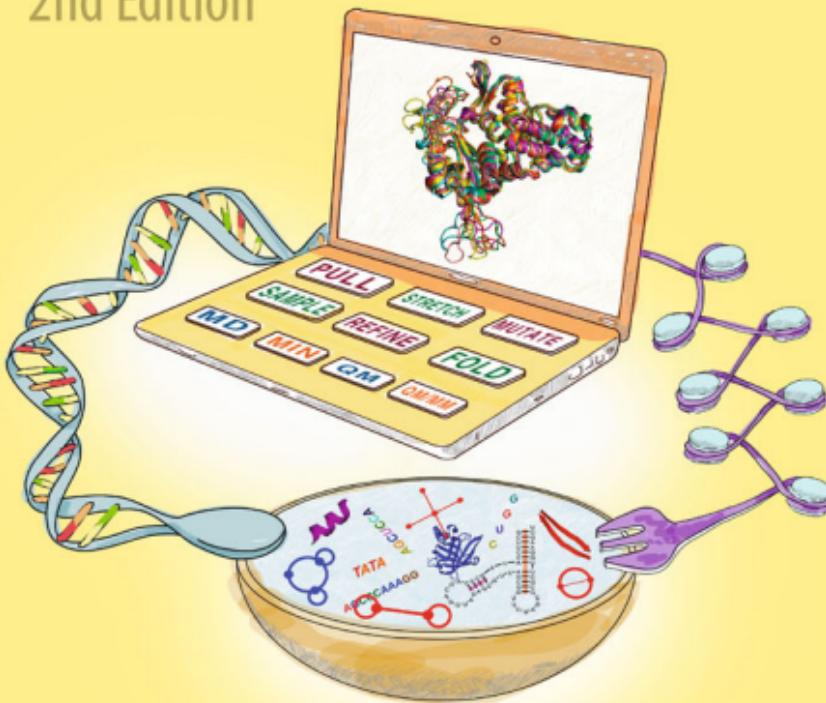


INTERDISCIPLINARY APPLIED MATHEMATICS

Tamar Schlick  
**Molecular Modeling  
and Simulation:  
An Interdisciplinary Guide**

2nd Edition



Springer

# **Interdisciplinary Applied Mathematics**

---

## **Volume 21**

### *Series Editors*

**S.S. Antman   J.E. Marsden  
L. Sirovich**

### *Series Advisors*

**C.L. Bris   L. Glass  
P. S. Krishnaprasad   R.V. Kohn  
J.D. Murray   S.S. Sastry**

### *Geophysics and Planetary Sciences*

### *Imaging, Vision, and Graphics*

**D. Geman**

### *Mathematical Biology*

**L. Glass, J.D. Murray**

### *Mechanics and Materials*

**R.V. Kohn**

### *Systems and Control*

**S.S. Sastry, P.S. Krishnaprasad**

Problems in engineering, computational science, and the physical and biological sciences are using increasingly sophisticated mathematical techniques. Thus, the bridge between the mathematical sciences and other disciplines is heavily traveled. The correspondingly increased dialog between the disciplines has led to the establishment of the series: *Interdisciplinary Applied Mathematics*.

The purpose of this series is to meet the current and future needs for the interaction between various science and technology areas on the one hand and mathematics on the other. This is done, firstly, by encouraging the ways that mathematics may be applied in traditional areas, as well as point towards new and innovative areas of applications; and, secondly, by encouraging other scientific disciplines to engage in a dialog with mathematicians outlining their problems to both access new methods and suggest innovative developments within mathematics itself.

The series will consist of monographs and high-level texts from researchers working on the interplay between mathematics and other fields of science and technology.

For further volumes:

<http://www.springer.com/series/1390>

## Reviews of First Edition (*continued* *from back cover)*

“The interdisciplinary structural biology community has waited long for a book of this kind which provides an excellent introduction to molecular modeling.” (Harold A. Scheraga, Cornell University)

“A uniquely valuable introduction to the modeling of biomolecular structure and dynamics. A rigorous and up-to-date treatment of the foundations, enlivened by engaging anecdotes and historical notes.” (J. Andrew McCammon, Howard Hughes Medical Institute, University of California at San Diego)

“The text is beautifully illustrated with many color illustrations. Even part of the text is typeset in color. Not only the illustrations interrupt the very readable text, there are also many box-insertions . . .” (Adhemar Bultheel, Bulletin of the Belgian Mathematical Society, Vol. 11 (4), 2004)

“This textbook evolved from a graduate course in molecular modeling, and was expanded to serve as an introduction to the field for scientists in other disciplines. . . . The book is unique in that it combines introductory molecular biology with advanced topics in modern simulation algorithms . . . . the author provides 1000+ references, and additionally includes reading lists complementing the main text. This is an excellent introductory text that is a pleasure to read.” (Henry van den Bedem, MathSciNet, September, 2004)

“This book provides an excellent introduction to the modeling of biomolecular structures and dynamics. . . . The book’s appendices complement the material in the main text through homework assignments, reading lists, and other information useful for teaching molecular modeling. The book is intended for students of an interdisciplinary graduate course in molecular modeling as well as for researchers (physicists, mathematicians and engineers) to get them started in computational molecular biology.” (Ivan Krivy, University of Ostrava, Czech Republic, Zentralblatt MATH, Issue 1011, 2003)

“The book . . . is the outcome of the author Tamar Schlick’s teaching experience at New York University. It is a fantastic graduate textbook to get into structural biology. . . . even the most sophisticated problems are part of a gradual approach . . . . The book will obviously be of great interest to students and teachers but it should also be very valuable for research scientists, especially newcomers to the field . . . as a reference book and a point of entry in the more specialised literature.” (Benjamin Audit, Bioinformatics, January, 2003)

“The basic goal of this new text is to introduce students to molecular modelling and simulation and to the wide range of biomolecular problems being attacked by computational techniques. . . . the text emphasises that the field is changing very rapidly and that it is full of exciting discoveries. . . . This book stimulates this excitement, while still providing students many computational details. . . . It contains detailed illustrations throughout . . . . It should appeal to beginning graduate students . . . in many scientific departments . . . .” (Biotech International, Vol. 15 (2), 2003)

Tamar Schlick

# Molecular Modeling and Simulation

An Interdisciplinary Guide

2nd edition



Springer

Prof. Tamar Schlick  
New York University  
Courant Institute of Mathematical Sciences  
and Department of Chemistry  
251 Mercer Street  
New York, NY 10012  
USA  
schlick@nyu.edu

*Editors*

S.S. Antman  
Department of Mathematics  
*and*  
Institute for Physical Science  
and Technology  
University of Maryland  
College Park, MD 20742, USA  
ssa@math.umd.edu

J.E. Marsden  
Control and Dynamical Systems  
Mail Code 107-81  
California Institute of Technology  
Pasadena, CA 91125, USA  
marsden@cds.caltech.edu

L. Sirovich  
Department of Biomathematics  
Laboratory of Applied Mathematics  
Mt. Sinai School of Medicine  
Box 1012  
New York, NY 10029  
USA  
Lawrence.Sirovich@mssm.edu

ISSN 0939-6047  
ISBN 978-1-4419-6350-5 e-ISBN 978-1-4419-6351-2  
DOI 10.1007/978-1-4419-6351-2  
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2010929799

Mathematics Subject Classification (2010): MSC 2010: 62P10, 65C05, 65C10, 65C20, 68U20, 92B05,  
92C05, 92C40, 92E10, 97M60

© Springer Science+Business Media, LLC 2010  
All rights reserved. This work may not be translated or copied in whole or in part without the written  
permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York,  
NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use  
in connection with any form of information storage and retrieval, electronic adaptation, computer  
software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.  
The use in this publication of trade names, trademarks, service marks, and similar terms, even if they  
are not identified as such, is not to be taken as an expression of opinion as to whether or not they are  
subject to proprietary rights.

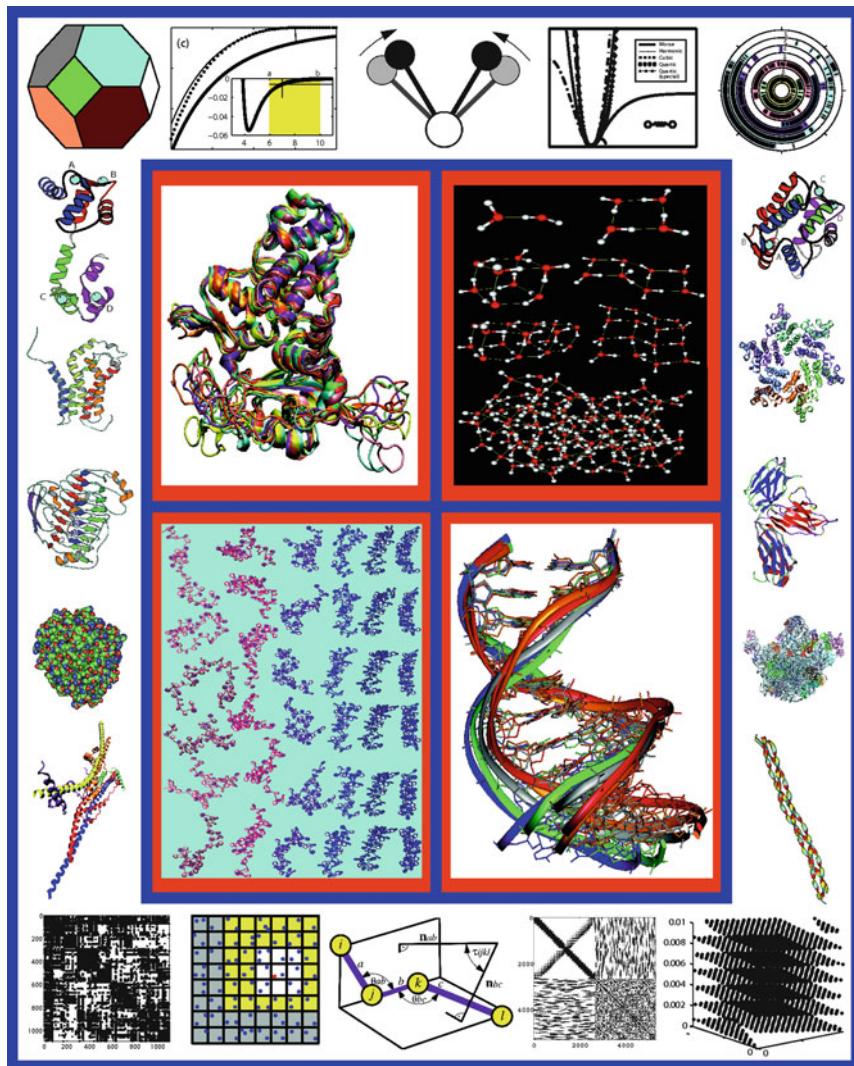
Printed on acid-free paper

Springer is part of Springer Science+Business Media ([www.springer.com](http://www.springer.com))

## About the Cover

Molecular modelers are artists in some respects. Their subjects are complex, irregular, multiscaled, highly dynamic, and sometimes multifarious, with diverse states and functions. To study these complex phenomena, modelers must apply computer programs based on precise algorithms that stem from solid laws and theories from mathematics, physics, and chemistry. Like innovative chefs, they also borrow their inspiration from other fields and blend the ingredients and ideas to create appealing inventive dishes.

The West-Coast-inspired landscape paintings of artist Wayne Thiebaud, whose work *Reservoir Study* decorated the cover of the first edition of this book, embodied that productive blend of nonuniformity with orderliness as well as the multiplicity in perspectives and interpretations central to molecular modeling. For this edition, the collage on the back cover (created with Shereef Elmetwaly) reflects such an amalgam of foundations, techniques, and applications. The computer salad image on the front cover (created with Namhee Kim and James Van Arsdale) further reflects a vision for the near future when modeling and simulation techniques will be reliable so as to compute folded structures and other desired aspects of biomolecular structure, motion, and function. I hope such creative blends will trigger readers' appetite for more creations to come.



*To the memory of my beloved aunt Cecilia, who filled my life with love, joy, beauty, and courage which I will forever carry with me.*

# Book URLs

**For Text:**

[www.biomath.nyu.edu/index/book.html](http://www.biomath.nyu.edu/index/book.html)

**For Course:**

[www.biomath.nyu.edu/index/course/IndexMM.html](http://www.biomath.nyu.edu/index/course/IndexMM.html)

# Preface

As I update parts of this textbook seven years after the original edition, I find the progress in the field to be overwhelming, almost unfitting to justify maintaining the same book. In fact, the sports analogy “Bigger, faster, stronger” seems most appropriate to the field of biomolecular modeling. Indeed, as modeling and simulation are used to probe more biological and chemical processes — with improved force fields and algorithms and faster computational platforms — new discoveries are being made that help interpret as well as extend experimental data. To experimentalists and theoreticians alike, modeling remains a valuable, albeit challenging, tool for probing numerous conformational, dynamic, and thermodynamic questions. We can certainly anticipate more exciting developments in biomolecular modeling as the first decade of this new century has ended and another began. At the same time, we should be reminded by the wisdom of the great French mathematician and scientist Pierre Simon de Laplace, who I quote more than once in this text, who also said: “*Ce que nous connaissons est peu de chose; ce que nous ignorons est immense*”. (What we know is little; what we do not know is immense).

Besides small additions and revisions made throughout the text and displayed materials to reflect the latest literature and field developments, some chapters have undergone more extensive revisions for this second edition. These include Chapters 1 and 2 that provide a historical perspective and an overview of current applications to biomolecular systems; Chapter 4, which reflects modified protein classification with new protein examples and sequence statistics; the chapter Topics in Nucleic Acids (now expanded into two chapters, 6 and 7), which includes recent developments in RNA structure and function; the force field chapters 8–10, which contain new sections on enhanced sampling methods; Chapter 15, which

includes an update on pharmacogenomics developments; and Appendices B and C which list key papers in the field and reference books, respectively.

As in the original book, the focus is on a broad and critical introduction to the field rather than a comprehensive view, though some algorithmic topics are presented in more depth. There are many books now since the first edition was written that provide more details on various aspects of biomolecular modeling and simulation (see Appendix C).

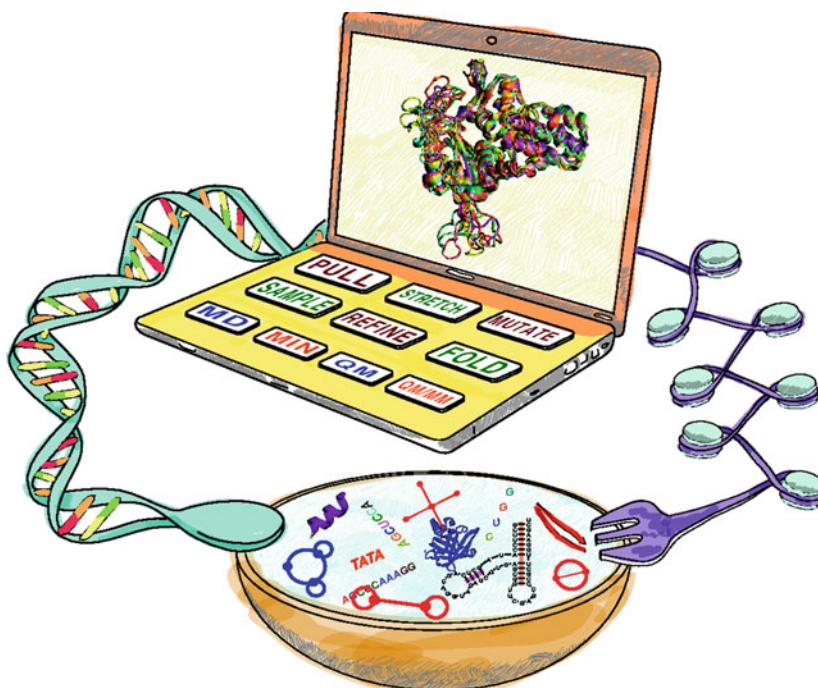
I would like to thank my many lab members and colleagues who have contributed to this effort, by providing scientific and technical information, making figures, and/or reading various versions of this text, including Lisa Chase, Rosana Collepardo, Ron Dror, Shereef Elmetwaly, Meredith Foley, Joachim Frank, Hin Hark Gan, Joe Izzo, Namhee Kim, Itzhak Krinsky, Christian Laing, Pierre L'Ecuyer, Connie Lee, Rubisco Li, Michael Overton, Vijay Pande, Ogi Perisic, Giulio Quarta, Klaus Schulten, Rick Solway, James Van Arsdale, Arieh Warshel, Michael Watters, Ada Yonath, and Yingkai Zhang.

As before, I invite readers to share their comments and thoughts with me directly via email; I enjoy reading them all.

Tamar Schlick

New York, NY

March 2, 2010



## Preface to the 2002 Edition

Science is a way of looking, reverencing. And the purpose of all science, like living, which amounts to the same thing, is not the accumulation of gnostic power, the fixing of formulas for the name of God, the stockpiling of brutal efficiency, accomplishing the sadistic myth of progress. The purpose of science is to revive and cultivate a perpetual state of wonder. For nothing deserves wonder so much as our capacity to experience it.

Roald Hoffman and Shira Leibowitz Schmidt, in *Old Wine, New Flasks: Reflections on Science and Jewish Tradition* (W.H. Freeman, 1997).

## Challenges in Teaching Molecular Modeling

This textbook evolved from a graduate course termed *Molecular Modeling* introduced in the fall of 1996 at New York University. The primary goal of the course is to stimulate excitement for molecular modeling research — much in the spirit of Hoffman and Leibowitz Schmidt above — while providing grounding in the discipline. Such knowledge is valuable for research dealing with many practical problems in both the academic and industrial sectors, from developing treatments for AIDS (via inhibitors to the protease enzyme of the human immunodeficiency virus, HIV-1) to designing potatoes that yield spot-free potato chips (via transgenic potatoes with altered carbohydrate metabolism). In the course of writing this text, the notes have expanded to function also as an introduction to the field for scientists in other disciplines by providing a global perspective into problems and approaches, rather than a comprehensive survey.

As a textbook, my intention is to provide a framework for teachers rather than a rigid guide, with material to be supplemented or substituted as appropriate for the audience. As a reference book, scientists who are interested in learning about biomolecular modeling may view the book as a broad introduction to an exciting new field with a host of challenging, interdisciplinary problems.

The intended audience for the course is beginning graduate students in medical schools and in all scientific departments: biology, chemistry, physics, mathematics, computer science, and others. This interdisciplinary audience presents a special challenge: it requires a broad presentation of the field but also good coverage of specialized topics to keep experts interested. Ideally, a good grounding in basic biochemistry, chemical physics, statistical and quantum mechanics, scientific computing (i.e., numerical methods), and programming techniques is desired. The rarity of such a background required me to offer tutorials in both biological and mathematical areas.

The introductory chapters on biomolecular structure are included in this book (after much thought) and are likely to be of interest to physical and mathematical scientists. Chapters 3 and 4 on proteins, together with Chapters 5 and 6 on nucleic acids, are thus highly abbreviated versions of what can be found in numerous texts specializing in these subjects. The selections in these tutorials also reflect some of my group's areas of interest. Because many introductory and up-to-date texts exist for protein structure, only the basics in protein structure are provided, while a somewhat more expanded treatment is devoted to nucleic acids.

Similarly, the introductory material on mathematical subjects such as basic optimization theory (Chapter 10) and random number generators (Chapter 11) is likely to be of use more to readers in the biological / chemical disciplines. General readers, as well as course instructors, can skip around this book as appropriate and fill in necessary gaps through other texts (e.g., in protein structure or programming techniques).

## Text Limitations

*By construction, this book is very broad in scope and thus no subjects are covered in great depth. References to the literature are only representative. The material presented is necessarily selective, unbalanced in parts, and reflects some of my areas of interest and expertise. This text should thus be viewed as an attempt to introduce the discipline of molecular modeling to students and to scientists from disparate fields, and should be taken together with other related texts, such as those listed in Appendix C, and the representative references cited.*

The book format is somewhat unusual for a textbook in that it is nonlinear in parts. For example, protein folding is introduced early (before protein basics are discussed) to illustrate challenging problems in the field and to interest more advanced readers; the introduction to molecular dynamics incorporates illustrations that require more advanced techniques for analysis; some specialized topics are also included throughout. For this reason, I recommend that students re-read certain parts of the book (e.g., first two chapters) after covering others (e.g., the biomolecular tutorial chapters). Still, I hope most of all to grab the reader's attention with exciting and current topics.

Given the many caveats of introducing and teaching such a broad and interdisciplinary subject as molecular modeling, the book aims to introduce selected biomolecular modeling and simulation techniques, as well as the wide range of biomolecular problems being tackled with these methods. Throughout these presentations, the central goal is to develop in students a good understanding of the inherent approximations and errors in the field so that they can adequately assess modeling results. Diligent students should emerge with basic knowledge in modeling and simulation techniques, an appreciation of the fundamental problems — such as force field approximations, nonbonded evaluation protocols, size and timestep limitations in simulations — and a healthy critical eye for research. A historical perspective and a discussion of future challenges are also offered.

## Dazzling Modeling Advances Demand Perspective

The topics I chose for this course are based on my own unorthodox introduction to the field of modeling. As an applied mathematician, I became interested in the field during my graduate work, hearing from Professor Suse Broyde — whose path I crossed thanks to Courant Professor Michael Overton — about the fascinating problem of modeling carcinogen/DNA adducts.

The goal was to understand some structural effects induced by certain compounds on the DNA (deduced by energy minimization); such alterations can render DNA more sensitive to replication errors, which in turn can eventually lead to mutagenesis and carcinogenesis. I had to roam through many references to obtain a grasp of some of the underlying concepts involving force fields and simulation protocols, so many of which seemed so approximate and not fully physically grounded. By now, however, I have learned to appreciate the practical procedures and compromises that computational chemists have formulated out of sheer necessity to obtain answers and insights into important biological processes that cannot be tackled by instrumentation. In fact, approximations and simplifications are not only tolerated when dealing with biomolecules; they often lead to insights that cannot easily be obtained from more detailed representations. Furthermore, it is often the neglect of certain factors that teaches us their importance, sometimes in subtle ways.

For example, when Suse Broyde and I viewed in the mid 1980s her intriguing carcinogen/modified DNA models, we used a large Evans and Sutherland computer while wearing special stereoviewers; the hard-copy drawings were ball and stick models, though the dimensionality projected out nicely in black and white. (Today, we still use stereo glasses, but current hardware stereo capabilities are much better, and marvelous molecular renderings are available). At that time, only small pieces of DNA could be modeled, and the surrounding salt and solvent environment was approximated. Still, structural and functional insights arose from those earlier works, many of which were validated later by more comprehensive computation, as well as laboratory experiments.

## Book Overview

The book provides an overview of three broad topics: (a) biomolecular structure and modeling: current problems and state of computations (Chapters 1–6); (b) molecular mechanics: force field origin, composition, and evaluation techniques (Chapters 7–9); and (c) simulation techniques: conformational sampling by geometry optimization, Monte Carlo, and molecular dynamics approaches (Chapters 10–13). Chapter 14 on the similarity and diversity problems in chemical design introduces some of the challenges in the growing field related to combinatorial chemistry.

Specifically, Chapters 1 and 2 give a historical perspective of biomolecular modeling, outlining progress in experimental techniques, the current

computational challenges, and the practical applications of this enterprise — to convey the immense interest in, and support of, the discipline. *Since these chapters discuss rapidly changing subjects (e.g., genome projects, disease treatments), they will be updated as possible on the text website.* General readers may find these chapters useful as an introduction to biomolecular modeling and its applications.

Chapters 3 and 4 review the basic elements in protein structure, and Chapter 5 similarly presents the basic building blocks and conformational flexibility in nucleic acids. Chapter 6 presents additional topics in nucleic acids, such as DNA sequence effects, DNA/protein interactions, departures from the canonical DNA helix forms, RNA structure, and DNA supercoiling.

The second part of the book begins in Chapter 7 with a view of the discipline of molecular mechanics as an offspring of quantum mechanics and discusses the basic premises of molecular mechanics formulations. A detailed presentation of the force field terms — origin, variation, and parameterization — is given in Chapter 8. Chapter 9 is then devoted to the computation of the nonbonded energy terms, including cutoff techniques, Ewald and multipole schemes, and continuum solvation alternatives.

The third part of the book, simulation algorithms,<sup>1</sup> begins with a description of optimization methods for multivariate functions in Chapter 10, emphasizing the tradeoff between algorithm complexity and performance. Basic issues of Monte Carlo techniques, appropriate to a motivated novice, are detailed in Chapter 11, such as pseudorandom number generators, Gaussian random variates, Monte Carlo sampling, and the Metropolis algorithm. Chapters 12 and 13 describe the algorithmic challenges in biomolecular dynamics simulations and present various categories of integration techniques, from the popular Verlet algorithm to multiple-timestep techniques and Brownian dynamics protocols. Chapter 14 outlines the challenges in similarity and diversity sampling in the field of chemical design, related to the new field of combinatorial chemistry.

The book appendices complement the material in the main text through homework assignments, reading lists, and other information useful for teaching molecular modeling.

Instructors may find the sample course syllabus in Appendix A helpful. Important also to teaching is an introduction to the original literature; a representative reading list of articles used for the course is collected in Appendix B. An annotated general reference list is given in Appendix C.

Selected biophysics applications are highlighted through the homework assignments (Appendix D). Humor in the assignments stimulates creativity in many students. These homeworks are a central component of learning molecular

---

<sup>1</sup>The word *algorithm* is named after the ninth-century Persian (Iranian in present-day terminology) mathematician al-Khwarizmi (nicknamed after his home town of Khwarizm, now Khiva in the Uzbek Republic), who stressed the importance of methodical procedures for solving problems in his algebra textbook. The term has evolved to mean the systematic process of solving problems by machine execution.

modeling, as they provide hands-on experience, extend upon subjects covered in the chapters, and expose the students to a wide range of current topics in biomolecular structure. Advanced students may use these homework assignments to learn about molecular modeling through independent research.

Many homework assignments involve a molecular modeling software package. I selected the *Insight* program in conjunction with our Silicon Graphics computer laboratory, but other suitable modeling programs can be used. Students also learn other basic research tools (such as programming and literature searches) through the homeworks.

Our memorable “force field debate” (see homework 7 in Appendix D) even brought the AMBER team to class in white lab coats, each accented with a name tag corresponding to one of AMBER’s original authors. The late Peter Kollman would have been pleased. Harold Scheraga would have been no less impressed by the long list of ECEPP successes prepared by his loyal troopers. Martin Karplus would not have been disappointed by the strong proponents of the CHARMM approach. I only hope to have as much spunk and talent in my future molecular modeling classes.

Extensive use of web resources is encouraged, while keeping in mind the caveat of lack of general quality control. I was amazed to find some of my students’ discoveries regarding interesting molecular modeling topics mentioned in the classroom, especially in the context of the term project, which requires them to find outstanding examples of the successes and/or failures of molecular modeling.

Interested readers might also want to glance at additional course information as part of my group’s home page, [monod.biomath.nyu.edu/](http://monod.biomath.nyu.edu/). Supplementary text information (such as program codes and figure files) can also be obtained.

To future teachers of molecular modeling who plan to design similar assignments and material, I share with you my following experience regarding student reactions to this discipline: what excited students the most about the subject matter and led to enthusiasm and excellent feedback in the classroom were the rapid pace at which the field is developing, its exciting discoveries, and the medical and technological breakthroughs made possible by important findings in the field.

In more practical terms, a mathematics graduate student, Brynja Kohler, expressed this enthusiasm succinctly in the introduction to her term project:

As I was doing research for this assignment, I found that one interesting article led to another. Communication via e-mail with some researchers around the world about their current investigations made me eagerly anticipate new results. The more I learned the more easy it became to put off writing a final draft because my curiosity would lead me on yet another line of inquiry. However, alas, there comes a time when even the greatest procrastinator must face the music, and evaluate what it is that we know and not linger upon what we hope to find out.

Future teachers are thus likely to have an enjoyable experience with any good group of students.

## Acknowledgments

I am indebted to Jing Huang for her devoted assistance with the manuscript preparation, file backups, data collection, and figure design. I also thank Wei Xu and Mulin Ding for important technical assistance. I am grateful to my other devoted current and former group members who helped read book segments, collect data, prepare the figures found throughout this book, and run to libraries throughout New York City often: Karunesh Arora, Danny Barash, Paul Batcho, Dan Beard, Mulin Ding, Hin Hark Gan, Jennifer Isbell, Joyce Noah, Xiaoliang Qian, Sonia Rivera, Adrian Sandu, Dan Strahs, Dexuan Xie, Linjing Yang, and Qing Zhang. Credits for each book figure and table are listed on the text's website.

I thank my colleagues Ruben Abagyan, Helen Berman, Dave Case, Jonathan Goodman, Andrej Sali, and Harold Scheraga, who gave excellent guest lectures in the course; and my course assistants Karunesh Arora, Margaret Mandziuk, Qing Zhang, and Zhongwei Zhu for their patient, dedicated assistance to the students with their homework and queries.

I am also very appreciative of the following colleagues for sharing reprints, information, and unpublished data and/or for their willingness to comment on segments of the book: Lou Allinger, Nathan Baker, Mike Beer, Helen Berman, Suse Broyde, John Board, Dave Beveridge, Ken Breslauer, Steve Burley, Dave Case, Philippe Derreumaux, Ron Elber, Eugene Fluder, Leslie Greengard, Steve Harvey, Jan Hermans, the late Peter Kollman, Robert Krasny, Michael Levitt, Xiang-Jun Lu, Pierre L'Ecuyer, Neocles Leontis, the late Shneior Lifson, Kenny Lipkowitz, Jerry Manning, Andy McCammon, Mihaly Mezei, Jorge NoCEDAL, Wilma Olson, Michael Overton, Vijay Pande, Dinshaw Patel, Harold Scheraga, Shulamith Schlick, Klaus Schulten, Suresh Singh, Bob Skeel, A.R. Srinivasan, Emad Tajkhorshid, Yuri Ushkaryov, Wilfred van Gunsteren, Arieh Warshel, Eric Westhof, Weitao Yang, and Darren York. Of special note are the extremely thorough critiques which I received from Lou Allinger, Steve Harvey, Jerry Manning, Robert Krasny, Wilma Olson, and Bob Skeel; their extensive comments and suggestions led to enlightening discussions and helped me see the field from many perspectives. I thank my colleague and friend Suse Broyde for introducing me to the field and for reading nearly every page of this book's draft.

To my family — parents Haim and Shula, sisters Yael and Daphne, aunt Cecilia, and especially Rick and Duboni — I am grateful for tolerating my long months on this project.

Finally, I thank my excellent students for making the course enjoyable and inspiring.

Tamar Schlick

New York, NY  
June 10, 2002

# Prelude

Every sentence I utter must be understood not as an affirmation but as a question.

Niels Bohr (1885–1962).

Only rarely does science undergo a dramatic transformation that can be likened to a tectonic rumble, as its character is transfigured under the weights of changing forces. *We are now in such an exciting time.* The discovery of the DNA double helix in the early 1950s prefigured the rise of molecular biology and its many offspring in the next half century, just as the rise of Internet technology in the 1980s and 1990s has molded, and is still reshaping, nearly every aspect of contemporary life. With completion of the first draft of the human genome sequence trumpeting the beginning of the twenty-first century, triumphs in the biological sciences are competing with geopolitics and the economy for prominent-newspaper headlines. The genomic sciences now occupy the center stage, linking basic to applied (medical) research, applied research to commercial success and economic growth, and the biological sciences to the chemical, physical, mathematical and computer sciences.

The subject of this text, molecular modeling, represents a subfield of this successful marriage. In this text, I attempt to draw to the field newcomers from other disciplines and to share basic knowledge in a modern context and interdisciplinary perspective. Though many details on current investigations and projects will undoubtedly become obsolete as soon as this book goes to press, the basic foundations of modeling will remain similar. Over the next decades, we will surely witness a rapid growth in the field of molecular modeling, as well as many success stories in its application.

# Contents

<b>About the Cover</b>	v
<b>Book URLs</b>	ix
<b>Preface</b>	xi
<b>Prelude</b>	xix
<b>Table of Contents</b>	xxi
<b>List of Figures</b>	xxxiii
<b>List of Tables</b>	xxxix
<b>Acronyms, Abbreviations, and Units</b>	xli
<b>1 Biomolecular Structure and Modeling: Historical Perspective</b>	1
1.1 A Multidisciplinary Enterprise . . . . .	2
1.1.1 Consilience . . . . .	2
1.1.2 What is Molecular Modeling? . . . . .	3
1.1.3 Need For Critical Assessment . . . . .	5
1.1.4 Text Overview . . . . .	6
1.2 The Roots of Molecular Modeling in Molecular Mechanics . . . . .	8
1.2.1 The Theoretical Pioneers . . . . .	8
1.2.2 Biomolecular Simulation Perspective . . . . .	11

1.3	Emergence of Biomodeling from Experimental Progress in Proteins and Nucleic Acids . . . . .	14
1.3.1	Protein Crystallography . . . . .	14
1.3.2	DNA Structure . . . . .	17
1.3.3	The Technique of X-ray Crystallography . . . . .	18
1.3.4	The Technique of NMR Spectroscopy . . . . .	20
1.4	Modern Era of Technological Advances . . . . .	22
1.4.1	From Biochemistry to Biotechnology . . . . .	22
1.4.2	PCR and Beyond . . . . .	23
1.5	Genome Sequencing . . . . .	25
1.5.1	Projects Overview: From Bugs to Baboons . . . . .	25
1.5.2	The Human Genome . . . . .	30
<b>2</b>	<b>Biomolecular Structure and Modeling: Problem and Application Perspective</b>	<b>41</b>
2.1	Computational Challenges in Structure and Function . . . . .	41
2.1.1	Analysis of the Amassing Biological Databases . . . . .	41
2.1.2	Computing Structure From Sequence . . . . .	46
2.2	Protein Folding – An Enigma . . . . .	46
2.2.1	‘Old’ and ‘New’ Views . . . . .	46
2.2.2	Folding Challenges . . . . .	48
2.2.3	Folding by Dynamics Simulations? . . . . .	49
2.2.4	Folding Assistants . . . . .	50
2.2.5	Unstructured Proteins . . . . .	52
2.3	Protein Misfolding – A Conundrum . . . . .	53
2.3.1	Prions and Mad Cows . . . . .	53
2.3.2	Infectious Protein? . . . . .	53
2.3.3	Other Possibilities . . . . .	54
2.3.4	Other Misfolding Processes . . . . .	55
2.3.5	Deducing Function From Structure . . . . .	56
2.4	From Basic to Applied Research . . . . .	57
2.4.1	Rational Drug Design: Overview . . . . .	58
2.4.2	A Classic Success Story: AIDS Therapy . . . . .	58
2.4.3	Other Drugs and Future Prospects . . . . .	65
2.4.4	Gene Therapy – Better Genes . . . . .	67
2.4.5	Designed Compounds and Foods . . . . .	69
2.4.6	Nutrigenomics . . . . .	72
2.4.7	Designer Materials . . . . .	74
2.4.8	Cosmeceuticals . . . . .	74
<b>3</b>	<b>Protein Structure Introduction</b>	<b>77</b>
3.1	The Machinery of Life . . . . .	77
3.1.1	From Tissues to Hormones . . . . .	77
3.1.2	Size and Function Variability . . . . .	78
3.1.3	Chapter Overview . . . . .	79

3.2	The Amino Acid Building Blocks . . . . .	82
3.2.1	Basic C <sup>α</sup> Unit . . . . .	82
3.2.2	Essential and Nonessential Amino Acids . . . . .	83
3.2.3	Linking Amino Acids . . . . .	85
3.2.4	The Amino Acid Repertoire: From Flexible Glycine to Rigid Proline . . . . .	85
3.3	Sequence Variations in Proteins . . . . .	89
3.3.1	Globular Proteins . . . . .	90
3.3.2	Membrane and Fibrous Proteins . . . . .	90
3.3.3	Emerging Patterns from Genome Databases . . . . .	92
3.3.4	Sequence Similarity . . . . .	92
3.4	Protein Conformation Framework . . . . .	97
3.4.1	The Flexible $\phi$ and $\psi$ and Rigid $\omega$ Dihedral Angles .	97
3.4.2	Rotameric Structures . . . . .	99
3.4.3	Ramachandran Plots . . . . .	99
3.4.4	Conformational Hierarchy . . . . .	103
<b>4</b>	<b>Protein Structure Hierarchy</b>	<b>105</b>
4.1	Structure Hierarchy . . . . .	106
4.2	Helices: A Common Secondary Structural Element . . . . .	106
4.2.1	Classic $\alpha$ -Helix . . . . .	106
4.2.2	$3_{10}$ and $\pi$ Helices . . . . .	107
4.2.3	Left-Handed $\alpha$ -Helix . . . . .	109
4.2.4	Collagen Helix . . . . .	110
4.3	$\beta$ -Sheets: A Common Secondary Structural Element . . . . .	110
4.4	Turns and Loops . . . . .	110
4.5	Formation of Supersecondary and Tertiary Structure . . . . .	113
4.5.1	Complex 3D Networks . . . . .	113
4.5.2	Classes in Protein Architecture . . . . .	113
4.5.3	Classes are Further Divided into Folds . . . . .	114
4.6	$\alpha$ -Class Folds . . . . .	114
4.6.1	Bundles . . . . .	114
4.6.2	Folded Leafs . . . . .	115
4.6.3	Hairpin Arrays . . . . .	115
4.7	$\beta$ -Class Folds . . . . .	115
4.7.1	Anti-Parallel $\beta$ Domains . . . . .	116
4.7.2	Parallel and Antiparallel Combinations . . . . .	116
4.8	$\alpha/\beta$ and $\alpha + \beta$ -Class Folds . . . . .	117
4.8.1	$\alpha/\beta$ Barrels . . . . .	117
4.8.2	Open Twisted $\alpha/\beta$ Folds . . . . .	118
4.8.3	Leucine-Rich $\alpha/\beta$ Folds . . . . .	118
4.8.4	$\alpha+\beta$ Folds . . . . .	118
4.8.5	Other Folds . . . . .	118
4.9	Number of Folds . . . . .	118
4.9.1	Finite Number? . . . . .	119

4.10	Quaternary Structure . . . . .	119
4.10.1	Viruses . . . . .	119
4.10.2	From Ribosomes to Dynamic Networks . . . . .	123
4.11	Protein Structure Classification . . . . .	126
<b>5</b>	<b>Nucleic Acids Structure Minitutorial</b>	<b>129</b>
5.1	DNA, Life's Blueprint . . . . .	130
5.1.1	The Kindled Field of Molecular Biology . . . . .	130
5.1.2	Fundamental DNA Processes . . . . .	132
5.1.3	Challenges in Nucleic Acid Structure . . . . .	133
5.1.4	Chapter Overview . . . . .	134
5.2	The Basic Building Blocks of Nucleic Acids . . . . .	135
5.2.1	Nitrogenous Bases . . . . .	135
5.2.2	Hydrogen Bonds . . . . .	136
5.2.3	Nucleotides . . . . .	137
5.2.4	Polynucleotides . . . . .	137
5.2.5	Stabilizing Polynucleotide Interactions . . . . .	140
5.2.6	Chain Notation . . . . .	140
5.2.7	Atomic Labeling . . . . .	141
5.2.8	Torsion Angle Labeling . . . . .	142
5.3	Nucleic Acid Conformational Flexibility . . . . .	142
5.3.1	The Furanose Ring . . . . .	143
5.3.2	Backbone Torsional Flexibility . . . . .	145
5.3.3	The Glycosyl Rotation . . . . .	148
5.3.4	Sugar/Glycosyl Combinations . . . . .	148
5.3.5	Basic Helical Descriptors . . . . .	150
5.3.6	Base-Pair Parameters . . . . .	151
5.4	Canonical DNA Forms . . . . .	155
5.4.1	B-DNA . . . . .	156
5.4.2	A-DNA . . . . .	157
5.4.3	Z-DNA . . . . .	160
5.4.4	Comparative Features . . . . .	161
<b>6</b>	<b>Topics in Nucleic Acids Structure: DNA Interactions and Folding</b>	<b>163</b>
6.1	Introduction . . . . .	164
6.2	DNA Sequence Effects . . . . .	165
6.2.1	Local Deformations . . . . .	165
6.2.2	Orientation Preferences in Dinucleotide Steps . . . . .	166
6.2.3	Orientation Preferences in Dinucleotide Steps With Flanking Sequence Context: Tetranucleotide Studies . . . . .	169
6.2.4	Intrinsic DNA Bending in A-Tracts . . . . .	169
6.2.5	Sequence Deformability Analysis Continues . . . . .	173

