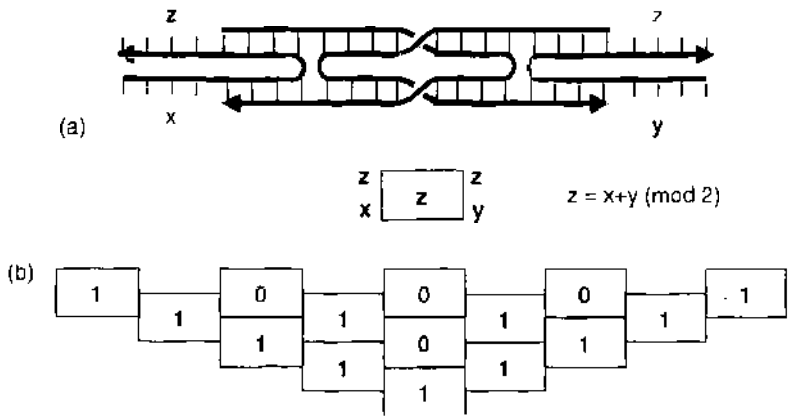


Figure 4. (a) DNA tiles for self-assembly of Sierpinski's triangle and their schematic representation. There are four combinations for x and y as 0 or 1. (b) Sierpinski's triangle generated by the DNA tiles in (a).

"bottom-up construction" of nanoscale structures directly. In this context, the role of molecular computation is clear. In order to construct complex nanoscale patterns in a bottom-up fashion, it is necessary for the process of self-assembly to perform some kind of computation autonomously. Recently, Seeman and others founded an academic society called the International Society of Nanoscale Science, Computation and Engineering, which is intended to contribute to nanotechnology via molecular computing.

Reif and LaBean's group have continued Seeman and Winfree's work. In their work on the self-assembly of triple-crossover tiles, they showed that by using triple-crossover tiles even one-dimensional self-assembly could perform interesting computations, such as bit-wise addition [18, 19]. More recently, they succeeded in the self-assembly of cross-shaped motifs, which can produce either planar structures or tubes (called DNA nanotubes), depending on the polarity of the motif [20]. As for DNA nanotubes, they coated their surface with silver and measured their conductivity as nanocables.

2.4. Nanorobotics

The focus of molecular computing on nanotechnology advocated by Winfree and Seeman has stimulated research on the construction of both static and dynamic structures using DNA molecules. Molecular structures with dynamic components will have more applications than simple static structures, and dynamic components will enhance the bottom-up construction of structures and increase their complexity. Molecular structures with dynamic components are simply called "molecular machines," and the line of research involved in constructing molecular machines is now called "nanorobotics."

The research on molecular machines made of DNA is still at a very primitive stage. Even the word "machine" is used ambiguously because it sometimes refers to molecular actuators, such as motors and transporters, while it also refers to information-processing units, such as finite-state machines or automata (or Turing machines).

For example, Yurke's DNA tweezers form a machine that closes and opens its arms depending on the existence of a so-called fuel molecule [21]. When the fuel molecule is put in a solution containing the machine, its arms close because the fuel hybridizes with the two arms. After a molecule complementary to the fuel is put in the solution, the complement removes the fuel from the machine by branch migration, so the machine reopens its arms. This machine can be regarded as an actuator that might be able to hold something, but it can also be regarded as a two-state machine that can store one bit of information in its state.

Yurke's idea of a DNA fuel for switching the state of a molecular machine has had a great effect on research in nanorobotics. The energy dissipated by the formation of double helices of DNA is probably the simplest form of energy that can be supplied to molecular machines.

By contrast, Shapiro's DNA automaton was designed as a pure information-processing machine but was recently applied as a machine that delivers a DNA drug. Shapiro's automaton consists of so-called transition molecules and input molecules [22]. One side of each

input molecule encodes the current state of the automaton and the next symbol of the input. Depending on the current state and the next symbol, one of the transition molecules hybridizes with the input molecule. A (IIS type) restriction enzyme then cleaves the input molecule to expose the next state and the next symbol. On iterating hybridization, which might be followed by ligation and digestion, the automaton successively changes its state until all the symbols in the input are processed.

Shapiro et al. recently proposed using their automaton to control the delivery of DNA drugs [23]. For this application, the original input molecule is not used as input but as part of the program controlling drug delivery. Instead, the expression levels of given genes are used as input to the machine. Each transition molecule could be prepared in such a way that it is active or inactive, depending on whether the corresponding mRNA is expressed. Overall, a simple binary tree program can be implemented to judge the combined expression levels of given genes. If the input molecule finishes a series of state transitions, then the output DNA molecule is dissipated. This DNA acts as an antisense drug against the target gene.

Shapiro's work has had an effect on research because it clearly shows the importance of molecular inputs and outputs in molecular computation and shows the applicability of molecular machines *in vivo*.

2.5. Future Perspectives of Molecular Computing

This small section briefly summarizes the future perspectives of molecular computing discussed above.

We have already mentioned the current status of the data-parallel molecular computation proposed by Adleman. We also touched on the possibility of applying data-parallel computation to multiplexing genomic analysis. In another direction, we can relate molecular evolution and Adleman's model. Remember that the second step of Adleman's model uses data-parallel computation to extract real solutions of the given problem. The model can be generalized easily to solve an optimization problem (i.e., to extract the best solution from among possible solutions of the given problem). Molecular evolution is also an optimization problem. The evaluation and extraction steps of molecular evolution correspond to the second step in Adleman's model, while the random generation step in molecular evolution corresponds to the first step in Adleman's model. With this correspondence, techniques that enhance the optimization of data-parallel computation should also enhance molecular evolution. In computer science, such techniques have been well studied in research on evolutionary computation, such as genetic algorithms and programming, which aim to solve optimization problems using evolutionary processes simulated in electronic computers. Recently, effort has been made to apply the techniques of evolutionary computation to increase the efficiency of molecular evolution.

The self-assembly of DNA has also been discussed extensively. It is now one of the main streams of "bottom-up" nanotechnology. Not only can DNA tiles be used as building blocks to construct nanoscale structures, but DNA molecules can also be used as glues to join nanoscale particles and wires. Using DNA molecules as glues, it is possible to control the self-assembly of particles and wires, just as the self-assembly of DNA tiles can be controlled.

Molecular machines have various applications, such as switches, sensors, and memory (both for storing information and for addressing). We have already touched on the possibility of using molecular machines to control the self-assembly of nanoscale structures. In other words, they are robots working in a "nanofactory." An even more promising application is in medicine; molecular machines working *in vivo* might be used to measure the cellular environment, make appropriate judgments from these measurements, and deliver a drug into the cell. Following this reasoning, we can even imagine cellular machines that are more sophisticated than molecular machines.

2.6. Molecular Programming

For the applications mentioned so far, it is necessary to establish systematic principles for designing and implementing molecular systems. In the early period of molecular computing, molecules and molecular reactions were designed and implemented in an ad hoc manner

for each molecular computation application. They were subsequently refined in an ad hoc manner. Therefore, we need to establish principles for designing molecules, for combining molecules into molecular systems, for adjusting reaction conditions, and for efficiently scheduling reactions.

We recently adopted the phrase "molecular programming" for such systematic design principles of molecular computation, to indicate that just as electronic computers are programmed, molecules and reactions should also be programmed in a systematic manner.

For DNA, designing molecules means designing DNA base sequences. In order to construct nanoscale structures or molecular machines, principles for designing base sequences that fold into given target structures are required. This problem is called "inverse folding" and has been studied for years by the Vienna group [24]. Furthermore, to design molecular machines that can change conformation dynamically, it is necessary to design sequences that fold into more than one structure (i.e., have multiple stable structures and allow conformational changes among the structures). For this purpose, methods for analyzing the energy landscape made of structures of a given base sequence are also necessary [25, 26].

Given these design principles, it will be easy to design various component molecules. For example, it will be possible to design molecules that behave as sensors for measuring environmental conditions, such as temperature and salt concentration. It will then be necessary to develop methods for combining such component molecules to form a molecular system that senses environmental conditions, makes judgments, and finally produces output to the environment. Providing such general-purpose component molecules for constructing programmable molecular systems will be the ultimate goal of molecular programming.

3. COMPUTATION BY HAIRPINS AND SECONDARY STRUCTURES

3.1. Computational Power of Secondary Structures

Just like proteins, DNA molecules fold into the three-dimensional conformation with the minimum free energy. For DNA, the free energy of a molecular conformation is mostly determined from the so-called secondary structure of the conformation.

Mathematically, the secondary structure is determined by the set of hydrogen bonds formed by pairs of bases belonging to the molecule. If a single-stranded DNA molecule consists of n bases, numbered from 1 to n, then a base pair is identified using a pair of numbers between 1 and n. A secondary structure is then defined as the set of pairs of numbers between 1 and n. Formally, since the order of numbers in a pair is insignificant and a base cannot hybridize with itself, a base pair is the set consisting of two distinct numbers between 1 and n. Furthermore, since a base can only hybridize with at most one base, if pairs $\{i, j\}$ and $\{i, k\}$ belong to a secondary structure, then j and k are identical.

For proteins, the folding problem (i.e., the problem of finding the minimum-energy structure of the given amino acid sequence) is NP-complete (i.e., computationally very difficult and requiring an exponential amount of computation), according to a suitable model of protein structures and their energy. This means that nature performs such high-performance computation when it folds a protein into its minimum-energy structure.

For DNA molecules, the problem of finding the minimum-energy secondary structure of the given base sequence is also NP-complete if we allow so-called pseudoknots in the secondary structure. A pseudoknot consists of two base pairs $\{i, j\}$ and $\{k, l\}$ such that $i < k < j < l$.

This section deals with much simpler secondary structures and shows that even if we restrict ourselves to such simple structures we can still achieve a variety of interesting molecular computations using DNA molecules.

Typical secondary structures include hairpins, stacks, bulges, and interior loops (Fig. 5). Since a hairpin consists of a single base pair $\{i, j\}$ and the bases between i and j, it is called a 1-loop (one-loop). Stacks, bulges, and interior loops consist of two base pairs $\{i, j\}$ and $\{k, l\}$, the bases between i and k, and the bases between l and j. In a stack, there are no bases between i and k and none between l and j. In a bulge, there are no bases between i and k or none between l and j.

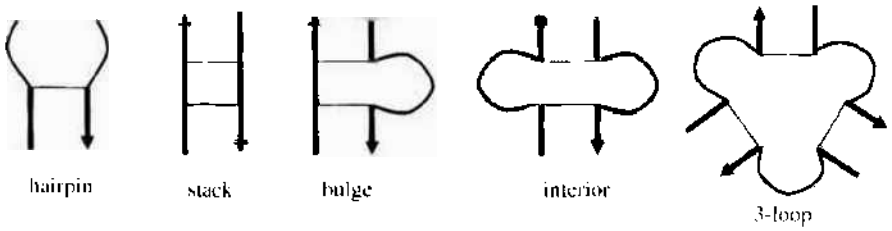


Figure 5. Various loop structures in DNA.

This section deals mainly with hairpins (i.e., the simplest secondary structures, and it also deals with bulges in Section 3.3). We show that even such simple secondary structures allow various molecular computations. Sections 3.2–3.4 focus on hairpin formation, hairpin dissociation, and iterated hairpin formation and dissociation, respectively. All the computational models in this section have been proposed and implemented by the author's research group in Japan. Although these models were developed independently, they are all based on the power of DNA molecules to form or dissociate hairpin structures. Interestingly, these models can be classified by the use of hairpins. The first is based on hairpin formation, while the second uses hairpin dissociation. The third model, the most sophisticated of the three, depends both on hairpin formation and hairpin dissociation.

In Section 3.2, we show that hairpin formation can be used to enhance Adleman's model of molecular computation by accelerating the extraction step of the model. During the process of hairpin formation within a single molecule, a subsequence of the molecule comprising a hairpin stem searches for its counterpart. This search is performed autonomously, as in the self-assembly of ordinary double-stranded DNA molecules in the first step of Adleman's model. Therefore, hairpin formation should make the second step of Adleman's model inherently more efficient. In the proposed model, the number of experimental operations required for the second step is, at least theoretically, independent of the size of the problem being solved.

In Section 3.3, we show that hairpin dissociation can be used to implement a molecular machine, whose initial conformation consists of a number of hairpins, each storing a certain amount of energy. The input to the machine is a singled-stranded DNA molecule, called an "opener," which hybridizes with a single-stranded part of the machine adjacent to one of the hairpins. Just as the complementary sequence removes the fuel molecule in Yurke's DNA tweezers using branch migration, an opener opens the hairpin adjacent to the singled-stranded part. After the hairpin is opened, another singled-stranded part is revealed, which allows another opener to hybridize and open the next hairpin of the machine. Therefore, the correct sequence of openers can produce successive conformational changes in the machine. By introducing bulges in addition to hairpins, we can even imagine a network of such molecular machines that can perform certain concurrent computations.

Finally, in Section 3.4, we show that more powerful molecular machines can be implemented by iterating hairpin formation and dissociation together with polymerase extension. The computational model is called "whiplash PCR," and we have investigated it for years. This model allows parallel computation called "multiple-program multiple-data" because each DNA molecule has its own program and can operate independently from the others. The program in each molecule consists of a set of transition rules that determine how the state of the molecule changes. The current state of the machine is represented by the subsequence of the molecule at its 3'-end. Hairpin formation and polymerase extension change the current state by attaching a new state to the 3'-end, and hairpin dissociation makes the next state change possible under isothermal conditions. This means that the machine successively and autonomously changes its state according to its own program.

3.2. Computation by Hairpin Formation

The power of a DNA molecule to form a hairpin structure can be used for various computations. As a concrete example of applying this power to computation, we solved an instance of the so-called SAT problem using DNA [27].

3.2.1. SAT Problem

Like the Hamiltonian path problem, the SAT problem is also classified as NP-complete and requires an amount of computation exponential to the size of the given instance of the problem.

SAT stands for "satisfiability." The SAT problem determines whether a given Boolean expression is satisfiable. A Boolean expression is called "satisfiable" if it can be made true by assigning appropriate truth values to its variables. For example, the Boolean expression "(x or y or not z) & (not x or y or z) & (not x or not y or not z)" can be made true if we assign truth to x, falsity to y, and truth to z, respectively. In addition to the symbol "&" meaning "and," we use the notation "~" to express "not" and "|" for "or," so the above Boolean formula is written "($x|y|\sim z$) & ($\sim x|y|z$) & ($\sim x|\sim y|\sim z$)."

This Boolean expression consists of three subexpressions linked using the "and" operator: "$x|y|\sim z$," "$\sim x|y|z$," and "$\sim x|\sim y|\sim z$." Each subexpression consists of variables, such as "x," or negations of a variable, such as "$\sim x$," combined using the "or" operator. These subexpressions are called "clauses." Since variables and negations of a variable are called "literals," a clause is a Boolean expression combining literals using the "or" operator (i.e., a clause is a "disjunctions" of literals). Finally, Boolean expressions that are obtained by combining clauses using the "and" operator are called "clausal forms" (or conjunctive normal forms) (i.e., a clausal form is a "conjunction" of clauses). Therefore, the Boolean expression "($x|y|\sim z$) & ($\sim x|y|z$) & ($\sim x|\sim y|\sim z$)" is a clausal form.

Even if we restrict ourselves to clausal forms the SAT problem is still difficult to solve (i.e., the SAT problem of clausal forms is still NP-complete). This means that molecular algorithms for solving a SAT problem of clausal forms are interesting, at least theoretically.

Since a clausal form is a conjunction of clauses, in order to make a clausal form true, it is necessary to make all its clauses true. Therefore, to show that a clausal form is satisfiable, we have to find an assignment that makes all its clauses true. Note that an "assignment" means a mapping of a truth value to each variable in the clausal form.

3.2.2. Molecular Algorithm

Our proposed algorithm is based on the observation that if a clause is made true by an assignment, then at least one of the literals of the clause is made true by the assignment because a clause is a disjunction of its literals. Therefore, satisfying a clausal form amounts to choosing one literal from each clause and making it true. For example, in order to satisfy the clausal form "($x|y|\sim z$) & ($\sim x|y|z$) & ($\sim x|\sim y|\sim z$)," we can choose "x" from "$x|y|\sim z$," "z" from "$\sim x|y|\sim z$," and "$\sim y$" from "$\sim x|\sim y|\sim z$." This process of selection does not always succeed. For example, if we choose "x" from "$x|y|\sim z$" and "$\sim x$" from "$\sim x|\sim y|\sim z$," we have to make both "x" and "$\sim x$" true, which is impossible. So, the selection of a literal from each clause should be "consistent" in the sense that a variable and its negation are not chosen simultaneously. Conversely, if we can choose a literal from each clause in such a way that both a variable and its negation are not chosen, then the assignment obtained by the selection makes the clausal form true, and the clausal form is shown to be satisfiable.

Given a clausal form, enumerating all such selections is similar to enumerating paths over a given graph in Adleman's model, which solves the Hamiltonian path problem using DNA. One can use a similar method to generate a pool of selections as DNA molecules randomly. Therefore, extracting a consistent selection from the random pool is the crucial step in our algorithm.

Incidentally, a variable "x" and its negation "$\sim x$" are called "complementary" to each other (i.e., "x" is complementary to "$\sim x$" and vice versa). Using this terminology, we came up with the idea of representing complementary literals using complementary base sequences in the sense of Watson and Crick. For example, if a variable is represented by the sequence 5'-ATGCA-3', then its negation "$\sim x$" is represented by 5'-TGCAT-3'. A selection of literals from clauses is simply represented by the concatenation of the sequences corresponding to the chosen literals. In practice, for random generation, we need to insert auxiliary sequences between the literals.

Now, imagine what happens if a selection is inconsistent (i.e., if a selection contains both a variable and its selection). Since a selection contains two subsequences that are

complementary in the sense of Watson and Crick, it should form a hairpin, if it is made single-stranded (Fig. 6).