6.3	DNA Hydration and Ion Interactions . . . . .	174
6.3.1	Resolution Difficulties . . . . .	175
6.3.2	Basic Patterns . . . . .	176
6.4	DNA/Protein Interactions . . . . .	180
6.5	Cellular Organization of DNA . . . . .	182
6.5.1	Compaction of Genomic DNA . . . . .	182
6.5.2	Coiling of the DNA Helix Itself . . . . .	184
6.5.3	Chromosomal Packaging of Coiled DNA . . . . .	185
6.6	Mathematical Characterization of DNA Supercoiling . . . . .	195
6.6.1	DNA Topology and Geometry . . . . .	195
6.7	Computational Treatments of DNA Supercoiling . . . . .	197
6.7.1	DNA as a Flexible Polymer . . . . .	198
6.7.2	Elasticity Theory Framework . . . . .	199
6.7.3	Simulations of DNA Supercoiling . . . . .	200
<b>7</b>	<b>Topics in Nucleic Acids Structure: Noncanonical Helices and RNA Structure</b>	<b>205</b>
7.1	Introduction . . . . .	205
7.2	Variations on a Theme . . . . .	206
7.2.1	Hydrogen Bonding Patterns in Polynucleotides . . . . .	206
7.2.2	Hybrid Helical/Nonhelical Forms . . . . .	210
7.2.3	Unusual Forms: Overstretched and Understretched DNA . . . . .	214
7.3	RNA Structure and Function . . . . .	216
7.3.1	DNA's Cousin Shines . . . . .	216
7.3.2	RNA Chains Fold Upon Themselves . . . . .	216
7.3.3	RNA's Diversity . . . . .	217
7.3.4	Non-Coding and Micro-RNAs . . . . .	221
7.3.5	RNA at Atomic Resolution . . . . .	222
7.4	Current Challenges in RNA Modeling . . . . .	225
7.4.1	RNA Folding . . . . .	225
7.4.2	RNA Motifs . . . . .	225
7.4.3	RNA Structure Prediction . . . . .	226
7.5	Application of Graph Theory to Studies of RNA Structure and Function . . . . .	229
7.5.1	Graph Theory . . . . .	229
7.5.2	RNA-As-Graphs (RAG) Resource . . . . .	230
<b>8</b>	<b>Theoretical and Computational Approaches to Biomolecular Structure</b>	<b>237</b>
8.1	The Merging of Theory and Experiment . . . . .	238
8.1.1	Exciting Times for Computationalists! . . . . .	238
8.1.2	The Future of Biocomputations . . . . .	240
8.1.3	Chapter Overview . . . . .	240

8.2	Quantum Mechanics (QM) Foundations of Molecular Mechanics (MM) . . . . .	241
8.2.1	The Schrödinger Wave Equation . . . . .	241
8.2.2	The Born-Oppenheimer Approximation . . . . .	242
8.2.3	Ab Initio QM . . . . .	242
8.2.4	Semi-Empirical QM . . . . .	244
8.2.5	Recent Advances in Quantum Mechanics . . . . .	244
8.2.6	From Quantum to Molecular Mechanics . . . . .	247
8.3	Molecular Mechanics: Underlying Principles . . . . .	251
8.3.1	The Thermodynamic Hypothesis . . . . .	251
8.3.2	Additivity . . . . .	252
8.3.3	Transferability . . . . .	254
8.4	Molecular Mechanics: Model and Energy Formulation . . . . .	256
8.4.1	Configuration Space . . . . .	258
8.4.2	Functional Form . . . . .	259
8.4.3	Some Current Limitations . . . . .	262
<b>9</b>	<b>Force Fields</b>	<b>265</b>
9.1	Formulation of the Model and Energy . . . . .	266
9.2	Normal Modes . . . . .	267
9.2.1	Quantifying Characteristic Motions . . . . .	267
9.2.2	Complex Biomolecular Spectra . . . . .	269
9.2.3	Spectra As Force Constant Sources . . . . .	269
9.2.4	In-Plane and Out-of-Plane Bending . . . . .	271
9.3	Bond Length Potentials . . . . .	272
9.3.1	Harmonic Term . . . . .	273
9.3.2	Morse Term . . . . .	274
9.3.3	Cubic and Quartic Terms . . . . .	275
9.4	Bond Angle Potentials . . . . .	276
9.4.1	Harmonic and Trigonometric Terms . . . . .	277
9.4.2	Cross Bond Stretch / Angle Bend Terms . . . . .	278
9.5	Torsional Potentials . . . . .	281
9.5.1	Origin of Rotational Barriers . . . . .	281
9.5.2	Fourier Terms . . . . .	281
9.5.3	Torsional Parameter Assignment . . . . .	282
9.5.4	Improper Torsion . . . . .	286
9.5.5	Cross Dihedral/Bond Angle and Improper/Improper Dihedral Terms . . . . .	287
9.6	The van der Waals Potential . . . . .	288
9.6.1	Rapidly Decaying Potential . . . . .	288
9.6.2	Parameter Fitting From Experiment . . . . .	289
9.6.3	Two Parameter Calculation Protocols . . . . .	289
9.7	The Coulomb Potential . . . . .	291
9.7.1	Coulomb's Law: Slowly Decaying Potential . . . . .	291
9.7.2	Dielectric Function . . . . .	292
9.7.3	Partial Charges . . . . .	294

9.8	Parameterization . . . . .	295
9.8.1	A Package Deal . . . . .	295
9.8.2	Force Field Comparisons . . . . .	295
9.8.3	Force Field Performance . . . . .	297
<b>10</b>	<b>Nonbonded Computations</b>	<b>299</b>
10.1	A Computational Bottleneck . . . . .	301
10.2	Approaches for Reducing Computational Cost . . . . .	302
10.2.1	Simple Cutoff Schemes . . . . .	302
10.2.2	Ewald and Multipole Schemes . . . . .	303
10.3	Spherical Cutoff Techniques . . . . .	304
10.3.1	Technique Categories . . . . .	304
10.3.2	Guidelines for Cutoff Functions . . . . .	305
10.3.3	General Cutoff Formulations . . . . .	306
10.3.4	Potential Switch . . . . .	307
10.3.5	Force Switch . . . . .	308
10.3.6	Shift Functions . . . . .	309
10.4	The Ewald Method . . . . .	311
10.4.1	Periodic Boundary Conditions . . . . .	311
10.4.2	Ewald Sum and Crystallography . . . . .	314
10.4.3	Mathematical Morphing of a Conditionally Convergent Sum . . . . .	316
10.4.4	Finite-Dielectric Correction . . . . .	320
10.4.5	Ewald Sum Complexity . . . . .	320
10.4.6	Resulting Ewald Summation . . . . .	321
10.4.7	Practical Implementation: Parameters, Accuracy, and Optimization . . . . .	322
10.5	The Multipole Method . . . . .	324
10.5.1	Basic Hierarchical Strategy . . . . .	324
10.5.2	Historical Perspective . . . . .	329
10.5.3	Expansion in Spherical Coordinates . . . . .	330
10.5.4	Biomolecular Implementations . . . . .	332
10.5.5	Other Variants . . . . .	333
10.6	Continuum Solvation . . . . .	333
10.6.1	Need for Simplification! . . . . .	333
10.6.2	Potential of Mean Force . . . . .	334
10.6.3	Stochastic Dynamics . . . . .	335
10.6.4	Continuum Electrostatics . . . . .	338
<b>11</b>	<b>Multivariate Minimization in Computational Chemistry</b>	<b>345</b>
11.1	Ubiquitous Optimization: From Enzymes to Weather to Economics . . . . .	347
11.1.1	Algorithmic Sophistication Demands Basic Understanding . . . . .	347
11.1.2	Chapter Overview . . . . .	347

11.2	Optimization Fundamentals . . . . .	348
11.2.1	Problem Formulation . . . . .	348
11.2.2	Independent Variables . . . . .	349
11.2.3	Function Characteristics . . . . .	349
11.2.4	Local and Global Minima . . . . .	351
11.2.5	Derivatives of Multivariate Functions . . . . .	353
11.2.6	The Hessian of Potential Energy Functions . . . . .	353
11.3	Basic Algorithmic Components . . . . .	356
11.3.1	Greedy Descent . . . . .	356
11.3.2	Line-Search-Based Descent Algorithm . . . . .	359
11.3.3	Trust-Region-Based Descent Algorithm . . . . .	361
11.3.4	Convergence Criteria . . . . .	362
11.4	The Newton-Raphson-Simpson-Fourier Method . . . . .	364
11.4.1	The One-Dimensional Version of Newton's Method . . . . .	364
11.4.2	Newton's Method for Minimization . . . . .	367
11.4.3	The Multivariate Version of Newton's Method . . . . .	368
11.5	Effective Large-Scale Minimization Algorithms . . . . .	369
11.5.1	Quasi-Newton (QN) . . . . .	370
11.5.2	Conjugate Gradient (CG) . . . . .	372
11.5.3	Truncated-Newton (TN) . . . . .	374
11.5.4	Simple Example . . . . .	376
11.6	Available Software . . . . .	378
11.6.1	Popular Newton and CG . . . . .	378
11.6.2	CHARMM's ABNR . . . . .	379
11.6.3	CHARMM's TN . . . . .	379
11.6.4	Comparative Performance on Molecular Systems . . . . .	379
11.7	Practical Recommendations . . . . .	380
11.8	Future Outlook . . . . .	383
<b>12</b>	<b>Monte Carlo Techniques</b>	<b>385</b>
12.1	MC Popularity . . . . .	386
12.1.1	A Winning Combination . . . . .	386
12.1.2	From Needles to Bombs . . . . .	387
12.1.3	Chapter Overview . . . . .	387
12.1.4	Importance of Error Bars . . . . .	388
12.2	Random Number Generators . . . . .	388
12.2.1	What is <i>Random</i> ? . . . . .	388
12.2.2	Properties of Generators . . . . .	389
12.2.3	Linear Congruential Generators (LCG) . . . . .	392
12.2.4	Other Generators . . . . .	396
12.2.5	Artifacts . . . . .	400
12.2.6	Recommendations . . . . .	401
12.3	Gaussian Random Variates . . . . .	403
12.3.1	Manipulation of Uniform Random Variables . . . . .	403
12.3.2	Normal Variates in Molecular Simulations . . . . .	403

12.3.3	Odeh/Evans Method . . . . .	404
12.3.4	Box/Muller/Marsaglia Method . . . . .	405
12.4	Means for Monte Carlo Sampling . . . . .	406
12.4.1	Expected Values . . . . .	406
12.4.2	Error Bars . . . . .	409
12.4.3	Batch Means . . . . .	410
12.5	Monte Carlo Sampling . . . . .	411
12.5.1	Density Function . . . . .	411
12.5.2	Dynamic and Equilibrium MC: Ergodicity, Detailed Balance . . . . .	411
12.5.3	Statistical Ensembles . . . . .	412
12.5.4	Importance Sampling: Metropolis Algorithm and Markov Chains . . . . .	413
12.6	Monte Carlo Applications to Molecular Systems . . . . .	418
12.6.1	Ease of Application . . . . .	418
12.6.2	Biased MC . . . . .	419
12.6.3	Hybrid MC . . . . .	420
12.6.4	Parallel Tempering and Other MC Variants . . . . .	421
<b>13</b>	<b>Molecular Dynamics: Basics</b>	<b>425</b>
13.1	Introduction: Statistical Mechanics by Numbers . . . . .	426
13.1.1	Why Molecular Dynamics? . . . . .	426
13.1.2	Background . . . . .	427
13.1.3	Outline of MD Chapters . . . . .	428
13.2	Laplace's Vision of Newtonian Mechanics . . . . .	429
13.2.1	The Dream Becomes Reality . . . . .	429
13.2.2	Deterministic Mechanics . . . . .	432
13.2.3	Neglect of Electronic Motion . . . . .	432
13.2.4	Critical Frequencies . . . . .	433
13.2.5	Hybrid Quantum/Classical Mechanics Treatments . . . . .	435
13.3	The Basics: An Overview . . . . .	435
13.3.1	Following the Equations of Motion . . . . .	435
13.3.2	Perspective on MD Trajectories . . . . .	436
13.3.3	Initial System Settings . . . . .	437
13.3.4	Sensitivity to Initial Conditions and Other Computational Choices . . . . .	440
13.3.5	Simulation Protocol . . . . .	442
13.3.6	High-Speed Implementations . . . . .	443
13.3.7	Analysis and Visualization . . . . .	445
13.3.8	Reliable Numerical Integration . . . . .	445
13.3.9	Computational Complexity . . . . .	446
13.4	The Verlet Algorithm . . . . .	448
13.4.1	Position and Velocity Propagation . . . . .	449
13.4.2	Leapfrog, Velocity Verlet, and Position Verlet . . . . .	451
13.5	Constrained Dynamics . . . . .	453

13.6	Various MD Ensembles . . . . .	455
13.6.1	Need for Other Ensembles . . . . .	455
13.6.2	Simple Algorithms . . . . .	456
13.6.3	Extended System Methods . . . . .	459
<b>14</b>	<b>Molecular Dynamics: Further Topics</b>	<b>463</b>
14.1	Introduction . . . . .	464
14.2	Symplectic Integrators . . . . .	465
14.2.1	Symplectic Transformation . . . . .	466
14.2.2	Harmonic Oscillator Example . . . . .	467
14.2.3	Linear Stability . . . . .	467
14.2.4	Timestep-Dependent Rotation in Phase Space . . . . .	469
14.2.5	Resonance Condition for Periodic Motion . . . . .	470
14.2.6	Resonance Artifacts . . . . .	471
14.3	Multiple-Timestep (MTS) Methods . . . . .	472
14.3.1	Basic Idea . . . . .	472
14.3.2	Extrapolation . . . . .	473
14.3.3	Impulses . . . . .	474
14.3.4	Vulnerability of Impulse Splitting to Resonance Artifacts . . . . .	475
14.3.5	Resonance Artifacts in MTS . . . . .	476
14.3.6	Limitations of Resonance Artifacts on Speedup; Possible Cures . . . . .	478
14.4	Langevin Dynamics . . . . .	479
14.4.1	Many Uses . . . . .	479
14.4.2	Phenomenological Heat Bath . . . . .	480
14.4.3	The Effect of $\gamma$ . . . . .	480
14.4.4	Generalized Verlet for Langevin Dynamics . . . . .	482
14.4.5	The LN Method . . . . .	482
14.5	Brownian Dynamics (BD) . . . . .	487
14.5.1	Brownian Motion . . . . .	487
14.5.2	Brownian Framework . . . . .	489
14.5.3	General Propagation Framework . . . . .	491
14.5.4	Hydrodynamic Interactions . . . . .	491
14.5.5	BD Propagation Scheme: Cholesky vs. Chebyshev Approximation . . . . .	494
14.6	Implicit Integration . . . . .	496
14.6.1	Implicit vs. Explicit Euler . . . . .	497
14.6.2	Intrinsic Damping . . . . .	498
14.6.3	Computational Time . . . . .	498
14.6.4	Resonance Artifacts . . . . .	499
14.7	Enhanced Sampling Methods . . . . .	503
14.7.1	Overview . . . . .	503
14.7.2	Harmonic-Analysis Based Techniques . . . . .	503
14.7.3	Other Coordinate Transformations . . . . .	505

14.7.4	Coarse Graining Models . . . . .	507
14.7.5	Biasing Approaches . . . . .	508
14.7.6	Variations in MD Algorithm and Protocol . . . . .	509
14.7.7	Other Rigorous Approaches for Deducing Mechanisms, Free Energies, and Reaction Rates . . . . .	511
14.8	Future Outlook . . . . .	513
14.8.1	Integration Ingenuity . . . . .	513
14.8.2	Current Challenges . . . . .	514
<b>15</b>	<b>Similarity and Diversity in Chemical Design</b>	<b>519</b>
15.1	Introduction to Drug Design . . . . .	520
15.1.1	Chemical Libraries . . . . .	520
15.1.2	Early Drug Development Work . . . . .	521
15.1.3	Molecular Modeling in Rational Drug Design . . . . .	523
15.1.4	The Competition: Automated Technology . . . . .	524
15.1.5	Chapter Overview . . . . .	526
15.2	Problems in Chemical Libraries . . . . .	526
15.2.1	Database Analysis . . . . .	526
15.2.2	Similarity and Diversity Sampling . . . . .	527
15.2.3	Bioactivity Relationships . . . . .	529
15.3	General Problem Definitions . . . . .	532
15.3.1	The Dataset . . . . .	532
15.3.2	The Compound Descriptors . . . . .	534
15.3.3	Characterizing Biological Activity . . . . .	535
15.3.4	The Target Function . . . . .	536
15.3.5	Scaling Descriptors . . . . .	536
15.3.6	The Similarity and Diversity Problems . . . . .	538
15.4	Data Compression and Cluster Analysis . . . . .	540
15.4.1	Data Compression Based on Principal Component Analysis (PCA) . . . . .	540
15.4.2	Data Compression Based on the Singular Value Decomposition (SVD) . . . . .	542
15.4.3	Relation Between PCA and SVD . . . . .	544
15.4.4	Data Analysis via PCA or SVD and Distance Refinement . . . . .	545
15.4.5	Projection, Refinement, and Clustering Example . . . . .	546
15.5	Future Perspectives . . . . .	551
<b>Epilogue</b>		<b>555</b>
<b>Appendices</b>		<b>556</b>
<b>A Molecular Modeling Sample Syllabus</b>		<b>557</b>
<b>B Article Reading List</b>		<b>559</b>

<b>C Supplementary Course Texts</b>	<b>563</b>
<b>D Homework Assignments</b>	<b>571</b>
<b>References</b>	<b>623</b>

# List of Figures

1.1	Crystal structure of two eubacterial ribosomal subunits solved by the Yonath group . . . . .	4
1.2	Cryo-EM view of 70S ribosome complex solved by the Frank group . . . . .	5
1.3	Simulation evolution (3D version) . . . . .	15
1.4	Simulation evolution (2D version) . . . . .	16
1.5	The progress of DNA sequencing technology. . . . .	34
2.1	Sequence and structure data . . . . .	44
2.2	Paracelsus' Janus . . . . .	47
2.3	GroEL/GroES chaperonin/co-chaperonin complex . . . . .	51
2.4	Prion protein . . . . .	57
2.5	AIDS drugs . . . . .	59
3.1	An amino acid . . . . .	79
3.2	Water clusters . . . . .	81
3.3	Dipeptide formation . . . . .	82
3.4	Peptide formula . . . . .	84
3.5	Aspartame . . . . .	84
3.6	The amino acid repertoire . . . . .	86
3.7	Amino acids structures . . . . .	87
3.8	Amino acid frequencies . . . . .	89
3.9	Fibrous proteins . . . . .	91
3.10	Rop . . . . .	93
3.11	EF proteins . . . . .	95

3.12	Protein-structure variants . . . . .	96
3.13	Gauche and trans orientations . . . . .	97
3.14	Dihedral angle . . . . .	98
3.15	Rotations in polypeptides . . . . .	99
3.16	Lysine rotamers . . . . .	99
3.17	Amino acids rotamers . . . . .	100
3.18	Ramachandran plots . . . . .	101
3.19	Further study of Ramachandran plots . . . . .	102
4.1	The $\alpha$ -helix and $\beta$ -sheet motifs . . . . .	107
4.2	$\alpha$ -helical proteins (a) . . . . .	108
4.3	$\alpha$ -helical proteins (b) . . . . .	109
4.4	$\beta$ -helical proteins (a) . . . . .	111
4.5	$\beta$ -helical proteins (b) . . . . .	112
4.6	$\alpha/\beta$ proteins . . . . .	120
4.7	$\alpha + \beta$ proteins . . . . .	121
4.8	Multi-domain proteins . . . . .	122
4.9	Membrane and cell surface proteins . . . . .	123
4.10	Other proteins . . . . .	124
4.11	Tomato bushy stunt virus . . . . .	125
5.1	The DNA double helix . . . . .	136
5.2	Nucleic acid components . . . . .	137
5.3	Watson-Crick base pairing . . . . .	138
5.4	The polynucleotide chain and labeling . . . . .	139
5.5	Sugar envelope and twist puckers . . . . .	143
5.6	Sugar pseudorotation cycle . . . . .	144
5.7	Common sugar puckers . . . . .	145
5.8	Sugar pucker clustering . . . . .	146
5.9	Torsion angle wheel . . . . .	147
5.10	Deoxyadenosine adiabatic map . . . . .	149
5.11	Base-pair coordinate system . . . . .	153
5.12	Base-pair step and base pair parameters . . . . .	155
5.13	Model A, B, and Z-DNA . . . . .	158
5.14	Model A, B, and Z-DNA, stereo side . . . . .	159
5.15	Model A, B, and Z-DNA, stereo top . . . . .	160
6.1	Bending in long DNA . . . . .	171
6.2	Net DNA bending examples . . . . .	172
6.3	A-tract DNA dodecamer . . . . .	173
6.4	Sequence-dependent local DNA hydration . . . . .	179
6.5	DNA/protein binding motifs . . . . .	183
6.6	Interwound and toroidal supercoiling . . . . .	184
6.7	Schematic view of DNA levels of folding . . . . .	187
6.8	Nucleosome core particle . . . . .	189

6.9	Tetranucleosome and other solved nucleosomes . . . . .	190
6.10	Chromatin fiber models . . . . .	191
6.11	Mesoscale chromatin model . . . . .	192
6.12	Polynucleosome modeling . . . . .	195
6.13	Supercoiling topology and geometry . . . . .	196
6.14	Brownian dynamics snapshots of DNA . . . . .	202
6.15	Site juxtaposition measurements . . . . .	203
7.1	Various hydrogen-bonding schemes . . . . .	208
7.2	DNA/protein complex with Hoogsteen bp . . . . .	209
7.3	Oligonucleotide analogues . . . . .	212
7.4	Various nucleotide-chain folding motifs . . . . .	218
7.5	RNA pseudoknot motif . . . . .	219
7.6	RNAs with pseudoknots . . . . .	220
7.7	Examples of solved riboswitches. . . . .	224
7.8	Seven major motifs of tertiary interactions in RNA . . . . .	226
7.9	RNA structure annotation . . . . .	227
7.10	RNA junctions annotation . . . . .	228
7.11	RNA as graphs . . . . .	230
7.12	RAG segments . . . . .	232
7.13	Candidate RNA motifs . . . . .	233
7.14	<i>In silico</i> pool selection . . . . .	234
8.1	QM/MM concept . . . . .	245
8.2	DNA quantum-mechanically derived electrostatic potentials . . . . .	248
8.3	Enolase active site . . . . .	249
8.4	DNA pol $\beta$ chemical synthesis reaction . . . . .	250
8.5	Molecular geometry . . . . .	251
8.6	CHARMM atom types . . . . .	257
9.1	Normal modes of a water molecule. . . . .	269
9.2	Computed protein and water spectra . . . . .	271
9.3	Vibrational modes types . . . . .	272
9.4	Bond-length potentials . . . . .	275
9.5	Bond angles . . . . .	278
9.6	Bond-angle potentials . . . . .	279
9.7	Stretch/bend cross terms . . . . .	280
9.8	Butane torsional orientations . . . . .	280
9.9	Torsion-angle potentials . . . . .	284
9.10	Model compounds for torsional parameterization . . . . .	285
9.11	Wilson angle . . . . .	287
9.12	Van der Waals potentials . . . . .	291
9.13	Coulomb potentials . . . . .	294

10.1	CPU time for nonbonded calculations . . . . .	302
10.2	Cutoff schemes . . . . .	306
10.3	Switch and shift functions . . . . .	309
10.4	Periodic domains . . . . .	312
10.5	Various periodic domains . . . . .	313
10.6	Space-filling polyhedra . . . . .	313
10.7	Ewald's trick of Gaussian masking . . . . .	319
10.8	CPU time for PME vs. fast multipole . . . . .	324
10.9	Fast multipole schemes . . . . .	326
10.10	Screened Coulomb potential . . . . .	341
10.11	Poisson-Boltzmann rendering of the 30S ribosome . . . . .	343
11.1	One-dimensional function . . . . .	351
11.2	2D Contour curves for quadratic functions . . . . .	354
11.3	3D curves for quadratic functions . . . . .	355
11.4	Sparse Hessians . . . . .	357
11.5	Sparse Hessians, continued . . . . .	358
11.6	Line search minimization . . . . .	360
11.7	Newton's method, simple illustration . . . . .	365
11.8	Newton's method, quadratic example output . . . . .	367
11.9	Newton's method, cubic example output . . . . .	368
11.10	Minimization paths . . . . .	377
11.11	Minimization progress . . . . .	381
12.1	Lattice structure for simple random number generators . . . . .	396
12.2	Structures for linear congruential generators . . . . .	397
12.3	MC computation of $\pi$ . . . . .	407
12.4	Boltzmann probabilities . . . . .	413
12.5	MC moves for DNA . . . . .	416
12.6	MC and BD DNA Distributions . . . . .	417
12.7	Bad MC protocol . . . . .	417
13.1	Sampling methods . . . . .	430
13.2	Equilibration . . . . .	439
13.3	Chaos in MD . . . . .	441
13.4	Butane's end-to-end distance . . . . .	442
13.5	Butane's end-to-end distance convergence . . . . .	444
13.6	Energy drift . . . . .	446
14.1	Effective Verlet phase space rotation . . . . .	470
14.2	Verlet resonance for a Morse oscillator . . . . .	472
14.3	Extrapolative vs. Impulse MTS . . . . .	474
14.4	Impulse vs. extrapolative force splitting . . . . .	476
14.5	Resonance from force splitting . . . . .	477
14.6	Harmonic oscillator Langevin trajectories . . . . .	481

14.7	BPTI means and variances by Langevin and Newtonian MTS . . . . .	483
14.8	LN algorithm . . . . .	484
14.9	Manhattan plots for polymerase/DNA . . . . .	486
14.10	Polymerase/DNA system . . . . .	487
14.11	BPTI spectral densities . . . . .	487
14.12	Polymerase/DNA spectral densities . . . . .	488
14.13	Polymerase/DNA geometry . . . . .	488
14.14	Cholesky vs. Chebyshev approaches for random force . . . . .	496
14.15	Implicit and explicit Euler . . . . .	498
14.16	Verlet and implicit-midpoint energies . . . . .	500
14.17	PCA examples . . . . .	506
14.18	Pol $\beta$ 's Closing Pathway. . . . .	512
14.19	Pol $\beta$ Kinetic Profile. . . . .	513
15.1	Sample drugs . . . . .	525
15.2	Related pairs of drugs . . . . .	527
15.3	Chemical library . . . . .	533
15.4	SVD/refinement performance . . . . .	547
15.5	SVD-based database projection in 2D and 3D . . . . .	548
15.6	Cluster analysis . . . . .	549
15.7	PCA projection in 2D, with similar pairs . . . . .	550
15.8	PCA projection in 2D, with diverse pairs . . . . .	551
D.1	Sample histogram for protein/DNA interaction analysis. . . . .	587
D.2	Biphenyl. . . . .	608
D.3	Structure for linear congruential generators . . . . .	613
D.4	Hydrogen bond geometry . . . . .	620

# List of Tables

1.1	Structural biology chronology . . . . .	7
1.2	Biomolecular simulation evolution . . . . .	9
2.1	Protein databases . . . . .	45
3.1	Amino acid frequency . . . . .	90
5.1	Genetic code . . . . .	132
5.2	Nucleic acid torsion angle definitions. . . . .	142
5.3	Mean properties of representative DNA forms. . . . .	151
5.4	Selected parameters for model DNA helices . . . . .	152
6.1	Base-pair step parameters for free and protein-bound DNA . . . . .	167
6.2	Protein/DNA complexes . . . . .	181
6.3	DNA content of representative genomes. . . . .	185
7.1	Some classes of non-coding RNA (ncRNA). . . . .	217
8.1	Some CHARMM atom types . . . . .	256
9.1	Characteristic stretching vibrational frequencies . . . . .	270
9.2	Characteristic bending and torsional vibrational frequencies . . . . .	270
9.3	Examples of torsional potentials . . . . .	286
10.1	CPU time for nonbonded calculations . . . . .	303

xl List of Tables

11.1 Optimization software . . . . .	378
11.2 Minimization comparisons . . . . .	382
12.1 MC calculations for $\pi$ . . . . .	408
13.1 Biomolecular sampling methods . . . . .	428
13.2 High-frequency modes . . . . .	434
13.3 Biomolecular timescales . . . . .	434
14.1 Verlet timestep restriction timescales . . . . .	468
14.2 Stability limits . . . . .	473

# Acronyms, Abbreviations, and Units

## A

A	adenine (purine nitrogenous base)
Å	angstrom ( $10^{-10}$ m)
AdMLP	adenovirus major late promoter (protein)
AIDS	acquired immune deficiency syndrome
Ala (A)	alanine
Arg (R)	arginine
Asn (N)	asparagine
Asp (D)	aspartic acid
AS	Altona/Sundaralingam (sugar description)
ATP	adenosine triphosphate (energy source)
AZT	zidovudine (AIDS drug)

## B

bp	base pair
bps	base pairs
BAC	bacterial artificial chromosome
BOES	Born-Oppenheimer energy surfaces
BPTI	bovine pancreatic trypsin inhibitor
BSE	bovine spongiform encephalopathy ('mad cow disease')

## C

cm	centimeter ( $10^{-2}$ m)
C	cytosine (pyrimidine nitrogenous base)

CAP	catabolite gene activator protein
CASP	Critical Assessment of Techniques for Protein Structure Prediction
CG	Conjugate gradient method (for minimization)
CJD	Creutzfeld-Jakob disease (brain disorder, human version of BSE)
CN	Crigler-Najjar (debilitating disease, gene therapy applications)
CP	Cremer/Pople (sugar description)
CPU	central processing units
Cys (C)	cysteine

## D

DFT	density functional theory (quantum mechanics approach)
DH	Debye-Hückel
DNA	deoxyribonucleic acid (also A-, B-, C-, D-, P-, S-, T-, and Z-DNA)
DOE	Department of Energy

## E

erg	energy unit ( $10^{-7}$ J)
EM	electron microscopy

## F

fs	femtosecond ( $10^{-15}$ s)
FFT	Fast Fourier Transforms

## G

G	guanine (purine nitrogenous base)
Gln (Q)	glutamine
Glu (E)	glutamic acid
Gly (G)	glycine
GSS	Gerstmann-Straussler-Scheinker disease (brain disorder similar to CJD)

## H

HDV	hepatitis delta helper virus
His (H)	histidine
HIV	human immunodeficiency virus
HMC	hybrid Monte Carlo
HTH	helix/turn/helix (motif)
Hz	hertz (inverse second)

## I

Ile (I)	isoleucine
IHF	integration host factor (protein)

**K**

kbp	kilobase pairs
kcal/mol	kilocalories per mole (energy unit)
kDa	kilodaltons (mass unit used for proteins)
KR	Kirkwood-Riseman

**L**

Leu (L)	leucine
Lys (K)	lysine
LCG	linear congruential generator

**M**

m	meter
mgr	minor groove
ms	millisecond ( $10^{-3}$ s)
$\mu$ s	microsecond ( $10^{-6}$ s)
mm	millimeter ( $10^{-3}$ m)
MAD	multiple isomorphous replacement (crystallography technique)
MC	Monte Carlo
MD	molecular dynamics
Met (M)	methionine
Mgr	major groove
MIR	multiwavelength anomalous diffraction (crystallography technique)
MLCG	multiplicative linear congruential generator
MTS	multiple-timestep methods (for MD)

**N**

nm	nanometer ( $10^{-9}$ m)
ns	nanosecond ( $10^{-9}$ s)
NCBI	National Center for Biotechnology Information
NASA	National Aeronautics and Space Administration
NDB	nucleic acid database
NIH	National Institutes of Health
NMR	nuclear magnetic resonance
NSF	National Science Foundation

**O**

OTC	ornithine transcarbamylase (chronic ailment, gene therapy applications)
-----	---

**P**

pn	picoNewton (force unit)
ps	picosecond ( $10^{-12}$ s)

PB	Poisson-Boltzmann
PBE	Poisson-Boltzmann equation
PC	principal component
PCA	principal component analysis
PCR	polymerase chain reaction
PDB	protein databank
Phe (F)	phenylalanine
PIR	Protein Information Resource
PME	particle-mesh Ewald
PNA	peptide nucleic acid (DNA mimic)
Pro (P)	proline
PrP <sup>C</sup>	prion protein cellular (harmless)
PrP <sup>Sc</sup>	harmful isoform of PrP <sup>C</sup> , causes scrapie in sheep
Pur	purine (base)
Pyr	pyrimidine (base)

**Q**

QM	quantum mechanics
QN	quasi Newton method (for minimization)
QSAR	quantitative structure/activity relationships

**R**

RCSB	Research Collaboratory for Structural Bioinformatics
RMS (rms)	root-mean-square
RMSD	root-mean-square deviations
RNA	ribonucleic acid (also cRNA, gRNA, mRNA, rRNA, snRNA, tRNA)
RT	reverse transcriptase (AIDS protein)

**S**

s	second
Ser (S)	serine
SAR	structure/activity relationships
SCF	self-consistent field (quantum mechanical approach)
SCOP	structural classification of proteins
SD	steepest descent method (for minimization)
SGI	Silicon Graphics Inc.
SNPs	single-nucleotide polymorphisms (“snips”)
SRY	sex determining region Y (protein)
STS	single-timestep methods (for MD)
SVD	singular value decomposition

**T**

T	thymine (pyrimidine nitrogenous base)
Thr (T)	threonine
Trp (W)	tryptophan
Tyr (Y)	tyrosine
TBP	TATA-box DNA binding protein (transcription regulator)
TE	transcription efficiency
TMD	targeted molecular dynamics
TN	truncated Newton method (for minimization)
2D	two-dimensional
3D	three-dimensional

**U**

U	uracil (pyrimidine nitrogenous base)
URL	uniform resource locator
UV	ultraviolet spectroscopy

**V**

Val (V)	valine
---------	--------

**W**

WC	Watson/Crick base pairing
----	---------------------------

# 1

## Biomolecular Structure and Modeling: Historical Perspective

### Chapter 1 Notation

SYMBOL	DEFINITION
<b>Vectors</b>	
$\mathbf{h}$	unit cell identifier (crystallography)
$\mathbf{r}$	position
$F_{\mathbf{h}}$	structure factor (crystallography)
$\phi_{\mathbf{h}}$	phase angle (crystallography)
<b>Scalars</b>	
$d$	distance between parallel planes in the crystal
$I_{\mathbf{h}}$	intensity, magnitude of structure factor (crystallography)
$V$	cell volume (crystallography)
$\theta$	reflection angle (crystallography)
$\lambda$	wavelength of the X-ray beam (crystallography)

... physics, chemistry, and biology have been connected by a web of causal explanation organized by induction-based theories that telescope into one another. ... Thus, quantum theory underlies atomic physics, which is the foundation of reagent chemistry and its specialized offshoot biochemistry, which interlock with molecular biology — essentially, the chemistry of organic macromolecules — and hence, through successively higher levels of organization, cellular,

organismic, and evolutionary biology. . . . Such is the unifying and highly productive understanding of the world that has evolved in the natural sciences.

Edward O. Wilson: “Resuming the Enlightenment Quest”, in *The Wilson Quarterly*, Winter 1998.

## 1.1 A Multidisciplinary Enterprise

### 1.1.1 *Consilience*

The exciting field of modeling molecular systems by computer has been steadily drawing increasing attention from scientists in varied disciplines. In particular, modeling large biological polymers — proteins, nucleic acids, and lipids — is a truly multidisciplinary enterprise. Biologists describe the cellular picture; chemists fill in the atomic and molecular details; physicists extend these views to the electronic level and the underlying forces; mathematicians analyze and formulate appropriate numerical models and algorithms; and computer scientists and engineers provide the crucial implementational support for running large computer programs on high-speed and extended-communication platforms. The many names for the field (and related disciplines) underscore its cross-disciplinary nature: computational biology, computational chemistry, *in silico* biology, computational structural biology, computational biophysics, theoretical biophysics, theoretical chemistry, and the list goes on.

As the pioneer of sociobiology Edward O. Wilson reflects in the opening quote, some scholars believe in a unifying knowledge for understanding our universe and ourselves, or *consilience*<sup>1</sup> that merges all disciplines in a biologically-grounded framework [1377]. Though this link is most striking between genetics and human behavior — through the neurobiological underpinnings of states of mind and mental activity, with shaping by the environment and lifestyle factors — such a unification that Wilson advocates might only be achieved by a close interaction among the varied scientists at many stages of study. The genomic era has such immense ramifications on every aspect of our lives — from health to technology to law — that it is not difficult to appreciate the effects of the biomolecular revolution on our 21st-century society. Undoubtedly, a more integrated synthesis of biological elements is needed to decode life [584].

In biomolecular modeling, a multidisciplinary approach is important not only because of the many aspects involved — from problem formulation to solution — but also since the best computational approach is often closely tailored to the

---

<sup>1</sup>*Consilience* was coined in 1840 by the theologian and polymath William Whewell in his synthesis *The Philosophy of the Inductive Sciences*. It literally means the *alignment*, or *jumping together*, of knowledge from different disciplines. The sociobiologist Edward O. Wilson took this notion further recently by advocating in his 1998 book *Consilience* [1377] that the world is orderly and can be explained by a set of natural laws that are fundamentally rooted in biology.

biological problem. In the same spirit, close connections between theory and experiment are essential: computational models evolve as experimental data become available, and biological theories and new experiments are performed as a result of computational insights.<sup>2</sup>

Although few theoreticians in the field have expertise in experimental work as well, the classic example of Werner Heisenberg's genius in theoretical physics but naiveté in experimental physics is a case in point: Heisenberg required the resolving power of the microscope to derive the uncertainty relations. In fact, an error in the experimental interpretations was pointed out by Niels Bohr, and this eventually led to the 'Copenhagen interpretation of quantum mechanics'.

If Wilson's vision is correct, the interlocking web of scientific fields rooted in the biological sciences will succeed ultimately in explaining not only the functioning of a biomolecule and the workings of the brain, but also many aspects of modern society, through the connections between our biological makeup and human behavior.

### 1.1.2 What is Molecular Modeling?

Molecular modeling is the science and art of studying molecular structure and function through model building and computation. The model building can be as simple as plastic templates or metal rods, or as sophisticated as interactive, animated color stereographics and laser-made wooden sculptures. The computations encompass *ab initio* and semi-empirical quantum mechanics, empirical (molecular) mechanics, molecular dynamics, Monte Carlo, free energy and solvation methods, structure/activity relationships (SAR), chemical/biochemical information and databases, and many other established procedures. The refinement of experimental data, such as from nuclear magnetic resonance (NMR) or X-ray crystallography, is also a component of biomolecular modeling.

I often remind my students of Pablo Picasso's statement on art: "*Art is the lie that helps tell the truth*". This view applies aptly to biomolecular modeling. Though our models represent a highly-simplified version of the complex cellular environment, systematic studies based on tractable quantitative tools can help discern patterns and add insights that are otherwise difficult to observe. *The key in modeling is to develop and apply models that are appropriate for the questions being examined with them*. Thus, the model's regime of applicability must be clearly defined and its predictability power demonstrated. A case in point is the use of limited historical data on home prices for extrapolative modeling of mortgage-backed securities and credit derivatives; the resulting mispricing of risk was a contributor to the U.S. subprime loan crisis that started in 2007.