On an electronic computer, such a selection would be represented by an array of chosen literals, say $s[i]$. If there were m clauses in the clausal form, i.e., if there were m literals in the array $s[i]$, we would need a computation of the following type to check whether the selection is inconsistent:

```
For each i from 1 to m do
  for each j from i + 1 to m do
    if s[i] and s[j] are complementary
    then report inconsistent.
```

Using the representation of selections introduced so far, we can skip the above iterative computation and let the molecules search for complementary literals themselves.

After inconsistent selections form hairpin structures, it is only necessary to choose those selections that have not formed hairpin structures. If there is at least one such selection, then the given clausal form is deemed satisfiable. Therefore, we need a method to remove those selections containing hairpins from the pool of selections.

3.2.3. Separating Hairpins

In our group, Sakamoto et al. have developed a number of experimental techniques for separating molecules according to whether they contain hairpins. In particular, the following two techniques were developed and actually employed.

First, a restriction site is inserted in the middle of a sequence representing a literal. Those single-stranded DNA molecules that form a hairpin are then cut by the corresponding restriction enzyme. If each single-stranded molecule has primers at both ends, then those molecules that are cut by the restriction enzyme cannot be amplified by the primers.

Second, we developed a variation of PCR that can hardly amplify molecules containing hairpins. In general, the efficiency of PCR amplification is different according to whether DNA molecules contain a hairpin (between primers). The variation of PCR, called *exclusive PCR* (or *ePCR*), intensifies this distinction. More precisely, if the concentration of the molecules is low, they form hairpins rather than allowing intermolecular hybridization, and if they form hairpins, the efficiency of PCR amplification is expected to be very small. In order to maintain the low concentration of the single-stranded molecules, the reaction mixture is diluted twice after each PCR cycle. This is what we call ePCR.

3.2.4. Experiment

In our experiment [27], we actually solved the following instance of the SAT problem:

$$(a|b|\sim c)\&(a|c|d)\&(a|\sim c|\sim d)\&(\sim a|\sim c|d)\&(a|\sim c|e)\&$$
$$(a|d|\sim f)\&(\sim a|c|d)\&(a|c|\sim d)\&(\sim a|\sim c|\sim d)\&(\sim a|c|\sim d)$$

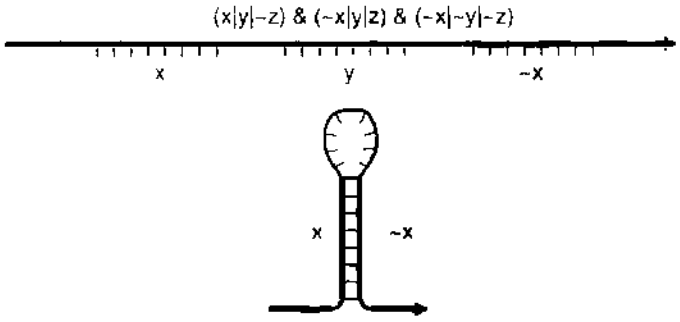


Figure 6. Selection of literals from clauses. If a selection contains complementary literals, it forms a hairpin because complementary literals are encoded by complementary sequences in the sense of Watson and Crick.

Since this problem has 10 clauses, and each clause has three literals, we should generate a random pool of selections consisting of 3^{10} species of molecules. Because the method for random generation is not a main topic of this chapter, we only note here that we used the restriction enzyme BstXI, whose recognition site is of the form CCAWNNNNWTGG, where W is either A or T, and N is any base. Each junction between literals contains this sequence, where NNNN is specific to the order of the junction. For example, the junction between a literal from the second clause and a literal from the third clause has its own NNNN. Since the sequence NNNN becomes sticky ends when it is cut by the restriction enzyme, beginning with a small number of selections, one can obtain a random pool of selections simply by digestion and ligation.

The sequences representing literals are 30 bases long, including the subsequences CCAW and WTGG of the recognition site of BstXI. They also contain the 5-base restriction site of BstNI in the middle for hairpin separation as mentioned above. So, 17 bases are specific to each literal.

The method for random generation as described briefly yields a random pool of double-stranded DNA molecules. One strand of those double-stranded molecules is biotinylated and immobilized on magnetic beads for alkali treatment, which separates it from the other strand. On the surface, the immobilized molecules immediately form hairpins if they contain complementary literals. Digestion by the restriction enzyme BstNI is also applied while they are on the surface because immobilization can also prevent intermolecular hybridization, which may digest molecules that do not contain hairpins.

By the experiment, we actually found five solutions of the SAT problem as reported in [27].

3.3. Computation by Hairpin Dissociation

In the previous section, we explained the computational power of hairpin formation using a concrete DNA algorithm that solves the SAT problem for clausal forms. In this section, we review how hairpin dissociation can be used for computation. The fundamental element of our computational scheme is hairpin dissociation using an oligomer called an "opener." As Fig. 7 shows, the stem of a hairpin is adjacent to a single-stranded part of the hairpin called the "lead section." An opener is a singled-stranded oligomer consisting of the "lead section" and the "invasion section." The lead section of the opener is complementary to that of the hairpin. The invasion section of the opener is complementary to the subsequence of the hairpin that is adjacent to its lead and forms its stem.

If the lead section of the hairpin hybridizes with that of the hairpin, the invasion section of the opener invades the hairpin stem and opens the hairpin via branch migration.

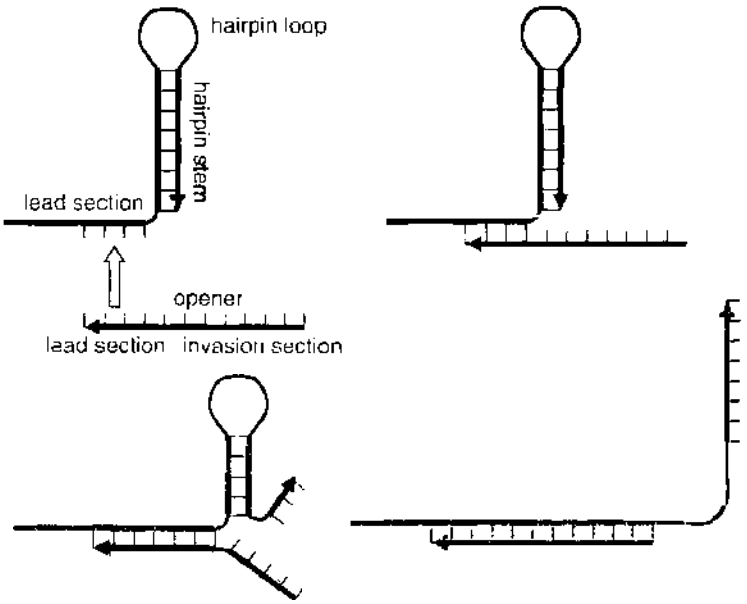


Figure 7. The interaction between a hairpin and an opener.

After the hairpin is opened in this way, the other subsequence of the hairpin stem, which is identical to the invasion section of the opener, is exposed as single-stranded. In addition, the subsequence of the hairpin loop is also exposed. Although the hairpin loop is already single-stranded when the hairpin is formed, its three-dimensional structure usually prohibits it from hybridizing with its complementary sequence. We discuss this in more detail later in this section. In summary, two new single-stranded subsequences are exposed when the opener opens the hairpin. They can initiate further hybridization and allow successive reactions comprising some computational steps.

3.3.1. Successive Hairpins and Molecular Memory

The simplest form of such successive reactions can be realized by simply concatenating hairpins, as in Fig. 8. After its opener opens the first hairpin, the subsequence adjacent to the stem of the second hairpin is exposed as single-stranded. Therefore, if the opener of the second hairpin is present, its lead section hybridizes with the subsequence, which acts as of the lead section of the second hairpin, and its invasion section opens the second hairpin. Note that while the first hairpin is formed, even if the opener of the second hairpin is present, it cannot open the second hairpin because its lead section cannot hybridize with the hairpin, so that its invasion section cannot invade the hairpin stem. Although hybridization between the second opener and second hairpin is favored energetically, the high-energy barrier to exposing the lead section of the second hairpin prohibits hybridization. Such hybridization prohibition allows various cascaded reactions for computation by hairpin dissociation.

As an application of this simple computational scheme, we advocate the use of DNA molecules with successive hairpins as elements of molecular memory. Each hairpin of a memory element can be regarded as one digit of the address by which the element is referenced. Only by putting the openers in the right order, can all the hairpins be opened. One can extract the memory element with the last opener, which can only hybridize with that element after all the preceding openers have hybridized with it.

This simple scheme has actually been implemented by our research group and shown to work well [28], although there are a few technical problems. First, it is not obvious how a DNA molecule with a number of hairpins could be synthesized. Recently, we actually synthesized a DNA molecule with four hairpins by decomposing it into four parts, as in Fig. 9 and concatenating them using ligase. In this implementation, each hairpin stem is 20 bases long, and each hairpin loop consists of 7 bases.

Second, the openers for successive hairpins have complementary subsequences and can hybridize. If many hairpins are concatenated together, their openers also hybridize to form a complex structure, which hinders the activity of the openers. The simplest way to avoid

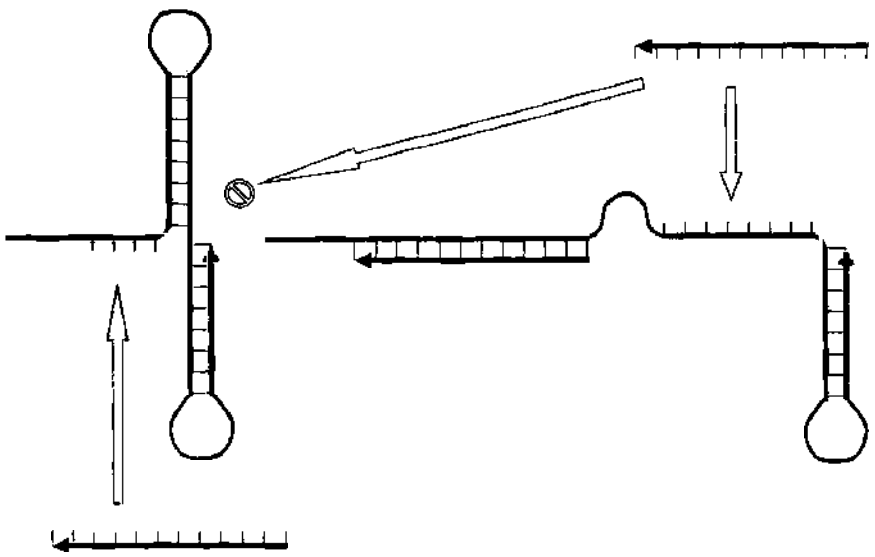


Figure 8. Two successive hairpins and openers. The second can only hybridize with the hairpin only after the first opener has opened its corresponding hairpin.

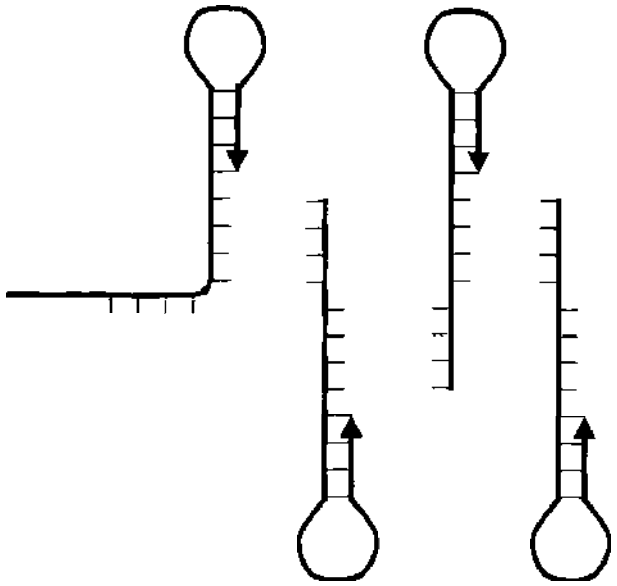


Figure 9. Generation of four successive hairpins.

this problem is to increase the concentration of the openers systematically or to remove the previous opener before adding the next one. Ideally, openers for different hairpins should not interact. Another way to solve the problem will be discussed later.

3.3.2. Branching Machines

We can imagine structures in which hairpins are concatenated in a more complex manner, as in Fig. 10. This structure has eight hairpins. By putting appropriate openers in the appropriate order, only one of the hairpins can be opened. Furthermore, once a hairpin is opened, other openers cannot interact with the structure. The simplest example of this kind of structure is shown in Fig. 11. It has two hairpins concatenated with a single-stranded sequence, which acts as the lead section for both hairpins. This structure accepts two openers, one for opening either hairpin. Since the two hairpins share the lead section, after one of the

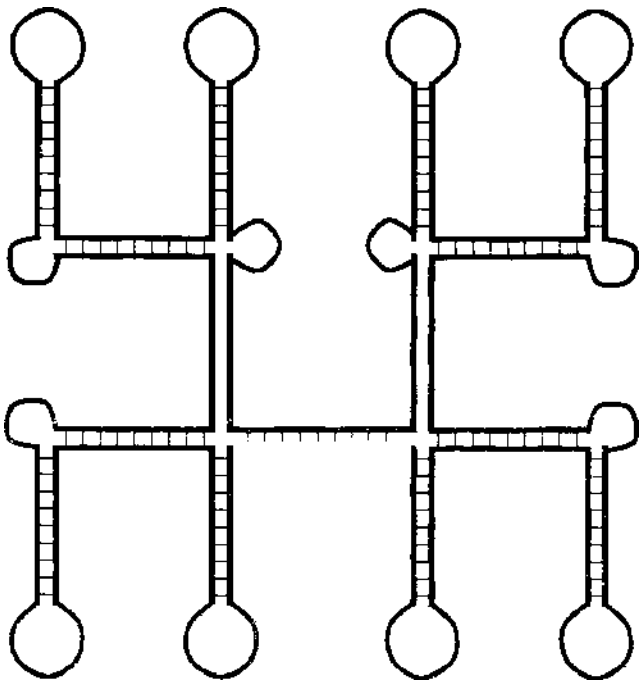


Figure 10. The imaginary branching machine.

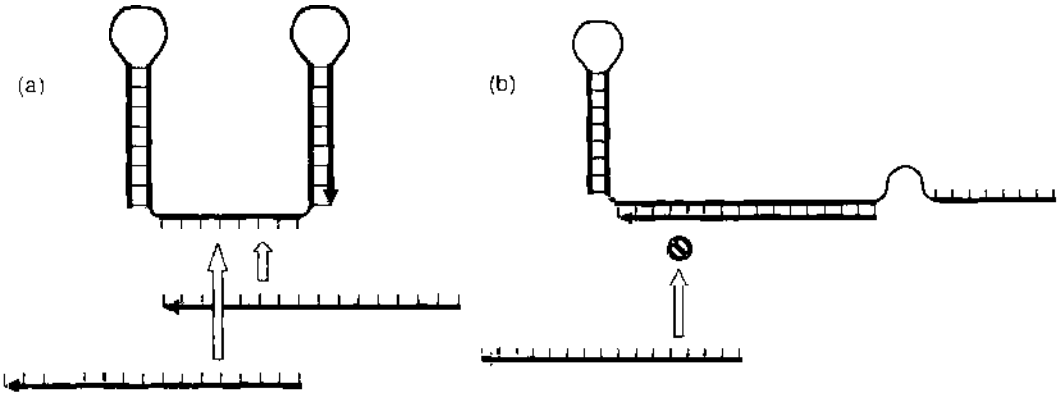


Figure 11. (a) A simple branching machine consisting of two hairpins and its openers. The openers share their lead section. (b) The interaction of the second opener and the machine after the first opener has hybridized. Because the lead section is occupied by the first opener, the second cannot hybridize.

openers is added and opens either hairpin, the other opener cannot hybridize with the lead section because of the high-energy barrier. Therefore, the result of adding the two openers depends on the order in which the openers are added.

Our group has reported an experimental result using this simple hairpin machine [29]. The actual sequences used in the experiment and the result of gel electrophoresis are shown in Fig. 12, which shows that different orders of openers result in different conformations.

3.3.3. Network of Machines

The molecular machines based on hairpins that have been introduced so far only accept openers as input and successively open their hairpins. They do not produce output that can be given to another machine as input.

To design molecular machines that have output, one can consider that it is difficult for a hairpin loop, although it is single-stranded, to hybridize with its complementary sequence. If a hairpin is opened, its hairpin loop is exposed as (completely) single-stranded and is then free to hybridize with its complementary sequence. Bulge loops have a similar property, which Tuberfield et al. reported in their method that prohibits hybridization between a bulge loop and its complementary sequence [30]. To increase this prohibition, it is desirable to concatenate a useless sequence to the sequence complementary to a bulge or hairpin loop.

In Fig. 13, an ordinary opener opens the first hairpin, exposing its hairpin loop as single-stranded. This hairpin loop and the adjacent single-stranded stem now act as an opener. In particular, the hairpin loop acts as the lead section of the opener and hybridizes with the lead section of the second hairpin. In summary, the first hairpin interacts with its opener and becomes the opener for the second hairpin. Although the first hairpin does not produce separate output, the result of its interaction is given to another machine as input.

Notice that the two hairpins do not interact with each other unless the opener is given to the first hairpin because the hairpin loop prohibits hybridization with the lead section of the second hairpin. Consequently, cascaded reactions can only occur after the opener is added.

We can view this process as one transforming a signal for the first hairpin into a signal for the second hairpin. The first hairpin transforms its signal X into signal Z for the second hairpin. We can prepare another hairpin that can be opened by another signal Y and produces the same signal Z. These two hairpins consist of a system that produces signal Z with the existence of signal X or signal Y. In other words, the system acts as an OR gate whose input is X and Y, and whose output is Z.