The questions being addressed by computational approaches today are as intriguing and as complex as the biological systems themselves. They range

---

<sup>2</sup>See [176,362,395,396,948], for example, in connection to the characterization of protein folding mechanisms.

from understanding the equilibrium structure of a small biopolymer subunit, to the energetics of hydrogen-bond formation in proteins and nucleic acids, to the kinetics of protein folding, to the complex functioning of a supramolecular aggregate. As experimental triumphs are being reported in structure determination — from ion channel proteins, signaling receptor proteins (receptors), membrane transport proteins (transporters), ribosomes (see Figs. 1.1 and 1.2), various nucleosomes (see figures in Chapter 6), and non-coding RNAs — including new methodologies for their solution, such as advanced NMR, cryo-electron microscopy, and single-molecule biochemistry techniques, modeling approaches are needed to pursue many fundamental questions concerning their biological motions and functions. Modeling provides a way to systematically explore structural/dynamical/thermodynamic patterns, test and develop hypotheses, interpret and extend experimental data, and help better understand and extend basic laws that govern molecular structure, flexibility, and function. In tandem with experimental advances, algorithmic and computer technological advances, especially concerning distributed, loosely-coupled computer networks, have made problems and approaches that were insurmountable a few years ago now possible.

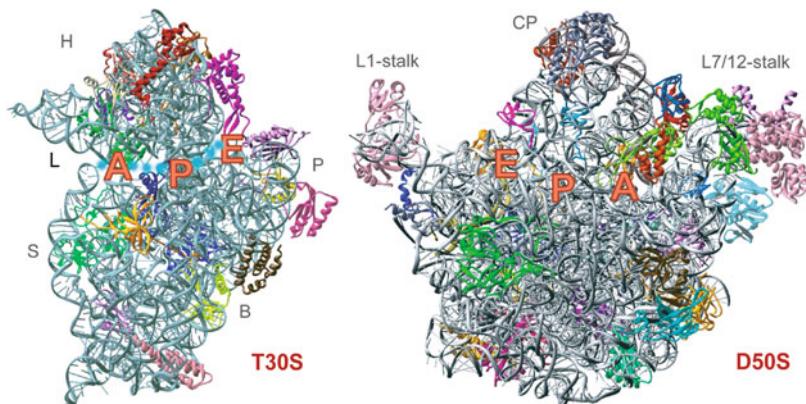


Figure 1.1. The inter-subunit interface of the two eubacterial ribosomal subunits at 3 Å resolution, showing their main architectural features. D50S is the large ribosomal subunit from *Deinococcus radiodurans* [516], and T30S is the small ribosomal subunit from *Thermus thermophilus* [1135], showing the head, platform, shoulder and latch (H,P,S,L, respectively). The cyan dots indicate the approximate mRNA channel; A, P, and E are the approximate positions of the anti-codon loops (on T30S) and the edges of the tRNAs acceptor stems (on D50S) of the three tRNA substrates: aminoacylated-tRNA (A), peptidyl-tRNA (P), and Exit tRNA (E). Image was kindly provided by Ada Yonath.

### 1.1.3 Need For Critical Assessment

The field of biomolecular modeling is relatively young, having started in the 1960s, and only gained momentum since the mid 1980s with the advent of supercomputers. Yet the field is developing with astonishing speed. Advances are driven by improvements in instrumental resolution and genomic and structural databases, as well as in force fields, algorithms for conformational sampling and molecular dynamics, computer graphics, and the increased computer power and memory capabilities. These impressive technological and modeling advances are steadily establishing the field of theoretical modeling as a partner to experiment and a widely used tool for research and development.

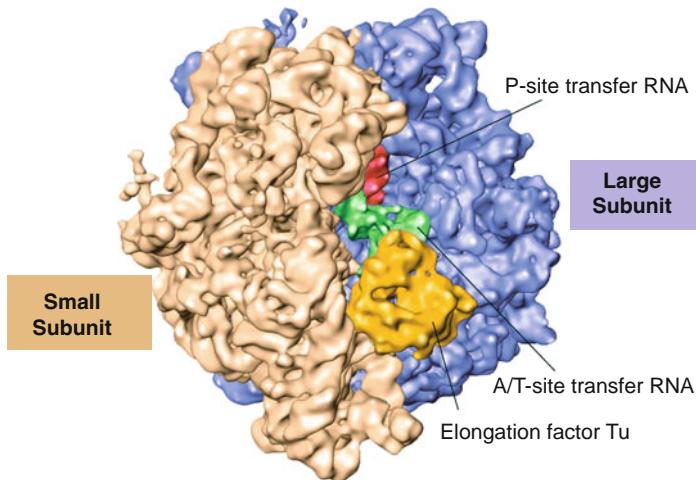


Figure 1.2. Cryo-EM view of the 70S ribosome particle solved by J. Frank's group at 6.7 Å resolution in a complex with the EF-Tu-aa-tRNA ternary complex, GDP, and the antibiotic kirromycin [710]. Images were kindly provided by Michael Watters and Joachim Frank.

Yet as we witness the tantalizing progress, a cautionary usage of molecular modeling tools as well as a critical perspective of the field's strengths and limitations are warranted. This is because the current generation of users and application scientists in the industrial and academic sectors may not be familiar with some of the caveats and inherent approximations in biomolecular modeling and simulation approaches that the field pioneers clearly recognized. Indeed, the tools and programs developed by a handful of researchers several decades ago have now resulted in extensive profit-making software for genomic information, drug design, and every aspect of modeling. More than ever, a comprehensive background in the methodology framework is necessary for sound studies in the exciting era of computational biophysics that lies on the horizon.

### 1.1.4 Text Overview

This text aims to provide this critical perspective for field assessment while introducing the relevant techniques. Specifically, following an overview of biomolecular structure and modeling with a historical perspective and a description of current applications in this chapter and the next chapter, the elementary background for biomolecular modeling will be introduced in the chapters to follow: protein and nucleic-acid structure tutorials (Chapters 3–7), overview of theoretical approaches (Chapter 8), details of force field construction and evaluation (Chapters 9 and 10), energy minimization techniques (Chapter 11), Monte Carlo simulations (Chapter 12), molecular dynamics and related methods (Chapters 13 and 14), and similarity/diversity problems in chemical design (Chapter 15).

As emphasized in this book’s Preface, given the enormously broad range of these topics, depth is often sacrificed at the expense of breadth. Thus, many specialized texts (e.g., in Monte Carlo, molecular dynamics, or statistical mechanics) are complementary, such as those listed in Appendix C; the representative articles used for the course (Appendix B) are important components. For introductory texts to biomolecular structure, biochemistry, and biophysical chemistry, see those listed in Appendix C, such as [163, 197, 275, 394, 1235]. For molecular simulations, a solid grounding in classical statistical mechanics, thermodynamic ensembles, time-correlation functions, and basic simulation protocols is important. Good introductory texts for these subjects, including biomolecular applications are [22, 165, 178, 428, 474, 494, 846, 853, 1038, 1067].

The remainder of this chapter and the next chapter provide a historical context for the field’s development. Overall, this chapter focuses on a historical account of the field and the experimental progress that made biomolecular modeling possible. Chapter 2 introduces some of the field’s challenges as well as practical applications of their solution.

Specifically, to appreciate the evolution of biomolecular modeling and simulation, we begin in the next section with an account of the milieu of growing experimental and technical developments. Following an introduction to the birth of molecular mechanics (Section 1.2), experimental progress in protein and nucleic-acid structure is described (Section 1.3). A selective reference chronology to structural biology is shown in Table 1.1.

The experimental section of this chapter discusses separately the early days of biomolecular instrumentation — as structures were emerging from X-ray crystallography — and the modern era of technological developments — stimulating the many sequencing projects and the rapid advances in biomolecular NMR and crystallography. Within this presentation, separate subsections are devoted to the techniques of X-ray crystallography and NMR and to the genome projects.

Chapter 2 continues this perspective by describing the computational challenges that naturally emerge from the overwhelming progress in genome projects and experimental techniques, namely deducing structure and function from sequence. Problems are exemplified by protein folding and misfolding. (Students unfamiliar with basic protein structure are urged to re-read Chapter 2 after the

protein minitutorial chapters). The sections that follow mention some of the exciting and important biomedical, industrial, and technological applications that lend enormous practical utility to the field. These applications represent a tangible outcome of the confluent experimental, theoretical, and technological advances.

Table 1.1. Structural Biology Chronology.

1865	Genes discovered by Mendel
1910	Genes in chromosomes shown by Morgan's fruitfly mutations
1920s	Quantum mechanics theory develops
1926	Early reports of crystallized proteins
1930s	Reports of crystallized proteins continue and stimulate Pauling & Corey to compile bond lengths and angles of amino acids
1944	Avery proves genetic transformation via DNA (not protein)
1946	Molecular mechanics calculations reported (Westheimer, others)
1949	Sickle cell anemia identified as 'molecular disease' (Pauling)
1950	Chargaff determines near-unity A:T and G:C ratios in many species
1951	Pauling & Corey predict protein $\alpha$ -helices and $\beta$ -sheets
1952	Hershey & Chase reinforce genetic role of DNA (phage experiments)
1952	Wilkins & Franklin deduce that DNA is a helix (X-ray fiber diffraction)
1953	<b>Watson &amp; Crick report the structure of the DNA double helix</b>
1959	Myoglobin & hemoglobin deciphered by X-ray (Kendrew & Perutz)
1960s	Systematic force fields developed (Allinger, Lifson, Scheraga, others)
1960s	Genetic code deduced (Crick, Brenner, Nirenberg, Khorana, Holley, coworkers)
1969	Levinthal paradox on protein folding posed
1970s	Biomolecular dynamics simulations develop (Stillinger, Karplus, others)
1970s	Site-directed mutagenesis techniques developed by M. Smith; restriction enzymes discovered by Arber, Nathans, and H. Smith
1971	Protein Data Bank established
1974	t-RNA structure reported
1975	First simulation of protein folding
1975	Fifty solved biomolecular structures available in the PDB
1977	DNA genome of the virus $\phi$ X174 (5.4 kb) sequenced; soon followed by human mitochondrial DNA (16.6 kb) and $\lambda$ phage (48.5 kb)
1980s	Dazzling progress realized in automated sequencing, protein X-ray crystallography, NMR, recombinant DNA, and macromolecular synthesis
1985	PCR devised by Mullis; numerous applications follow
1985	NSF establishes five national supercomputer centers
1990	International Human Genome Project starts; spurs others
1994	RNA hammerhead ribozyme structure reported; other RNAs follow
1995	First non-viral genome completed ( <i>bacterium H. influenzae</i> ), 1.8 Mb
1996	Yeast genome ( <i>Saccharomyces cerevisiae</i> ) completed, 13 Mb
1997	Chromatin core particle structure reported; confirms earlier structure
1998	Roundworm genome ( <i>C. elegans</i> ) completed, 100 Mb
1998	Crystal structure of ion channel protein reported
1998	Private Human Genome initiative competes with international effort
1999	Fruitfly genome ( <i>Drosophila melanogaster</i> ) completed (Celera), 137 Mb

Table 1.1 (continued)

1999	Human chromosome 22 sequenced (public consortium)
1999	IBM announces petaflop computer to fold proteins by 2005
2000	<b>First draft of human genome sequence announced</b> , 3300 Mb
2000	Moderate-resolution structures of ribosomes reported
2000	ENCODE project consortium established to characterize the human genome
2001	First annotation of the human genome (February)
2002	First draft of rice genome sequence, 430 Mb (April)
2003	Human genome sequence completed (April)
2006	All human chromosomes sequenced
2007	Craig Venter's DNA solved and analyzed
2008	James Watson's DNA solved and analyzed in a triumph of sequencing technology
Ongoing	Many projects continue to interpret findings of the HGP, including ENCODE, 1000 genomes, HapMap, Cancer Atlas, Personal Genome Project, and the sequencing of many organisms to allow comparative studies

## 1.2 The Roots of Molecular Modeling in Molecular Mechanics

The roots of molecular modeling began with the notion that molecular geometry, energy, and various molecular properties can be calculated from mechanical-like models subject to basic physical forces. A molecule is represented as a mechanical system in which the *particles* — atoms — are connected by *springs* — the bonds. The molecule then rotates, vibrates, and translates to assume favored conformations in space as a collective response to the inter- and intramolecular forces acting upon it.

The forces are expressed as a sum of harmonic-like (from Hooke's law) terms for **bond-length** and **bond-angle** deviations from reference equilibrium values; trigonometric **torsional terms** to account for *internal rotation* (rotation of molecular subgroups about the bond connecting them); and **nonbonded van der Waals and electrostatic potentials**. See Chapter 9 for a detailed discussion of these terms, as well as of more intricate cross terms.

### 1.2.1 The Theoretical Pioneers

Molecular mechanics arose naturally from the concepts of molecular bonding and van der Waals forces. The Born-Oppenheimer approximation assuming fixed nuclei (see Chapter 8) followed in the footsteps of quantum theory developed in the 1920s. While the basic idea can be traced to 1930, the first attempts of molecular

Table 1.2. The evolution of molecular mechanics and dynamics.

Period	System and Size <sup>a</sup>	Trajectory Length <sup>b</sup> [ns]	CPU Time/Computer <sup>c</sup>
1973	Dinucleoside (GpC) in vacuum (8 flexible dihedral angles)	—	—
1977	BPTI, vacuum (58 residues, 885 atoms)	0.01	
1983	DNA, vacuum, 12/24 bp (754/1530 atoms)	0.09	several weeks each, Vax 780
1984	GnRH, vacuum (decapeptide, 161 atoms)	0.15	
1985	Myoglobin, vacuum (1423 atoms)	0.30	50 days, VAX 11/780
1985	DNA, 5 bp (2800 atoms)	0.50	20 hrs, Cray X-MP
1989	Phospholipid Micelle ( $\approx$ 7,000 atoms)	0.10	
1992	HIV protease (25,000 atoms)	0.10	100 hrs., Cray Y-MP
1997	Estrogen/DNA (36,000 atoms, multipoles)	0.10	22 days, HP-735 (8)
1998	DNA, 24 bp (21,000 atoms, PME)	0.50	1 year, SGI Challenge
1998	$\beta$ -heptapeptide in methanol ( $\approx$ 5000/9000 atoms)	200	8 months, SGI Challenge (3)
1998	Villin headpiece (36 residues, 12,000 atoms, cutoffs)	1000	4 months, 256-proc. Cray T3D/E
1999	$bc_1$ complex in phospholipid bilayer (91,061 atoms, cutoffs)	1	75 days, 64 450-MHz-proc. Cray T3E
2001	C-terminal $\beta$ -hairpin of protein-G (177 atoms, implicit solvent)	38000 <sup>b</sup>	$\sim$ 8 days, 5000 proc. <b>Folding@home</b> megacluster
2002	Channel protein in lipid mem- brane (106,189 atoms, PME)	5	30 hrs, 500 proc. LeMieux terascale system; 50 days, 32 proc. Linux (Athlon)
2006	Complete satellite tobacco mosaic virus (1 million atoms)	50	55 days ( $\approx$ Ins/day), 256 Altix nodes, NCSA Athlon 2600+, NAMD program
2007	B-DNA dodecamer in solvent, PME, AMBER parm98 (15,774 atoms)	1200	130 days, 32 PowerPC BladeCenter proc., MareNostrum Supercomputer, Barcelona
2007	Villin headpiece (9,684 atoms) AMBER-2003	1000	6 months, <b>Folding@home</b> X86 megacluster, GROMACS/MPI
2008	Ubiquitin protein, explicit solvent OPLS-AA/SPC forcefield, (19,471 atoms)	1200	14 days (87ns/day), 32 processors Operon cluster, Desmond program
2008	Fip35 protein, explicit solvent NAMD/CHARMM	10000	14 weeks, NCSA Abe cluster, NAMD program
2009	$\beta_2$ AR protein mutants (50,000-99,000 atoms) CHARMM27 forcefield	2000	28 days, 32 (2.66 GHz) E5430 processors Desmond program

<sup>a</sup>The examples for each period are representative. The first five systems are modeled in vacuum and the others in solvent. Except for the dinucleoside, simulations refer to molecular dynamics (MD). The two system sizes for the  $\beta$ -heptapeptide [285] reflect two (temperature-dependent) simulations. See text for definitions of abbreviations and further entry information.

<sup>b</sup>The 38  $\mu$ s  $\beta$ -hairpin simulation in 2001 represents an ensemble (or aggregate) dynamics simulation, as accumulated over several short runs, rather than one long simulation [1428].

<sup>c</sup>The computational time is given where possible; estimates for the vacuum DNA, heptapeptide,  $\beta$ -hairpin, and channel protein simulations [285, 746, 1247, 1428] were kindly provided by M. Levitt, W. van Gunsteren, V. Pande, and K. Schulten, respectively.

mechanics calculations were recorded in 1946. Frank Westheimer's calculation of the relative racemization rates of biphenyl derivatives illustrated the success of such an approach. However, computers were not available at that time, so it took several more years for the field to gather momentum.

In the early 1960s, pioneering work on development of systematic force fields — based on spectroscopic information, heats of formation, structures of small compounds sharing the basic chemical groups, other experimental data, and quantum-mechanical information — began independently in the laboratories of the late Shneior Lifson at the Weizmann Institute of Science (Rehovot, Israel) [747], Harold Scheraga at Cornell University (Ithaca, New York), and Norman Allinger at Wayne State University (Detroit, Michigan) and then the University of Georgia (Athens). These researchers and their talented coworkers (notably Warshel and Levitt) began to develop force field parameters for families of chemical compounds by testing calculation results against experimental observations regarding structure and energetics. In 1969, following the pioneering Cartesian coordinate treatment described by Lifson and Warshel a year earlier [766], Levitt and Lifson reported the first energy calculation on entire protein molecules (myoglobin and lysozyme), in which molecular potentials and experimental constraints defined the target energy function minimized in Cartesian coordinates by the steepest descent method to refine low-resolution experimental coordinates [748]. Such formulations in Cartesian coordinates paved the way for all subsequent energy minimization and molecular dynamics calculations of biomolecules. In fact, Warshel's recognition in the mid 1960s that programming molecular force fields in Cartesian coordinates rather than internal coordinates [766] led to efficient evaluation of the functions along with analytic first and second derivatives and to program segments in many current macromolecular modeling programs [747].

In the early 1970s, Rahman and Stillinger reported the first molecular dynamics work of a polar molecule, liquid water [1034, 1035]; results offered insights into the structural and dynamic properties of this life sustaining molecule. Rahman and Stillinger built upon the simulation technique described much earlier (1959) by Alder and Wainwright but applied to hard spheres [19].

In the late 1970s, the idea of using molecular mechanics force fields with energy minimization as a tool for refinement of crystal structures was presented [598] and developed [670]. This led to the modern versions employing simulated annealing and related methods [248, 676].

It took a few more years, however, for the field to gain some 'legitimacy'.<sup>3</sup> In fact, these pioneers did not receive much general support at first, partly because their work could not easily be classified as a traditional discipline of chemistry (e.g., physical chemistry, organic chemistry). In particular, spectroscopists criticized the notion of transferability of the force constants, though at the same time

---

<sup>3</sup>Personal experiences shared by Norman L. Allinger on those early days of the field form the basis for the comments in this paragraph. I am grateful for him sharing these experiences with me.

they were quite curious about the predictions that molecular mechanics could make. In time, it indeed became evident that force constants are not generally transferable; still, the molecular mechanics approach was sound since nonbonded interactions are included, terms that spectroscopists omitted.

Ten to fifteen more years followed until the first generation of biomolecular force fields was established. The revitalized idea of molecular dynamics in the late 1970s propagated by Martin Karplus and colleagues at Harvard University sparked a flame of excitement that continues with full force today with the fuel of supercomputers. Most programs and force fields today, for both small and large molecules, are based on the works of the pioneers cited above (Allinger, Lifson, and Scheraga) and their coworkers. The water force fields developed in the late 1970s and early 1980s by Berendsen and coworkers (e.g., [1079]) and by Jorgensen and coworkers [617] (SPC and TIP3P/TIP4P, respectively) laid the groundwork for biomolecular simulations in solution. Important concepts in protein electrostatics and enzyme/substrate complexes in solution laid by Warshel and colleagues [1344, 1346] paved the way to quantitative modeling of enzymatic reactions and hybrid quantum/molecular mechanics methods [1342].

Peter Kollman's legacy is the development and application of force field methodology and computer simulation to important biomolecular, as well as medicinal, problems such as enzyme catalysis and protein/ligand design [1335]; his group's free energy methods and combined quantum/molecular mechanics approaches have opened many new doors of applications. With Kollman's untimely death in May 2001, the community mourned the loss of a great leader and innovator.

Modern versions of these and other molecular simulation packages have led to competition for “better, bigger, and faster” program design. For example, free software at the University of Illinois at Urbana-Champaign by Klaus Schulten and coworkers called NAMD for nanoscale MD (see the NAMD homepage) can be run on hundreds of parallel microprocessors with various force fields [996]. David Shaw group's program Desmond is fast even on a small number of processors [156]. GROMACS [124, 770] is also widely used for long MD simulations (e.g., [364]). And Anton, a specialized computer hard-wired for long MD simulations, has been launched [1169]. With these excellent teams of software and hardware engineers and biomolecular scientists, the average MD user can look forward to improved and faster applications.

### 1.2.2 Biomolecular Simulation Perspective

Table 1.2 and Figures 1.3 and 1.4 provide a perspective of biomolecular simulations. Specifically, the selected examples illustrate the growth in time of system complexity (size and model resolution) and simulation length. The three-dimensional (3D) rendering in Figure 1.3 shows ‘blocks’ with heights

proportional to system size. Figure 1.4 offers molecular views of the simulation subjects and extrapolations for long-time simulations of proteins and cells based on [339].

### Representative Progress

Starting from the first entry in the table, **dinucleoside GpC** (guanosine-3', 5'-cytidine monophosphate) posed a challenge in the early 1970s for finding all minima by potential energy calculations and model building [1222]. Still, clever search strategies and constraints found a correct conformation (dihedral angles in the range of helical RNA and sugar in C3'-endo form) as the lowest energy minimum. *Global optimization remains a difficult problem!* (See Chapter 11).

Following the first MD simulation of a biological process of duration 100 fs [1344], the small protein **BPTI** (Bovine Pancreatic Trypsin Inhibitor) was simulated **1977** [845], showing substantial atomic fluctuations on the picosecond timescale.

The 12 and 24-base-pair (bp) **DNA** simulations in **1983** [746] were performed in vacuum without electrostatics, and that of the DNA pentamer system in 1985, with 830 water molecules and 8 sodium ions and full electrostatics [1158]. Stability problems for nucleic acids emerged in the early days — unfortunately, in some cases the strands untwisted and separated [746]. Stability became possible with the introduction of scaled phosphate charges in other pioneering nucleic-acid simulations [523, 1013, 1260] and the introduction a decade later of more advanced treatments for solvation and electrostatics; see, for example, [220], for a discussion.

The linear **decapeptide GnRH** (gonadotropin-releasing hormone) was studied in **1984** for its pharmaceutical potential, as it triggers LH and FSH hormones [1234].

The 300 ps dynamics simulation of the protein **myoglobin** in **1985** [752] was considered three times longer than the longest previous MD simulation of a protein. The results indicated a slow convergence of many thermodynamic properties.

The large-scale **phospholipid** aggregate simulations in **1989** [1361] was an ambitious undertaking: it incorporated a hydrated micelle (i.e., a spherical aggregate of phospholipid molecules) containing 85 LPE molecules (lysophosphatiadyl-ethanolamine) and 1591 water molecules.

The **HIV protease** system simulated in solution in **1992** [518] captured an interesting flap motion at the active site. See also Figure 2.5 and a discussion of this motion in the context of protease inhibitor design.

The **1997 estrogen/DNA simulation** [679] sought to understand the mechanism underlying DNA sequence recognition by the protein. It used the multipole electrostatic treatment, crucial for simulation stability, and also parallel processing for speedup [1133].

The **1998 DNA** simulation [1417] used the alternative, Particle Mesh Ewald (PME) treatment for consideration of long-range electrostatics (see Chapter 10) and uncovered interesting properties of A-tract sequences.

The **1998 peptide** simulation in methanol used periodic boundary conditions (defined in Chapter 10) and captured reversible, temperature-dependent folding [285]; the 200 ns time reflects four 50 ns simulations at various temperatures.

The **1998 1  $\mu$ s villin-headpiece** simulation (using periodic boundary conditions) [338] was considered longer by three orders of magnitude than prior simulations. A folded structure close to the native state was approached; see also [340].

The solvated protein ***bc*<sub>1</sub> embedded in a phospholipid bilayer** [597] was simulated in **1999** for over 1 ns by a ‘steered molecular dynamics’ algorithm (45,131 flexible atoms) to suggest a pathway for proton conduction through a water channel. As in villin, the Coulomb forces were truncated.

In **2002**, an aquaporin membrane channel protein in the glycerol conducting subclass (*E. coli* **glycerol channel, GlpF**) in a lipid membrane (106,189 total atoms) was simulated for 5 ns (as well as a mutant) with all nonbonded interactions considered, using the PME approach [1247]. The simulations suggested details of a selective mechanism by which water transport is controlled; see also [606] for simulations examining the glycerol transport mechanism.

By early 2002, the longest simulation published of 38 $\mu$ s reflected aggregate (or ensemble) dynamics — usage of many short trajectories to simulate the microsecond timescale — for the C-terminal  **$\beta$ -hairpin from protein G** (16 residues) in **2001** [1428]. Whereas the continuous 1  $\mu$ s villin simulation required months of dedicated supercomputing, the  $\beta$ -hairpin simulation (177 atoms, using implicit solvation and Langevin dynamics) was performed to analyze folding kinetics on a new distributed computing paradigm which employs personal computers from around the world (see **Folding@home** at [folding.stanford.edu](http://folding.stanford.edu) and [1177]). About 5000 processors were employed and, with the effective production rate of 1 day per nanosecond per processor, about 8 days were required to simulate the 38  $\mu$ s aggregate time. See also [1205] for a later set of simulations (reviewed in [177]) and a large-scale molecular dynamics study of a variant of the villin headpiece that consisted of hundreds of 1  $\mu$ s simulations [364].

Several years later, longer and larger simulations, though not yet routine, are clearly possible using specialized programs that exploit high-speed multiprocessor systems, like NAMD, GROMACS, and/or specialized computing resources like Anton [1169].

While trends continue to simulate larger biomolecular systems (e.g., entire **satellite mosaic virus** in 2006 with one million atoms) [425] and longer time frames (e.g., **B-DNA dodecamer** [987], **ubiquitin** protein [827], **Fip35 protein** [424], and  **$\beta_2$ AR protein receptor** [337] for over one microsecond, and small proteins for milliseconds [1462]) with specialized MD programs and dedicated supercomputers, coarse-grained models and combinations of enhanced sampling methods are emerging as the way to go for simulating complex biomolecular systems (recently reviewed in [655, 729, 1116, 1117]). This is because computer power alone is not likely to solve the folding problem in general. For example, the 10  $\mu$ s simulation of Fip35 [424] did not provide the anticipated folded conformation nor the folding trajectory from the extended state, as expected from

experimental measurements; this long simulation also raised force field and algorithmic stability questions, which were explored later [426]. Still, for other proteins, folding simulations can be very successful (e.g., [337, 427, 637]).

### Trends

Note from the table and figure the transition from simulations in vacuum (first five entries) to simulations in solvent (remaining items). Observe also the steady increase in simulated system size, with a leap increase in simulation lengths made more recently.

Large system sizes or long simulation times can generally be achieved by sacrificing other simulation aspects. For example, truncating long-range electrostatic interactions makes possible the study of large systems over short times [597], or small systems over long times [338]. Using implicit solvent and cutoffs for electrostatic interactions also allows the simulation of relatively small systems over long times [1428]. And simplified, coarse-grained models with effective potentials also allow simulations over longer time frames, with the correct physical behavior. (These topics are discussed in later chapters). In fact, with the increased awareness of the sampling problem in dynamic simulation (see Chapter 13), long single simulations are often replaced by several trajectories, leading to overall better sampling statistics, and coarse graining is being applied to biological systems and problems of greater complexity [655].

Duan *et al.* make an interesting ‘fanciful’ projection on the computational capabilities of modeling in the coming decades [339]: they suggest the feasibility, in 20 years, of simulating a second in the life-time of medium-sized proteins and, in 50–60 years, of following the entire life cycle of an *E. Coli* cell (1000 seconds or 20 minutes, for 30 billion atoms). This estimate was extrapolated on the basis of two data points — the 1977 BPTI simulation [845] and the 1998 villin simulation [338, 340] discussed above — and relied on the assumption that computational power increases by a factor of 10 every 3–4 years (Even better progress was actually realized). These projections are displayed by entries for the years 2020 and 2055 in Figure 1.4.

## 1.3 Emergence of Biomodeling from Experimental Progress in Proteins and Nucleic Acids

At the same time that molecular mechanics developed, tremendous progress on the experimental front also began to trigger further interest in the theoretical approach to structure determination.

### 1.3.1 Protein Crystallography

The first records of crystallized polypeptides or proteins date back to the late 1920s / early 1930s (1926: urease, James Sumner; 1934: pepsin, J. D. Bernal and

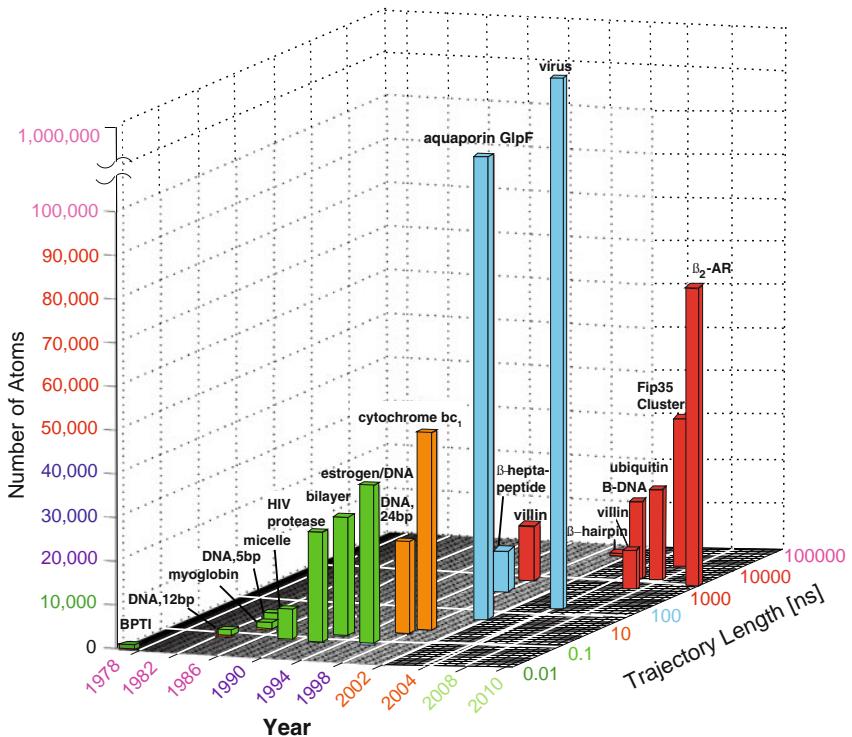


Figure 1.3. The evolution of molecular dynamics simulations with respect to system sizes and simulation lengths (see also Table 1.2).

Dorothy Crowfoot-Hodgkin; 1935: insulin, Crowfoot-Hodgkin). However, only in the late 1950s did John Kendrew (Perutz' first doctoral student) and Max Perutz succeed in deciphering the X-ray diffraction pattern from the crystal structure of the protein (1958: myoglobin, Kendrew; 1959: hemoglobin, Perutz). This was possible by Perutz' crucial demonstration (around 1954) that structures of proteins can be solved by comparing the X-ray diffraction patterns of a crystal of a native protein to those associated with the protein bound to heavy atoms like mercury (i.e., by ‘isomorphous replacement’). The era of modern structural biology began with this landmark development.

As glimpses of the first X-ray crystal structures of proteins came into view, Linus Pauling and Robert Corey began in the mid 1930s to catalogue bond lengths and angles in amino acids. By the early 1950s, they had predicted the two basic structures of amino acid polymers on the basis of hydrogen bonding patterns:  $\alpha$  helices and  $\beta$  sheets [974, 976]. As of 1960, about 75 proteins had been crystallized, and immense interest began on relating the sequence content to catalytic activity of these enzymes.

By then, the exciting new field of molecular biology was well underway. Perutz, who founded the Medical Research Council Unit for Molecular Biology at the

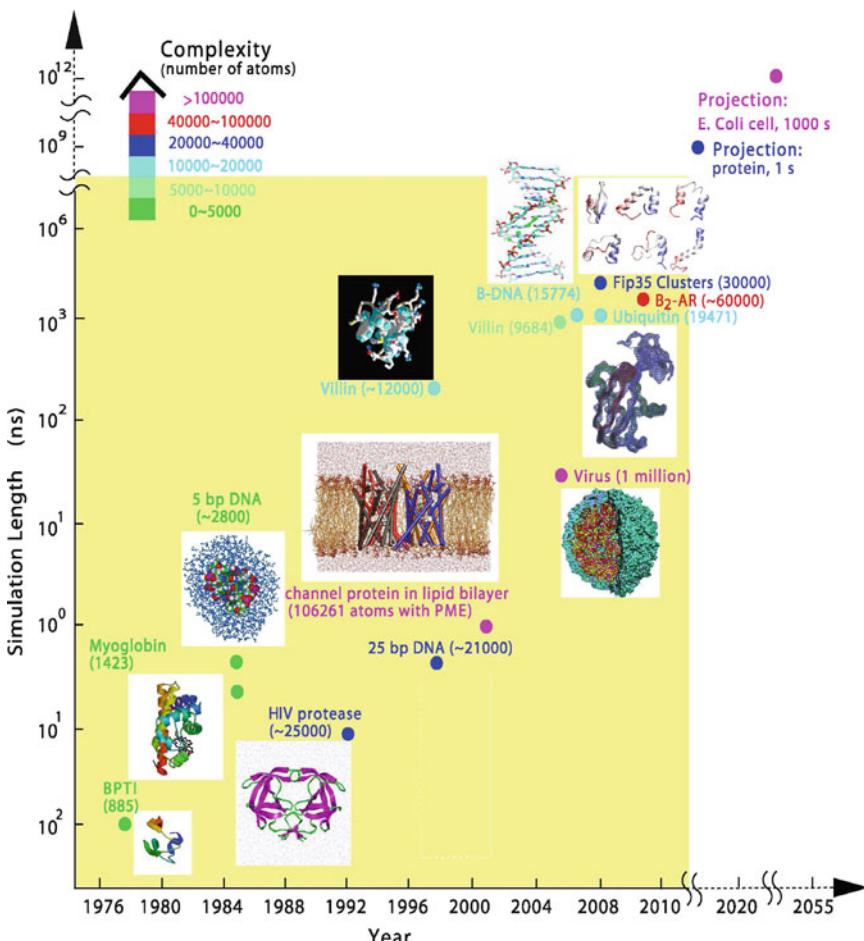


Figure 1.4. The evolution of molecular dynamics simulations with respect to simulation lengths (see also Table 1.2 and Figure 1.3). The data points for 2020 and 2055 represent extrapolations from the 1977 BPTI [845] and 1998 villin [338,340] simulations, assuming a computational power increase by a factor of 10 every 3–4 years, as reported in [339].

Cavendish Laboratory in Cambridge in 1947, also created the Laboratory of Molecular Biology there in 1962. Perutz and Kendrew received the Nobel Prize in Chemistry for their accomplishments in 1962.<sup>4</sup>

<sup>4</sup>See the formidable electronic museum of science and technology, with related lectures and books that emerged from Nobel-awarded research, on the website of the Nobel Foundation ([nobelprize.org](http://nobelprize.org)). This virtual museum was recently constructed to mark the 100th anniversary in 2001 of Alfred B. Nobel's legacy. See also a marvelous account in [1258] of Perutz (who died at the age of 87 in 2002) as scientist and person, including his relationship with Kendrew.

### 1.3.2 DNA Structure

Momentum at that time came in large part from parallel experimental work that began in 1944 in the nucleic acid world and presaged the discovery of the DNA double helix.

Inspired by the 1928 work of the British medical officer Fred Griffith, Oswald Avery and coworkers Colin MacLeod and Maclyn McCarty studied pneumonia infections. Griffith's intriguing experiments showed that mice became fatally ill upon infection from a live but harmless (coatless) strain of pneumonia-causing bacteria mixed with the DNA from heat-killed pathogenic bacteria; thus, the DNA from heat-killed pathogenic bacteria transformed live harmless into live pathogenic bacteria. Avery and coworkers mixed DNA from virulent strains of pneumococci with harmless strains and used enzymes that digest DNA but not proteins. Their results led to the cautious announcement that the ‘transforming agent’ of traits is made exclusively of DNA.<sup>5</sup>

Their finding was held with skepticism until the breakthrough, Nobel prize-winning phage experiments of Alfred Hershey and Martha Chase eight years later, which demonstrated that only the nucleic acid of the phage entered the bacterium upon infection, whereas the phage protein remained outside.<sup>6</sup>

Much credit for the transforming agent evidence is due to the German theoretical physicist and Nobel laureate Max Delbrück, who brilliantly suggested to use bacterial viruses as the model system for the genome demonstration principle. Delbrück shared the Nobel Prize in Physiology or Medicine in 1969 with Hershey and Salvador Luria for their pioneering work that established bacteriophage as the premier model system for molecular genetics.

In 1950, Erwin Chargaff demonstrated that the ratios of adenine-to-thymine and guanine-to-cytosine bases are close to unity, with the relative amount of each kind of pair depending on the DNA source.<sup>7</sup> These crucial data, together with the X-ray fiber diffraction photographs of hydrated DNA taken by Rosalind Franklin<sup>8</sup> and Raymond Gosling (both affiliated with Maurice Wilkins who was engaged in related research [622]), led directly to Watson and Crick’s ingenious proposal of the structure of DNA in 1953. The photographs were crucial as they suggested a helical arrangement.

---

<sup>5</sup> Interested readers can visit the virtual gallery of Profiles in Science at [wwwprofiles.nlm.nih.gov/](http://wwwprofiles.nlm.nih.gov/) for a profile on Avery.

<sup>6</sup> A wonderful introduction to the rather recluse Hershey, who died at the age of 88 in 1997, can be enjoyed in a volume edited by Franklin W. Stahl titled *We can sleep later: Alfred D. Hershey and the origins of molecular biology* (Cold Spring Harbor Press, New York, 2000). The title quotes Hershey from his letter to contributors of a volume on *bacteriophage λ* which he edited in 1971, urging them to complete and submit their manuscripts!

<sup>7</sup> Chargaff died in June 2002 at the age of 96. Sadly, he was a sardonic man who did not easily fit into the sharply focused world of most scientists; he further isolated himself when he denounced the molecular biology community in the late 1950s.

<sup>8</sup> See an outstanding study on “the dark lady of DNA” in a recent biography [810] and [358].

Although connecting these puzzle pieces may seem straightforward to us now that the DNA double helix is a household word, these two ambitious young Cambridge scientists deduced from the fiber diffraction data and other evidence that the observed base-pairing specificity, together with steric restrictions, can be reconciled in an *anti-parallel* double-helical form with a sugar-phosphate backbone and nitrogenous-bases interior. Their model also required a key piece of information from the organic chemist Jerry Donahue regarding the *tautomeric* states of the bases.<sup>9</sup> Though many other DNA forms besides the classic Crick and Watson (B-DNA) form are now recognized, including triplexes and quadruplexes, the B-form is still the most prevalent under physiological conditions. Indeed, the 50th anniversary in April 2003 of Watson and Crick's seminal paper was celebrated with much fanfare throughout the world.