Using bulge loops in addition to hairpin loops, we can build more complex networks of molecular machines. Bulge loops also prohibit hybridization with their complementary sequences. In Fig. 14, two single-stranded DNA molecules form a structure containing two bulge loops. The first input (input 1) hybridizes with the single-stranded section of the structure and opens the first bulge loop via branch migration. Since the bulge loop is exposed

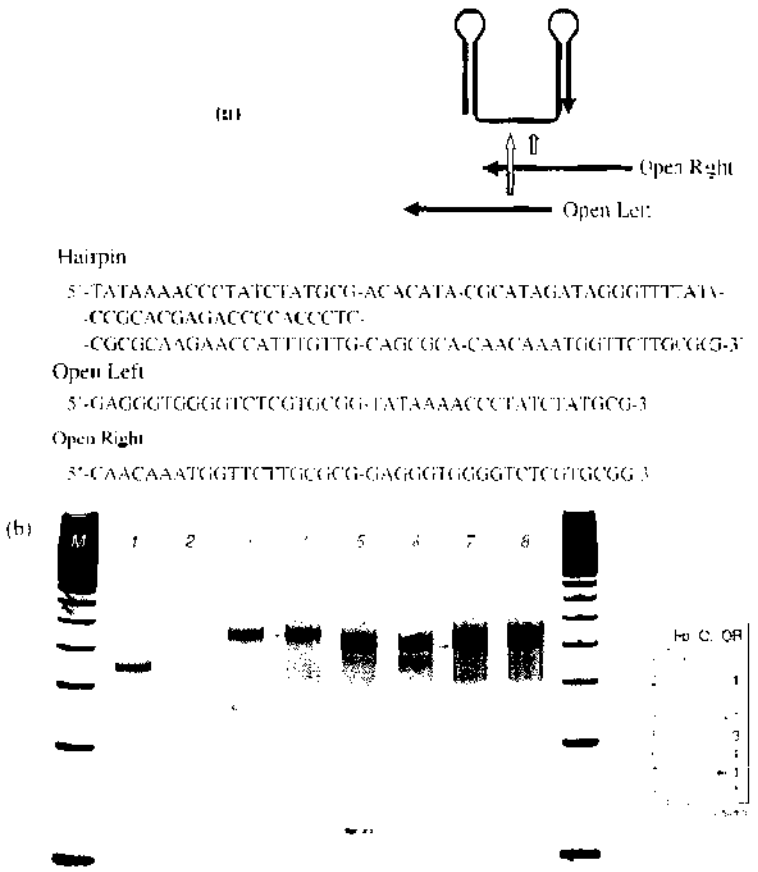


Figure 12. (a) Sequences of the branching machine and its openers. Hairpin: the branching machine consisting of two hairpins. Open Left: the opener that opens the left hairpin. Open Right: the opener that opens the right hairpin. (b) Experimental results of the simple branching machine. Lane 1: the branching machine only. Lane 2: the left opener and the right opener together. Lane 3: the result of adding the left opener to the machine. Lane 4: the result of adding the right opener to the solution of Lane 3 after 5 minutes. Lanes 5 and 6: the result of adding the right opener to the machine. The quantity of the right opener in Lane 5 is three times larger than that of Lane 6. Since the right opener forms a certain secondary structure, it is somewhat inefficient in hybridizing with the machine. The band in Lane 1 can also be seen in Lane 5 and Lane 6. Lane 7: the result of adding the left opener to the solution of Lane 6 after 5 minutes. Note that the uppermost band in Lane 6 is maintained, while the second band is replaced with the band of Lane 3. This means that those machines that could not hybridize with the right opener now hybridize with the left opener. Lane 8: the left and right openers are added simultaneously. Except in Lane 5, 1.5 pmol of each oligomer is put in a total 10 μl reaction mixture.

as single-stranded, it can hybridize with the second input (input 2), which opens the second bulge loop via branch migration and frees the lower strand of the structure as output.

Note that the two inputs do not interact with each other without the bulge structure, and the bulge structure can only interact with the first input. The output cannot interact with any of the inputs or the bulge structure. Since the output can only be produced in the presence of both inputs, this system comprises an AND gate.

Now that we have designed AND gates and OR gates, we can form more complex logical circuits as networks of interacting molecular machines.

3.3.4. Sensors

So far, single-stranded DNA molecules are given as inputs to molecular machines. By preparing molecules that sense environmental conditions, such as temperature, salt concentration, or light, one can also design molecular machines that can respond to their environment.

Figure 15 shows a sensor molecule that opens its hairpin when the surrounding temperature is raised. The molecule can then open another hairpin molecule by acting as its opener. Note that the second hairpin has a longer stem that cannot be opened even at the raised temperature. As a whole, these two hairpins form a sensor for high temperature. As we already explained, the second hairpin can interact with another machine. We can also replace the hairpin with a more complex machine, possibly one with bulges.

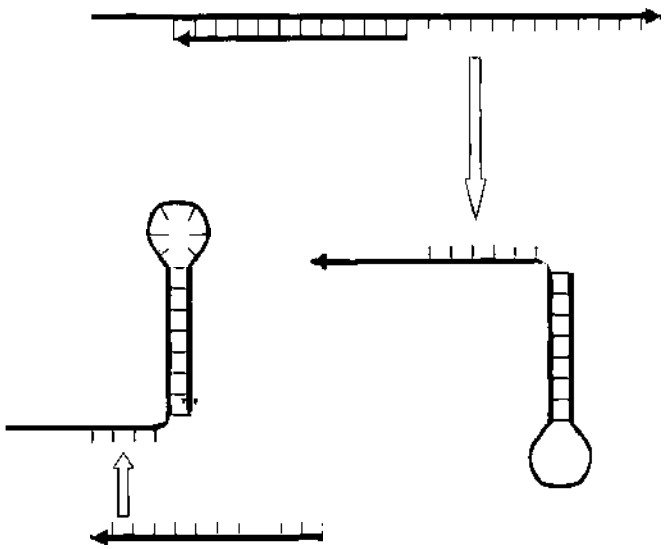


Figure 13. The interaction between a hairpin loop and another hairpin. If a hairpin is opened by an opener, its loop sequence is exposed as completely single-stranded, so it can hybridize with another hairpin as an opener.

3.3.5. Toward General-Purpose Molecular Systems

As the ultimate goal of this line of research, we hope to prepare a set of general-purpose molecular components and establish a systematic principle for designing a network of such molecular components comprising a complex molecular system. The components should include various molecular sensors, logical gates, and actuators and can be combined with no constraints. In other words, one would be able to program a molecular system by simply combining these building blocks. Establishing such a general-purpose framework for molecular machines is our ultimate goal.

3.4. Computation by Iterated Hairpin Formation and Dissociation

Near its melting temperature, a hairpin undergoes formation and dissociation with thermal fluctuation of the molecule. With the help of polymerase, the enzyme that copies the DNA sequence of a template, this autonomous iterated formation and dissociation of a hairpin can be used to realize an extremely interesting computational model using DNA. Historically, our group began to use hairpins with this computational model and has been working on

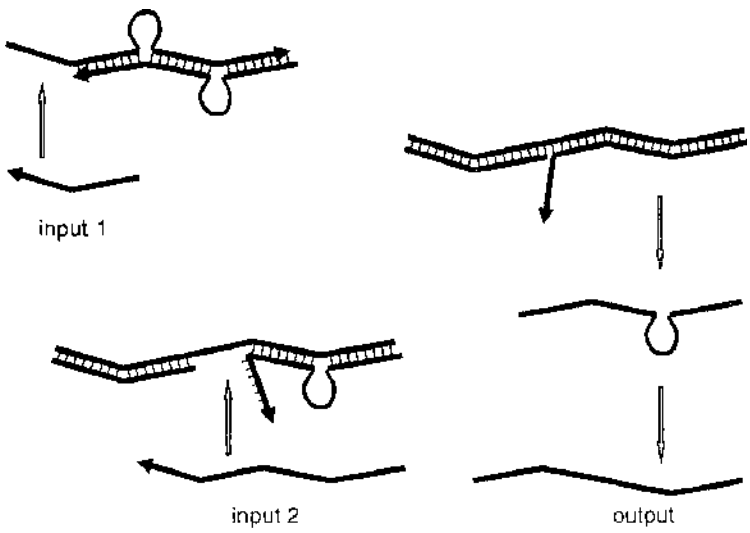


Figure 14. An AND gate consisting of two bulge loops. After the first input is given, the left bulge is opened and allows the second input to hybridize with it. The second input then invades the right bulge and releases the lower strand, which is considered as output of this AND gate.

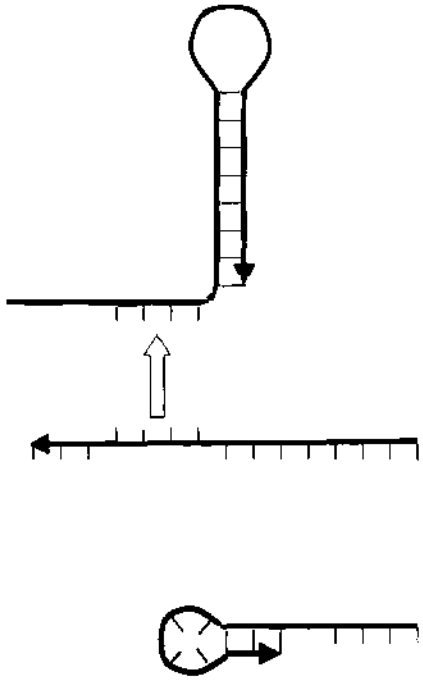


Figure 15. A thermal sensor by DNA. An opener for a hairpin itself forms a small hairpin. When the surrounding temperature is raised, the small hairpin is denatured, so the opener can hybridize with the lead section of the hairpin and opens it.

the model under the name of "Whiplash PCR." (According to Winfree, Adleman coined this terminology.)

3.4.1. Polymerization Stop

As in Fig. 16, when a hairpin is formed with the 3'-end of the DNA molecule at the foot of the hairpin stem, polymerase extends the 3'-end. This extension usually continues until the 3'-end encounters the 5'-end of the molecule and does not lead to any interesting computation. Constructing a computational model using polymerase extension of a hairpin requires stopping polymerization somewhere in the middle of extension. In order to halt polymerization, we insert a subsequence called a "stopper."

There are several different implementations for stoppers. For example, unnatural bases that cannot hybridize with any natural base can be used as a stopper. PNA molecules,

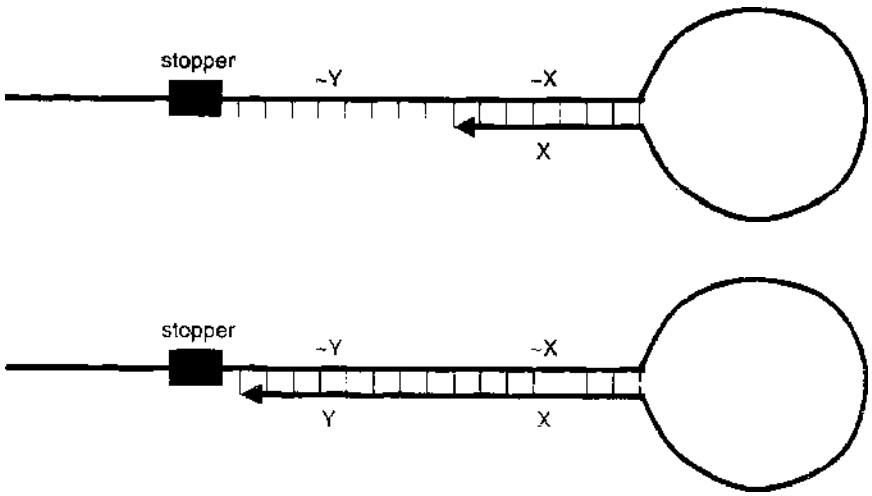


Figure 16. A state transition in Whiplash PCR. The sequence denoted by X hybridizes with its complementary sequence by forming a hairpin. With the existence of polymerase, the 3'-end of the molecule is extended and the sequence Y is newly added to the 3'-end. The extension stops at the stopper sequence as explained in the text. This extension is interpreted as a transition from state X to state Y.

which can hybridize with specific DNA sequences, can also be used as a stopper. The method we actually adopted simply restricts the kinds of bases in the polymerization buffer. We only put three bases in the polymerization buffer, say, A, G, and C. Because T is lacking in the polymerization buffer, the polymerase cannot copy sequences consisting of A. Therefore, a stopper is implemented as a sequence consisting of a reasonable number of A's, say AAA. Other sequences are coded using only three bases: T, C, and G. We call this technique of implementing a stopper "polymerization stop" [31, 32].

3.4.2. States and State Transitions

Given implementation of a stopper, let us look at Fig. 16 again. Using polymerase extension, a new sequence, denoted Y, is added to the 3'-end of the DNA molecule. Let X denote the old sequence at the 3'-end, which first hybridized with its complementary sequence and formed a hairpin. If the sequence at the 3'-end is considered to represent the "current state" of the molecule, this phenomenon can be interpreted as follows. The molecule changed its current state from X to Y via hairpin formation and polymerase extension. Note that the rule for this state change (or state transition) is also stored in the molecule itself (i.e., the molecule contains the sequence complementary to YX).

This interpretation of polymerase extension is the basic idea of Whiplash PCR. A single-stranded DNA molecule is considered a state machine (i.e., a machine that can change its state). The state is coded as the subsequence at the 3'-end of the molecule. The machine changes its state using a rule stored in the molecule itself. The machine can have an arbitrary number of such "state transition" rules, provided that the 3'-end can hybridize with those rules by forming a hairpin. Each state transition rule takes the form 3'-stopper-~Y-~X-5'; that is, it consists of a stopper, the complement of the new state Y, and the complement of the old state X. This state transition rule can be written as follows.

$$X \rightarrow Y$$

3.4.3. Evaluation of Boolean Expressions

Using the simple state transitions, it is possible to realize various interesting computations. As an example, let us implement a program for evaluating Boolean expressions [31]. In this example, we consider a restricted class of Boolean expressions called "read-once" Boolean expressions. In this kind of Boolean expression, each Boolean variable is referred to only once. The expression "x & $\sim y|z$" is a read-once Boolean expression, while "x & $\sim y|\sim x$ & z" is not. We can extend the set of read-once expressions by introducing another Boolean operator, "if-then-else." The expression "if x then y else z" is true if x is true and y is true, or if x is false and z is true. Using an if-then-else expression, one can rewrite "x & $\sim y|\sim x$ & z" as "if x then $\sim y$ else z," which is a read-once expression because all the variables, x, y, and z, are referred to only once. Note that "x & $\sim y|\sim x$ & z" and "if x then $\sim y$ else z" are logically equivalent.

Given a read-once expression, one can construct a set of state transition rules for computing the read-once expression. Take "if x then $\sim y$ else z" as an example. In this expression, x is referred to first. Let us denote the initial state of the machine as "Begin," and the state for referring to x as "Check-x." Then, the first state transition rule is as follows:

$$\text{Begin} \rightarrow \text{Check-}x.$$

If the value of variable x is true, then one goes to the state for referring to y, and if the value is false, then one goes to the state for referring to z. Therefore, we have the following two rules:

$$\text{True-}x \rightarrow \text{Check-}y,$$

$$\text{False-}x \rightarrow \text{Check-}z.$$

We assume that the machine enters the state "True-x" if the value of x is true and "False-x" if the value is false. We explain how these states are entered later. If the value of x is true and the value of y is true, then the entire expression is evaluated as false. In other words,

the final result is false. We denote the state for returning the false value as "Return-false." Therefore, we have the following rule.

$$\text{True-}y \rightarrow \text{Return-false.}$$

We have a similar rule for the case in which y is false:

$$\text{False-}y \rightarrow \text{Return-true.}$$

In addition, we have rules for z:

$$\text{True-}z \rightarrow \text{Return-true,}$$

$$\text{False-}z \rightarrow \text{Return-false.}$$

These seven rules compose the program for evaluating the Boolean expression "if x then y else z." They can be put into a single DNA molecule. An input to the Boolean expression can also be represented using a set of state transition rules. Remember that an input to a Boolean expression is an assignment of truth values (true or false) to the Boolean variables in it. Such an assignment can also be represented by a set of state transition rules, one for each variable. For example, if the true value is assigned to the variable x, then we prepare the following rule.

$$\text{Check-}x \rightarrow \text{True-}x.$$

Similarly, if y is false, we prepare the following one:

$$\text{Check-}y \rightarrow \text{False-}y.$$

So, an assignment to n variables is represented by a set of n state transition rules. These state transition rules can be encoded into DNA sequences and concatenated.

A program for evaluating a Boolean expression and an input to the program can be prepared separately. A single program may be run on many inputs, and a single input may be given to many programs. When a program is run on a given input, it is concatenated as a single-stranded DNA molecule and the initial state is attached at the 3′-end of the molecule. This will start the evaluation of the Boolean expression on the input.

3.4.4. Checking Conjunctions

A conjunction of variables is a simple kind of read-once Boolean expression. The conjunction "x & y & z" means that the variables x, y, and z are all true. One can check such a conjunction more easily. In this example, if x is true, then one can directly go to the state that checks y, and if y is true, one goes to the state to check z. Finally, if z is true, then one returns the final true result. We assume that we are only interested in getting the final true result and need not obtain the false result explicitly. We start with the rule

$$\text{Begin} \rightarrow \text{Check-}x.$$

The respective true values of the variables, x, y, and z, are represented as

$$\text{Check-}x \rightarrow \text{Check-}y,$$

$$\text{Check-}y \rightarrow \text{Check-}z,$$

$$\text{Check-}z \rightarrow \text{Return-true.}$$

Only if all the variables are true (i.e., if all these rules are stored in a DNA molecule), can the molecule make a state transition from Begin to Return-true.