RNA crystallography is at an earlier stage, but has recently made quantum leaps with the solution of several ribosomes (see Fig. 1.1), other significant RNA molecules, and newly-identified novel roles for RNA, including, most prominently, non-coding RNAs, aptamers and riboswitches, with many potential benefits to biomedicine and nanotechnology (see Chapter 7). These developments followed the exciting discoveries in the 1980s that established that RNA, like protein, can act as an enzyme in living cells, and discoveries in recent years — that RNA has numerous regulatory roles in biological processes — which have transformed our understanding of RNA's functional repertoire. Sidney Altman and Thomas Cech received the 1989 Nobel Prize in Chemistry for their discovery of RNA biocatalysts, *ribozymes*, and twice in 2006 were RNA discoveries recognized by Nobel Prizes, including for uncovering gene silencing, which can be exploited to selectively knock out protein functions for disease analysis. RNA made headlines again in 2009 when the Nobel Prize in Chemistry was aptly awarded to three scientists who independently uncovered the atomic-level detail of the magnificent RNA/protein machine that makes up the ribosome: Ada Yonath, Venkatraman Ramakrishnan, and Thomas Steitz. Their X-ray crystallographic views and functional analyses have also led to antibiotic design.

The next two subsections elaborate upon the key techniques for solving biomolecular structures: X-ray crystallography and NMR. We end this section on experimental progress with a description of modern technological advances and the genome sequencing projects they inspired.

### 1.3.3 The Technique of X-ray Crystallography

Much of the early crystallographic work was accomplished without computers and was inherently very slow. *Imagine calculating the Fourier series by hand!*

---

<sup>9</sup>Proton migrations within the bases can produce a *tautomer*. These alternative forms depend on the dielectric constant of the solvent and the pH of the environment. In the bases, the common *amino* group ( $-N-H_2$ ) can tautomerize to an *imino* form ( $=N-H$ ), and the common *keto* group ( $-C=O$ ) can adopt the *enol* state ( $=C-O-H$ ); the fraction of bases in the rare imino and enol tautomers is only about 0.01% under regular conditions.

Only in the 1950s were direct methods for the phase problem developed, with a dramatic increase in the speed of structure determination occurring about a decade later.

Structure determination by X-ray crystallography involves analysis of the X-ray diffraction pattern produced when a beam of X-rays is directed onto a well-ordered crystal. Crystals form by vapor diffusion from purified protein solutions under optimal conditions. See [163, 930] for overviews.

The diffraction pattern can be interpreted as a reflection of the primary beam source from sets of parallel planes in the crystal. The diffracted spots are recorded on a detector (electronic device or X-ray film), scanned by a computer, and analyzed on the basis of Bragg's law<sup>10</sup> to determine the unit cell parameters.

Each such recorded diffraction spot has an associated amplitude, wavelength, and phase; all three properties must be known to deduce atomic positions. Since the phase is lost in the X-ray experiments, it must be computed from the other data. This central obstacle in crystal structure analysis is called the *phase problem* (see Box 1.1). Together, the amplitudes and phases of the diffraction data are used to calculate the electron density map; the greater the resolution of the diffraction data, the higher the resolution of this map and hence the atomic detail derived from it.

Both the laborious crystallization process [852] and the necessary mathematical analysis of the diffraction data limit the amount of accurate biomolecular data available. Well-ordered crystals of biological macromolecules are difficult to grow, in part because of the disorder and mobility in certain regions. Crystallization experiments must therefore screen and optimize various parameters that influence crystal formation, such as temperature, pH, solvent type, and added ions or ligands.

The phase problem was solved by direct methods for small molecules (roughly  $\leq 100$  atoms) by Jerome Karle and Herbert Hauptman in the late 1940s and early 1950s; they were recognized for this feat with the 1985 Nobel Prize in Chemistry. For larger molecules, biomolecular crystallographers have relied on the method pioneered by Perutz, Kendrew and their coworkers termed *multiple isomorphous replacement* (MIR).

MIR introduces new X-ray scatters from complexes of the biomolecule with heavy elements such as selenium or heavy metals like osmium, mercury, or uranium. The combination of diffraction patterns for the biomolecule, heavy elements or elements or metals, and biomolecule/heavy-metal complex offers more information for estimating the desired phases. The differences in diffracted

---

<sup>10</sup>The Braggs (father William-Henry and son Sir William-Lawrence) observed that if two waves of electromagnetic radiation arrive at the same point in phase and produce a maximal intensity, the difference between the distances they traveled is an integral multiple of their wavelengths. From this they derived what is now known as *Bragg's law*, specifying the conditions for diffraction and the relation among three key quantities:  $d$  (distance between parallel planes in the crystal),  $\lambda$  (the wavelength of the X-ray beam), and  $\theta$  (the reflection angle). Bragg's condition requires that the difference in distance traveled by the X-rays reflected from adjacent planes is equal to the wavelength  $\lambda$ . The associated relationship is  $\lambda = 2d \sin \theta$ .

intensities between the native and derivative crystals are used to pinpoint the heavy atoms, whose waves serve as references in the phase determination for the native system.

To date, advances in the experimental, technological, and theoretical fronts have dramatically improved the ease of crystal preparation and the quality of the obtained three-dimensional (3D) biomolecular models [163, last chapter]. Techniques besides MIR to facilitate the phase determination process — by analyzing patterns of heavy-metal derivatives using *multi-wavelength anomalous diffraction* (MAD) or by molecular replacement (deriving the phase of the target crystal on the basis of a solved related molecular system) [540, 541] have been developed. Very strong X-ray sources from synchrotron radiation (e.g., with light intensity that can be 10,000 times greater than conventional beams generated in a laboratory) have become available. New techniques have made it possible to visualize short-lived intermediates in enzyme-catalyzed reactions at atomic resolution by time-resolved crystallography [433, 870, 994]. And improved methods for model refinement and phase determination are continuously being reported [1287]. Such advances are leading to highly refined biomolecular structures<sup>11</sup> (resolution  $\leq 2 \text{ \AA}$ ) at much greater numbers [110], even for nucleic acids [894].

### Box 1.1: The Phase Problem

The mathematical phase problem in crystallography [531, 634] involves resolving the phase angles  $\phi_{\mathbf{h}}$  associated with the structure factors  $F_{\mathbf{h}}$  when only the intensities (squares of the amplitudes) of the scattered X-ray pattern,  $I_{\mathbf{h}} = |F_{\mathbf{h}}|$ , are known. The structure factors  $F_{\mathbf{h}}$ , defined as

$$F_{\mathbf{h}} = |F_{\mathbf{h}}| \exp(i\phi_{\mathbf{h}}), \quad (1.1)$$

describe the scattering pattern of the crystal in the Fourier series of the electron density distribution:

$$\rho(\mathbf{r}) = \frac{1}{V} \sum_{\mathbf{h}} F_{\mathbf{h}} \exp(-2\pi i \mathbf{h} \cdot \mathbf{r}). \quad (1.2)$$

Here  $\mathbf{r}$  denotes position,  $\mathbf{h}$  identifies the three defining planes of the unit cell (e.g.,  $h, k, l$ ),  $V$  is the cell volume, and  $\cdot$  denotes a vector product. See [1047], for example, for details.

#### 1.3.4 The Technique of NMR Spectroscopy

The introduction of NMR as a technique for protein structure determination came much later (early 1960s), but since 1984 both X-ray diffraction and NMR have been valuable tools for determining protein structure at atomic resolution. Kurt

<sup>11</sup>The resolution value is similar to the quantity associated with a microscope: objects (atoms) can be distinguished if they are separated by more than the resolution value. Hence, the lower the resolution value the more molecular architectural detail that can be discerned.

Wütrich was awarded the 2002 Nobel Prize in Chemistry<sup>12</sup> for his pioneering efforts in developing and applying NMR to biological macromolecules.

Nuclear magnetic resonance is a versatile technique for obtaining structural and dynamic information on molecules in solution. The resulting 3D views from NMR are not as detailed as those that can result from X-ray crystallography, but the NMR information is not static and incorporates effects due to thermal motions in solution.

In NMR, powerful magnetic fields and high-frequency radiation waves are applied to probe the magnetic environment of the nuclei. The local environment of the nucleus determines the frequency of the resonance absorption. The resulting NMR spectrum contains information on the interactions and localized motion of the molecules containing those resonant nuclei.

The absorption frequency of particular groups can be distinguished from one another when high-frequency NMR devices are used (“*high resolution NMR*”). Until recently, this requirement for nonoverlapping signals to produce a clear picture has limited the protein sizes that can be studied by NMR to systems with masses in the range of 50 to 100 kDa. However, dramatic increases (such as a tenfold increase) have been possible with novel strategies for isotopic labeling of proteins [1423] and detection of signals from disordered residues with fast internal motions by cross correlated relaxation-enhanced polarization transfer [397]. For example, the Horwich and Wütrich labs collaborated in 2002 to produce a high resolution solution NMR structure of the chaperonin/co-chaperonin GroEL/GroES complex (~900 kDa) [397]. Advances in solid-state NMR techniques may be particularly valuable for structure analysis of membrane proteins.

As in X-ray crystallography, advanced computers are required to interpret the data systematically. NMR spectroscopy yields a wealth of information: a network of distances involving pairs of spatially-proximate hydrogen atoms. The distances are derived from Nuclear Overhauser Effects (NOEs) between neighboring hydrogen atoms in the biomolecule, that is, for atom pairs separated by less than 5–6 Å.

To calculate the 3D structure of the macromolecule, these NMR distances are used as conformational restraints in combination with various supplementary information: primary sequence, reference geometries for bond lengths and bond angles, chirality, steric constraints, spectra, and so on. A suitable energy function must be formulated and then minimized, or surveyed by various techniques, to find the coordinates that are most compatible with the experimental data. See [248] for an overview. Such deduced models are used to back calculate the spectra inferred from the distances, from which iterative improvements of the model are pursued to improve the matching of the spectra. Indeed, the difficulty of using

---

<sup>12</sup>The other half of the 2002 Chemistry prize was split between John B. Fenn and Koichi Tanaka who were recognized for their development of ionization methods for analysis of proteins using mass spectrometry.

NMR data for structure refinement in the early days can be attributed to this formidable refinement task, formally, an over-determined or under-determined global optimization problem.<sup>13</sup>

The pioneering efforts of deducing peptide and protein structures in solution by NMR techniques were reported between 1981 and 1986; they reflected year-long struggles in the laboratory. Only a decade later, with advances on the experimental, theoretical, and technological fronts, 3D structures of proteins in solution could be determined routinely for monomeric proteins with less than 200 amino acid residues. See [368, 1396] for texts by modern NMR pioneers, [487] for a historical perspective of biomolecular NMR, and [248, 249, 1184] for recent advances.

Today's clever methods have been designed to facilitate such refinements, from formulation of the target energy to conformational searching, the latter using tools from distance geometry, molecular dynamics, simulated annealing, and many hybrid search techniques [181, 203, 248, 487]. The ensemble of structures obtained is not guaranteed to contain the "best" (global) one, but the solutions are generally satisfactory in terms of consistency with the data. The recent technique of residual dipolar coupling also has great potential for structure determination by NMR spectroscopy without the use of NOE data [1188, 1265].

## 1.4 Modern Era of Technological Advances

### 1.4.1 *From Biochemistry to Biotechnology*

The discovery of the elegant yet simple DNA double helix not only led to the birth of molecular biology; it led to the crucial link between biology and chemistry. Namely, the genetic code relating triplets of RNA (the template for protein synthesis) to the amino acid sequence was decoded ten years later, and biochemists began to isolate enzymes that control DNA metabolism.

One class of those enzymes, restriction endonucleases, became especially important for recombinant DNA technology. These molecules can be used to break huge DNA into small fragments for sequence analysis. Restriction enzymes can also cut and paste DNA (the latter with the aid of an enzyme, ligase) and thereby create spliced DNA of desired transferred properties, such as antibiotic-resistant bacteria that serve as informants for human insulin makers. The discovery of these enzymes was recognized by the 1978 Nobel Prize in Physiology or Medicine to Werner Arber, Daniel Nathans, and Hamilton O. Smith.

Very quickly, X-ray, NMR, recombinant DNA technology, and the synthesis of biological macromolecules improved. The 1970s and 1980s saw steady advances

---

<sup>13</sup>Solved NMR structures are usually presented as sets of structures since certain molecular segments can be over-determined while others under-determined. The better the agreement for particular atomic positions among the structures in the set, the more likely it is that a particular atom or component is well determined.

in our ability to produce, crystallize, image, and manipulate macromolecules. Site-directed mutagenesis developed in 1970s by Canadian biochemist Michael Smith (1993 Nobel laureate in Chemistry) has become a fundamental tool for protein synthesis and protein function analysis.

### 1.4.2 PCR and Beyond

The polymerase chain reaction (PCR) devised in 1985 by Kary Mullis (winner of the 1993 Nobel Prize in Chemistry, with Michael Smith) and coworkers [884] revolutionized biochemistry: small parent DNA sequences could be amplified exponentially in a very short time and used for many important investigations. DNA analysis has become a standard tool for a variety of practical applications. Noteworthy classic and current examples of PCR applications are collected in Box 1.2. See also [1044] for stories on how genetics teaches us about history, justice, diseases, and more.

Beyond amplification, PCR technology made possible isolation of gene fragments and their usage to clone whole genes; these genes could then be inserted into viruses or bacterial cells to direct the synthesis of biologically active products. With dazzling speed, the field of bioengineering was born. Automated sequencing efforts continued during the 1980s, leading to the start of the International Human Genome Project in 1990, which spearheaded many other sequencing projects (see next section).

Macromolecular X-ray crystallography and NMR techniques are also improving rapidly in this modern era of instrumentation, both in terms of obtained structure resolution and system sizes [874]. Stronger X-ray sources, higher-frequency NMR spectrometers, and refinement tools for both data models are leading to these steady advances. The combination of instrumental advances in NMR spectroscopy and protein labeling schemes suggests that the size limit of protein NMR may soon reach 100 kDa [1404, 1423].

In addition to crystallography (see Fig. 1.1) and NMR, cryogenic electron microscopy (cryo-EM) contributes important macroscopic views at lower resolution for proteins that are not amenable to NMR or crystallography [418, 1252, 1286]. This technique involves imaging rapidly-frozen samples of randomly-oriented molecular complexes at low temperatures and reconstructing 3D views from the numerous planar EM projections of the structures. Adequate particle detection imposes a lower limit on the subject of several hundred kDa, but cryo-EM is especially good for large molecules with symmetry, as size and symmetry facilitate the puzzle gathering (3D image reconstruction) process. Though the resolution is low compared to crystallography and NMR, the resolution is becoming better and better with optimization of parameters and protocols used in the reconstruction process, as demonstrated for the 70S ribosome [710] shown in Figure 1.2.

Together with recombinant DNA technology, automated software for structure determination, and supercomputer and graphics resources, structure determination at a rate of one biomolecule per day (or more) is on the horizon.

**Box 1.2: PCR Application Examples**

- **Medical diagnoses of diseases and traits.** DNA analysis can be used to identify gene markers for many maladies, like cancer (e.g., BRCA1/2, *p53* mutations), schizophrenia, late Alzheimer's or Parkinson's disease. A classic story of cancer markers involves Vice President Hubert Humphrey, who was tested for bladder cancer in 1967 but died of the disease in 1978. In 1994, after the invention of PCR, his cancerous tissue from 1976 was compared to a urine sample from 1967, only to reveal the same mutations in the *p53* gene, a cancer suppressing gene, that escaped the earlier recognition. Sadly, if PCR technology had been available in 1967, Humphrey may have been saved.
- **Historical analysis.** DNA is now being used for genetic surveys in combination with archaeological data to identify markers in human populations.<sup>14</sup> Such analyses can discern ancestors of human origins, migration patterns, and other historical events [1070]. These analyses are not limited to humans; the evolutionary metamorphosis of whales has been unraveled by the study of fossil material combined with DNA analysis from living whales [1389].

Historical analysis by French and American viticulturists also showed that the entire gene pool of 16 classic wines can be conserved by growing only two grape varieties: *Pinot noir* and *Gouais blanc*. Depending on your occupation, you may either be comforted or disturbed by this news . . . .

PCR was also used to confirm that the fungus that caused the Irish famine (since potato crops were devastated) in 1845–1846 was caused by the fungus *P. infestans*, a water mold (infected leaves were collected during the famine) [1053]. Studies showed that the Irish famine was not caused by a single strain called US-1 which causes modern plant infections, as had been thought. Significantly, the studies taught researchers that further genetic analysis is needed to trace recent evolutionary history of the fungus spread.

- **Forensics and crime conviction.** DNA profiling — comparing distinctive DNA sequences, aberrations, or numbers of sequence repeats among individuals — is a powerful tool for proving with extremely high probability the presence of a person (or related object) at a crime, accident, or another type of scene. In fact, in the two decades since DNA evidence began to be used in court (1988), about 250 prisoners have been exonerated in the U.S., including from death row and one after 35 years behind bars, and many casualties from disasters (like airplane crashes and the 11 September 2001 New York World Trade Center terrorist attacks) were identified from DNA analysis of assembled body parts. In this connection, personal objects analyzed for DNA — like a black glove or blue dress — made headlines as crucial ‘imaginary witnesses’<sup>15</sup> in the O.J. Simpson and Lewinsky/Clinton affairs. In fact,

---

<sup>14</sup>Time can be correlated with genetic markers through analysis of mitochondrial DNA or segments of the Y-chromosome. Both are genetic elements that escape the usual reshuffling of sexual reproduction; their changes thus reflect random mutations that can be correlated with time.

<sup>15</sup>George Johnson, “OJ’s Blood and The Big Bang, Together at Last”, *The New York Times*, Sunday, May 21, 1995.

a new breed of high-tech detectives is emerging with modern scientific tools, for example, by using bugs in crime research. See also the Anthrax attack case solved in 2008 described at the end of this chapter.

- **Family lineage / paternity identification.** DNA fingerprinting can also be used to match parents to offspring. In 1998, DNA from the grave confirmed that President Thomas Jefferson fathered at least one child by his slave mistress, Sally Hemmings, 200 years ago. The mystery concerning the remains of Tsar Nicholas II's family, executed in 1918, was solved after nine decades, by DNA analysis of bone shards found in a forest; the latest remnants analyzed in 2008 were determined to be those of his two children Alexksei and Maria, therefore completing the findings of the entire family. In April 2000, French historians with European scientists solved a 205-year-old mystery by analyzing the heart of Louis XVII, preserved in a crystal urn, confirming that the 10-year old boy died in prison after his parents Marie Antoinette and Louis XVI were executed, rather than spirited out of prison by supporters (Antoinette's hair sample is available). Similar post-mortem DNA analysis proved false a paternity claim against Yves Montand. In 2008, Egypt announced DNA analysis projects of 3500-year-old mummies, including fetuses of the mummies, to determine lineage connections to King Tutankhamun.

See also the book by Reilly [1044] for many other examples.

---

## 1.5 Genome Sequencing

### 1.5.1 Projects Overview: From Bugs to Baboons

Spurred by this dazzling technology, thousands of researchers worldwide have been, or are now, involved in hundreds of sequencing projects for species like the cellular slime mold, roundworm, zebrafish, cat, rat, pig, cow, and baboon. Limited resources focus efforts into the seven main categories of genomes besides *Homo sapiens*: viruses, bacteria, fungi, *Arabidopsis thaliana* ('the weed'), *Drosophila melanogaster* (fruitfly), *Caenorhabditis elegans* (roundworm), and *M. musculus* (mouse). For an overview of sequence data, see [www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome) and for genome landmarks, readers can search the online collection available on [www.sciencemag.org/feature/plus/sfg/special/index.shtml](http://www.sciencemag.org/feature/plus/sfg/special/index.shtml). The Human Genome Project is described in the next section.

The first completed genome reported was of the bacterium *Haemophilus influenzae* in 1995 (see also Box 1.3). Soon after came the yeast genome (*Saccharomyces cerevisiae*) (1996, see [www.yeastgenome.org](http://www.yeastgenome.org)), the bacterium *Bacillus subtilis* (1997), and the tuberculosis bacterium (*Mycobacterium tuberculosis*) in 1998. Reports of the worm, fruitfly, mustard plant, and rice genomes (described below) represent important landmarks, in addition to the human genome (next section).

### Roundworm, *C. elegans* (1998)

The completion of the genome deciphering of the first multicellular animal, the one-millimeter-long soil worm *C. elegans*, made many headlines in 1998 (see the 11 December 1998 issue of *Science*, volume 282, and [www.wormbase.org/](http://www.wormbase.org/)). It reflects a triumphant collaboration of more than eight years between Cambridge and Washington University laboratories.

The nematode genome paves the way to obtaining many insights into genetic relationships among different genomes, their functional characterization, and associated evolutionary pathways. A comparison of the worm and yeast genomes, in particular, offers insights into the genetic changes required to support a multicellular organism.

A comparative analysis between the worm and human genome is also important. Since it was found that roughly one third of the worm's proteins (>6000) are similar to those of mammals, automated screening tests are already in progress to search for new drugs that affect worm proteins that are related to proteins involved in human diseases. For example, diabetes drugs can be tested on worms with a mutation in the gene for the insulin receptor.

Opportunities for studying and treating human obesity (by targeting relevant proteins) also exist: in early 2003 [66] biologists have identified the genes that regulate fat storage and metabolism in the roundworm (i.e., 305 that reduce and 112 that increase body fat) in a landmark experiment that inactivated nearly all of the animal's genes (i.e., expression was disrupted for 16,757 worm genes out of the predicted total of 19,757 that code for proteins) in a single experiment using new RNA interference technology [626]. Such studies are now routine (e.g., to study regulation of immunity to pathogenic bacterial infections [273], or of programmed cell death (or *apoptosis*) [229]). (see Chapter 7 on RNA).

The remarkable roundworm also made headlines in the Fall of 2002 when the Nobel Prize in Physiology or Medicine was awarded to Sydney Brenner, Robert Horvitz, and John Sulston for their collective work over 30 years on *C. elegans* on programmed cell death, *apoptosis*. This process by which healthy cells are instructed to kill themselves is vital for proper organ and tissue development and also leads to diseases like cancer and neurodegenerative diseases. Better knowledge of what leads to cell death and how it can be blocked helps to identify agents of many human disorders and eventually to develop treatments.

### Fruitfly, *Drosophila* (1999)

The deciphering of most of the fruitfly genome in 2000 by Celera Genomics, in collaboration with academic teams in the Berkeley and European *Drosophila* Genome projects, made headlines in March 2000 (see the 24 March 2000 issue of *Science*, volume 287, and [www.fruitfly.org](http://www.fruitfly.org)), in large part due to the groundbreaking "annotation jamboree" employed to assign functional guesses to the identified genes.

Interestingly, the million-celled fruitfly genome has fewer genes than the tiny, 1000-celled worm *C. elegans* (though initial reports of the number of worm's genes may have been over-estimated) and only twice the number of genes as the unicellular yeast. This is surprising given the complexity of the fruitfly — with wings, blood, kidney, and a powerful brain that can compute elaborate behavior patterns. Like some other eukaryotes, this insect has developed a nested set of genes with alternate splicing patterns that can produce more than one meaning from a given DNA sequence (i.e., different mRNAs from the same gene). Indeed, the number of core proteins in both fruitflies and worms is roughly similar (8000 vs. 9500, respectively).

Fly genes with human counterparts may help to develop drugs that inhibit encoded proteins. Already, one such fly gene is *p53*, a tumor-suppressor gene that, when mutated, allows cells to become cancerous. The humble baker's yeast proteins are also being exploited to assess activity of cancer drugs.

#### Mustard Plant, *Arabidopsis* (2000)

*Arabidopsis thaliana* is a small plant in the mustard family, with the smallest genome and the highest gene density so far identified in a flowering plant (125 million base pairs and roughly 25,000 genes). Two out of the five chromosomes of *Arabidopsis* were completed by the end of 1999, and the full genome (representing 92% of the sequence) published one year later, a major milestone for genetics. See the 14 December 1999 issue of *Nature*, volume 408, and [www.arabidopsis.org/](http://www.arabidopsis.org/), for example.

This achievement is important because gene-dense plants (25,000 genes versus 19,000 and 13,000 for brain and nervous-system containing roundworm and fruitfly, respectively) have developed an enormous and complex repertoire of genes for the needed chemical reactions involving sunlight, air, and water. Understanding these gene functions and comparing them to human genes will provide insights into other flowering plants, like corn and rice, and will aid in our understanding of human life. Plant sequencing analysis should lead to improved crop production (in terms of nutrition and disease resistance) by genetic engineering and to new plant-based ingredients in our medicine cabinets. For example, engineered crops that are larger, more resistant to cold, and that grow faster have already been produced.

*Arabidopsis*'s genome is also directly relevant to human biological function, as many fundamental processes of life are shared by all higher organisms. Some common genes are related to cancer and premature aging. The much more facile manipulation and study of those disease-related genes in plants, compared to human or animal models, is a boon for medical researchers.

Interestingly, scientists found that nearly two-thirds of the *Arabidopsis* genes are duplicates, but it is possible that different roles for these apparently-duplicate genes within the organism might be eventually found. Others suggest that duplication may serve to protect the plants against DNA damage from solar

radiation; a ‘spare’ could become crucial if a gene becomes mutated. Intriguingly, the plant also has little “junk”<sup>16</sup> (i.e., not gene coding) DNA, unlike humans.

The next big challenge for *Arabidopsis* aficionados is to determine the function of every gene by precise experimental manipulations that deactivate or overactivate one gene at a time. For this purpose, the community launched a 10-year gene-determination project (a “2010 Project”) in December 2000. Though guesses based on homology sequences with genes from other organisms have been made (for roughly one half of the genes) by the time the complete genome sequence was reported, much work lies ahead to nail down each function precisely. This large number of “mystery genes” promises a vast world of plant biochemistry awaiting exploration.

### Mouse (2001, 2002)

The international Mouse Genome Sequencing Consortium (MGSC) formed in late fall of 2000 followed in the footsteps of the human genome project [288]. The mouse represents one of five central model organisms that were planned at that time to be sequenced. Though coming in the backdrop of the human genome, draft versions of the mouse genome were announced by both the private and public consortia in 2001 and 2002, respectively.

In the end of 2002, the MGSC published a high-quality draft sequence and analysis of the mouse genome (see the 5 December 2002 issue of *Nature*, volume 420). The 2.5 billion size of the mouse genome is slightly smaller than the human genome (3 billion in length), and the number of estimated mouse genes, around 30,000, is roughly similar to the number believed for humans. Intriguingly, the various analyses reported in December 2002 revealed that only a small percentage (1%) of the mouse’s genes has no obvious human counterpart. This similarity makes the mouse genome an excellent organism for studying human diseases and proposed treatments. But the obvious dissimilarity between mice and men and women also begs for further comparative investigations; why are we not more like mice? Part of this question may be explained through an understanding of how mouse and human genes might be regulated differently.

Related to this control of gene activation and function are newly-discovered mechanisms for transcription regulation. Specifically, the mouse genome analyses suggested that a novel class of genes called *RNA genes* — RNA transcripts that do not code for proteins — has other essential regulatory functions that may play significant roles in each organism’s survival (see discussion on RNAs in Chapter 7 on non-coding RNAs). As details of these mechanisms, as well as comparisons among human and other closely-related organisms, will emerge, explanations may arise. In the mean time, genetic researchers have a huge boost of resources and

---

<sup>16</sup>The term “junk DNA”, coined early in the genomics game, turned out to be misleading. Non-coding DNAs are now known to have important functions, far from useless DNA and more like a reservoir for rearranging genes. These repetitive elements are thus essential components of eukaryotic organisms [817].

are already exploiting similarities to generate expression patterns for genes of entire chromosomes as a way to research specific human diseases like Down's syndrome, which occurs when a person inherits three instead of two copies of chromosome 21.

#### Rice (2002)

The second largest genome sequencing project — for the rice plant (see a description of the human genome project below) — has been underway since the late 1990s in many groups around the world. The relatively small size of the rice genome makes it an ideal model system for investigating the genomic sequences of other grass crops like corn, barley, wheat, rye, and sugarcane. Knowledge of the genome will help create higher quality and more nutritious rice and other cereal crops. Significant impact on agriculture and economics is thus expected.

By May 2000, a rough draft (around 85%) of the rice genome (430 million bases) was announced, another exemplary cooperation between the St. Louis-based agrobiotechnology company Monsanto (now part of Pharmacia) and a University of Washington genomics team.

In April 2002, two groups (a Chinese team from the Beijing Genomics Institute led by Yang Huanming and the Swiss agrobiotech giant Syngenta led by Stephen Goff) published two versions of the rice genome (see the 5 April 2002 issue of *Science*, volume 296), for the rice subspecies *indica* and *japonica*, respectively. Both sequences contain many gaps and errors, as they were solved by the whole-genome 'shotgun' approach (see Box 1.3), but represent the first detailed glimpse into a world food staple. The complete sequences of chromosomes 1 and 4 of rice were reported in late 2002 (see the 21 November 2002 issue of *Nature*, volume 420).

#### Pufferfish, *Fugu* (2002)

The highly poisonous delicacy from the tiger pufferfish prepared by trained Japanese chefs has long been a genomic model organism for Sydney Brenner, a founder of molecular biology and recipient of the 2002 Nobel Prize in Physiology or Medicine for his work on programmed cell death (see above, under Roundworm). The compact *Fugu rubripes* genome is only one-ninth the size of the human genome but contains approximately the same number of genes: shorter introns and less repetitive DNA account for this difference. The whole-genome shotgun approach (see Box 1.3) was used to sequence *Fugu* (see the 23 August 2002 issue of *Science*, volume 297). Through comparative genomics, analyses of this ancient vertebrate genome and of many others help understand the extent of protein evolution (through common and divergent genes) and help interpret many human genes.

#### Homo Sapiens (2003)

See the separate section.

## Other Organisms

Complete sequences and drafts of many genomes are now known. (Check websites such as [www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome) for status reports). Included are bacterial genomes of a microbe that can survive environments lethal for most organisms and might turn useful as a metabolizer of toxic waste (*D. radiodurans* R1); a nasty little bacterium that causes diseases in oranges, grapes, and other plants (*Xylella fastidiosa*, decoded by a Brazilian team); and the bugs for human foes like the common cold, anthrax, cholera, syphilis, tuberculosis, malaria, the plague, typhus, and SARS (severe acute respiratory syndrome). Proteins unique to these pathogens are being studied, and disease treatments will likely follow (e.g., cholera vaccine).

The third mammalian genome, that of the rat (Brown Norway rat) was completed in early 2004 (see the 1 April 2004 issue of *Nature*, volume 428), allowing us to explore characteristics that are specific to rodents but also common to all mammals.

By early 2010, the genomes of several mammals have been characterized, including the human, chimp, Rhesus macaque, dog, cow, horse, opossum, platypus, giant panda, and Tibetan antelope. The platypus genome, for example, reveals both reptilian and mammalian features and helps define the ancestral line of animal evolution. The cow genome is useful for studying human diabetes, since bovine insulin is a model for studying many human endocrine diseases. The giant panda genome helps explain the animal's bamboo-chomping habit (malfunction of a gene related to digestion). The genome of the Tibetan antelope, also an endangered species, may guide researchers in understanding the animal's ability to adapt to harsh environments (extreme cold and low oxygen) and the pathogenesis of chronic plateau sickness.

### 1.5.2 *The Human Genome*

The International Human Genome Project was launched in 1990 to sequence all three billion bases in human DNA [288]. The public consortium has contributions from many parts of the world (such as the United States, United Kingdom, Japan, France, Germany, China, and more) and is coordinated by academic centers funded by NIH and the Wellcome Trust of London, headed by Francis Collins and Michael Morgan (with groups at the Sanger Institute near Cambridge, UK, and four centers in the United States).

In 1998, a competing private enterprise led by Craig Venter's biotechnology firm Celera Genomics and colleagues at The Institute for Genomic Research (TIGR), both at Rockville, Maryland (owned by the PE Corporation; see [www.celera.com](http://www.celera.com)), entered the race. Eventually, this race to decode the human genome turned into a collaboration in part, not only due to international pressure but also because the different approaches for sequencing taken by the public and private consortia are complementary (see Box 1.3 and related articles [2,268,476, 888,1351,1352] comparing the approaches for the human genome assembly).

## Milestones

A first milestone was reached in December 1999 when 97% of the second smallest chromosome, number 22, was sequenced by the public consortium (the missing 3% of the sequence is due to 11 gaps in contiguity); see the 2 December 1999 issue of *Nature*, volume 402. Though small (43 million bases, <2% of genomic DNA), chromosome 22 is gene rich and accounts for many genetic diseases (e.g., schizophrenia).

Chromosome 21, the smallest, was mapped soon after (11 May 2000 issue of *Nature*, volume 405) and found to contain far fewer genes than the 545 in chromosome 22. This opened the possibility that the total number of genes in human DNA is less than the 100,000 previously estimated. Chromosome 21 is best known for its association with Down's syndrome; affected children are born with three rather than two copies of the chromosome. Learning about the genes associated with chromosome 21 may help identify genes involved in the disease and, eventually, develop treatments.

Completion of the first draft of the human genome sequence project broke worldwide headlines on 26 June 2000 (see, for example, the July 2000 issue of *Scientific American*, volume 283). This draft reflects 97% of the genome cloned and 85% of it sequenced accurately, that is, with 5 to 7-fold redundancy.

Actually, the declaration of the ‘draft’ status was arbitrary<sup>17</sup> and even fell short of the 90% figure set as target. Still, there is no doubt that the human genome represents a landmark contribution to humankind, joined to the ranks of other ‘Big Science’ projects like the Manhattan project and the Apollo space program. The June 2000 announcement also represented a ‘truce’ between the principal players of the public and private human genome efforts and a commitment to continue to work together for the general cause.

A New York Times editorial by David Baltimore on the Sunday before the Monday announcement was expected underscored this achievement, but also emphasized the work that lies ahead:

The very celebration of the completion of the human genome is a rare day in the history of science: an event of historic significance is recognized not in retrospect, but as it is happening . . . While it is a moment worthy of the attention of every human, we should not mistake progress for a solution. There is yet much work to be done. It will take many decades to fully comprehend the magnificence of the DNA edifice built over four billion years of evolution and held in the nucleus of each cell of the body of each organism on earth.

David Baltimore, *New York Times*, 25 June 2000.

---

<sup>17</sup>It has been said [1238] that this day happened to be free in the diaries of U.S. President Bill Clinton and Britain’s Prime Minister Tony Blair, who proclaimed victory over the genome along with leading scientists.

Baltimore further explains that the number of proteins, not genes, determines the complexity of an organism. The gene number should ultimately explain the complexity of humans. In June 2000, the estimated number of total human genes was 50,000, compared to 14,000 in a fly or 18,000 in a worm. Several months after the June announcement, this estimate was reduced to 30,000–40,000 (see the 15 February 2001 issue of *Nature*, volume 409, and the 16 February 2001 issue of *Science*, volume 291). This implies an ‘equivalence’ of sorts between each human and roughly two flies . . . . However, the estimated number has declined since then to around 20,000–25,000.

This astonishingly low number suggests that hidden levels of complexity exist in the human genome from networks of genes rather than individual genes. Clearly, the final word on the number of human genes and the conserved genes that humans share with flies, mice or other organisms awaits further studies.

Three years after the ‘working draft’ of the human genome sequence was announced with much fanfare, the Human Genome Project as originally devised was declared complete, to an accuracy of 99.9%; the international consortium of genome sequencing centers put all the fragments of the 3.1-billion DNA units of the human genome in order, and closed nearly all of the gaps. The month of April 2003 for this declaration was timed to coincide with the 50th anniversary of Watson and Crick’s report of the structure of the DNA double helix.

Some chromosome segments of the human genome, like in chromosome Y, are more difficult to characterize as they are highly repetitive; fortunately, these segments may be relatively insignificant for the genome’s overall function.

Of course, we still have little clue on what to make of the DNA book of life presented before us in terms of greater predispositions for specific diseases and individualized response to therapy. These tasks will undoubtedly occupy us in the coming decades.

With the determination of the human genome sequence for several individuals, including Craig Venter (in 2007) [753], James Watson (in 2008) [1368], a 4000-year-old man from Greenland and five southern Africans including Archbishop Desmond Tutu (in 2010), variations (polymorphisms) in the DNA sequence that contribute to disease in different populations are being investigated and analyzed. The ancient DNA analysis also sheds light on migration trends and ancestry by revealing unsuspected movements from Siberia to Greenland about 5500 years ago. The African genomes also help understand human genetic history because the number of variations is relatively high. The establishment of the Personal Genome and 1000 Genomes Projects promises to deliver more such insights.

In addition to identifying these variations, the next task is to define the proteins produced by each gene and understand the cellular interactions of those proteins. This is opening new avenues for disease diagnostics and development of designer drugs. Undoubtedly, the determination of sequences for 1000 major species in the next decade will shed further insights into the human genome, but clearly we are only beginning to understand what it all means. Until then, knowing and interacting with an individual may be more informative than sorting through her/his DNA, as M. Olson suggests in his commentary on Watson’s genome sequencing [936].

### A Triumph of Technology

It should be emphasized that none of these studies would be possible without remarkable advances in sequencing technology in terms of speed and efficiency [1323]. The HGP took 13 years to complete in 2003 at a cost of \$2.7 billion. Four years later, Venter's DNA was sequenced after only four years of work at a cost of \$100 million [753]. Just one year later, in 2008, Watson's DNA was determined after about 4 months of work at a fraction of the cost (<\$1.5 million) [936, 1368].

In 2009, BioNanomatrix designed a nanofluidic chip approach to sequencing that could lower DNA sequencing costs down to <\$100 within the next 5 years (see Figure 1.5). Another innovative 'DNA Sudoku' approach to sequencing was soon after reported — using combinatorial pooling strategies to sequence multiple genomes for the purpose of analyzing rapidly specific regions of sequence variants, such as associated with disease [366]. This strategy of mixing many genome samples and using logic and combinatorial rules as used in number puzzles to search for specific patterns in the pool of samples is best suited for genotype analysis of short segments to diagnose genetic diseases such as Tay-Sachs or cystic fibrosis that tend to occur in certain ethnic groups.

Many other companies (e.g., Illumina, Life Technologies, Oxford Nanopore Technologies, Pacific Biosciences, and IBM) have entered the personal sequencing service with innovative approaches; efficiency will certainly increase and cost dramatically decrease in the very near future. Next-generation machines could also make possible rapid sequencing of disease-causing microbes to allow infection diagnosis or tracking, as well as lead to important engineering applications (e.g., bacteria that produce more hydrogen). The increasing challenge of handling and processing large amounts of sequence data may also be alleviated by the use of cloud computing — computational resources distributed over the Internet.

It is no surprise that the era of large-scale sequencing projects has been supplanted by many private, for-profit companies like Navigenics, 23andMe, GeneWize, Knome, and others who are offering individuals personal atlases of their DNA. Will this type of direct-to-consumer genetics be the norm in the near future?

#### **Box 1.3: Different Sequencing Approaches**

Two synergistic approaches have been used for sequencing. The public consortium's approach relies on a 'clone-by-clone' approach: breaking DNA into large fragments, cloning each fragment by inserting it into the genome of a bacterial artificial chromosome (BAC), sequencing the BACs once the entire genome is spanned, and then creating a physical map from the individual BAC clones. The last part — rearranging the fragments in the order they occur on the chromosome — is the most difficult. It involves resolving the overlapped fragments sharing short sequences of DNA ('sequence-tagged sites').

The alternative approach pioneered by Venter's Celera involves reconstructing the entire genome from small pieces of DNA without a prior map of their chromosomal positions.