Since the number of experimental operations for Whiplash PCR does not depend on the number of state transitions, Whiplash PCR provides an efficient way to check such a conjunction. For example, in the second step of Adleman's work, in order to check that a path is Hamiltonian, one has to check that all the nodes in the directed graph appear on the

path. This check requires the same number of affinity separation steps as that of nodes in the graph. If the representation of each node contains a state transition rule, then one can use Whiplash PCR for this check. More concretely, each node n contains the following rule:

$$\text{Check-}n \rightarrow \text{Check-}n+1$$

Only those paths that contain all the nodes in the graph allow state transitions from the first node to the last. Komiya recently conducted an actual experiment in which he solved an instance of the Hamiltonian path problem using Whiplash PCR.

3.4.5. Isothermal Implementation

Our group has actually implemented Whiplash PCR using the polymerization stop technique explained in the previous section [31, 32]. One of the crucial points in implementing Whiplash PCR is the thermal control allowing successive state transitions.

As in Fig. 17, immediately after one state transition, the DNA molecule forms a hairpin structure whose stem is twice as long as a sequence representing a state. For example, if states are encoded by 15-base sequences, then the molecule forms a hairpin whose stem is 30 bases long after it makes one state transition. For the molecule to make the next state transition, the current hairpin should be dissociated and a new hairpin should be formed for the next polymerase extension. In other words, it should be possible to form a 15-base double helix at the cost of dissociating a 30-base double helix. This is energetically impossible at room temperature. Near the melting temperature of the hairpins, however, they iterate formation and dissociation because of thermal fluctuation, and the molecule forms different hairpins from time to time.

Of course, the probability that 15-base double helices form is much smaller than the probability of 30-base double helices. However, once a 15-base helix that allows polymerase extension is formed, the 3'-end is extended and the molecule makes the next state transition. This is an irreversible process in the presence of polymerase. Conversely, a 30-base double helix does not allow polymerase extension. Because of thermal fluctuation, such a 30-base double helix is eventually dissociated and replaced with another double helix. With a certain probability, the molecule forms a 15-base helix and makes the next transition. Therefore, we anticipated that the molecule could make successive state transitions under isothermal conditions.

The formation of a 30-base double helix that prohibits the next state transition is called "back-hybridization." To make successive state transitions by preventing back-hybridization, one should carefully adjust the reaction temperature. If the temperature is low, the

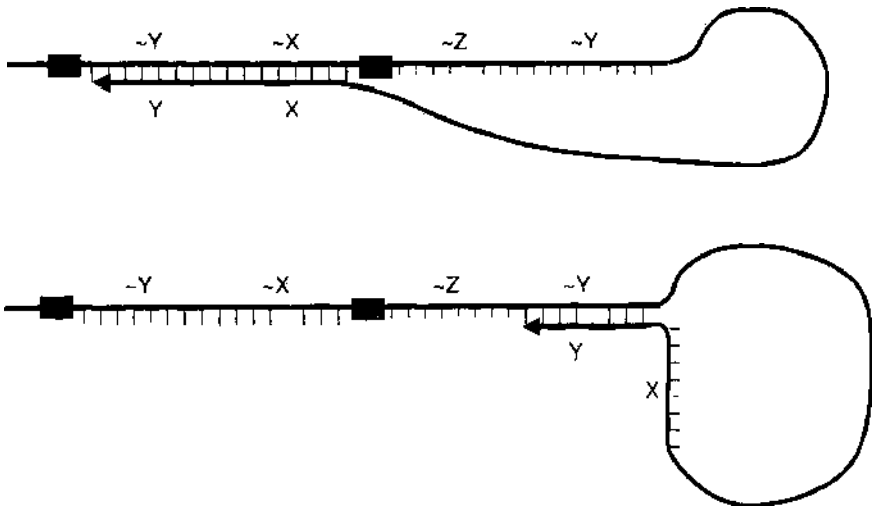


Figure 17. Successive transitions in Whiplash PCR. For the next state transition, the old hairpin structure should be dissociated. However, this structure is energetically more favored because its stem is twice as long as the new hairpin. The formation of the old hairpin is called back-hybridization. Although back-hybridization is energetically favored, it does not extend the 3'-end due to the stopper. Therefore, the next transition is eventually achieved.

back-hybridization form dominates. Conversely, if the temperature is too high, the hairpin that allows the next state transition does not form sufficiently well.

Figure 18 shows a DNA sequence that allows two successive state transitions. We used this sequence to verify that isothermal conditions actually allow successive state transitions and found the most appropriate temperature (Fig. 19). In this experiment, 78 °C was the best temperature.

Rose made a Markov process model of Whiplash PCR based on equilibrium thermodynamics, and predicted the behavior of successive state transitions [33]. According to his model, for two successive transitions, 78 °C is the best temperature, and this matches the experimental result well.

Komiya recently conducted an experiment in which eight successive state transitions were observed to be possible. In this experiment, he also tried various temperatures and found the best temperature program. This result also matches Rose's prediction.

Rose proposed another technique for preventing back-hybridization—the use of PNA molecules that bind between two state sequences [34]. For example, when the state of the molecule is changed from X to Y, the sequence of Y is attached to the sequence X at the 3'-end. Then, a PNA molecule binds between X and Y. This greatly lowers the stability of the back-hybridization form and enhances the next state transition.

3.4.6. Applications

In contrast to data-parallel computation proposed by Adleman, Whiplash PCR allows program-parallel computation by molecules. Molecules with different programs can be executed in parallel. Because each molecule can also have its own data and program, the parallelism realized by Whiplash PCR can be classified as MPMD (multiple-program multiple-data). For example, each molecule can have its own Boolean expression and input (assignment of truth values to variables) and evaluate the Boolean expression on the input in parallel.

Using program-parallel computation, it is possible to search for programs satisfying certain conditions. As a typical example of a search for programs, we proposed the idea of learning Boolean expressions using examples of inputs and outputs [31]. This problem involves finding a Boolean expression that satisfies the given constraints on inputs and the corresponding outputs. For example, try to find a Boolean expression of two variables x and y that satisfies the following constraints:

$$x = \text{true} \quad y = \text{false} \quad \text{output} = \text{true}$$

$$x = \text{false} \quad y = \text{false} \quad \text{output} = \text{false}$$

This is a typical instance of a learning problem from positive and negative data and is very difficult to solve computationally.

In Section 2.5, we pointed out the relationship between data-parallel computation and evolutionary computation. Because Whiplash PCR realizes program-parallel computation, it allows program evolution in the sense of "genetic programming." Unlike a "genetic algorithm," which is software simulation of evolution, genetic programming allows the evolution of programs via the evolutionary processes mutation, crossover, evaluation, and selection. Rose proposed the idea of applying Whiplash PCR to genetic programming [35]. He devised a program for generating protein-coding DNA sequences as a set of state transition rules. Each new sequence attached at the 3'-end in one state transition encodes a certain protein

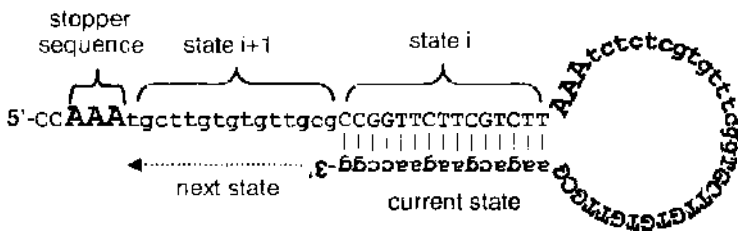


Figure 18. The sequence used for the experiment of two successive transitions.

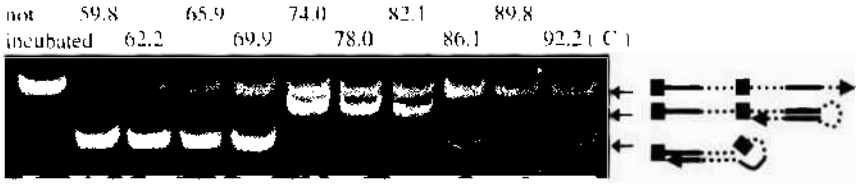


Figure 19. The results of two successive transitions under various temperatures. The solution is first kept at 94 °C for 1 minute and then incubated at each specified temperature for 5 minutes. The gel is 8 M urea 8% PAGE, and the solution is 1XPfx buffer (the composition unknown) with 1 mM $MgSO_4$, 0.2 mM dATP, dCTP, dGTP, and 1.5 units Platinum Pfx DNA polymerase. Each band is predicted to consist of the drawn structure.

unit, such as an alpha helix. The entire sequence generated by successive state transitions encodes a new protein molecule. Using the well-known technique of protein evolution, also known as an "*in vitro* virus," generated protein molecules are concatenated with their programs and evaluated using the given selection criterion. Rose argues that Whiplash PCR allows more sophisticated and efficient protein evolution.