The reconstruction is accomplished through sophisticated data-processing equipment. Essentially, this gargantuan jigsaw puzzle is assembled by matching sequence pieces as the larger picture evolves.

The first successful demonstration of this piecemeal approach was reported by Celera for decoding the genome of the bacterium *Haemophilus influenzae* in 1995. This bacterium has a mere 1.8 million base pairs with estimated 1700 genes, versus three billion base pairs for human DNA with at least 30,000 genes. The sequence of *Drosophila* followed in 1998 (140 million base pairs, 13,000 estimated genes) and was released to the public in early 2000 (see the 24 March 2000 issue of *Science*, volume 287).

This ‘shotgun’ approach has been applied to the human genome, more challenging than the above organisms for two reasons. The human genome is larger — requiring the puzzle to be formed from ~70 million pieces — and has many more repeat sequences, complicating accurate genome assembly. For this reason, the public data were incorporated into the whole genome assembly [476]. The whole-genome shotgun approach has also been applied to obtain a draft of the mouse (2001), rice (2002) and pufferfish (2002) genomes, for example.

The two approaches are complementary, since the rapid deciphering of small pieces by the latter approach relies upon the larger picture generated by the clone-by-clone approach for overall reconstruction. See a series of articles in 2002 [476, 888, 1351] that scrutinized those approaches, focusing on the extent of public-database information utilized in the

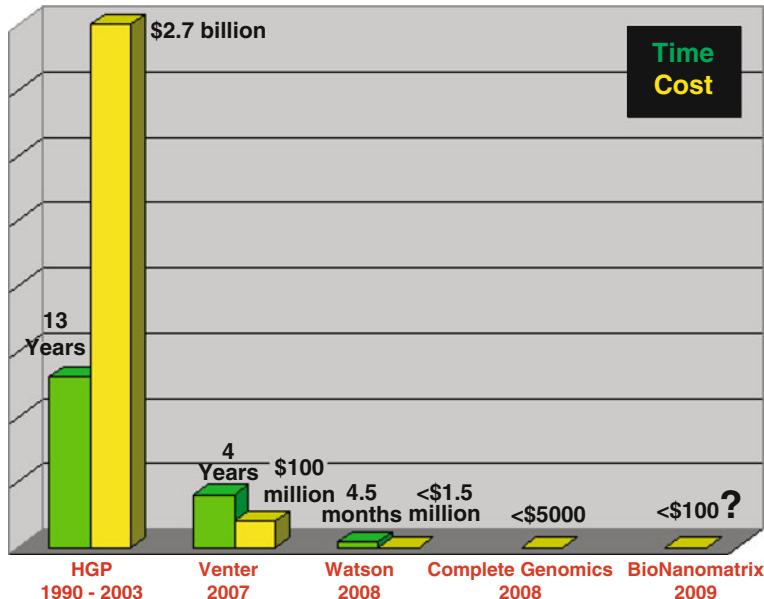


Figure 1.5. The progress of DNA sequencing technology. The sequencing time and cost associated with the human genome project, determination of Venter’s and Watson’s DNA, and biotechnology company prospects are given.

whole-genome shotgun approach to the human genome assembly, and the second round of debate in 2003 [2, 268, 1352]. The recent success of individual DNA sequencing involves rapid-sequencing methodologies that cut DNA into tiny segments of only 250 base pairs long. This makes genome assembly more technically challenging but the the entire sequencing process much faster and cheaper.

---

### A Gold Mine of Biodata

The most up-to-date information on sequencing projects can be obtained from the U.S. National Center for Biotechnology Information (NCBI) at the U.S. National Library of Medicine, which is developing a sophisticated analysis network for the human genome data.

For information, see the Human Genome Resources Guide [www.ncbi.nlm.nih.gov/genome/guide/human](http://www.ncbi.nlm.nih.gov/genome/guide/human) (click on Map Viewer), the U.S. National Human Genome Research Institute's site [www.nhgri.nih.gov/](http://www.nhgri.nih.gov/), that of Department of Energy (DOE) at [genomics.energy.gov](http://genomics.energy.gov), the site of the University of California at Santa Cruz at [genome.ucsc.edu/](http://genome.ucsc.edu/), and others.<sup>18</sup>

Since 1992, NCBI has maintained the GenBank database of publicly available nucleotide sequences ([www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)). A typical GenBank entry includes information on the gene locus and its definition, organism information, literature citations, and biological features like coding regions and their protein translations. Many search and analysis tools are also available to serve researchers.

### Implications – Some Application Examples

The genomic revolution and the comparative genomics enterprises now underway will not only provide fundamental knowledge about the organization and evolution of biological systems in the decades to come [672] but will also lead to medical breakthroughs.

Already, some practical benefits of genomic deciphering have emerged (e.g., [529, 892]). A dramatic demonstration in 2000 was the design of the first vaccine to prevent a deadly form of bacterial meningitis using a two-year gene-hunting process at Chiron Corporation. Researchers searched through the computer database of all the bacterium's genes and found several key proteins that in laboratory experiments stimulated powerful immune responses against all known strains of the *Neisseria meningitidis* Serogroup B Strain MC58 bug [1003].

---

<sup>18</sup>Some useful web sites for genomic data include [www.arabidopsis.org](http://www.arabidopsis.org), [www.ncbi.nlm.nih.gov/Sitemap/](http://www.ncbi.nlm.nih.gov/Sitemap/), plant genomes for the specialist, the Agricultural Genome Information System, *Caenorhabditis elegans* Genetics and Genomics, Crop Genome Databases at Cornell University, FlyBase, The Genome Database, Genome Sequencing Center (Washington University), GenomeNet, U.S. National Agricultural Library, Online Mendelian Inheritance in Man, *Pseudomonas aeruginosa* Community Annotation Project, *Saccharomyces* Genome Database, The Sanger Institute, Taxonomy Browser, and UniGene.

In April 2003, just two months after the first inklings of a deadly disease called SARS emerged from Asia, a global effort coordinated by the World Health Organization announced that it had mapped the coronavirus genome that causes this highly infectious disease (see resulting papers in the 30 May 2003 issue of *Science*, volume 300). This was made possible by the new high-tech science era of internet links and sequencing methods. Soon after, Affymetrix released a SARS resequencing array encompassing the entire 30-kb genome of the virus, allowing analysis of variation among various virus versions and thus an understanding of how rapidly the virus changes and spreads. Having the viral genome sequence, work continues on understanding the roles of viral proteins in SARS pathogenesis, crucial for developing suitable drugs and vaccines [42]. Finding a treatment for SARS certainly presents a challenge, since the devastating economic losses due to the SARS outbreak underscored our vulnerabilities to infectious diseases. Yet the global virus hunt is a model *par excellence* for the potential of genomics initiatives. Indeed, three years later, a promising inhibitor of the SARS virus was identified by computer-aided molecular design [329].

Full genome sequencing and genomics association studies were once again celebrated when years of secret forensic investigations led in August 2008 to the source of the 2001 anthrax attacks in the U.S. It was then that the U.S. Federal Bureau of Investigation (FBI) finally produced conclusive evidence against Army Institute scientist Bruce Ivins (in Fort Detrick, MD) as the propagator of the anthrax murders in 2001. Not only does this case reveal the dangers of genomic information of killer organisms like *Bacillus anthracis*; it underscores the importance of full-genome sequencing and related technological breakthroughs as forensic tools.

The story begins in 2001, when in the wake of the horrid Al Qaeda attacks on New York and Washington, D.C., mysterious white powders mailed to several individuals caused several illnesses and five eventual deaths from various forms of anthrax. For years, the public was frustrated at the FBI's lack of resolution of the culprit, though Army scientist Steven Hatfill was implicated (and ultimately vindicated and compensated by a large sum of money). From the start, scientists specializing in biowarfare were suspected since using anthrax as a killer requires expert knowledge.

In July 2008, Ivins died from a drug overdose, apparently a suicide, raising public suspicions and re-igniting the public's desire for a conclusion to this affair. Little did we know that the FBI has been busy for years with meticulous scientific investigations, most of which was commissioned to The Institute for Genomic Research (TIGR) in MD. When the story was finally unveiled, it was concluded that the anthrax species at Ivins' possession, flask RMR-1027, matched conclusively the victims' anthrax. The technical complications also became clear: the nonuniform anthrax genome contained variants or *spores* that required special cultivation followed by sequencing to identify, isolate, and finally match perfectly to one source: only the Ivins source and seven directly-related isolates contained all the four mutations that were identified from the victims' anthrax genome (see [363] and N. Wade in *The New York Times*, Aug. 21, 2008). The dangerous nature

of the genome also required the FBI to make piecemeal requests for the scientific work involved. Though Ivins' motive for these deadly acts remains unclear, the possibility that Ivins propagated the scare in an effort to ensure continued government funding for anthrax vaccine development and other biological warfare developments remains viable.

The anthrax story emphasizes the long-recognized dangers of having genome information in the wrong hands. However, in theory, genomic information on deadly agents, from anthrax to the influenza virus, can bring about new treatments. In practice, this has turned to be a greater challenge than anticipated. For example, in our ongoing search for new potent antibiotics to fight pernicious infections that are resistant to many known antibiotics (like MRSA, or Methicillin-resistant *Staphylococcus aureus*), the availability of hundreds of sequenced bacterial genomes has hardly led to new antibiotic targets. Genomics may simply be an insufficient basis to serve as a foundation for developing better infection treatments since even a minute amount of resistant bodies can trigger resurgence of disease and/or drug resistance.

Indeed, our susceptibility to viruses, for example, was made clear in the 2009 pandemic of the swine flu (H1N1). This virus resulted in millions of infections throughout the world in a very short time and hundreds of deaths, despite genomics advances and modern screening and treatment tools.

Of course, one of the primary hopes for genomics applications has been the expectation for better diagnosis and treatment of human disease. However, genome association studies have generally shown limited value in predicting human diseases because genetic variations only explain a small part of the genetic/disease link. In other words, the genetic link to disease remains largely unclear and it is possible that rare variants account for most of the genetic basis for disease. Indeed, dissenting voices have suggested that our efforts to pinpoint the genetics of common diseases may not work. Still, successful examples of using a patient's DNA to help tailor medication has shown some promise, as in the case of Tamoxifen for breast cancer, Erbitux for colon cancer, or Iressa for lung cancer — all of which work only for patients with a particular genetic mutation. Nonetheless, these findings are complicated by the fact that genetic tests are not sufficiently accurate at present, as indicated by the breast cancer drug Herceptin, which is only effective on certain subtypes of the disease. See Chapter 15 for an expanded discussion of pharmacogenomics, including Herceptin.

Many still remain generally hopeful because exploiting genomic information for new medical breakthroughs is a new field that will take time to mature and succeed. There is no doubt that the many ideas and tools not available to us several decades ago but now at our disposal will ultimately lead to significant medical advances. And given the deadly infectious diseases, like HIV/AIDS, Ebola, Marbourg, and antibiotics-resistant 'super-bugs', that are responsible for more than 25% of the world's deaths, it is hoped that the new tools, together with new ideas and improved techniques, will lead to these needed biomedical developments.

### Ongoing Challenges and Ramifications

As gene products are being identified, the biological revolution is beginning to affect many aspects of our lives [259], perhaps not too far away from Wilson's vision of consilience. A 'gold mine' of biological data is now amassing, likened to "orchards . . . just waiting to be picked".<sup>19</sup> This rich resource for medicine and technology also provides new foundations, as never before, for computational applications.

Consequently, in fifty years' time, we anticipate breakthroughs in protein folding, medicine, cellular mechanisms (regulation, gene interactions), development and differentiation, history (population genetics, origin of life), and perhaps new life forms, through analysis of conserved and vital genes as well as new gene products. See the 5 October 2001 issue of *Science* (volume 294) for a discussion of new ideas, projects, and scientific advances that followed since the sequencing of the human genome.

Among the promising medical leaps are personalized and molecular medicine, perhaps in large part due to the revolutionary DNA microarray technology (see [400] and Box 1.4) and gene therapy. Of course, *information is not knowledge*, but rather a road that can lead to perception. Therefore, these aforementioned achievements will require concerted efforts to extract information from all the sequence data concerning gene products.

Many initiatives are underway to process genetic data in the goal of understanding, and eventually treating, human diseases. For example, in 2003 Britain launched a 'genetic census' *Biobank* project — assembly of a database of medical information based on 500,000 Britons representing Britain's demographics aimed at quantifying the combined genetic and environmental (e.g., pollution, smoking, exercise, diet) influence on common human ailments.

Other national genetic database projects (with corresponding numbers of participants) are underway in Iceland (275,000), Sweden (80,000), Estonia (1 million), and Latvia (50,000). Private genomic database projects are also being assembled by the American Cancer Society (110,000), Mayo Clinic (200,000), and CARTaGENE (50,000). Companies like the pioneering Icelandic DeCode Genetics went on a hunt to search for disease genes in these genealogies.<sup>20</sup> Many other international consortia and large-scale projects have been formed, including ENCODE, 1000 genomes project, Cancer Atlas, HapMap, Personal Genome Project, and much more, to interpret the human genome, and many private companies have started to exploit many biomedical and technological issues of genomic sciences.

---

<sup>19</sup>B. Sinclair, in *The Scientist*, 19 March 2001.

<sup>20</sup>In November 2009, DeCode Genetics, which was founded in 1996, filed for bankruptcy. Though quickly becoming the world leader in the race to identify genetic connections to common diseases like cancer, diabetes and schizophrenia, experts believe that — regardless of business strategies used by the company — the genetic nature of human disease has turned out to be much more complex than originally envisioned. In January 2010, the company announced that it emerged from bankruptcy and will continue its genetics research as a private company, though it will abandon its drug development efforts.

When leading scientists were queried in 2002 by the publisher of the website **Edge.org** devoted to science to advise on action to take concerning the most pressing scientific issues in the world, physicist Freeman Dyson boldly suggested “a planetary genome sequencing project to identify all the segments of the genomes of all the millions of species that live together in the planet”. Dyson’s vision for completing the sequencing of the biosphere within less than half a century aims to profoundly increase our understanding of the ecology of the planet, which could lead to environmental improvements and cures for human diseases. There is no doubt that creative and well engineered projects combining technological innovations with biological data can have enormous ramifications on our lives. See, for example, a vision for the future of genomics research by Collins and coworkers [258].

Some of the ongoing challenges that face us now include establishing gene number, location, and function; understanding the interaction of protein networks; understanding non-coding DNA (amount, distribution, function); determining protein structure and function evolution; correlating single nucleotide polymorphisms to health and disease predisposition; establishing evolutionary trends among organisms; and exploiting genome information for environmental restoration via designer organisms.

Many societal, ethical, economic, legal, and political issues will also have to be addressed with these developments. Still, like the relatively minor Y2K (Year 2000) anxiety, these problems could be resolved in stride through multidisciplinary networks of expertise.

In a way, sequencing projects make the giant leap directly from *sequence to function* (possible only when a homologous sequence is available whose function is known). However, the crucial middle aspect — *structure* — must be relied upon to make systematic functional links. This systematic interpolation and extrapolation between sequence and structure relies and depends upon advances in biomolecular modeling, in addition to high-throughput structure technology (‘the human proteomics project’).

The next chapter introduces some current challenges in modeling macromolecules and mentions important applications in medicine and technology.

#### **Box 1.4: Genomics & Microarrays**

DNA microarrays — also known as gene chips, DNA chips, and biochips — are becoming marvelous tools for linking gene sequence to gene products. They can provide, in a single experiment, an expression profile of many genes [400]. As a result, they have important applications to basic and clinical biomedicine. Particularly exciting is the application of such genomic data to *personalized medicine* or *pharmacogenomics* — prescribing medication based on genotyping results of both patient and any associated bacterial or viral pathogen [370]. Prescribing specific diets to affect health based on genetic responses to diet (*nutritional genomics* or *nutrigenomics*) is another application gaining momentum.

Essentially, each microarray is a grid of DNA oligonucleotides (called *probes*) prepared with sequences that represent various genes. These probes are directed to a specific gene or mRNA samples (called *targets*) from tissues of interest (e.g., cancer cells). Binding between probe and target occurs if the RNA is complementary to the target nucleic acid. Thus, probes can be designed to bind a target mRNA if the probe contains certain mutations. Single nucleotide polymorphisms or SNPs, which account for 0.1% of the genetic difference among individuals, can be detected this way [848].

The hybridization event — amount of RNA that binds to each cell grid — reflects the extent of gene expression (gene activity in a particular cell). Such measurements can be detected by fluorescence tagging of oligonucleotides. The color and intensity of the resulting base-pair matches reveal gene expression patterns.

Different types of microarray technologies are now used (e.g., using different types of DNA probes), each with strengths and weaknesses. The technique of principal component analysis (PCA, see Chapter 15) has shown to be useful in analyzing microarray data (e.g., [1009]). Technical challenges remain concerning verification of the DNA sequences and ensuring their purity, amplifying the DNA samples, and assessing the results. For example, false positives or false negatives can result from irregular target/probe binding (e.g., mismatches) or from self-folding of the targets, respectively. The problem of accuracy of the oligonucleotides has stimulated various companies to develop appropriate design techniques. Affymetrix Corporation, for example, has developed technology for designing silicon chips with oligonucleotide probes synthesized directly onto them, with thousands of human genes on one chip. All types of DNA microarrays rely on substantial computational analysis of the experimental data to determine absolute or relative patterns of gene expression.

Such patterns of gene expression (induction and repression) can prove valuable in drug design. An understanding of the affected enzymatic pathway by proven drugs, for example, may help screen and design novel compounds with similar effects. This potential was demonstrated for the bacterium *M. tuberculosis*, based on experimental profiles obtained before and after exposure to the tuberculosis drug Isoniazid [1378].

---

# 2

## Biomolecular Structure and Modeling: Problem and Application Perspective

All things come out of the one, and the one out of all things. Change,  
that is the only thing in the world which is unchanging.

Heraclitus of Ephesus (550–475 BC).

### 2.1 Computational Challenges in Structure and Function

#### 2.1.1 *Analysis of the Amassing Biological Databases*

The experimental progress described in the previous chapter has been accompanied by an increasing desire to relate the complex three-dimensional (3D) shapes of biomolecules to their biological functions and interactions with other molecular systems. Structural biology, computational biology, genomics, proteomics, bioinformatics, chemoinformatics, and others are natural partner disciplines in such endeavors.

*Structural biology* focuses on obtaining a detailed structural resolution of macromolecules and relating structure to biological function.

*Computational biology* was first associated with the discipline of finding similarities among nucleotide strings in known genetic sequences, and relating

these relationships to evolutionary commonalities; the term has grown, however, to encompass virtually all computational enterprises to molecular biology problems [747].

*Comparative genomics* — the search and comparison of sequences among species — is a natural outgrowth of the sequencing projects [672]. So are *structural and functional genomics*, the characterization of the 3D structure and biological function of these gene products [149, 167, 212, 635, 867].

In the fall of 2000, the U.S. National Institute of General Medical Sciences (NIGMS) launched a five-year structural genomics initiative (also called PSI for Protein Structure Initiative) by funding seven research groups aiming to solve collectively the 3D structures of 10,000 proteins, each representing a protein family, over the next decade. This goal of assembling a protein fold library required improvements in both structural biology's technology and methodology, so the goals included development of methodology and technology to enable high-throughput structure determination and subsequent automation of unique protein structures. After five years, it was realized that those ambitious goals were not met, so both the methodology and structure-determination aims were scaled down significantly in the next phase of funding (2005–2010).

However, despite steady progress [214], by 2008 it became apparent to some leaders in the community, both within and outside these initiatives, that the task at large is much more difficult than previously imagined, perhaps even unattainable, because of the infinitely-large size of the fold space; in addition, the very tight state of funding to NIH-supported scientists in the wake of the Iraq war prompted many scientists to re-evaluate funding such large-scale cataloging projects at the expense of individual, hypothesis-driven research labs who desperately needed the funding. See [993], for example, urging termination of PSI, and an opposing view from some involved scientists [87]. Since then, efforts are mostly shifting to new functional/structural annotations, which may be more meaningful to our goal of understanding the function of genome products.

*Proteomics* is another current buzzword defining the related discipline of protein structure and function (see [373] for an introduction), and even *cellomics* has been introduced.<sup>1</sup> Cellomics reflects the expanded interest of gene sequencers in integrated cellular structure and function. The *Human Proteomics Project* — a collaborative venture to churn out atomic structures using high-throughput and robotics-aided methods based on NMR spectroscopy and X-ray crystallography, rather than sequences — may well be on its way.

New instruments that have revolutionized genomics known as DNA microarrays, biochips, or gene expression chips (introduced in Chapter 1 and Box 1.4) allow researchers to determine which genes in the cell are active and to identify gene networks.

---

<sup>1</sup>A glossary of biology disciplines coined with “ome” or “omic” terms can be found at <http://www.genomicglossaries.com/content/omes.asp>.

The range of genomic sciences also extends [935] to the *metabolome*, the endeavor to define the complete set of metabolites (low-molecular cellular intermediates) in cells, tissues, and organs. Experimental techniques for performing these integrated studies are continuously being developed. For example, yeast geneticists have developed a clever technique for determining whether two proteins interact and, thereby by inference, participate in related cellular functions [1283]. Such approaches to proteomics provide a powerful way to discover functions of newly identified proteins. DNA chip technology is also thought to hold the future of individualized health care now coined personalized medicine or *pharmacogenomics*; see Chapter 15 and Box 1.4. Additionally, as mentioned in the first chapter, genomics and its disciples have already led to drug discovery, as in the notable case of a SARS virus inhibitor [329]; see [191] for the impact of systems biology on drug discovery.

It has been said that current developments in these fields are *revolutionary rather than evolutionary*. This view reflects the clever exploitation of biomolecular databases with computing technology and the many disciplines contributing to the biomolecular sciences (biology, chemistry, physics, mathematics, statistics, and computer science). *Bioinformatics* is an all-embracing term for many of these exciting enterprises [571, 881] (structural bioinformatics is an important branch); *chemoinformatics* has also followed (see Chapter 15) [507]. Some genome-technology company names are indicative of the flurry of activity and grand expectations from our genomic era. Consider titles like Genetics Computer Group, Genetix Ltd., Genset, Protana, Protein Pathways, Inc., Pyrosequencing AB, Sigma-Genosys, or Transgenomic Incorporated. With many companies now in the business of personal genetics (like Navigenics, 23andMe, Knome), even our approach to health, disease prevention, and treatment may be changing.

This excitement in the field's developments and possibilities is echoed by the chief executive of the software giant Oracle Corp., Lawrence Ellison, who surrounds himself by molecular biologists — the scientists, board members, and fellows of his Ellison Medical Foundation; explaining to a *Wall Street Journal* reporter his preference of molecular biology over racing sailboats, Ellison said: “*The race is more interesting, the people in the race are more interesting and the prize is bigger.*” (*Wall Street Journal*, January 9, 2003). This means a lot from the owner of a multi-million-dollar 90-foot wonder-yacht!

When a new “game”, named Foldit, developed by researchers at the University of Washington, based on the Rosetta@home software, was introduced to the general public, a *ScienceDaily* report featured the headline: “*Computer Game’s High Score Could Earn The Nobel Prize in Medicine*” (May 9, 2008). Whether the serious business of protein folding can be turned into a competitive sport remains to be seen, but the software has surely caught the attention of gamers at large.

Although the number of sequence databases has grown very rapidly and exceeds the amount of structural information, the 1990s saw an exponential rise of structural databases as well. From only 50 solved 3D structures in the Protein Data Bank (PDB) in 1975, the number rose to 500 in 1988; another order of magnitude was reached in 1996 (around 5000 entries), and 50,000 entries were

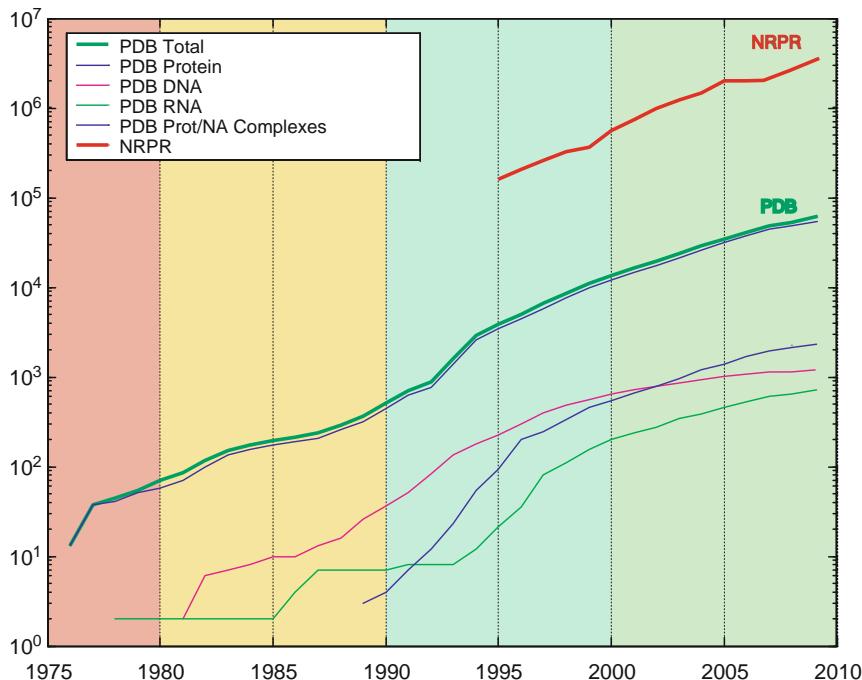


Figure 2.1. The growth of the protein sequence database, NRPR, versus structural database of macromolecules (PDB). See Table 2.1 and [www.dna.affrc.go.jp/growth/P-history.html](http://www.dna.affrc.go.jp/growth/P-history.html). The NRPR database represents merged, non redundant protein database entries from several databases: PIR, SWISS-PROT, GenPept, and PDB.

reported before the end of 2008. In fact, the rate of growth of structural information is approaching the rate of increase of amino acid sequence information (see Figure 2.1 and Table 2.1). It is no longer a rare event to see a new crystal structure on the cover of *Nature* or *Science*. Now, on a weekly basis, groups compete to have their newly-solved structure on the cover of prominent journals. The number of NMR-deduced structures deposited in the PDB has risen slowly, reflecting roughly 13% of the total solved structures by the end of 2009. (For updated information, check the holdings on RCSB).

This trend, coupled with tremendous advances in genome sequencing projects [695], argues strongly for increased usage of computational tools for analysis of sequence/structure/function relationships and for structure prediction and design applications. Thus, besides genomics-based analyses and comparisons, accurate, reliable, and rapid theoretical tools for describing structural and functional aspects of gene products are important. (See [635], for example, for mathematical and computational challenges in genomics).

Table 2.1. Growth of protein sequence databases.

Year	PDB <sup>a</sup>					NRPR <sup>b</sup>
	Protein	DNA	RNA	Pr/NA	Total	
1976	13				13	
1977	37				37	
1978	41		2		43	
1979	51		2		53	
1980	57		2		59	
1981	71	2	2		75	
1982	99	6	2		107	
1983	133	7	2		142	
1984	154	8	2		164	
1985	172	10	2		184	
1986	188	10	4		202	
1987	206	13	7		226	
1988	256	16	7		279	
1989	317	26	7	3	353	
1990	449	36	7	4	496	
1991	616	52	8	7	683	
1992	772	83	8	12	875	
1993	1404	134	8	23	1569	
1994	2606	181	12	55	2854	
1995	3457	227	21	92	3797	159808
1996	4422	308	35	197	4962	204123
1997	5794	404	81	242	6521	258272
1998	7637	484	111	339	8571	324237
1999	9753	560	156	450	10919	360674
2000	12146	644	202	545	13537	560973
2001	14745	717	241	656	16359	744991
2002	17517	785	279	781	19362	990928
2003	21356	862	343	958	23519	1289979
2004	26190	935	392	1195	28712	1472200
2005	31217	1022	454	1385	34078	1988730
2006	37289	1072	536	1675	40572	1988730
2007	44126	1134	615	1934	47809	3638747
2008	50738	1172	694	2225	54829	4456326
2009	57469	1241	760	2528	61998	5097840

<sup>a</sup>From the Protein Data Bank (PDB).

Pr/NA denotes protein/nucleic acid complexes.

<sup>b</sup>From the ‘Non Redundant Proteins database merged Regular release’, [www.dna.affrc.go.jp/growth/P-history.html](http://www.dna.affrc.go.jp/growth/P-history.html).

### 2.1.2 Computing Structure From Sequence

One of the most successful approaches to date on structure prediction comes from *homology modeling* (also called comparative modeling) [16, 79].

In general, a large degree of sequence similarity often suggests similarity in 3D structure. It has been reported, for example, that a sequence identity of greater than 40% usually implies more than 90% 3D-structure overlap (defined as percentage of C<sup>α</sup> atoms of the proteins that are within 3.5 Å of each other in a rigid-body alignment; see definitions in Chapter 3) [1087]. Thus, sequence similarity of at least 50% suggests that the associated structures are similar overall. Conversely, low sequence similarity generally implies structural diversity. This argument of the poor performance of homology modeling when sequence similarity is <50% has been used against large-scale initiatives like the PSI [993].

There are many exceptions to these homology/structure similarity relationships, however, as demonstrated humorously in a contest presented to the protein folding community (see Box 2.1). The *myoglobin* and *hemoglobin* pair is a classic example where large structural, as well as evolutionary, similarity occurs despite little sequence similarity (20%). Other exceptional examples and various sequence/structure relationships are discussed separately in Chapter 3, as well as Homework 6; see also [436] for examples.

More general than prediction by sequence similarity is structure prediction *de novo* [79], a *Grand Challenge* of the field, as described next.

## 2.2 Protein Folding – An Enigma

### 2.2.1 ‘Old’ and ‘New’ Views

There has been much progress on the protein folding challenge since Cyrus Levinthal first posed the well-known “paradox” named after him; see [395, 564] for historical perspectives. Levinthal suggested that well-defined folding pathways might exist [743] since real proteins cannot fold by exhaustively traversing through their entire possible conformational space [744]. (See [318, 1294, 1295] for a related discussion on whether the number of protein conformers depends exponentially or non-exponentially on chain length). Levinthal’s paradox led to the development of two views of folding — the ‘old’ and the ‘new’ — which have since merged [313, 314, 362, 707].

The former accents the existence of a specific folding pathway characterized by well-defined intermediates. The latter emphasizes the rugged, heterogeneous multidimensional energy landscape governing protein folding, with many competing folding pathways [1385]. Yet, the boundary between the two views is pliant and the intersection substantial [362, 395, 564]. This integration has resulted from a variety of information sources: theories on funnel-shaped energy landscape (e.g., [179, 314, 949]); folding and unfolding simulations of simplified models (e.g., [396, 430, 509, 660, 749, 861, 1170]), at high temperatures or low pH

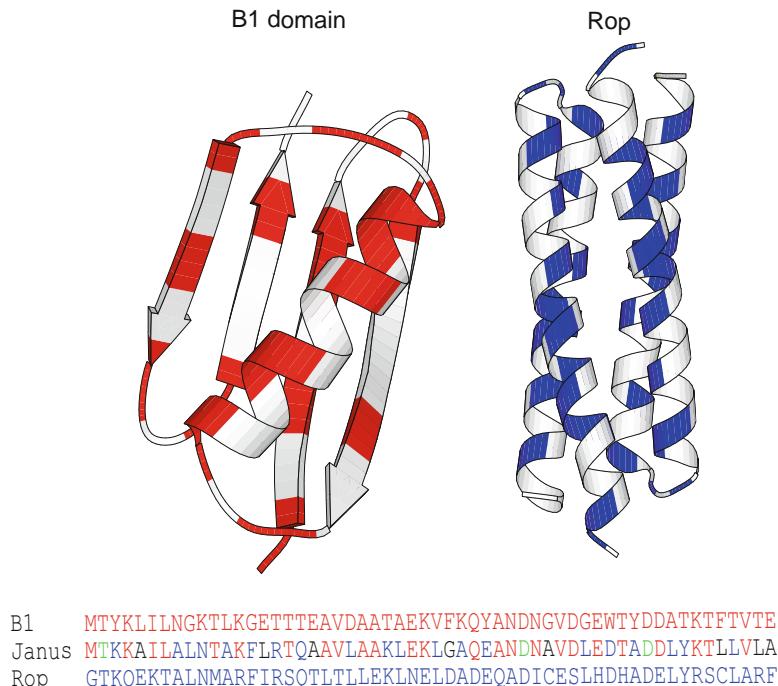


Figure 2.2. Ribbon representations of the B1 domain of IgG-binding protein G and the Rop monomer (first 56 residues), which Janus resembles [281], with corresponding sequences. Half of the protein G  $\beta$  domain (B1) residues were changed to produce Janus in response to the Paracelsus challenge (see Box 2.1 and [1066]). The origin of the residues is indicated by the following color schemes: residues from B1: red; residues from Rop: blue; residues in both: green; residues in neither: black. While experimental coordinates of protein G and Rop are known, the structure of Janus was deduced by modeling. The single-letter amino acid acronyms are detailed in Table 3.1, Chapter 3.

concentrations (e.g., [707, 1430]); NMR spectroscopic experiments that monitor protein folding intermediates (e.g., [345, 948]); predictions of secondary and/or tertiary structure on the basis of evolutionary information [1086]; and statistical mechanical theories.

Such studies suggest that while wide variations in folding pathways may occur, there exists in general a unifying pattern for the evolution of native-structure contacts, which are encoded in the amino acid sequence of the protein [362]. Thus, while independent pathways can result from occasional misfolding errors that can block certain pathway points and affect intermediates, protein folding is generally an ordered process based on native-like *foldon units* — cooperative structural units of the native protein — and intermediates. Protein folding landscapes and the different views can therefore be reconciled and interpreted in terms of the combined factors of cooperativity of these structural units, their stepwise stabilization, and the chance occurrence of folding errors.

Interestingly, a recent experimental work focusing on protein folding and unfolding kinetics [1362] confirmed long-standing theoretical hypotheses that protein landscape roughness causes slow folding. Wensley et al. probed the reasons for different folding profiles of the protein domains of  $\alpha$ -spectrin, a protein of the intracellular matrix of red blood cells important for membrane elasticity. Prior experimental studies have shown that the R15 domain folds very quickly, roughly 3000 times faster than its homologues R16 and R17. Because the structures are similar, the reason for this behavior was not apparent. By swapping domains through chimeric constructs, the researchers demonstrated that protein landscape roughness, or internal friction resulting from residue-specific interactions that can lead to mis docking of helices, causes slow folding and unfolding of the R16 and R17 domains of  $\alpha$ -spectrin. For this membrane protein, slow unfolding kinetics are advantageous because they imply fewer rearrangements during the cell's lifetime; this, in turn, decreases potential degradation. Thus, besides producing important experimental evidence and insights into folding/unfolding pathways, this work underscores the value of theory in understanding biomolecular structures, dynamics, and pathways. Biomolecular modeling is well on its way to becoming a full partner to experiment and a field on its own right [1464].

### 2.2.2 *Folding Challenges*

The great progress in the field can also be seen by evaluations of the highly successful biannual prediction exercises (termed CASP for Critical Assessment of Techniques for Protein Structure Prediction) and associated meetings conducted since 1994. See [predictioncenter.org/](http://predictioncenter.org/) for the latest meeting developments, including detailed Proceedings, such as [683, 880, 1270]. Though these events have become high profile endeavors for researchers in the field because success in CASP leads to great recognition, the scientific lessons learned year to year and over time have been invaluable to the protein community. From participation of 35 groups in CASP1 (1994), the number has increased to several hundred.

The goals of CASP are to assess capabilities and limitations of current protein structure prediction, highlight promising areas, pinpoint specific difficulties, and thereby stimulate progress in the field. Specifically, the CASP organizers assign certain proteins for theoretical prediction that protein crystallographers and NMR spectroscopists expect to complete by the next CASP meeting. Prediction assessors then consider several categories of structural prediction tools, for example: template based modeling for tertiary structure prediction; template free modeling for tertiary structure prediction; side chain, loop, and active-site prediction for high resolution models; high accuracy modeling; disordered protein-region identification; domain-boundary identification; function prediction; and more. Evaluators assess how well various *in silico* approaches such as comparative (homology) modeling or *ab initio* prediction (i.e., using first principles) perform.

The meetings are important not only for motivating progress in protein prediction but also for revealing important trends concerning the strategies that work well and those that may not be as promising. In particular, the meetings

demonstrated that comparative modeling approaches can produce reasonably good structural models, with notably more accurate predictions becoming possible, but that it is still difficult to predict the structure of regions that are substantially different from the target. For example, when the quality of the prediction is characterized in terms of  $C^\alpha$  root-mean-square (RMS) deviations, the best values obtained — in the lower part of the range of 2–6 Å — are from the best comparative modeling approaches.

It has also become evident that by combining information from two or more templates and by following homology modeling by clever all-atom refinement, prediction accuracy and quality can be enhanced; all-atom refinement, in particular, has been a stumbling block for a long time, so it is gratifying to finally see progress in this area. Furthermore, the accuracy of models predicted by automatic servers is approaching that of manual manipulation, lending promise to the notion that ultimately every interested individual might be able to automate such protein folding predictions on her/his desktop. Automation and rapid folding simulations can make possible applications to enzyme design, such as done with the *Rosetta* program; see [609], for example. Still, recognizing entirely novel folds remains a challenge, as well as predicting secondary structures and long-range contacts. Template-free or *ab initio* modeling has shown more modest general improvements, pointing to needed new ideas; nonetheless it is encouraging that such methods can occasionally perform very well on some targets.

Modeling work in the field is invaluable because it teaches us to ask, and seek answers to, systematic questions about sequence/structure/function relationships and about the underlying forces that stabilize biomolecular structures, especially when using *ab initio* methods. Still, given the rapid improvements in the experimental arena, the pace at which modeling predictions improve must be expeditious to make a significant contribution to protein structure prediction from sequence.

### 2.2.3 *Folding by Dynamics Simulations?*

While molecular dynamics simulations are beginning to approach the timescales necessary to fold small peptides [285] or small proteins [338], we are far from finding the Holy Grail, if there is one [121]. Indeed, ambitious goals declared in the late 1990s like IBM’s desire to fold proteins with the ‘Blue Gene’ *petaflop* computer (i.e., capable of  $10^{15}$  floating-point operations per second) depends on the computational models guiding this ubiquitous cellular process. One of the major computational initiatives to come in its wake is ‘Blue Waters’ by the University of Illinois, NCSA, IBM, and their partners who will launch a petascale computing system in 2011 (see [ncsa.uiuc.edu](http://ncsa.uiuc.edu)). This effort will likely bring unprecedented computing power that could be exploited to tackle numerous computation-intensive applications involving large and complex systems like biomolecular dynamics; planetary, star and galaxy motion; economical modeling; cyberinfrastructure networks; and more.

For protein folding applications, computational power alone may hardly be sufficient; the well recognized force field approximation remains an issue, as does the need to account for all key factors that dictate folding *in vivo*. For example, some proteins require active escorts to assist in their folding *in vivo*. These *chaperone* molecules assist in the folding and rescue misbehaved polymers. Though many details are not known about the mechanisms of chaperone assistance (see below), we recognize that chaperones help by guiding structure assembly and preventing aggregation of misfolded proteins. For an overview of chaperones, see [519, 762, 1326], for example.

---

### **Box 2.1: Paracelsus Challenge**

In 1994, George Rose and Trevor Creamer posed a challenge, named after a 16th-century alchemist: change the sequence of a protein by 50% or less to create an entirely different 3D global folding pattern [1066]. Though this challenge might sound not particularly difficult, imagine altering at most 50% of the ingredients for a chocolate cake recipe so as to produce bouillabaisse instead! Rose and Creamer offered a reward of \$1000 to entice entrants.