REFERENCES

1. L. M. Adleman, *Science* 266, 1021 (1994).
2. R. J. Lipton, *Science* 268, 542 (1995).
3. Q. Liu, L. Wang, A. Frutos, A. E. Condon, and L. M. Smith, *Nature* 413, 175 (2000).
4. D. Faulhammer, A. R. Cukras, R. J. Lipton, and L. F. Landweber, *Proc. Natl. Acad. Sci. U.S.A.* 97, 1385 (2000).
5. E. Winfree, X. Yang, and N. C. Seeman, *Math. Theor. Comput. Sci.* 44, 191 (1999).
6. M. Arita and S. Kobayashi, *New Gen. Comput.* 20, 263 (2002).
7. S. Kobayashi, T. Kondo, K. Okuda, and M. Arita, Lecture Notes in Computer Science, Vol. 2568, p. 205, Springer, Berlin, 2002.
8. M. Andronescu, R. Aguirre-Hernandez, A. Condon, and H. H. Hoos, *Nucleic Acids Res.* 31, 3416 (2003).
9. R. M. Dirks, M. Lin, E. Winfree, and N. A. Pierce, *Nucleic Acid Research* 32, 1392 (2004).
10. R. S. Braich, N. Chelyapov, C. Johnson, P. W. K. Rothemund, and L. Adleman, *Science* 296, 499 (2002).
11. H. Yoshida and A. Suyama, *DIMACS Ser. Discrete Math. Theor. Comput. Sci.* 54, 9 (2000).
12. Y. Sakakibara and A. Suyama, *Genome Informatics* 11, 33 (2000).
13. E. Winfree, X. Yang, and N. C. Seeman, *DIMACS Ser. Discrete Math. Theor. Comput. Sci.* 44, 191 (1999).
14. N. C. Seeman, H. Wang, B. Liu, J. Qi, X. Li, X. Yang, F. Liu, W. Sun, Z. Shen, Y. Wang, R. Sha, C. Mao, S. Zhang, T.-J. Fu, S. M. Du, J. E. Mueller, Y. Zhang, and J. Chen, *DIMACS Ser. Discrete Math. Theor. Comput. Sci.* 44, 215 (1999).
15. N. C. Seeman, *Nat. Comput.* 1, 53 (2002).
16. E. Winfree, F. Liu, L. A Wenzler, and N. C. Seeman, *Nature* 394, 539 (1998).
17. P. W. K. Rothemund, N. Papadakis, and E. Winfree1, *PLoS Biol.* 2, e424 (2004).
18. T. H. LaBean, H. Yan, J. Kopatsch, F. Liu, E. Winfree, J. H. Reif, and H. C. Seeman, *J. Am. Chem. Soc.* 122, 1848 (2000).
19. D. Liu, J. H. Reif, and T. H. LaBean, *Proc. Natl. Acad. Sci. U.S.A.* 101, 717 (2004).
20. H. Yan, S. H. Park, G. Finkelstein, J. H. Reif, and T. H. LaBean, *Science* 301, 1882 (2003).
21. B. Yurke, A. J. Turberfield, A. P. Mills, Jr., F. C. Simmel, and J. L. Neumann, *Nature* 406, 605 (2000).
22. Y. Benenson, T. Paz-Elizur, R. Adar, E. Keinan, Z. Livneh, and E. Shapiro, *Nature* 414, 430 (2001).
23. Y. Benenson, B. Gil, U. Ben-Dor, R. Adar, and E. Shapiro, *Nature* 429, 423 (2004).
24. I. L. Hofacker, *Nucleic Acids Res.* 31, 3429 (2003).
25. C. Flamm, W. Fontana, I. L. Hofacker, and P. Schuster, *RNA* 6, 325 (2000).
26. M. Kubota and M. Hagiya, Lecture Notes in Computer Science, Vol. 3384, p. 202, Springer, Berlin, 2005.
27. K. Sakamoto, H. Gouzu, K. Komiya, D. Kiga, S. Yokoyama, T. Yokomori, and M. Hagiya, *Science* 288, 1223 (2000).
28. A. Kameda, M. Yamamoto, H. Uejima, M. Hagiya, K. Sakamoto, and A. Ohuchi, *Nat. Comput.* 4, 103 (2005).
29. M. Kubota, K. Ohtake, K. Komiya, K. Sakamoto, and M. Hagiya, in "Proceedings of the 2003 Congress on Evolutionary Computation (CEC'03)" Canberra, December 8–12 (2003), p. 2542.
30. A. J. Turberfield, J. C. Mitchell, B. Yurke, A. P. Mills, Jr., M. I. Blakey, and F. C. Simmel, *Phys. Rev. Lett.* 90, 11, 118 (2003).
31. M. Hagiya, M. Arita, D. Kiga, K. Sakamoto, and S. Yokoyama, *DIMACS Ser. Discrete Math. Theor. Comput. Sci.* 48, 57 (1999).
32. K. Sakamoto, D. Kiga, K. Komiya, H. Gouzu, S. Yokoyama, S. Ikeda, H. Sugiyama, and M. Hagiya, *Biosys.* 52, 81 (1999).
33. J. A. Rose, R. J. Deaton, M. Hagiya, and A. Suyama, *Phys. Rev. E* 65, 021910 (2002).
34. J. A. Rose, R. J. Deaton, M. Hagiya, and A. Suyama, Lecture Notes in Computer Science, Vol. 2340, p. 104, Springer, Berlin, 2002.
35. J. A. Rose, M. Takano, M. Hagiya, and A. Suyama, *J. Gen. Program. Evolvable Mach.* 139 (2003).

Index

A

ISBN: 1-58883-048-9

Handbook of Theoretical and Computational Nanotechnology
Edited by Michael Rieth and Wolfram Schommers
Volume 6: Pages (841–872)

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

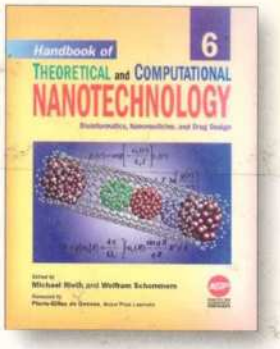
V

W

Y

Z

Handbook of THEORETICAL and COMPUTATIONAL NANOTECHNOLOGY



Edited by **Michael Rieth** and **Wolfram Schommers**

The future applications of nanotechnology in high-tech industries require deep understanding of the theoretical and computational aspects of all kinds of materials and devices on a nanometer scale. ***Handbook of Theoretical and Computational Nanotechnology*** is the first single reference source ever published in the field that offers such an unified approach, covering all of the major topics dealing with theory, modeling, design, and simulations of nanostructured materials and nanodevices, quantum computing, computational chemistry, physics, and biology, nanomechanics, nanomachines, nanoelectronics, nanoprocesses, nanomagnetism, nanooptics, nanomedicines, nanobiotechnology, etc. This 10-volume handbook provides the first ideal introduction and an up-to-date survey of the fascinating new developments and interdisciplinary activities in the whole field presented by scientists working in different subject areas of science, engineering, and medicine. This handbook is the most profound publication on this topic-the first treatment of computational nanotechnology. This outstanding handbook, presented by the world's leading scientists, is the most significant academic title ever published in this research field. This handbook has been divided into 10 thematic volumes by documenting computational treatment of nanomaterials and nanodevices.

Volume 1: Basic Concepts, Nanomachines, and Medical Nanodevices
Volume 2: Atomistic Simulations—Algorithms and Methods
Volume 3: Quantum and Molecular Computing, Quantum Simulations
Volume 4: Nanomechanics and Multiscale Modeling
Volume 5: Transport Phenomena and Nanoscale Processes
Volume 6: Bioinformatics, Nanomedicine, and Drug Design
Volume 7: Magnetic Nanostructures and Nanooptics
Volume 8: Functional Nanomaterials, Nanoparticles, and Polymer Design
Volume 9: Nanocomposites, Nano-Assemblies, and Nanosurfaces
Volume 10: Nanodevice Modeling and Nanoelectronics

KEY FEATURES

- The World's first handbook ever published in the field of theoretical and computational nanotechnology.
- The first comprehensive reference dedicated to all disciplines of science, engineering, and medicine.
- Most up-to-date reference source drawing on the past two decades of pioneering research.
- About 140 Review chapters written by world leading scientists familiar with the current trends of nanotechnology.
- Over 8,000 pages written by 265 authors from 30 countries, truly international.
- 26,000 references, 4124 figures, 374 tables, and thousands of mathematical equations and formula.
- Clearly written, self-contained, timely, authoritative, and most comprehensive contributions.
- Extensive cross-refereeing in each chapter provides reader with a broader range of knowledge.
- Multidisciplinary reference source for scientists, engineers, biologists, medical experts and related professionals.

READERSHIP

This handbook is an invaluable reference source for scientists, engineers, and biologists working in the field of theoretical and computational nanotechnology. The handbook is intended for a broad audience working in the fields of quantum chemistry, physics, biology, materials science, electrical and electronics engineering, mechanical engineering, optical science, ceramic and chemical engineering, device engineering, aerospace engineering, computer science and technology, information technology, bioinformatics, biotechnology, medical sciences, medicine, surface science, and polymer science and technology.

ISBN 1-58883-048-9
9 781588 830487 90000

AMERICAN SCIENTIFIC PUBLISHERS
Los Angeles, California, USA

Printed in the United States of America