The transmutation was accomplished four years later by Lynne Regan and coworkers [281], who converted the four-stranded  $\beta$ -sheet B1 domain of protein G — which has the  $\beta$  sheets packed against a single helix — into a four-helix bundle of two associating helices called Janus (see Figure 2.2). These contestants achieved this wizardry by replacing residues in a  $\beta$ -sheet-encoding domain (i.e., those with high  $\beta$ -sheet-forming propensities) with those corresponding to the four-helix-bundle protein Rop (repressor of primer). Other modifications were guided by features necessary for Rop stability (i.e., internal salt bridge), and the combined design was guided by energy minimization and secondary-structure prediction algorithms.

The challenge proposers, though delighted at the achievement they stimulated, concluded that in the future only t-shirt prizes should be offered rather than cash!

---

#### **2.2.4 Folding Assistants**

Current studies on chaperone-assisted folding, especially of the archetypal chaperone duo, the *E. coli* bacterial chaperonin GroEL and its cofactor GroES, are providing insights into the process of protein folding [387, 762, 1257] (see Figure 2.3 and Box 2.2). The rescue acts of chaperones depend on the subclass of these escorts and the nature of the protein being aided. Some chaperones can assist a large family of protein substrates, while others are more restrictive (see Box 2.2); detailed structural explanations remain unclear. Many families of chaperones are also known, varying in size from small monomers (e.g., 40 or 70 kDa for DnaJ and DnaK of Hsp70) to large protein assemblies (e.g., 810 kDa for GroEL or 880 kDa for the GroEL/GroES complex).

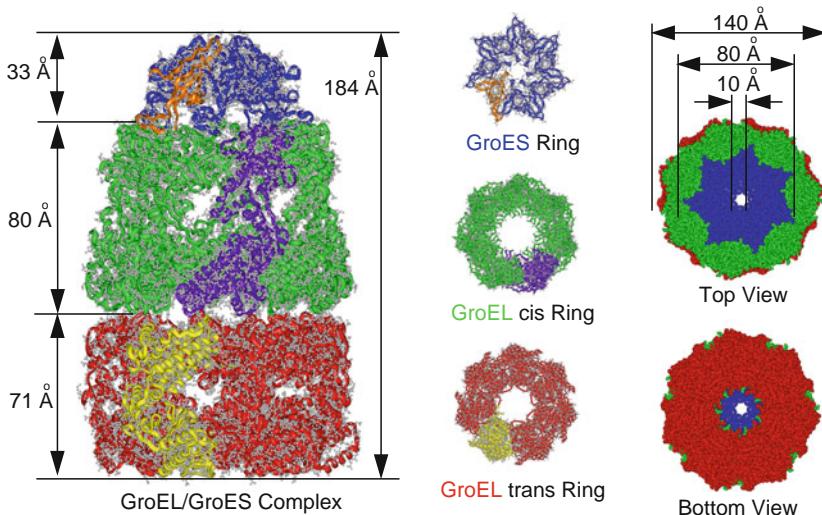


Figure 2.3. The bullet-shaped architecture of the GroEL/GroES chaperonin/co-chaperonin complex sequence. Overall assembly and dimensions are shown from a side view (left). The top ring is the GroES ‘cap’, and the other layers are GroEL rings. Sidechains are shown in grey. As seen from the top and bottom views (right), a central channel forms in the interior, conducive to protein folding. The protein is organized as three rings that share a 7-fold rotational axis of symmetry (middle), where GroEL contains 14 identical protein subunits assembled in two heptameric rings, and GroES contains 7 smaller identical subunits in its heptamer ring.

The small assistants bind to short runs of hydrophobic residues<sup>2</sup> to delay premature folding and prevent aggregation. Larger chaperones are likely needed to prevent aggregation of folded compact intermediates in the cell termed ‘molten globules’, requiring a complex trap-like mechanism involving co-chaperones (see also Box 2.2).

Such protein aggregation can occur due to even minor changes in intracellular physiochemical conditions, such as temperature and pressure. Chaperones can rescue active proteins from forming these disrupting aggregates by isolating, unfolding, and translocating them as needed. Together with the cellular machinery for removing damaged proteins, the work of chaperones maintains the pool of active proteins critical to an organism’s life. Misfolded proteins can be the root cause of many debilitating human disorders like Alzheimer’s Disease and Cystic Fibrosis (see separate section). Studies of misfolding are helping to investigate these complex phenomena (e.g., [637]).

---

<sup>2</sup>The terms *hydrophobic* (‘water-hating’) and *hydrophilic* (‘water-loving’) characterize water-insoluble and water-soluble molecular groups, respectively.

**Box 2.2: Studies on Protein Escorts**

The archetypal chaperone GroEL is a member of a chaperone class termed *chaperonins*; hsp60 of mitochondria and chloroplasts is another member of this class. These chaperones bind to partially-folded peptide chains and assist in the folding with the consumption of ATP. The solved crystal structure of GroEL/GroES [1405] suggests beautifully, in broad terms, how the large central channel inside a barrel-shaped chaperone might guide protein compaction in its container and monitor incorrect folding by diminishing aggregation (see Figure 2.3). The two-ringed GroEL chaperonin (middle and bottom levels in Figure 2.3) attaches to its partner chaperonin GroES (top ring) upon ATP binding, causing a major conformational change; the size of GroEL nearly doubles, and it assumes a cage shape, with GroES capping over it. This capping prevents the diffusion of partially folded compact intermediates termed ‘molten globules’ and offers them another chance at folding correctly.

Experiments that track hydrogen exchange in unfolded rubisco protein by radioactive tritium (a hydrogen isotope) labeling suggest how misfolded proteins fall into this cavity and are released: a mechanical stretching force triggered by ATP binding partially or totally unfolds the misfolded proteins, eventually releasing the captive protein [1180].

These results also support an *iterative annealing* (or *network* model) for chaperone-guided folding, in which the process of forceful unfolding of misfolded molecules, their trapping in the cavity, and their subsequent release is iterated upon until proper folding.

The identification of preferential substrates for GroEL *in vivo* [570], namely multidomain *E. coli* proteins with complex  $\alpha\beta$  folds, further explains the high-affinity interactions formed between the misfolded or partially folded proteins and binding domains of GroEL. These proteins require the assistance of a chaperone because assembly of  $\beta$ -sheet domains requires long-range coordination of specific contacts (not the case for formation of  $\alpha$ -helices). Natural substrates for other chaperones, like Eukaryotic type II chaperonin CCT, also appear selective, favoring assistance to proteins like actin [782].

However, such insights into folding kinetics are only the tip of the iceberg. Chaperone types and mechanisms vary greatly, and the effects of macromolecular crowding (not modeled by *in vitro* experiments) complicate interpretations of folding mechanisms *in vivo*. Unlike chaperonins, members of another class of chaperones that includes the heat-shock protein Hsp70 bind to exposed hydrophobic regions of newly-synthesized proteins and short linear peptides, reducing the likelihood of aggregation or denaturation. These are classified as ‘stress proteins’ since their amount increases as environmental stresses increase (e.g., elevated temperatures) [762]. Other chaperones are known to assist in protein translocation across membranes.

---

### 2.2.5 Unstructured Proteins

Though our discussion has focused on the concept of native folds, not all proteins are intrinsically structured [346]. The intrinsic lack of structure can be advantageous, for example in binding versatility to different targets or in ability to

adapt different conformations. Unfolded or non-globular structures are recognized in connection with regulatory functions, such as binding of protein domains to specific cellular targets [346, 1391]. Examples include DNA and RNA-binding regions of certain protein complexes (e.g., basic region of leucine zipper protein GCN4, DNA-binding domain of NFATC1, RNA recognition regions of the HIV-1 Rev protein). Here, the unstructured regions become organized only upon binding to the DNA or RNA target. This folding flexibility offers an evolutionary advantage, which might be more fully appreciated in the future, as more gene sequences that code for unstructured proteins are discovered and analyzed.

## 2.3 Protein Misfolding – A Conundrum

### 2.3.1 Prions and Mad Cows

Further clues into the protein folding enigma are also emerging from another puzzling discovery involving certain proteins termed *prions*. These misfolded proteins — triggered by a conformational change rather than a sequence mutation — appear to be the source of infectious agent in fatal neurodegenerative diseases like bovine spongiform encephalopathy (BSE) or ‘mad cow disease’ (identified in the mid 1980s in Britain), and the human equivalent Creutzfeld-Jacob disease (CJD).<sup>3</sup> The precise mechanism of protein-misfolding induced diseases is not known, but connections to neurodegenerative diseases, which include Alzheimer’s, are growing and stimulating much interest in protein misfolding [251, 325, 326, 791].

Stanley Prusiner, a neurology professor at the University of California at San Francisco, coined the term prion to emphasize the infectious source as the protein (‘proteinaceous’), apparently in contradiction to the general notion that nucleic acids must be transferred to reproduce infectious agents. Prusiner won the 1998 Nobel Prize in Physiology or Medicine for this “*pioneering discovery of an entirely new genre of disease-causing agents and the elucidation of the underlying principles of their mode of action*”.

Prions add a new symmetry to the traditional roles long delegated to nucleic acids and proteins! Since the finding in the 1980s that nucleic acids (catalytic RNAs) can *catalyze* reactions — a function traditionally attributed to proteins only — the possibility that certain proteins, prions, *carry genetic instructions* — a role traditionally attributed to nucleic acids — completes the duality of functions to both classes of macromolecules.

### 2.3.2 Infectious Protein?

Is it possible for an ailment to be transmitted by ‘infectious proteins’ rather than viruses or other traditional infectious agents? The prion interpretation for the

---

<sup>3</sup>See information from the UK Department of Health on [www.doh.gov.uk](http://www.doh.gov.uk), the UK CJD Surveillance Unit at [www.cjd.ed.ac.uk](http://www.cjd.ed.ac.uk), and the CJD Disease Foundation at [cjdfoundation.org](http://cjdfoundation.org).

infection mechanism remains controversial for lack of clear molecular explanation. In fact, one editorial article stated that “*whenever prions are involved, more open questions than answers are available*” [9]. Yet the theory is winning more converts with laboratory evidence that an infectious protein that causes mad cow disease also causes a CJD variant in mice [1151]. These results are somewhat frightening because they suggest that the spread of this illness from one species to another is easier than has been observed for other diseases.

The proteinaceous theory suggests that the prion protein (see Figure 2.4) in the most studied neurodegenerative prion affliction, *scrapie* (long known in sheep and goats), becomes a pathologic agent upon conversion of one or more of its  $\alpha$ -helical regions into  $\beta$ -regions (e.g., parallel  $\beta$ -helix [1371]); once this conformational change occurs, the conversion of other cellular neighbors proceeds by a domino-like mechanism, resulting in many abnormally-folded molecules which eventually reap havoc in the mammal. This protein-only hypothesis was first formulated by J.S. Griffith in 1967, but Prusiner first purified the hypothetical abnormal protein thought to cause BSE. New clues are rapidly being added to this intriguing phenomenon (see Box 2.3).

Both the BSE and CJD anomalies implicated with prions have been linked to unusual deposits of protein aggregates in the brain. (Recent studies on mice also open the possibility that aberrant proteins might also accumulate in muscle tissue). It is believed that a variant of CJD has caused the death of dozens of people in Britain (and a handful in other parts of the world) since 1995 who ate meat infected with BSE, some only teenagers. Recent studies also suggest that deaths from the human form of mad cow disease could be rising significantly and spreading within Europe as well as to other continents.

Since the incubation period of the infection is not known — one victim became a vegetarian 20 years before dying of the disease — scientists worry about the extent of the epidemic in the years to come. The consequences of these deaths have been disastrous to the British beef industry and have led indirectly to other problems (e.g., the 2001 outbreak of foot-and-mouth disease, a highly infectious disease of most farm animals except horses). The panic has not subsided, as uncertainties appear to remain regarding the safety of various beef parts, as well as sheep meat, and the possible spread of the disease to other parts of the world.

### 2.3.3 Other Possibilities

Many details of this intriguing prion hypothesis and its associated diseases are yet to be discovered and related to normal protein folding. Some scientists believe that a lurking virus or virino (small nonprotein-encoding virus) may be involved in the process, perhaps stimulating the conformational change of the prion protein, but no such evidence has yet been found. Only creation of an infection *de novo* in the test tube is likely to convince the skeptics, but the highly unusual molecular transformation implicated with prion infection is very difficult to reproduce in the test tube.

---

**Box 2.3: Prion: Structural Evidence**

The detailed structural picture associated with the prion conformational change is only beginning to emerge as new data appear [10]. In 1997, Kurt Wüthrich and colleagues at the Swiss Federal Institute of Technology in Zurich reported the first NMR solution structure of the 208-amino acid glycoprotein “prion protein cellular” PrP<sup>C</sup> anchored to nerve cell membranes. The structure reveals a flexibly disordered assembly of helices and sheets (see Fig. 2.4). This organization of the harmless protein might help explain the conversion process to its evil isoform PrP<sup>Sc</sup>. It has been suggested that chaperone molecules may bind to PrP<sup>C</sup> and drive its conversion to PrP<sup>Sc</sup> and that certain membrane proteins may also be involved in the transformation.

In early 1998, a team from the University of California at San Francisco discovered a type of prion, different from that associated with mad cow disease, that attaches to a major structure in neuron cells and causes cells to die by transmitting an abnormal signal. This behavior was observed in laboratory rats who quickly died when a mutated type of prion was placed into the brains of newborn animals; their brains revealed the abnormal prions stuck within an internal membrane of neuron cells. The researchers believe that this mechanism is the heart of some prion diseases. They have also found such abnormal prions in the brain tissue of patients who died from a rare brain disorder called Gerstmann-Straussler-Scheinker disease (GSS) — similar to Creutzfeld-Jacob disease (CJD) — that destroys the brain.

Important clues to the structural conversion process associated with prion diseases were further offered in 1999, when a related team at UCSF, reported the NMR structure of the core segment of a prion protein rPrP that is associated with the scrapie prion protein PrP<sup>Sc</sup> [602, 777]. The researchers found that part of the prion protein exhibits multiple conformations. Specifically, an intramolecular hydrogen bond linking crucial parts of the protein can be disrupted by a single amino acid mutation, leading to different conformations. This compelling evidence on how the molecule is changed to become infectious might suggest how to produce scrapie-resistant or BSE-resistant species by animal cloning.

Prion views from several organisms (including human and cow) have been obtained [1429], allowing analyses of species variations, folding, and misfolding relationships; see [1371], for example. This high degree of similarity across species is shown in Figure 2.4.

Still, until prions are demonstrated to be infectious *in vivo*, the proteinaceous hypothesis warrants reservation. Clues into how prions work may emerge from parallel work on yeast prions, which unlike their mammalian counterparts do not kill the organism but produce transmitted heritable changes in phenotype; many biochemical and engineering studies are underway to explore the underlying mechanism of prion inheritance.

---

### 2.3.4 Other Misfolding Processes

There are other examples of protein misfolding diseases (e.g., references cited in [325, 326, 505, 791]). The family of amyloid diseases includes Alzheimer’s,

Parkinson's, and type II (late-onset) diabetes. For example, familial amyloid polyneuropathy is a heritable condition caused by the misfolding of the protein transthyretin. The amyloid deposits that result interfere with normal nerve and muscle function.

Dobson [325] intriguingly suggests that understanding the evolution of proteins holds the key to protein misfolding diseases. Namely, he argues that since evolutionary processes have selected sequences of amino acids that form close-packed, globular proteins, the effectively irreversible formation of amyloid fibrils reflects a conversion of proteins to their 'primordial' rather than evolved states, possibly from aging-induced mutations that destabilize native proteins.

Indeed, many protein misfolding diseases are strongly associated with aging, suggesting that the cell's ability to monitor misfolding and prevent aggregation deteriorate with age. Fortunately, recent biophysical and computational techniques are leading to an increased understanding of what triggers protein misfolding and what the intrinsic and extrinsic factors that contribute to the process *in vivo*, though we are far from rational design of therapeutic intervention [791]. Computational models of protein misfolding, in particular, can help relate systematically changes in temperature-dependent pathways and aggregation to observed phenomena.

As in mad cow disease, a molecular understanding of the misfolding process may lead to treatments of the disorders. In the case of familial amyloid polyneuropathy, research has shown that incorporating certain mutant monomers in the tetramer protein transthyretin reduces considerably the formation of amyloid deposits (amyloid fibrils); moreover, incorporating additional mutant monomers can prevent misfolding entirely [505]. These findings suggest potential therapeutic strategies for amyloid and related misfolding disorders. See also [983] for a pharmacological approach for treating human amyloid diseases by using a small-molecule drug that targets a protein present in amyloid deposits; the drug links two pentamers of that protein and leads to its rapid clearance by the liver.

Studies also suggest that misfolded proteins generated in the pathway of protein folding can be dangerous to the cell and cause harm (whether or not they convert normal chains into misfolded structures, as in prion diseases) [183, 1325]. The cellular mechanisms associated with such misfolded forms and aggregates are actively being pursued, including by modeling [637].

### 2.3.5 Deducing Function From Structure

Having the sequence and also the 3D structure at atomic resolution, while extremely valuable, is only the beginning of understanding biological function. How does a complex biomolecule accommodate its varied functions and interactions with other molecular systems? How sensitive is the 3D architecture of a biopolymer to its constituents?

Despite the fact that in many situations protein *structures* are remarkably stable to tinkering (mutations), their *functional* properties can be quite fragile. In other words, while a protein often finds ways to accommodate substitutions of a few

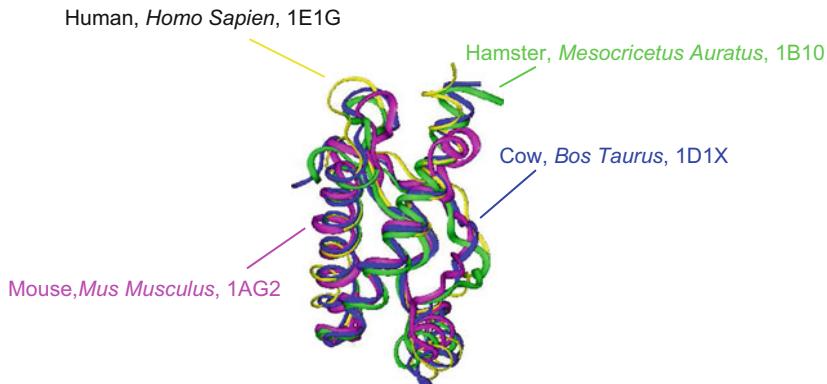


Figure 2.4. Structure of the prion protein.

amino acids so as not to form an entirely different overall folding motif [205], even the most minute sequence changes can alter biological activity significantly. Mutations can also influence the *kinetics* of the folding pathway.

An example of functional sensitivity to sequence is the altered transcriptional activity of various protein/DNA complexes that involve single base changes in the TATA-box recognition element and/or single protein mutations in TBP (TATA-Box binding protein) [971]. For example, changing just a single residue in the common nucleotide sequence of TATA-box element, TATAAAAG, to TAAAAAAAG impairs binding to TBP and hence disables transcriptional activity.

In principle, theoretical approaches should be able to explain these relations between sequence and structure from elementary physical laws and knowledge of basic chemical interactions. In practice, we are encountering immense difficulty pinpointing what Nature does so well. After all, the notorious “*protein folding*” problem is a challenge to us, not to Nature.

Much work continues on this active front.

## 2.4 From Basic to Applied Research

An introductory chapter on biomolecular structure and modeling is aptly concluded with a description of the many important practical applications of the field, from food chemistry to material science to drug design. A historical perspective on drug design is given in Chapter 15. Here, we focus on the current status of drug development as well as other applied research areas that depend strongly on progress in molecular modeling. Namely, as biological structures and functions are being resolved, natural disease targets that affect the course of disease can be proposed. Such new treatments can be approached both from the traditional drug design model which seeks inhibitors to specific targets (e.g., reviewed in [453, 913, 1193, 1447]) or from a systems biology approach which

attempts to modify response of genes, proteins, and metabolites by integrating organ and system-level modeling [191, 278, 649]. Other biological and polymer targets, such as the ripening genes of vegetables and fruit or strong materials, can also be manipulated to yield benefits to health, technology, and industry.

### 2.4.1 *Rational Drug Design: Overview*

The concept of systematic drug design, rather than synthesis of compounds that mimic certain desired properties, is only about 50 years old (see Chapter 15). Gertrude Elion and George Hitchings of Burroughs Wellcome, who won the 1988 Nobel Prize in Physiology or Medicine, pioneered the field by creating analogues of the natural DNA bases in an attempt to disrupt normal DNA synthesis. Their strategies eventually led to a series of drugs based on modified nucleic-acid bases targeted to cancer cells. Today, huge compound libraries are available for systematic screening by various combinatorial techniques, robotics, other automated technologies, and various modeling and simulation protocols (see Chapter 15).

Rational pharmaceutical design has now become a lucrative enterprise. The sales volume for the world's best seller prescription drug in 1999, *Prilosec* (for ulcer and heartburn), exceeded six billion dollars. A vivid description of the climate in the pharmaceutical industry and on Wall Street can be found in *The Billion-Dollar Molecule: One Company's Quest for the Perfect Drug* [1363]. This thriller describes the racy story of a new biotech firm for drugs to suppress the immune system, specifically the discovery of an alternative treatment to *Cyclosporin*, medication given to transplant patients. Since many patients cannot tolerate cyclosporin, an alternative drug is often needed.

Tremendous successes in 1998, like Pfizer's anti-impotence drug *Viagra* and Entre-Med's drugs that reportedly eradicated tumors in mice, have generated much excitement and driven sales and earnings growth for drug producers. A glance at the names of biotechnology firms is an amusing indicator of the hope and prospects of drug research: Biogen, Cor Therapeutics, Genetech, Genzyme, Immunex, Interneuron Pharmaceuticals, Liposome Co., Millennium Pharmaceuticals, Myriad Genetics, NeXstar Pharmaceuticals, Regeneron Pharmaceuticals, to name a few. Other success stories involve a small-molecule inhibitor of the SARS virus [329], glutamate nanosensors to monitor neurologic functions whose malfunction can lead to neurodegenerative disorders [933], and agonists to treat anxiety and depression [111]. Yet, both the monetary cost and development time required for each successful drug remains very high [39, 160], and great successes are now few and far between; see end of chapter for further discussion.

### 2.4.2 *A Classic Success Story: AIDS Therapy*

#### HIV Enzymes

A spectacular example of drugs made famous through molecular modeling successes are inhibitors of the two viral enzymes *HIV protease* (HIV: human

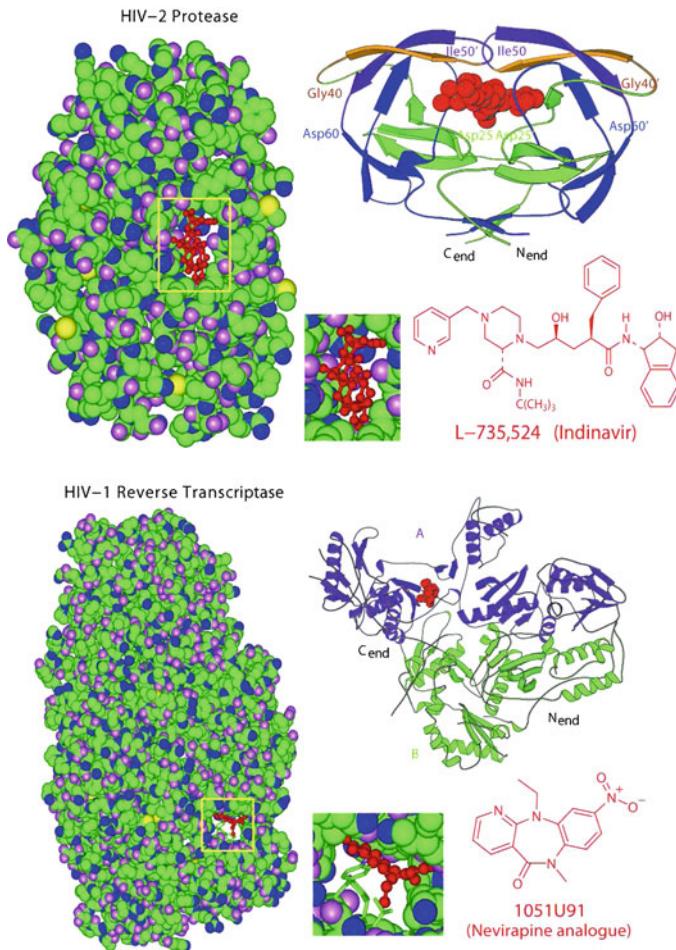


Figure 2.5. Examples of AIDS drug targets — the HIV protease inhibitor and reverse transcriptase (RT) — with corresponding designed drugs. The protease inhibitor *Indinavir* (crizivian) binds tightly to a critical area of the dimer protease enzyme (HIV-2, 198 residues total shown here [227]), near the flaps (residues 40 to 60 of each monomer), inducing a conformational change (flap closing) that hinders enzyme replication; intimate interactions between the ligand and enzyme are observed in residues 25 and 50 in each protease monomer. The non-nucleoside RT inhibitor 1051U91 (a nevirapine analogue), approved for use in combination with nucleoside analogue anti-HIV drugs like AZT, binds to a location near the active site of RT that does not directly compete with the oligonucleotide substrate. The large RT protein of 1000 residues contains two subdomains (A and B).

immunodeficiency virus) and *reverse transcriptase* for treating AIDS, acquired immune deficiency syndrome.

First hints of AIDS were reported in the summer of 1981, in clusters of gay men in large American cities; these groups exhibited severe symptoms of infection

by certain pneumonia combined with those from Kaposi's sarcoma (KS) cancer. Now considered among the most catastrophic pandemics to strike humankind, this infectious disease is caused by an insidious retrovirus. (See perspectives on the evolution of this pandemic, including treatment and prevention in [253,611,1221] and a personal reflection by Robert Gallo who was instrumental in identifying the retrovirus culprit [434]). Such a virus can convert its RNA genome into DNA, incorporate this DNA into the host cell genome, and then spread from cell to cell. To invade the host, the viral membrane of HIV must attach and fuse with the victim's cell membrane; once entered, the viral enzymes reverse transcriptase and integrase transform HIV's RNA into DNA and integrate the DNA into that of the host [529].

Current drugs inhibit enzymes that are key to the life cycle of the AIDS virus (see Figure 2.5). **Protease inhibitors** like *Indinavir*, *Saquinavir*, *Ritonavir*, *Nelfinavir*, etc. block the activity of proteases, protein-cutting enzymes that help a virus mature, reproduce, and become infectious [227]. **Reverse transcriptase** (RT) inhibitors block the action of an enzyme required by HIV to make DNA from its RNA [1045]. However, eradication of the disease by a preventive HIV vaccine has so far been largely unsuccessful due to the complex biology and life cycle of the virus [91, 253, 375, 611, 612]. Still, existing medications can give AIDS patients a life.

### AIDS Drug Development

One of the most commonly used drug cocktails is the triplet drug combination of a protease inhibitor like indinavir with the two nucleoside analogues like AZT (*Zidovudine*, or 3'-azido-3'-deoxythymidine) and 3TC. Another commonly prescribed regimen utilizes two nucleoside analogues and one non-nucleoside RT inhibitors (see below). More than one drug is needed because mutations in the HIV enzymes can confer drug resistance; thus, acting on different sites as well as on different HIV proteins increases effectiveness of the therapy.

The two types of RT blocker mentioned above are *nucleoside analogues* and *non-nucleoside* inhibitors. Members of the former group (*Zidovudine* or AZT, *Didanosine*, *Zalcitabine*, *Stavudine*, etc.) interfere with the HIV activity by replacing a building block used to make DNA from the HIV RNA virus with an inactive analog and thereby prevent accurate decoding of the viral RNA. Non-nucleoside RT inhibitors (e.g., *Nevirapine*, *Delavirdine*, and *Efavirenz*) are designed to bind with high affinity to the active site of reverse transcriptase and therefore physically interfere with the enzyme's action.

Design of such drugs was made possible in part by molecular modeling due to the structure determination of the HIV protease by X-ray crystallography in 1989 and RT a few years later [1218]. Figure 2.5 shows molecular views of these HIV enzymes complexed with drugs.

Besides the HIV protease and reverse transcriptase, a third target is the HIV integrase, which catalyzes the integration of a DNA copy of the viral genome into the host cell chromosomes. Scientists at Merck identified several years ago

1,3-diketo acid integrase inhibitors that block strand transfer, one of the two specific catalytic functions of HIV-1 integrase [536]; this function has not been affected by previous inhibitors. This finding paved the way for developing effective integrase inhibitors: *Raltegravir* was approved in this class in 2007.

### AIDS Drug Limitations

Much progress has been made in this area since the first report of the rational design of such inhibitors in 1990 [1054] (see [253, 434, 611, 612, 1221] for reviews). In fact, the dramatic decline of AIDS-related deaths by such drug cocktails can be attributed in large part to these new generation of designer drugs (see Box 2.4) since the first introduction of protease inhibitors in 1996. Indeed, the available triplet drug cocktails, of protease inhibitors and nucleoside analogues RT inhibitors, or nucleoside analogues and non-nucleoside RT inhibitors, have been shown to virtually suppress HIV, making AIDS a manageable disease.

However, the cocktails are not a cure. The virus returns once patients stop the treatment, and the enormous genetic diversity of mutations that occur enable HIV to reduce the effectiveness of treatment. Indeed, in very heavily treated patients, as many as one quarter of the amino acids (25 out of 99) of the viral protein HIV-1 protease can be mutated, but the enzyme continues to function. Moreover, the window of opportunity for the immune system to clear the initial infection is very narrow, because the virus quickly integrates itself into the host. The mechanisms of drug resistant mutations and the interactions among them are still not well understood despite enormous amount of research, and fundamental questions about the progression of HIV disease and the host response to the virus remain unanswered [91, 612].

In addition, few countries in the developing world, like Africa, can afford the virus suppressing drugs; the drug-cocktail regimen is complex, requiring many daily pills taken at multiple times and separated from eating, most likely for life; serious side effects also occur. For example, we now know that nucleoside analogues inhibit a variety of DNA polymerization reactions, in addition to those of the HIV-1 RT, and are thus associated with serious side effects.

In certain parts of the world, the situation is profoundly distressing: the life expectancy of patients living with HIV/AIDS in many African countries has fallen to 40 years of age today, a drastic difference from the age in the pre-AIDS era, and the number continues to drop.<sup>4</sup> Though in the developed world, AIDS is no longer a death sentence, the incidence of new infections is alarming in certain urban areas. For example, the U.S. Center for Disease Control and Prevention reported in late August 2008 that HIV is spreading in New York City at three times the national rate (72 versus 23 new cases per 100,000 people).

---

<sup>4</sup>In Swaziland, which has one of the worst rates of HIV infection in the world, life expectancy has fallen from 60 years in 1997 to less than half of that in 2008.

### Lurking Virus

As mentioned, even available treatments cannot restore the damage to the patient's immune system; the number of T-cell (white blood cells), which HIV attaches itself to, is still lower than normal (which lowers the body's defenses against infections), and there remain infected immune cells that the drugs cannot reach because of integration. Thus, new drugs are being sought to interrupt the first step in the viral life cycle — binding to a co-receptor on the cell surface to rid the body of the cell's latent reservoirs of the HIV virus, to chase the virus out of cells where it hides for subsequent treatment, or to drastically reduce the HIV reservoir so that the natural immune defenses can be effective. New structural and mechanistic targets are currently being explored (see Box 2.4). Some of the newest drugs under development include low-cost microbicidal drugs which can be topically applied prior to sexual contact to prevent, or directly destroy HIV [84].

A better understanding of the immune-system mechanism associated with AIDS, for example, may help explain how to prime the immune system to recognize an invading AIDS virus. Unlike traditional AIDS drug cocktails which inhibit division of already infected cells, fusion (or entry) inhibitors define another class of drugs that seek to prevent HIV from entering the cell membrane. This entry, called fusion, releases the virus's genetic material and allows it to replicate. The promising drug T-20 or *Enfuvirtide* (which must be injected into the skin) is a member of fusion-inhibitor or entry-inhibitor drugs that, when added to a combination of standard drugs, can significantly reduce HIV levels in the blood. Another such entry-inhibitor is *Maraviroc*, a CCR5 antagonist (see Box 2.4).

As manifested by its complex components of invasion that include the fusion apparatus, the AIDS virus has developed a complex, tricky, and multicomponent-protection infection machinery, as well as drug-resistant defense.

Besides integrase and fusion inhibitors, among the newer drugs to fight AIDS being developed are immune stimulators and antisense drugs. The former stimulate the body's natural immune response, and the latter mimic the HIV genetic code and prevent the virus from functioning.

---

#### Box 2.4: Fighting AIDS

AIDS drugs attributed to the success of molecular modeling include *AZT* (Zidovudine) sold by Bristol-Myers Squibb, and the newer drugs *Viracept* (Nelfinavir) made by Agouron Pharmaceuticals, *Crixivan* (Indinavir) by Merck & Company, and *Amprenavir* discovered at Vertex Pharmaceuticals Inc. and manufactured by Glaxo Wellcome. Amprenavir, in particular, approved by the U.S. Government in April 1999, is thought to cross the 'blood-brain barrier' so that it can attack viruses that lurk in the brain, where the virus can hide. This general class of inhibitors has advanced so rapidly that drug-resistant AIDS viruses have been observed.

Structural investigations are probing the structural basis for the resistance mechanisms, which remain mysterious, particularly in the case of nucleoside analogue RT inhibitors

like AZT [700]. The solved complex of HIV-1 reverse transcriptase [576] offers intriguing insights into the conformational changes associated with the altered viruses that influence the binding or reactivity of inhibitors like AZT and also suggests how to construct drug analogues that might impede viral resistance.

Basic research on the virus's process of invading host cells — by latching onto receptors (e.g., the CD4 glycoprotein, which interacts with the viral envelope glycoprotein, gp120, and the transmembrane component glycoprotein, gp41), and co-receptors (e.g., CCR5 and CXCR4) — may also offer treatments, since developments of disease intervention and vaccination are strongly aided by an understanding of the complex entry of HIV into cells; see [687] for example.

The HIV virus uses a spear-like agent on the virus's protein coat to puncture the membrane of the cells which it invades; vaccines might be designed to shut the chemical mechanism or stimuli that activate this invading harpoon of the surface protein. The solved structure of a subunit of gp41, for example, has been exploited to design peptide inhibitors that disrupt the ability of gp41 to contact the cell membrane [393]. A correlation has been noted, for example, between co-receptor adaptation and disease progression.

Novel techniques for gene therapy for HIV infections are also under development, such as internal antibodies (*intrabodies*) against the Tat protein, a vehicle for HIV infection of the immune cells; it is hoped that altered T-cells that produce their own anti-Tat intrabody will lengthen the survival time of infected cells or serve as an HIV 'dead-end'.

Other clues to AIDS treatments may come from the finding that HIV-1 originally came from a subspecies of chimpanzees [440]. Since chimps have likely carried the virus for hundreds of thousands of years but have not become ill from it, understanding this observation might help fight HIV-pathogeny in humans. Help may also come from the interesting finding that a subset of humans have a genetic mutation (32 bases deleted from the 393 of gene CCR5) that creates a deficient T-cell receptor; this mutation intriguingly slows the onset of AIDS. Additionally, a small subset of people is endowed with a huge number of helper (CD4) T-cells which can coordinate an attack on HIV and thus keep the AIDS virus under exquisite control for many years; such people may not even be aware of the infection for years.

---

## Vaccine?

Still, many believe that only an AIDS vaccine offers true hope against this deadly disease. Yet the research on vaccines trails behind the development of drugs, which offer much greater financial incentives and lower risks than vaccines. The vaccine AIDSVAX by the California-base company VaxGen, a genetically amplified version of a single protein from the outer shell of the AIDS virus, offered only limited protection.

Another vaccine under development by an Oxford team (part of the International AIDS Vaccine Initiative) is exploiting for vaccine development the immunological data gleaned from Nairobi women who have remained unaffected

by AIDS despite many years of high-risk sexual behavior. These women's T-cells were found to fight off the disease by attacking two particular proteins produced by the AIDS virus. The DNA sequences making those proteins were subsequently identified and used to create a vaccine specific to viral infections in East Africa; besides the DNA component associated with the relevant genes, the vaccine was amplified with a benign virus copy with same DNA sequences inserted.

Early attempts to target the outer protein envelope of HIV, gp120, turned disappointing, likely because not all virus particles were neutralized. Other vaccines have also been developed, but response is far from ideal. Thus, the announcement in September 2009 that, after 20 years of constant failure, a vaccine which blends two experimental vaccines that had previously failed to work on their own — Sanofi-Pasteur's ALVAC canary pox/HIV vaccine and VaxGen's AIDSVAX — offered some protection by reducing the rate of infection by 30% generated great excitement. However, results puzzle researchers because, while reducing infection, the combination vaccine does not reduce the virus levels in the blood. Research is ongoing.

In general, vaccine research experience suggests that a constant level of exposure (e.g., booster shots) is needed to yield immunity, and this defeats the main vaccine advantage of convenience and low cost. Observations also suggest that combinations of vaccines may be needed, since the HIV virus mutates and replicates quickly.<sup>5</sup> Still, it is hoped that therapeutic vaccination in combination with anti-HIV-1 drug treatment, even if it fails to eradicate infection, will suppress AIDS infection and the rate of transmission, and ultimately decrease the number of AIDS deaths substantially. One of the recent vaccine initiatives includes inducing primary T-cell mediated response to decrease the probability of initial infection [91, 612].

Besides focusing on the role of T cells in the control of the HIV disease progression, other current efforts are attempting to understand the complex immune-response behavior by various participants in the vaccination trials and to broaden the field of HIV vaccine research from new perspectives [375]. Very recently, a novel approach using RNA silencing has shown promise, by suppressing certain host viral genes crucial to the virus's replication; the small RNA molecules were delivered to the T cells via a small peptide [686].

However, it is becoming apparent that large resources and enormous leaps — in many fields like genetics, cellular and systems biology — are needed to succeed in preventing this devastating disease.

---

<sup>5</sup>For example, there is an enormous variation in the HIV-1 envelope protein. It has also been found that nearly all of non-nucleoside reverse transcriptase inhibitors can be defeated by site-directed mutation of tyrosine 181 to cysteine in reverse transcriptase. For this reason, the derivatives of *Calanolide A* under current development are attractive drug targets because they appear more robust against mutation [661].

### 2.4.3 Other Drugs and Future Prospects

#### Success Stories

Another example of drug successes based on molecular modeling is the design of potent *thrombin inhibitors*. Thrombin is a key enzyme player in blood coagulation, and its repressors are being used to treat a variety of blood coagulation and clotting-related diseases. Merck scientists reported [161] how they built upon crystallographic views of a known thrombin inhibitor to develop a variety of inhibitor analogues. In these analogues, a certain region of the known thrombin inhibitor was substituted by hydrophobic ligands so as to bind better to a certain enzyme pocket that emerged crucial for the fit. Further modeling helped select a subset of these ligands that showed extremely compact thrombin/enzyme structures; this compactness helps oral absorption of the drug. The most potent inhibitor that emerged from these modeling studies has demonstrated good efficacy on animal models [161].

Other examples of drugs developed in large part by computational techniques include the SARS virus inhibitor [329], glutamate nanosensors to monitor neurologic functions [933], agonists to treat anxiety and depression [111], the *antibacterial agent Norfloxacin* of Kyorin Pharmaceuticals (noroxin is one of its brand names), *glaucoma treatment Dorzolamide* (“Trusopt”/Merck), *Alzheimer’s disease treatment Donepezil* (“Aricept”/Eisai), and *migraine medicine Zolmitriptan* (“Zomig”) discovered by Wellcome and marketed by Zeneca [160]. The headline-generating drug that combats impotence (*Viagra*) was also found by a rational drug approach. It was interestingly an accidental finding: the compound had been originally developed as a drug for hypertension and then angina.

There are also notable examples of *herbicides and fungicides* that were successfully developed by statistical techniques based on linear and nonlinear regression and classical multivariate analysis (or QSAR, see Chapter 15): the herbicide metamitron — bestseller in 1990 in Europe for protecting sugar beet crops — was discovered by Bayer AG in Germany.

#### Impact of Technology and Modeling

With these new discoveries, we are enjoying improved treatments for cancer, AIDS, heart disease, Alzheimer and Parkinson’s disease, migraine, arthritis, and many more ailments. As new drug targets are being identified — such as new potential sites for antibiotics on the ribosome revealed by a combination of crystallography and bioinformatics, and new protein interfaces within the influenza virus’s RNA polymerase that might be targeted to disrupt polymerase assembly and thus viral replication, as revealed by crystallographic views of RNA polymerase — new opportunities for drug design by modeling become available.

In fact, high-throughput technologies that rely on progress in many fields from genomics to proteomics to imaging can now be processed through the new fields of knowledge-based biological information, like *bioinformatics* [571, 881] and *chemoinformatics* [507]. Improved modeling and library-based techniques,

coupled with robotics and high-speed screening, are also likely to increase the demand for faster and larger-memory computers. *“In a marriage of biotech and high tech,”* wrote the New York Times reporter Andrew Pollack in 1998, *“computers are beginning to transform the way drugs are developed, from the earliest stage of drug discovery to the late stage of testing the drugs in people”*.

### Declining Productivity

However, since the above statement was made, progress in drug development has not exhibited the growth hoped for by emerging technologies. In fact, the industry has actually contracted from a peak of around 50 new approved pharmaceutical agents, also known as new molecular entities (NMEs), in 1996 to half that value in 2008 and 2009 [885]. This slump is even more serious considering that Research and Development (R&D) costs have increased dramatically during this period. Thus, the average cost of \$500-800 million and time of 12–15 years required to develop a single drug remain extremely high.

There are many reasons for this disappointing trend.

First, due to safety issues discovered after drugs were approved,<sup>6</sup> the FDA has implemented continuously rising risk-averse requirements for drug approval, and these modified protocols affect all stages of drug development: discovery and preclinical testing, clinical studies, and registration/approval process.

Second, discovery of new drugs may be more difficult since many of the simple targets/strategies were already considered; this is not unlike the search for new protein folds, which has turned out to be more challenging than originally expected. This difficulty is also reflected by the smaller percentage of truly innovative new drugs among the NMEs.

Third, the “patent cliff” is also affecting this reduction in major pharmaceutical R&D productivity. This cliff refers to loss of revenue when patents for blockbuster drugs expire. These expirations are hitting many companies in a relatively short period around 2010.<sup>7</sup> These patent expirations lead to sharp profit declines if the company’s drug labs are barren, with no blockbuster substitutes coming out of the pipeline by patent expiration time; in turn, these losses reduce investments in new drug development.

Though a handful of new *biologics* — biomolecules derived from living cells instead of traditional small-molecule drugs (e.g., *Enbrel* for rheumatoid arthritis, *Herceptin* for breast cancer) — are being approved and these help make up for the dip in traditional small-molecule drugs, the long-anticipated breakthroughs

---

<sup>6</sup>One of the largest drug recalls involves Merck’s widely used arthritis drug *Vioxx*. Approved in 1999, *Vioxx* was withdrawn in 2004 after demonstrated increases in the risk of stroke, heart attack, and death.

<sup>7</sup>For example, patents for the migraine drug *Imitrex tablets* expired in 2009; *Advair* for asthma, *Levaquin* for bacterial infections, and *Lipitor* for cholesterol expire in 2010; *Actos* for type-2 diabetes, *Aprovel* for high blood pressure, and *Zomig tablets* for migraines expire in 2011; and *Avandia* for diabetes, *Crestor* for cholesterol, *Lexapro* for depression, *Singulair* for asthma, and *Zometa* for cancer expire in 2012.

in drug development due to high-throughput, genomics-based approaches and biotech agents have not yet been realized. See Chapter 15 for examples of biologics and further discussion of the computational challenges in drug design.

Perhaps, as the new director of NIH exclaimed in January 2010, “*The power of the molecular approach to health and disease has steadily gained momentum over the past several decades and is now poised to catalyze a revolution in medicine.*” [257]. However, it is becoming clear that such revolutionary advances in drug development, anticipated in the next decade from a combination of high-throughput approaches, biologics, pharmacogenomics, and other innovations, require new integrated paradigms to manage the complex scientific, technological, economic, and business factors involved and reverse the ebbing trends. A better yield of innovative and cost-effective pharmaceutical agents might also alleviate the industry’s political challenges, associated with inadequate availability of drugs to the world’s poor population.

#### 2.4.4 Gene Therapy – Better Genes

Looking beyond drugs, gene therapy is another approach that is benefiting from key advances in biomolecular structure/function studies. Gene therapy attempts to compensate for defective or missing genes that give rise to various ailments — like hemophilia, the severe combined immune deficiency SCID, sickle-cell anemia, cystic fibrosis, and Crigler-Najjar (CN) syndrome — by trying to coerce the body to make new, normal genes. This regeneration is attempted by inserting replacement genes into viruses or other vectors and delivering those agents to the DNA of a patient (e.g., intravenously). However, delivery control, biological reliability, as well as possible unwelcome responses by the body against the foreign invader, remain serious technical hurdles.

One of the classic gene therapy strategies involves direct injection of the thymidine kinase (TK) gene vector into tumors of cancer patients to control cell replication. When the TK gene is expressed, cancer cells can be killed after administration of *Gancyclovir*, which is converted by TK into a toxic nucleotide. This approach was initially used in aggressive brain tumors (glioblastoma multiforme) and more recently for locally recurrent prostate, breast, and colon tumors, among others. See Box 2.5 for other examples of gene therapy.

The first death in the fall of 1999 of a gene therapy patient treated with the common fast-acting weakened cold virus adenovirus led to a barrage of negative publicity for gene therapy.<sup>8</sup> However, the first true success of gene therapy was

---

<sup>8</sup>The patient of the University of Pennsylvania study was an 18-year old boy who suffered from ornithine transcarbamylase (OTC) deficiency, a chronic disorder stemming from a missing enzyme that breaks down dietary protein, leading to accumulation of toxic ammonia in the liver and eventually brain and kidney failure. The teenager suffered a fatal reaction to the adenovirus vector used to deliver healthy DNA rapidly. Autopsy suggests that the boy might had been infected with a second cold virus, parvovirus, which could have triggered serious disorders and organ malfunction that ultimately led to brain death.

reported four months later: the lives of most infants who would have died of the severe immune disorder SCID (and until then lived in airtight bubbles to avoid the risk of infection) were not only saved, but able to live normal lives following gene therapy treatments that restore the ability of a gene essential to make T cells [208]. Unfortunately, complications arose in several of the treated infants by late 2002, including deaths from gene therapy and as well as acquired leukemia [624]. (see Box 2.5).

Though such medical advances appear just short of a miracle, it remains to be seen how effective gene therapy will be on a wide variety of diseases and over a long period. Still, by early 2010, gene therapy treatments may have turned the corner. Small successes have accumulated, for treating children with a fatal brain disease (X-linked adrenoleukodystrophy or ADL) by inserting a corrective gene into the blood cells [890]; a rare form of inherited blindness that strikes at infancy (Leber's congenital amaurosis or LCA), by injecting the eye with a harmless virus carrying a gene coding for an enzyme necessary for making a light-sensing pigment [814]; and the severe immune disease SCID or “Bubble Boy”, by replacement of the enzyme adenosine deaminase [14]. Thus, cautious optimism is certainly warranted. And for the patients who gained site or normal function after living with serious genetic disorders, gene therapy can be short of a miracle.

A related technique for designing better genes is another relatively new approach known as *directed molecular evolution*. Unlike protein engineering, in which natural proteins are improved by making specific changes to them, directed evolution involves mutating genes in a test tube and screening the resulting ('fittest') proteins for enhanced properties. Companies specializing in this new Darwinian mimicking (e.g., Maxigen, Diversa, and Applied Molecular Evolution) are applying such strategies in an attempt to improve the potency or reduce the cost of existing drugs, or improve the stain-removing ability of bacterial enzymes in laundry detergents. Beyond proteins, such ideas might also be extended to evolve better viruses to carry genes into the body for gene therapy or evolve metabolic pathways to use less energy and produce desired nutrients (e.g., carotenoid-producing bacteria).

---

### Box 2.5: Gene Therapy Examples

A prototype disease model for gene therapy is hemophilia, whose sufferers lack key blood-clotting protein factors. Specifically, Factor VIII is missing in hemophilia A patients (the common form of the disease); the much-smaller Factor IX is missing in hemophilia B patients (roughly 20% of hemophiliacs in the United States).

Early signs of success in treatment of hemophilia B using adeno-associated virus (a vector not related to adenovirus, which is slower acting and more suitable for maintenance and prevention) were reported in December 1999. However, introducing the much larger gene needed for Factor VIII, as required by the majority of hemophiliacs, is more challenging. Here, the most successful treatments to date only increase marginally this protein's level.

Yet even those minute amounts are reducing the need for standard hemophilia treatment (injections of Factor IX) in these patients.

Larger vectors to stimulate the patient's own cells to repair the defective gene are thus sought, such as retroviruses (e.g., lentiviruses, the HIV-containing subclass), or non-virus particles, like chimeroplasts (oligonucleotides containing a DNA/RNA blend), which can in theory correct point mutations by initiating the cell's DNA mismatch repair machinery.

An interesting current project involving chimeroplasts is being tested in children of Amish and Mennonite communities to treat the debilitating Crigler-Najjar (CN) syndrome. Sufferers of this disease lack a key enzyme which break down the toxic waste product bilirubin, which in the enzyme's absence accumulates in the body and causes jaundice and overall toxicity. Children with CN must spend up to 18 hours a day under a blue light to clear bilirubin and seldom reach adulthood, unless they are fortunate to receive and respond to a liver transplant. Chimeroplasty offers these children hope, and might reveal to be safer than the adenovirus approach, but the research is preliminary and the immune response is complex and mysterious.

Recent success was reported for treating children suffering from the severe immune disorder SCID type XI [208]. Gene therapy involves removing the bone marrow from infants, isolating their stem-cells, inserting the normal genes to replace the defective genes via retroviruses, and then re-infusing the stem cells into the blood stream. As hoped, the inserted stem cells were able to generate the cells needed for proper immune functioning in the patients, allowing the babies to live normal lives. Though successful for 2–3 years for most infants, complications arose when several infants developed leukemia-like conditions and one child even died. Scientists believe that the retrovirus vectors lodged near a cancer-causing gene and activated it. Therefore, alternative vectors for carrying the genes into the body have been under development, including the HIV virus, modified so it could not cause the disease. Clearly, weighing the overall benefits against the risks remains an issue for gene therapy. In addition, questions regarding the long-term behavior of the children's new immune systems remain open.

Though clearly many bumps in the road are expected when new therapies are developed, scientists remain hopeful. Indeed, success in such gene therapy endeavors would lead to enormous progress in treating inherited diseases caused by point mutations.

---

#### 2.4.5 *Designed Compounds and Foods*

From our farms to medicine cabinets to supermarket aisles, designer foods are big business.

As examples of these practical applications, consider the transgenic organisms designed to manufacture medically-important compounds: bacteria that produce *human insulin*, goats whose milk contains *proteins to make silk* for use in surgical thread or bulletproof clothing, silkworms that produce *mammalian-type collagen and silk* for use in tissue engineering and other medical applications, and the food product *chymosin to make cheese*, a substitute for the natural rennet enzyme

traditionally extracted from cows' stomachs. Genetically-modified bacteria, more generally, hold promise for administering drugs and vaccines more directly to the body (e.g., the gut) without the severe side effects of conventional therapies. For example, a strain of the harmless bacteria *Lactococcus lactis* modified to secrete the powerful anti-inflammatory protein interleukin-10 (IL-10) has shown to reduce bowel inflammation in mice afflicted with inflammatory bowel disease (IBD), a group of debilitating ailments that includes Crohn's disease and ulcerative colitis.

The production of drugs in genetically-altered plants — “biopharming” or “molecular pharming” — represents a growing trend in agricultural biotechnology. The goal is to alter gene structure of plants so that medicines can be grown on the farm, such as to yield an edible vaccine from a potato plant against hepatitis B, or a useful antibody to be extracted from a tobacco plant. As in bioengineered foods, many obstacles must be overcome to make such technologies effective as medicines, environmentally safe, and economically profitable. Proponents of molecular pharming hope eventually for far cheaper and higher yielding drugs.

Genetically-engineered crops are also helping farmers and consumers by improving the taste and nutritional value of food, protecting crops from pests, and enhancing yields. Examples include the roughly one-half of the soybean and one-third of the corn grown in the United States, sturdier salad tomatoes,<sup>9</sup> corn pollen that might damage monarch butterflies, papaya plants designed to withstand the papaya ringspot virus, and caffeine-free plants (missing the caffeine gene) that produce decaffeinated cups of java.

The general public (first in Europe and then in the United States) has resisted genetically-modified or biotech crops, and this was followed by several blockades of such foods by leading companies, as well as global biosafety accords to protect the environment. Protesters have painted these products as unnatural, hazardous, evil, and environmentally dangerous ('Frankenfoods').<sup>10</sup>

---

<sup>9</sup>The *Flavr Savr* tomato that made headlines when introduced in 1993 contained a gene that reduces the level of the ripening enzyme polygalacturonase. However, consumers were largely disappointed: though beautiful, the genetically engineered fruit lacked taste. This is because our understanding of fruit ripening is still limited; a complex, coordinated series of biochemical steps is involved — modifying cell wall structure, improving texture, inducing softening, and producing compounds in the fruit that transform flavor, aroma, and pigmentation. Strawberries and other fruit are known to suffer similarly from the limitations of our understanding of genetic regulation of ripening and, perhaps, also from the complexity of human senses! See [1316], for example, for a recent finding that a tomato plant whose fruit cannot ripen carries a mutation in a gene encoding a transcription factor.

<sup>10</sup>Amusing Opinion/Art ads that appeared in The New York Times on 8 May 2000 include provocative illustrations with text lines like “GRANDMA'S MINI-MUFFINS are made with 100% NATURAL irradiated grain and other ingredients”; “TOTALLY ORGANIC Biomatter made with Nucleotide Resequencing”; “The Shady Glen Farms Promise: Our Food is **fresh from the** research labs buried deep under an abandoned farm”. [Note: the font size and form differences here are intentional, mimicking the actual ads].

With the exception of transferred allergic sensitivities — as in Brazil nut allergies realized in soybeans that contained a gene from Brazil nuts — most negative reactions concerning *food safety* may not be scientifically well-grounded. In fact, not only do we abundantly use various sprays and chemicals to kill flies, bacteria, and other organisms in our surroundings and on the farm; each person consumes around 500,000 kilometers of DNA on an average day! Furthermore, there are many potential benefits from genetically-engineered foods, like higher nutrients and less dependency on pesticides, and these considerations might win in the long run. Still, environmental effects must be carefully monitored so that genetically-altered food will succeed in the long run (see Box 2.6 for possible problems).

Perhaps to counter fear of introduced allergens, bioengineering is also being used to reduce or remove compounds that cause allergic reactions in people. Though at a relatively early stage, various companies worldwide are using genetic engineering to try to reduce allergies from foods like wheat, rice, soybean, ryegrass, and peanuts. Genes responsible for producing allergenic proteins can be removed (i.e., *knocked out*), as done for soybeans, or the associated proteins redesigned, as in peanuts, so that allergenicity is lost but other nut characteristics are retained. As above, care must be taken to retain flavor, freshness, and looks of the original product, and not to introduce other possible allergens.

In addition to tampering with plants to remove allergens, such biotech companies are also expanding effort on the removal of genes associated with natural toxins. For example, companies (with support of national security organizations) are attempting to remove the toxin *ricin* — one of the deadliest substances known — from castor plants. Castor beans have been cultivated for centuries, and the plant's natural oils (which lack toxicity) are widely used as laxatives and as component in brake fluid, dyes, soaps, and cosmetics. However, the toxic protein ricin can also be extracted from the castor plant, and has been associated with terrorist groups like Al Qaeda, with production of weapons “for mass destruction” in Iraq, and with an infamous killing of the spy Georgi Markov on a London sidewalk in 1978 by Bulgarian agents who injected ricin from an umbrella tip into the defector’s leg. Once removed, ricin-free castor plants can become more attractive to growers.

---

#### Box 2.6: Nutraceuticals Examples

The concept of fortified food is not new. Vitamin-D supplemented milk has eradicated rickets, and fortified breakfast cereals have saved many poor diets. In fact, classic bio-engineering has been used for a long time to manipulate genes through conventional plant and animal inter-breeding. But the new claims — relying on our increased understanding of our body’s enzymes and many associated vital processes — have been making headlines. (“Stressed Out? Bad Knee? Try a Sip of These Juices.”, J.E. Barnes and G. Winter, *New York Times*, Business, 27 May 2001). Tea brews containing sedative roots like kava promise to tame tension and ease stress. Fruit-flavored tonics with added glucosamine (building block of cartilage) and calcium are claimed to soothe stiff knees of aging bodies.

(See Chapter 3 on the fibrous protein collagen). Fiber-rich grains are now touted as heart-disease reducers, and fiber-rich foods have appeared in items well beyond cereals. Herb-coated snacks, like potato corn munchies coated with ginkgo biloba, are advertised as memory and alertness boosters.

With this growing trend of designer foods, the effect of these manipulations on our environment demands vigilant watch. This is because it is possible to create ‘super-resistant weeds’ or genetically-improved fish that win others in food or mate competitions. This potential danger emerges since, unlike conventional cross-breeding (e.g., producing a tangelo from a tangerine and grapefruit), genetic engineering can overcome the species barrier — by inserting nut genes in soybeans or fish genes in tomatoes, for example. This newer type of tinkering can have unexpected results in terms of toxins or allergens which, once released to the environment, cannot be stopped easily. For example, the first genetically-modified animal to reach American dinner plates is likely to be a genetically-altered salmon endowed with fortified genes that produce growth hormones, making the fish grow twice as fast as normal salmon. The effect of these endowed fish on the environment is yet unknown.

Popular examples of fortified food products with added vitamins and minerals (e.g., calcium and vitamin E) that also help protect against osteoporosis are orange juice, specialty eggs, and some vegetarian burritos. Other designer disease-fighting foods include drinks enriched with echinacea to combat colds; juices filled with amino acids and herbs claimed to boost muscle and brain function; margarines containing plant stanol esters (from soybean or pine trees) to fight heart disease and cancer (by blocking cholesterol absorption from the digestive tract), as well as green teas enriched with ginseng and other herbs; super-yogurts to enhance the immune system; and tofu and yams to combat hot flashes. Such functional foods are also touted to lower cholesterol, provide energy, fight off depression, or to protect against salmonella and *E. coli* poisoning (e.g., yogurt fortified with certain bacteria). Many other enriched food products are under design, for example fruit with increased vitamin C levels using a recently-isolated gene in strawberries (GalUR) that plays an important role in the production of vitamin C.

Will Ginkgo Biloba chips, Tension Tamer cocktails, or Quantum Punch juice become part of our daily diet (and medicine cabinet) in this millennium?

---

#### 2.4.6 Nutrigenomics

Closer to the supermarket, one of the fastest growing categories of foods today is *nutraceuticals* (a.k.a. functional foods or pharmaceuticals), no longer relegated only to health-food stores. These foods are designed to improve our overall nutrition as well as to help ward off disease, from cancer prevention to improved brain function. See Box 2.6 for examples.

However, while nutraceuticals in general may characterize the many products that flood our supermarket aisles with health claims concerning enhanced

cartilage support, cholesterol maintenance, relief of stress and tension, or maintenance of healthy lung function, the emerging field of *nutrigenomics* is a serious and well-grounded discipline. Nutrigenomics, at the interface of genomics, nutrition, and health, was made possible by recent developments in high-throughput transcriptomics, proteomics, and metabolomics technologies. Nutrigenomics integrates the genomics sciences with nutrition by studying how nature (the presence of particular genes or mutations) and nurture (our food intake, given environmental and behavioral factors) interact to manifest disease or protect us from it.

In its simplest form, diets low in certain proteins can be recommended for patients with phenylketonuria, or diets high in liver, broccoli, and other folic-acid rich foods can be a remedy for people with a genetic variation that produces a less efficient enzyme involved in processing folic acid. More generally, nutrition modifies the extent to which certain genes are expressed because macro-nutrients like proteins, micro-nutrients like vitamins, and naturally-occurring bioactive molecules like flavonoids regulate gene expression. Some of these compounds like *resveratrol* in red wine are ligands for transcription factors, and others like the natural amine nutrient *choline* — found in the lipids that make up cell membranes and in the neurotransmitter acetylcholine — alter signal transduction pathways and chromatin structure, thereby also affecting gene expression epigenetically. Because single nucleotide polymorphisms (SNPs) can alter gene functions, much of the focus in nutrigenomics has been on how the interaction of nutrients with SNPs increase or decrease disease risk.

Folate, for example, is among the nutrients critical to genome stability because it can cause DNA damage. More generally, key nutrients like folate, vitamin E, vitamin B<sub>12</sub>, niacin, or calcium are associated with a reduction in DNA damage, while riboflavins and biotin tend to increase such damage. The familiar advice to lower fat intake and increase amounts of cruciferous vegetables can be rationalized by the lowering by these agents of oxidative DNA damage, which occurs from environmental factors like tobacco smoke and dietary factors like ultra high-fat diets. Thus, folate and other antioxidants and phytochemicals are recommended because they enhance DNA repair and reduce oxidative DNA damage. Such dietary modifications can help compensate for inherited mutations that may impair DNA damage repair. Because of this connection between DNA damage/repair and nutrition, some cancer researchers have become particularly interested in nutrigenomics.

In addition to cancer, diabetes, obesity, and cardiovascular disease have been researched in connection with food intake. Genetic susceptibility to these diseases (e.g., APOE- $\epsilon$ 4 polymorphism, associated with elevated total cholesterol and increased risk of type-2 diabetes and Alzheimer's disease) can be counteracted in part by dietary modifications that include plant-rich, high-fiber and low-fat diets in combination with regular exercise. Thus, nutrigenomics is leading to customized diet ingredients and supplements that are tailored to genetic variations, but the field is only beginning.

### 2.4.7 Designer Materials

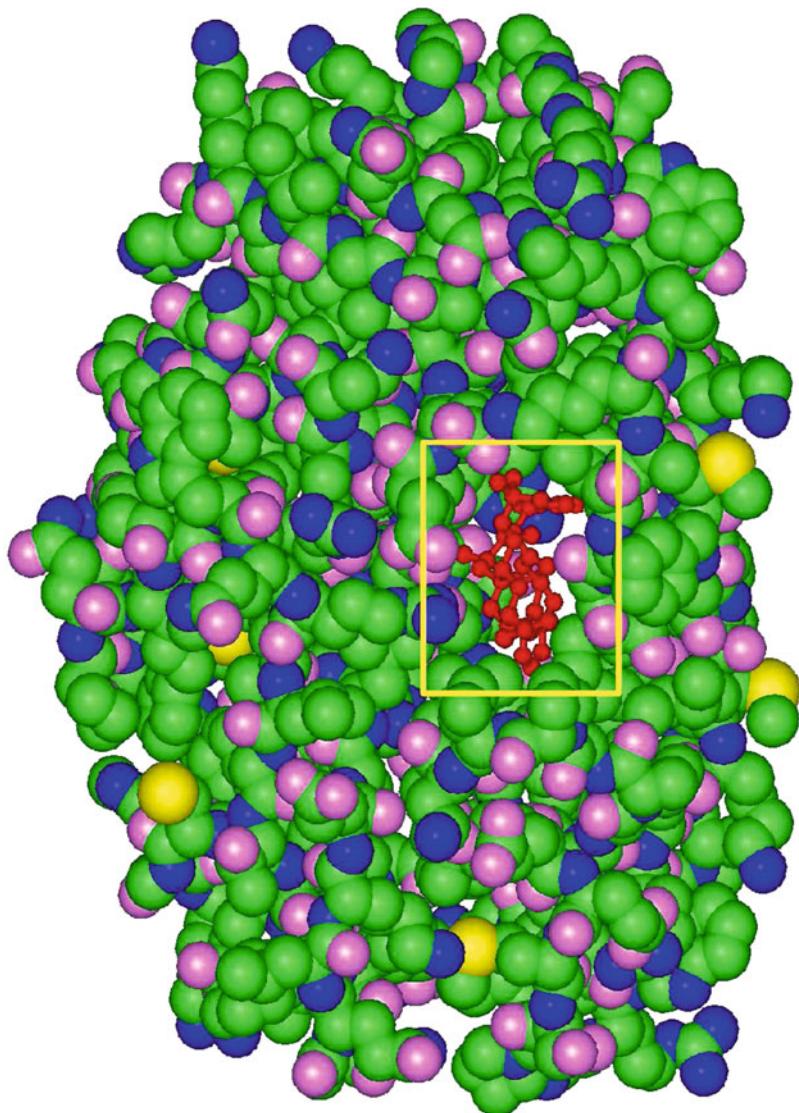
New specialty materials are also being developed in industry with the needed thermochemistry, stereochemistry (e.g., compounds that bind to one chemical but not its mirror image), and kinetic properties. Examples are enzymes for manufacturing detergents, adhesives and coatings, photography film, or biosensors for explosives. Fullerene nanotubes (giant linear fullerene chains that can sustain enormous elastic deformations [1406]), formed from condensed carbon vapor, have many potential applications. These range from architectural components of bridges and buildings, cars, and airplanes to heavy-duty shock absorbers, to components of computer processors, scanning microscopes, and semiconductors.

Long buckyball nanotube fibers have even been proposed as elements of ‘elevators’ to space in the new millennium [1406]. These applications arise from their small size (their thickness is five orders of magnitude smaller than human hair), amazing electronic properties, and enormous mechanical strength of these polymers. In particular, these minuscule carbon molecules conduct heat much faster than silicon, and could therefore replace the silicon-based devices used in microelectronics, possibly overcoming current limitations of computer memory and speed. Far from science fiction, NASA scientists believe that the first space elevator, to carry cargo, might be built in the not-too-distant future.

### 2.4.8 Cosmeceuticals

Cosmeceutical companies are also rising — companies that specialize in design of cosmetics with bioactive ingredients (such as designer proteins and enzymes), including cosmetics that are individually customized (by *pharmacogenomics*) based on genetic markers, such as single nucleotide polymorphisms (SNPs). Most popular are products for sun or age-damaged skin containing alpha hydroxy acids (mainly glycolic and lactic acid), beta hydroxy acids (e.g., salicylic acid), and various derivatives of vitamin A or retinol (e.g., the tretinoin-containing *Retin-A* and *Renova* topical prescriptions). Besides reducing solar scars and wrinkling, products can also combat various skin diseases. Many of these compounds work by changing the metabolism of the epidermis, for example by increasing the rate of cell turnover, thereby enhancing exfoliation and the growth of new cells. New cosmeceuticals contain other antioxidants, analogues of various vitamins (A, D, and E), and antifungal agents.

The recent information gleaned from the Human Genome Project can help recognize changes that age and wrinkle skin tissue, or make hair or teeth gray. This in turn can lead to the application of functional genomics technology to develop agents that might help rejuvenate the skin, or color only target gray hair or tooth enamel. Computational methods have an important role in such developments by screening and optimizing designer peptides or proteins. Such biotechnology research to produce products for personal care will likely rise sharply in the coming years.



# 3

## Protein Structure Introduction

Chapter 3 Notation

SYMBOL	DEFINITION
$C^\alpha$	$\alpha$ -Carbon
$\tau$	dihedral angle
$\phi$	$\{N-C^\alpha\}$ rotation about peptide bond
$\chi_1-\chi_4$	rotamer dihedral angles in amino acid sidechains
$\psi$	$\{C^\alpha-C\}=O$ rotation about peptide bond
$\omega$	$C_1^\alpha-\{C-N\}-C_2^\alpha$ rotation

Life is the mode of existence of proteins, and this mode of existence essentially consists in the constant self-renewal of the chemical constituents of these substances.

Friedrich Engels, 1878 (1820–1895).

### 3.1 The Machinery of Life

#### 3.1.1 From Tissues to Hormones

The term “protein” originates from the Greek word *proteios*, meaning “primary” or “of first rank”. The name was adapted by Jöns Berzelius in 1838 to emphasize the importance of this class of molecules. Indeed, proteins play

crucial, life-sustaining biological roles, both as constituent molecules and as triggers of physiological processes for all living things. For example, proteins provide the architectural support in muscle tissues, ligaments, tendons, bones, skin, hair, organs, and glands. Their environment-tailored structures make possible the coordinated function (motion, regulation, etc.) in some of these assemblies.

Proteins also provide the fundamental services of transport and storage, such as of oxygen and iron in muscle and blood cells. The first pair of solved protein structures **hemoglobin** and **myoglobin**, serve as the crucial oxygen carriers in vertebrates. Hemoglobin is found in red blood cells and is the chief oxygen carrier in the blood (it also transports carbon dioxide and hydrogen ions). Myoglobin is found in muscle cells, where it stores oxygen and facilitates oxygen movement in muscle tissue. The sperm whale depends on myoglobin in its muscle cells for large amounts of oxygen supplies during long underwater journeys.

Proteins further play crucial regulatory roles in many basic processes fundamental to life, such as reaction catalysis (e.g., digestion); immunological and hormonal functions; and the coordination of neuronal activity, cell and bone growth, and cell differentiation.

Given this enormous repertoire, Berzelius could not have coined a better name!

### 3.1.2 Size and Function Variability

Protein molecules come in a wide range of sizes and have evolved many functions. The major classes of proteins include *globular*, *fibrous*, and *membrane* proteins. Globular proteins are among the most commonly studied group. Newly found ribosomal proteins form a characteristic class of proteins that can be ordered as globular proteins, with disordered extensions.

To suit their environment and function, fibrous proteins (e.g., the collagen molecule in skin and bones), which are generally insoluble in aqueous environments, are extended in shape, whereas globular proteins tend to be compact. **Collagen** is a left-handed helix with a quaternary structure made of collagen fibrils aggregated in a parallel superhelical arrangement. See [680] for the crystal structure of a collagen-like peptide with a biologically relevant sequence (also shown in Figure 3.9) and summary of collagen structures elucidated to date. The globular protein **myoglobin** (see Figure 3.12) is highly compact, organized as 75%  $\alpha$ -helices. Similarly, **hemoglobin** is a tetramer composed of four polypeptide chains held by noncovalent interactions; each subunit of hemoglobin in humans is very similar to myoglobin. Both proteins bind oxygen molecules through a central heme group.

There certainly are some very large proteins such as the muscle protein **titin** of about 27000 amino acids (and mass of 3000 kDa), but the average protein contains several hundred residues. The size of polypeptides can be determined from gel electrophoresis experiments: the rate of migration of the molecule is inversely proportional to the logarithm of its length. The mass of a polypeptide or protein can be estimated from mobility-to-mass relationships established for reference proteins and by mass spectrometry measurements. Equilibrium ultracentrifuga-

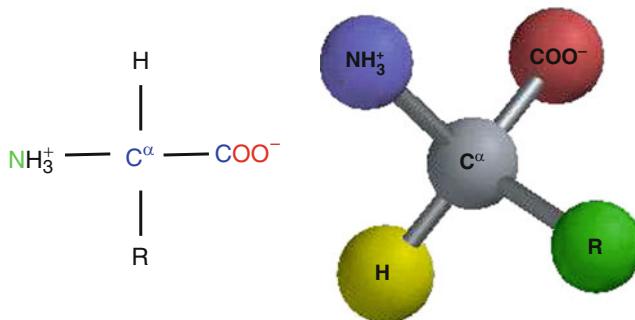


Figure 3.1. (left) The general formula for an amino acid, and (right) the spatial tetrahedral arrangement of an L-amino acid. The mirror image, a D-amino acid, is rare in proteins in Nature, if it exists.

tion [197] is another favored technique for determining various macromolecular features, including molecular weight, on the basis of transport properties.

### 3.1.3 Chapter Overview

This chapter introduces the bare basics in protein structure: the amino acid building blocks, primary sequence variations, and the framework for describing conformational flexibility in polypeptides. Included also is an introduction to the more advanced topic of sequence similarity and relation to structure, in the section on variations in protein sequences. *Students are encouraged to return to this subsection after reading Chapters 3 and 4.* Chapter 4 continues to describe secondary, supersecondary, and tertiary structural elements in proteins, as well as protein classification.

The protein treatment in these chapters is brief in comparison to the minitutorial on nucleic acids. Readers should consult the many excellent texts on protein structure (see books listed in Appendix C), like that by Branden and Tooze, and review introductory chapters in biochemistry texts like Stryer's. The 1999 text by Fersht [394], in particular, is a comprehensive description of the state-of-the-art in protein structure, and also reviews recent advances and insights from theoretical approaches.

#### Box 3.1: Water Structure

Water — that deceptively simple molecule composed of two hydrogens and one oxygen — displays highly unusual and complex properties that are far from fully understood [788]. Perhaps because of those properties — like the contraction of ice when it melts, large heat of vaporization, and large specific heat — water is the best of all solvents and a fundamental substance to sustain life. Solvent organization and reorganization are crucial to the stability of proteins, nucleic acids, saccharides, and other molecular systems and

affects functional motions profoundly [384]. The energetic and kinetic aspects of water structure are difficult to pinpoint by experiment and simulation because of the range of timescales associated with water motions, from the fast perturbations of order 0.1 ps to the slow proton exchanges of millisecond order.

Important to the understanding of solvation structure and dynamics in the vicinity of macromolecules is the tendency of water to form *hydrogen bonds* [1018] (see also Box 3.2 for a definition of a hydrogen bond). In ice, the ordered crystal structure of water molecules, each oxygen is surrounded by a tetrahedron of four other oxygen atoms, with one hydrogen between each oxygen pair. In liquid water, many water molecules are engaged in such a hydrogen-bonded network, but the network is highly dynamic, with hydrogen-bonded partners changing rapidly.

This local organization of liquid water can easily be observed from experimental and computed radial distribution functions (e.g., O–O and O–H distances), which reflect the degree of occupancy of neighbors from a central oxygen or hydrogen molecule. Thus, for example, the highest peak in the O–O radial distribution function at a distance of about 2.9 Å corresponds to the first solvation shell, in which the four near-neighbor oxygens of the central oxygen molecule can be found at room temperature.

Figure 3.2 illustrates the structure of water clusters as computed by minimizing the potential energy composed of bond length, bond angle, and intermolecular Coulomb and van der Waals terms (see Chapter 9 for energy terms discussion). Such hydrogen bonds form ubiquitously in the environment of biomolecules. Water molecules penetrate into the grooves of nucleic acid helices, aggregate around hydrophilic, or water-soluble, segments of proteins (which cluster at the protein surface) and stabilize solute conformations through various hydrogen bonds and bridging arrangements. The dynamic nature of both water structure and biomolecules gives rise to the concept of hydration shells; see Chapter 6 in the context of DNA. That is, the solvent structure around the solute is multilayered, with the first hydration shell associated with water molecules in direct contact with the solute and the outermost layer as the bulk solvent.

---

---

### Box 3.2: Hydrogen Bonds

A hydrogen bond is an attractive, weak electrostatic (noncovalent) bond [1018]. It forms when a hydrogen atom covalently binds to an electronegative atom and is electrostatically attracted to another (typically electronegative) atom. The atom to which the hydrogen atom (H) is covalently bound is considered the hydrogen *donor* (D), and the other atom is the hydrogen *acceptor* (A). Thus, the hydrogen bond is stabilized by the Coulombic attraction between the partial negative charge of the A atom and the partial positive charge of H. See Figure 3.2 for examples in water clusters.

In biological polymers, the donor and acceptor atoms are either nitrogens or oxygens. For example, in protein helices and sheets, the D–H ··· A sequence is N–H ··· O=C. In the nucleic-acid base pair of adenine–thymine, the two D–H ··· A sequences are N–H ··· O and N–H ··· (see Nucleic Acid chapters for details). Non-classical, weaker hydrogen

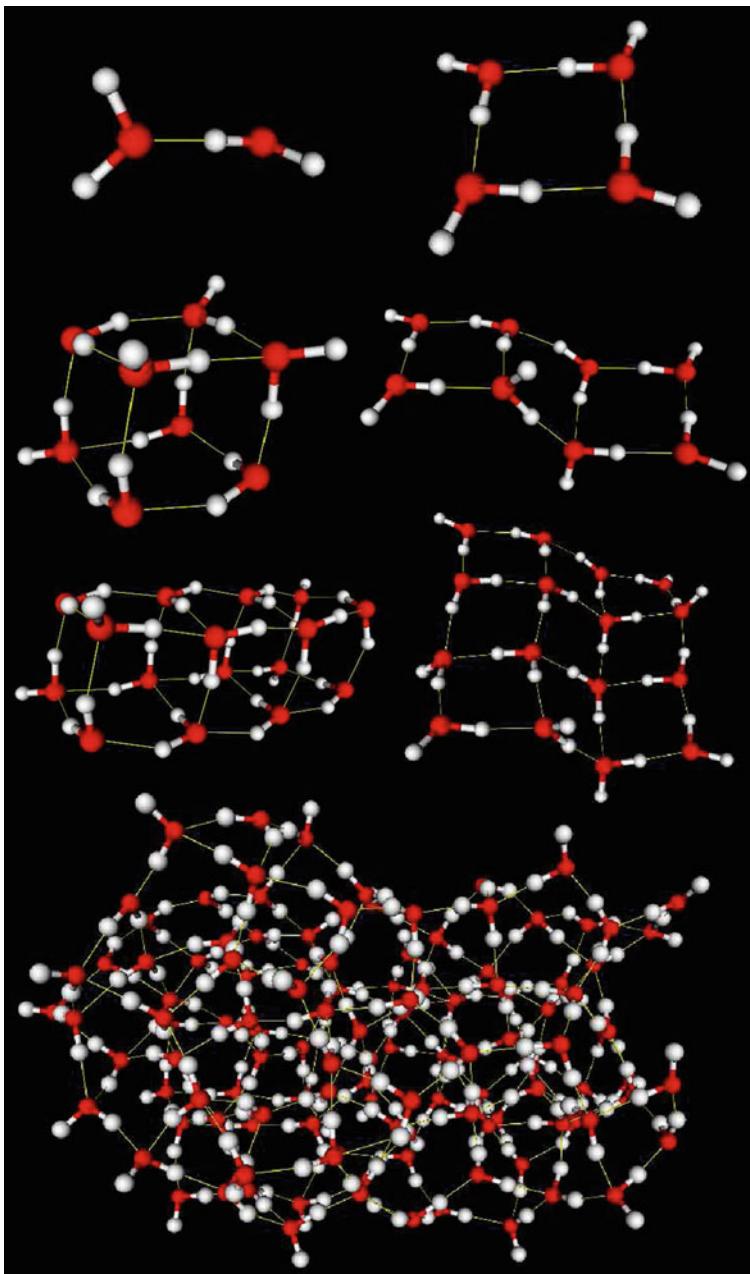


Figure 3.2. Structures of water clusters of 2, 4, 8, 16, and 125 molecules as minimized with the CHARMM force field from initial coordinates computed in [1120]. Two geometries are shown for both the 8 and 16-molecule clusters; the lower energy structures (by roughly 4 and 6%, respectively) are associated with the more compact, cube-like shapes (left side in both cases). The tetrahedral structure of water is apparent in the inner molecules of the larger systems, where each molecule is hydrogen bonded to four others.

bonds have been noted in biological systems (e.g., protein/DNA complexes), involving a carbon instead of one electronegative atom: C–H · · · O [820].

The strength of a hydrogen bond can be characterized by two geometric quantities which govern the hydrogen bond energy: colinearity of the D–H · · · A atoms, and optimal H · · · A (or D · · · A) distance. The ideal, strongest hydrogen bond often has its three atoms colinear. The strength of a hydrogen bond is several kilocalories per mole, compared to about 0.6 kcal/mol for thermal energy at room temperature, but the exact value remains uncertain (e.g., [319]). However, the formation of a network of hydrogen bonds in macromolecular systems leads to a cooperative effect that enhances stability considerably [1018].

---

## 3.2 The Amino Acid Building Blocks

Proteins and polypeptides are composed of linked amino acids. That amino acid composition of the polymer is known as the *primary structure* or sequence for short.

### 3.2.1 Basic $C^\alpha$ Unit

Each amino acid consists of a central tetrahedral carbon known as the alpha ( $\alpha$ ) carbon ( $C^\alpha$ ) attached to four units: a hydrogen atom, a protonated amino group ( $NH_3^+$ ), a dissociated carboxyl group ( $COO^-$ ), and a distinguishing sidechain, or R group (see Figure 3.1).

This dipolar or *zwitterionic* form of the amino acid ( $COO^-$  and  $NH_3^+$ ) is typical for neutral pH (pH of 7). The un-ionized form of an amino acid corresponds to COOH and NH<sub>2</sub> end groups. Different combinations involving ionized/un-ionized forms for each of the side groups can occur for the amino acid depending on the pH of the solution.

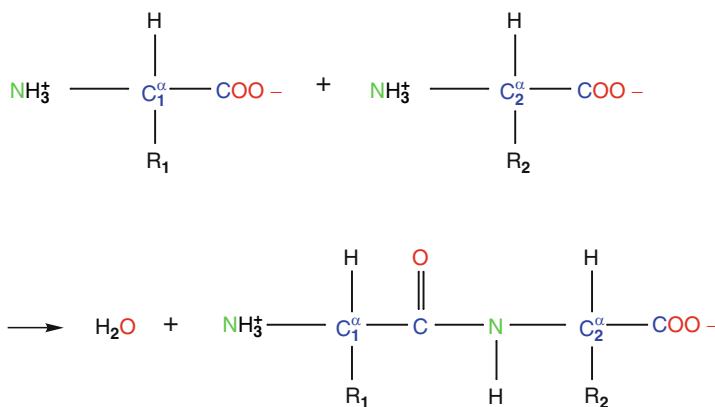


Figure 3.3. Formation of a dipeptide by joining two amino acids.

The tetrahedral arrangement about  $C^\alpha$  makes possible two mirror images for the molecule. Only the L-isomer (“left-handed” from the Latin word *levo*) is a constituent of proteins on earth (see Figure 3.1). This asymmetry is not presently understood, but one explanation is that this imbalance is related to an asymmetry in elementary particles.

### 3.2.2 Essential and Nonessential Amino Acids

There are 20 naturally-occurring amino acids.<sup>1</sup> Among them, humans can synthesize about a dozen. The remaining 9 amino acids must be ingested through our diet; these are termed *essential amino acids*. They are histidine, isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan, and valine.

Meat eaters need not be too concerned about a balanced diet of those nutrients, since animal flesh is a complete source of the essential amino acids. In contrast, vegetarians, particularly vegans — who omit all animal products like eggs and dairy in addition to meat, poultry, and fish — must perform a delicate balancing act to ensure that their bodies can synthesize all the basic proteins essential to good health. See Box 3.3 for a discussion of essential amino acids and balanced diets.

---

#### Box 3.3: Protein Chemistry and Vegetarian Diets

Vegetarians must take care to combine foods from three basic groups which are complementary with respect to their supply of these amino acids: (a) rice and grains like oats, wheat, corn, cereals, and breads; (b) legumes and soy-products; and (c) nut products (cashews, almonds, and various nut butters). Peanuts are technically members of the legume family rather than nuts.

Notable vegetarian combinations are: rice or grains (low in or lacking isoleucine, lysine, and threonine) with beans (rich in isoleucine and leucine and, in the case of lima beans, also lysine); cereals with leafy vegetables; corn, wheat, or rye (low in or lacking tryptophan and lysine) with soy protein/soybeans (rich in isoleucine, tryptophan, lysine, methionine, and valine); corn with nuts or seeds (rich in methionine, isoleucine, and leucine); bread/wheat with peanut butter (rich in valine and tryptophan); and potatoes (limited methionine and leucine) with onions, garlic, lentils and egg or fish (if permitted), all of which are rich in methionine.

A classic Native-American dish of acorn squash stuffed with wild rice, quinoa, and black beans is a superior mixture of nutrients. The plant quinoa, prepared like a grain, is actually a fruit and moreover a complete protein (rich in lysine and other amino acids), as well as rich in vitamins E and B, fiber, and the minerals calcium, phosphorus, and iron. Nuts also contain several vitamins and minerals that protect against heart disease, like folate, and calcium, magnesium, and potassium, which also protect against high blood pressure.

---

<sup>1</sup> At least two nonstandard by naturally occurring amino acids in certain Archaea and eubacteria are known, pyrrolysine and selenocysteine [68, 511].

Contrary to an existing myth, such food group combinations need not be eaten at the same meal to guarantee a complete source of the essential amino acids; a daily approach suffices. Given all these food sources, vegetarians — even vegans who carefully comply — will not be deficient in protein. However, nutrients that present a challenge to vegans and lacto-ovo vegetarians are **vitamin B<sub>12</sub>** (deficiencies of which can cause nerve damage), found in fortified cereals, and **zinc** (needed for protein synthesis, healing wounds, and immunity), available in fortified cereals, soy-based foods, and dairy products.

*Given the intimate relationship between protein chemistry and good nutrition, readers of this text should be healthy as well as smart!*

---

Interestingly, the requirements for vegetarian diets are now of crucial interest to NASA researchers: future astronauts who will spend extended periods of time in space stations (on Mars, Jupiter, or the Moon) will have to depend on hydroponically-grown plant crops for nearly all their protein requirements, as well as vitamins, minerals, and fiber. Research is now in progress on how best to select a limited set of plants that can adapt to growing in nutrient-enriched water (rather than soil) and in small spaces. At the same time, this selection must meet the basic dietary requirements of space-station scientists, as well as satisfy their culinary taste and demand for variety [173].

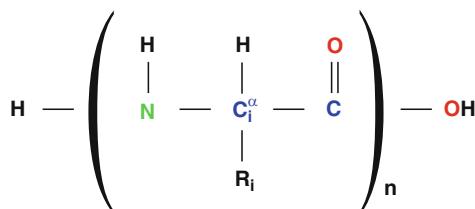


Figure 3.4. The repeating formula for a polypeptide.

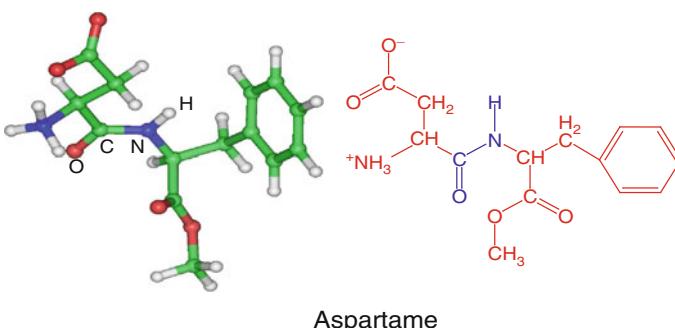


Figure 3.5. The dipeptide aspartame.

### 3.2.3 Linking Amino Acids

A polypeptide is formed when amino acids join together. Namely, the carboxyl carbon of one amino acid joins the amino nitrogen of another amino acid to form the peptide (C–N) bond with the release of one water molecule (Figure 3.3). The general repeating formula for a polypeptide is shown in Figure 3.4. When the amino acid residue is proline, its C<sup>α</sup> is linked to the nitrogen of the peptide backbone through the proline ring.

A model of **aspartame**, a dipeptide of aspartic acid and phenylalanine, is shown in Figure 3.5. It was discovered accidentally in 1965 by a ‘careless’ chemist who licked his fingers during his laboratory work. To his surprise, a substance 100–200 times sweeter than sucrose was discovered. Because it is a kind of protein, aspartame is metabolized in the body like proteins and is a source of amino acids. (*This should not, however, be taken as an endorsement for diet soft drinks as a source of nutrients!*)

The synthesis of polypeptides and proteins occurs in a cellular structure, the ribosome, *in vivo*. Synthesis *in vitro* is facile for 100–150 residues but much more involved for longer chains.

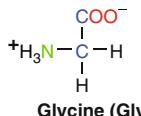
### 3.2.4 The Amino Acid Repertoire: From Flexible Glycine to Rigid Proline

The chemical formulas of the twenty L-amino acids are shown in Figure 3.6, with the corresponding space-filling models shown in Figure 3.7. The commonly used three-letter abbreviation for each acid is illustrated, as well as a grouping into amino acid subfamilies. A one-letter mnemonic is also used to identify sequences of amino acids, as shown in Table 3.1.

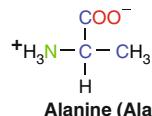
A broader classification than indicated in the figures consists of the following three groups:

- **NPo:** amino acids with strictly *nonpolar* (hydrophobic, or water insoluble) side chains:  
Ala, Val, Leu, Ile, Phe, Pro, Met, Gly, Trp, Tyr;
- **CPo:** amino acids with *charged polar* residues:  
Asp, Glu, His, Lys, Arg; and
- **UPo:** amino acids with *uncharged polar* side chains:  
Ser, Thr, Cys, Asn, Gln.

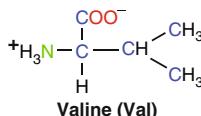
Each amino acid has a unique combination of properties — size, polarity, cyclic constituents, sulfur constituents, etc. — that critically affects the noncovalent and covalent (i.e., disulfide bonds) interactions that form and stabilize protein three-dimensional (3D) architecture. These interactions originate from electrostatic, van der Waals, hydrophobic, and hydrogen bonding forces. These properties are described in turn for these amino acid classes.

**ALIPHATIC SIDE CHAINS (NPo)**

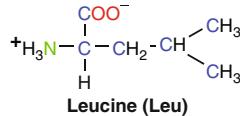
Glycine (Gly)



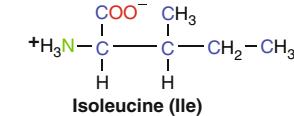
Alanine (Ala)



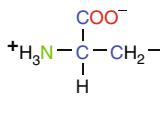
Valine (Val)



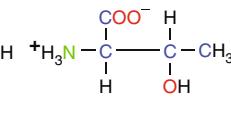
Leucine (Leu)



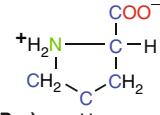
Isoleucine (Ile)

**ALIPHATIC HYDROXYL SIDE CHAINS (UPo)**

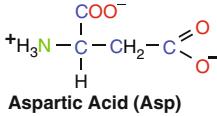
Serine (Ser)



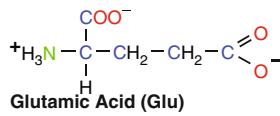
Threonine (Thr)



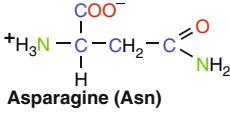
Proline (Pro)

**ACIDIC SIDE CHAINS AND THEIR AMIDE DERIVITIVES (CPo - Asp, Glu; UPo - Asn, Gln)**

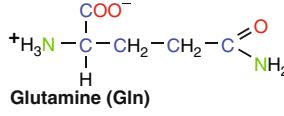
Aspartic Acid (Asp)



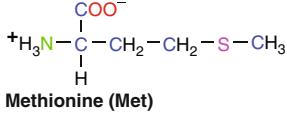
Glutamic Acid (Glu)



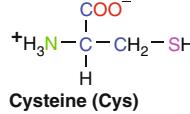
Asparagine (Asn)



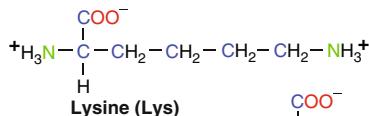
Glutamine (Gln)

**SULFUR-CONTAINING SIDE CHAINS (NPo)**

Methionine (Met)



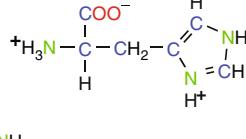
Cysteine (Cys)

**BASIC SIDE CHAINS (CPo)**

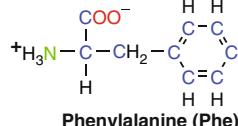
Lysine (Lys)



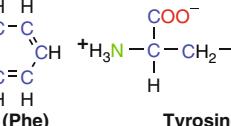
Arginine (Arg)



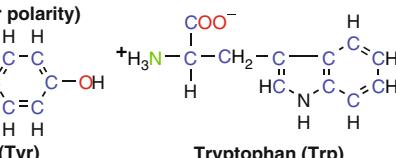
Histidine (His)

**AROMATIC SIDE CHAINS (NPo but potential for polarity)**

Phenylalanine (Phe)



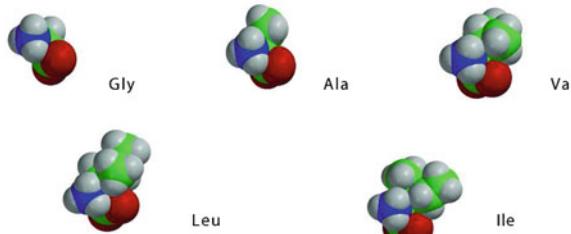
Tyrosine (Tyr)



Tryptophan (Trp)

Figure 3.6. The chemical formulas of the 20 natural amino acids as found in neutral pH (pH of 7). The acronyms **NPo**, **UPo**, **CPo** denote, respectively, nonpolar, uncharged polar, and charged polar amino acids.

## ALIPHATIC SIDE CHAINS



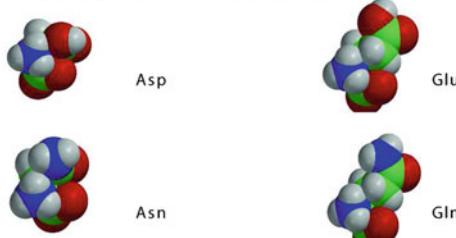
## ALIPHATIC HYDROXYL SIDE CHAINS



## SECONDARY AMINO GROUP



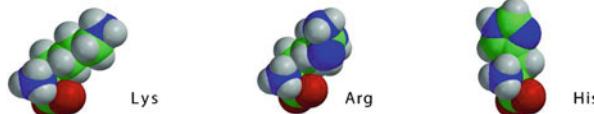
## ACIDIC SIDE CHAINS AND THEIR AMIDE DERIVATIVES



## SULFUR-CONTAINING SIDE CHAINS



## BASIC SIDE CHAINS



## AROMATIC SIDE CHAINS

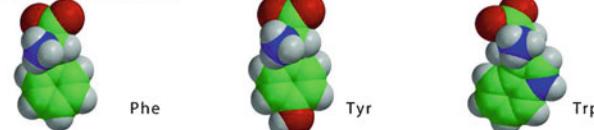


Figure 3.7. Space-filling models of the 20 amino acids.

Aliphatic R: Gly, Ala, Val, Leu, Ile

Glycine, alanine, valine, leucine, and isoleucine can be classified as nonpolar. Glycine, the simplest of the amino acids, is first in the aliphatic-sidechain subgroup. Each member in this family has a sidechain (R) which increases in

bulk and branching design. Glycine is most flexible and hence an important constituent of proteins. For example, glycine is a major component of the  $\alpha$ -helix of the protein  $\alpha$ -keratin, which makes up hair and wool, as well as the  $\beta$ -sheet of the polypeptide  $\beta$ -keratin, which is silk (see Figure 3.9). Since the increasing aliphatic substitutions in this family increases the bulkiness of the amino acid, the overall conformational flexibility correspondingly decreases within a polypeptide. However, the conformational variability of each of these amino acids increases due to the *rotameric* variations of the amino acid (roughly, different 3D arrangements about central bonds within the sidechain — see Section 3.4).

### Rigid Proline

Proline is a nonpolar amino acid as well. In contrast to glycine residues, which allow a great deal of conformational flexibility about the backbone (i.e., wide range of sterically-permissible rotations  $\phi$  and  $\psi$  about the peptide bond — see Section 3.4), flexibility in proline residues is largely limited, due to the cyclic nature of its sidechain.

### Aliphatic Hydroxyl R: Ser, Thr

Serine and threonine contain aliphatic hydroxyl groups and are considered uncharged polar, capable of forming hydrogen bonds (see Box 3.2).

### Acidic R and Amide Derivatives: Asn, Gln, Asp, Glu

Similarly, asparagine and glutamine possess amide groups and are also considered uncharged polar with potential for hydrogen bond formation. Their acidic analogs, aspartic acid and glutamic acid, are negatively charged (intrinsic pH of around 4) and thus considered charged polar, but the polar end of their sidechains is separated from  $C^\alpha$  by hydrophobic  $CH_2$  groups.

### Basic R: Lys, Arg, His

Lysine, arginine, and histidine have basic sidechains and are thus in the charged polar category of amino acids. Lysine and arginine — the longest amino acids — are positively charged at physiological concentrations (that is, sidechain pH of 10–12), whereas histidine's charge can be both positive or negative depending on its environment. This duality in histidine stems from its imidazole ring, which is in the physiological range of pH. For this reason, histidine residues serve as good metal binders and are often found in the active sites of proteins.

### Aromatic R: Phe, Tyr, Trp

The amino acids with aromatic sidechains — phenylalanine, tyrosine, and tryptophan — have significant potential for electrostatic interactions due to an electron deficit in the ring hydrogen atoms. Phenylalanine is highly hydrophobic while the

other two can be considered mildly hydrophobic, since their aromaticity is juxtaposed with polar properties (hydroxyl group of tyrosine and indole-ring nitrogen of tryptophan). The aromatic rings of this amino acid family also have potential for electron transfer. They can all be classified as nonpolar, though the mild hydrophobicity of tyrosine often warrants its classification as an uncharged polar amino acid.

### Sulfur-Containing R: Met, Cys

Finally, nonpolar cysteine and methionine contain sulfur in their sidechains and are thus hydrophobic. Cysteine, in particular, is very reactive and binds to heavy metals. It has an important role in protein conformations through its unique ability to form *disulfide bonds* between two cysteine residues. Disulfide bonds are covalent but reversible and are thought to be important in many cases by directing a protein to its native structure and maintaining this functionally-important state.

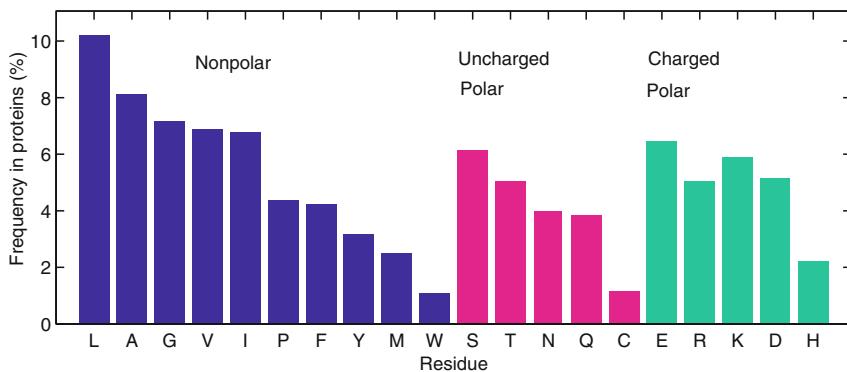


Figure 3.8. Amino acid frequencies as computed from 45,137 proteins collected from 15 taxa representing the three kingdoms of life (Bacteria, Archaea, and Eukaryota) [615]. See Table 3.1 for the frequencies and key to the one-letter amino acid abbreviations.

## 3.3 Sequence Variations in Proteins

From the constituent library of twenty natural amino acids,  $20^N$  sequence combinations for an  $N$ -residue peptide are possible, an enormous number when  $N$  is several hundred. However, natural evolution has favored certain sequences more than others. Sequence similarity is an important factor in indicating common evolutionary ancestry of proteins, as discussed below. It is therefore widely used as a tool for classifying proteins into families, as well as for relating sequence to structure and predicting structure from sequence (*homology modeling* [16, 79], as introduced in Chapter 2).

### 3.3.1 Globular Proteins

In most proteins, the twenty amino acids occur at roughly similar frequencies. Notable exceptions occur for certain amino acids like methionine, which is frequently found at the N-terminus of the peptide since it serves as the amino acid initiator of synthesis, or special groups of proteins, such as membrane or fibrous proteins.

Table 3.1 shows the frequency of occurrence of amino acid residues in the PDB40 dataset of 971 domains of unrelated proteins with a sequence identity of 40% or less [1959], and Figure 3.8 displays the data as histograms. We see that nonpolar Ala and Leu (boldfaced entries in the table) have the highest percentages (above 8%) within the representative protein database. The lowest frequencies (4% and below) occur for Trp (aromatic sidechain), Cys (sulfur-containing sidechain), His and Met (the other sulfur-containing sidechain), Tyr and Phe (aromatic sidechains also), and Gln.

Table 3.1. Amino acid frequencies in proteins based on the data of [615] which analyzed 45,137 proteins from 15 taxa. Bold and italicics types are used, respectively, for the highest ( $>8\%$ ) and lowest ( $\leq 2.5\%$ ) frequencies.

Amino Acid	Freq. [%]
<b>Alanine</b> (Ala, A)	<b>8.1</b>
Arginine (Arg, R)	5.1
Asparagine (Asp, D)	5.2
Aspartic acid (Asn, N)	4.0
<i>Cysteine</i> (Cys, C)	1.2
Glutamine (Gln, Q)	3.8
Glutamic acid (Glu, E)	6.5
Glycine (Gly, G)	7.2
<i>Histidine</i> (His, H)	2.2
Isoleucine (Ile, I)	6.8
<b>Leucine</b> (Leu, L)	<b>10.3</b>
Lysine (Lys, K)	5.9
<i>Methionine</i> (Met, M)	2.5
Phenylalanine (Phe, F)	4.2
Proline (Pro, P)	4.3
Serine (Ser, S)	6.2
Threonine (Thr, T)	5.1
<i>Tryptophan</i> (Trp, W)	1.1
Tyrosine (Tyr, Y)	3.2
Valine (Val, V)	6.9

### 3.3.2 Membrane and Fibrous Proteins

Membrane proteins are embedded in a dynamic lipid bilayer environment, where mobility is more restricted. They therefore have more hydrophobic residues than

globular proteins, which favor polar groups on the exterior surface. Since membrane proteins are particularly difficult to crystallize, simulation work is especially important in this area to understand their detailed function.

Fibrous, or structural, proteins tend to have repetitive sequences. The triple-stranded collagen helix, for example, is composed of repeating triplets which include glycine as the first residue and often proline as one or both of the other residues of the triplet. A model of collagen is shown in Figure 3.9.

Since collagen is needed to rebuild joint cartilage, there are important practical applications to skin and bone ailments. For example, a gelatin-containing (glucosamine and calcium-enriched) powdered drink mix called *Knox NutraJoint* is being touted as a dietary supplement that helps maintain healthy joints and bones. ('Juice Your Joints' touts an ad featuring an athletic sexagenarian water skier). Gelatin is rich in two amino acids, glycine and proline, that make up collagen.) Even though our bodies make these two amino acids, manufacturers claim that this gelatin-containing supplement may be helpful in decreasing the progression of osteoarthritis, a condition caused by cartilage deterioration.

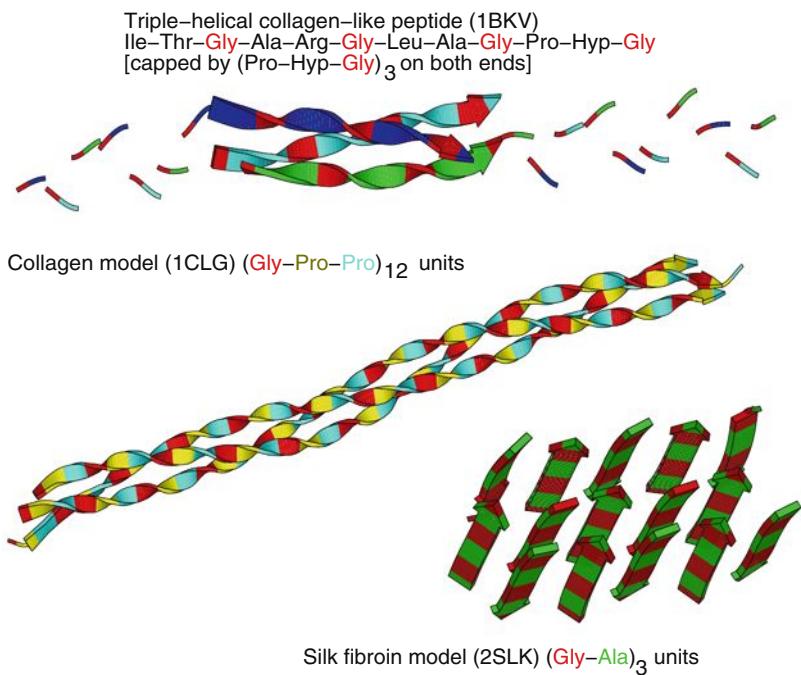


Figure 3.9. Models of the fibrous proteins collagen (triple helix) and silk, along with a crystallographically-determined collagen-like peptide (Hyp denotes hydroxyproline).

Another use of collagen is in a skin product used to heal wounds such as from venous skin ulcers, burns, and skin surgery. In May 1998 the FDA approved Apilgraf, a product for treating venous skin ulcers made of human skin cells mixed with collagen cells from cattle.

Silk is another example of a fibrous protein with a repetitive sequence. The product of many insects and spiders, silk is the polypeptide  $\beta$ -keratin composed largely of glycine, alanine, and serine residues, with smaller amounts of other amino acids such as glutamine, tyrosine, leucine, valine, and proline. The softness, flexibility, and high tensile strength of silk stems from its unique arrangement of loose hydrogen bonding networks in the form of  $\beta$ -sheets connected by  $\beta$ -turns, a mixture of both highly-ordered and less densely-packed regions. Figure 3.9 shows a model of the repetitive  $\beta$ -sheet network of silk (without connecting regions).

### 3.3.3 Emerging Patterns from Genome Databases

As genome sequencing projects are completed, interesting findings on enzyme sequences also emerge. For example, the genome of the tuberculosis bacterium (completed in 1998 by the Wellcome Trust Genome Campus of the Sanger Institute in collaboration with the Institut Pasteur in Paris) revealed surprisingly that, unlike other bacteria, repetitive gene families of glycine-rich proteins exist in *M. tuberculosis*; these approximately 10% of the enzyme-coding sequences are associated with gene families involved in anaerobic respiratory functions.

### 3.3.4 Sequence Similarity

#### Sequence Similarity Generally Implies Structure Similarity

As mentioned above, sequence similarity generally implies structural, functional, and evolutionary commonality. Thus, for example, if we were to scan the Protein Databank (PDB) randomly and find two proteins with low sequence identity (say less than 20%), we could reasonably propose that they also have little structural similarity. Such an example is shown in Figure 3.12 for the **cytochrome/barstar pair**. Similarly, large sequence similarity generally implies structural similarity (see introduction in 2.1.2 of Chapter 2).

In general, small mutations (e.g., single amino acid substitutions) are well tolerated by the native structure, even when they occur at critical regions of secondary structure. The small protein **Rop** (Repressor of primer), which controls the mechanism of plasmid replication, provides an interesting subject to both this *sequence-implies-structure* paradigm, and to exceptions to this rule (discussed below).

Rop is a dimer, with each monomer consisting of two antiparallel  $\alpha$ -helices connected by a short turn; it dimerizes to form a four-helix bundle as active form, as shown in Figure 3.10. (Fold details and motifs are discussed in the next chapter). Recall that Rop was used as the basis for solving Paracelsus challenge (Chapter 2) because the  $\alpha$ -helix motif was thought to be quite stable.

The high stability of Rop emerged surprisingly from experiments of Castagnoli *et al.* [205]. When these researchers deleted just a few residues in a key turn region that produces the overall bundle fold in the native Rop structure, they expected one long contiguous helix to form. Instead, their tinkering produced a small variation

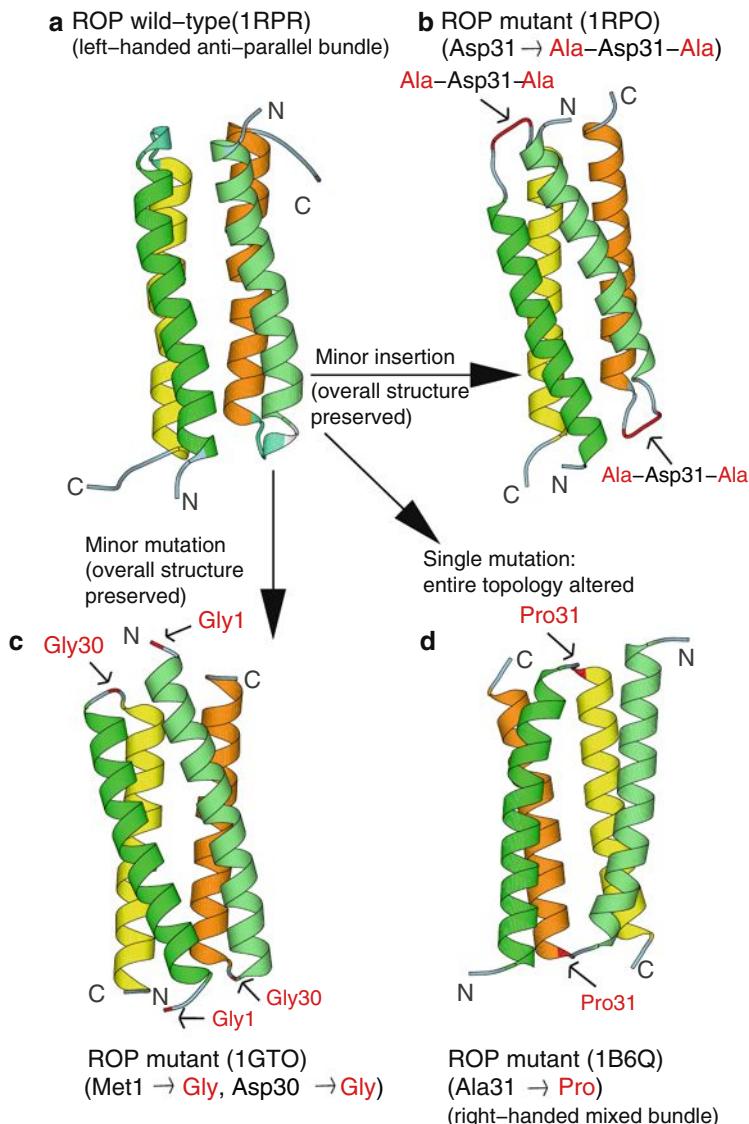


Figure 3.10. The protein **Repressor of Primer** (63 residues per monomer) provides interesting examples of the paradigm of structure inference by sequence similarity: the four-helix bundle motif of the wildtype (a) can be both *structurally stable*, i.e., resistant to mutations — as shown by the two variants in (b) and (c) — or *structurally fragile* and highly sensitive to mutations, caused by proline substitution at the turn region — as shown in (d), a mutant with an entirely different topology [462]. In (b), two Ala residues were inserted at both sides of the amino acid Asp in the loop region. In (c), the Asp residue connecting the two  $\alpha$ -helices of each monomer was mutated to Gly, and Met1 was changed to Gly [205].

of the original bundle motif. Apparently, the four-helix bundle motif is so stable that a *new turn* was formed from residues that used to be part of the  $\alpha$ -helix backbone! Thus, the original bundle motif, though slightly smaller, was maintained in the mutants. This is seen in Figure 3.10, which displays the wildtype enzyme structure (a) and those of two mutants in the above cited study (b and c).

Though this experiment supports the general notion that protein *structures* are remarkably stable to tinkering (mutations), we emphasize that *functional* properties of proteins are fragile and quite sensitive to sequence changes.

### Exceptions Exist

There are many exceptions, however, to this simple sequence/structure/function relationship.

Namely, examples exist where despite *large sequence similarity* there is *small structural and functional similarity*. A classic example of this relationship is the disease sickle-cell anemia, where a minute substitution in sequence leads to altered function with devastating consequences. This abnormality results from the replacement of the highly-polar glutamate residue in **hemoglobin** by the nonpolar amino acid valine. This key substitution at the surface of the protein leads to an entirely different quaternary structure for this multidomain red-blood pigment protein. This is because the markedly altered structure affects the solubility of oxygenated hemoglobin and leads to a clumping of the deoxygenated form of the molecule (HbS instead of HbA).

Conversely, examples exist where despite *small sequence similarity* there is *large structural*, and even functional and evolutionary similarity. A classic example for this relationship is the **myoglobin/hemoglobin** pair of proteins (see Figure 3.12). These proteins only share 20% of the sequence. However, as oxygen-carrying molecules, they share structural, functional, and evolutionary similarity. Proteins in the **calmodulin** family are also known to display a great deal of structural variability for similar sequences [664] (see Figure 3.11).

More generally, changes in 3D architecture (despite a nontrivial degree of sequence similarity) can result from a variety of factors, as follows.

- Mutations in *critical regions* of the proteins, such as active sites and ligand binding sites, can change 3D structures dramatically. Such an example is shown for the pair of **immunoglobulins** in Figure 3.12.
- Mutations in regions that *connect two secondary-structural elements* can also be responsible for structural divergence, as in the helix-loop-helix motif of the **EF-hand family**, and the connecting loops in helix bundles.

Figure 3.11 illustrates this principle for the two EF-hand calcium-binding proteins **calmodulin** and **sarcoplasmic calcium-binding** protein: one is overall extended in shape while the other is more compact [979].

Helix bundles are sensitive to mutations in loop or turn regions that connect different helices, to the extent that a single amino acid substitution (alanine

to proline) can change the topology of a homodimeric 4-helical bundle protein from the canonical left-handed all-antiparallel form to a right-handed mixed parallel and antiparallel bundle [462]. Figure 3.10(d) shows this different resulting topology of the **Rop four-helix bundle** subject to the single mutation Ala31→Pro at the turn region.

- Structural variations can be observed in the same system determined at different *environmental conditions* such as solvent or crystal packing. The same **T4-lysozyme** mutants in Figure 3.12 (100% sequence similarity) display intriguing mobility, adopting 5 different crystal conformations [374] due to a hinge bending motion.
- *Multidomain proteins* can adapt quaternary structures that depend sensitively on the number of subunits and/or on the sequence.

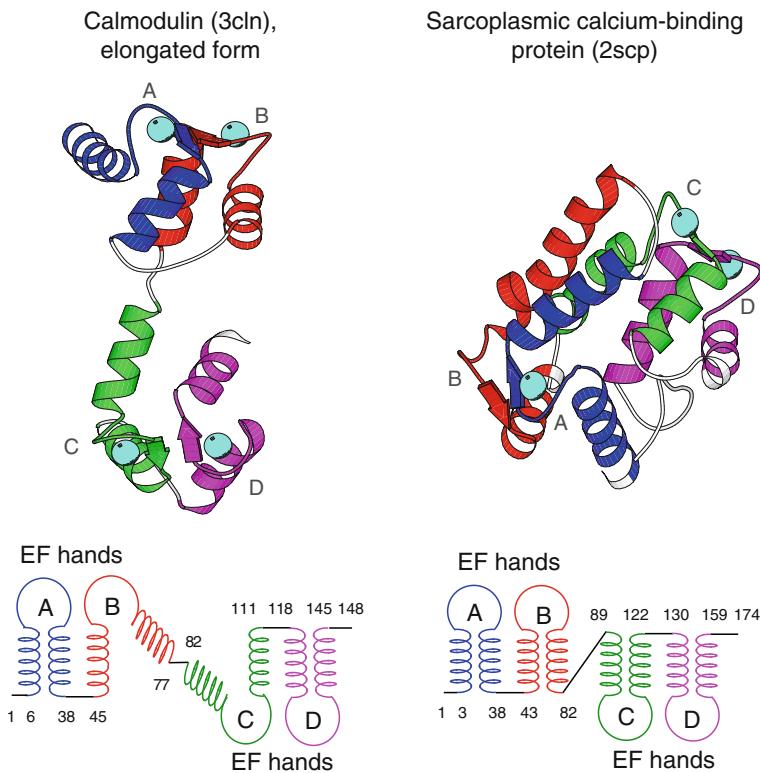


Figure 3.11. Structural variability despite large similarity in the protein secondary-structural elements is illustrated for two calcium-binding proteins — **calmodulin** (148 residues) and **sarcoplasmic calcium binding protein** (174 residues) [979] — due to different overall 3D arrangement of the shared motifs. Though sharing only 30% of the sequence, both proteins are made of two repeating units, each consisting of two EF hand motifs. Each hand motif contains two helical regions surrounding a calcium binding loop (crystal-bound calcium atoms are rendered as large spheres; only three are bound to 2scp).

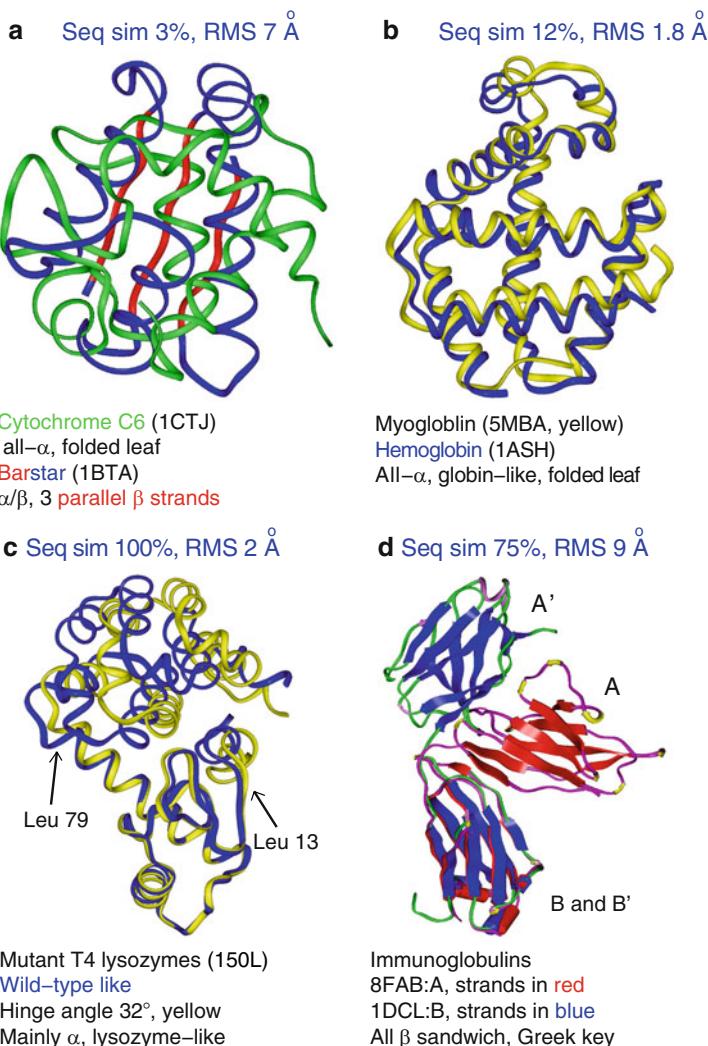


Figure 3.12. Various examples of sequence/structure relationships in proteins: (a) Low sequence similarity (3% for alignment of 72% of the residues) generally implies low structure similarity (**cytochrome C6** versus **barstar**). Still, exceptions are found. For example, in (b), despite low (12%) sequence similarity, there is large structure and function similarity (**hemoglobin** and **myoglobin**); conversely, despite high sequence similarity, there can be structural diversity, due to (c) hinge bending in two **lysozyme mutants** (Met → Ile in residue 6) or (d) different orientation of one of the two subunits in two **immunoglobulins**. The lysozyme mutant displays 5 different crystal conformations, one similar to the wild-type (shown in **blue**) and others overall very similar except for a different hinge-bending angle (see defining arrows); the form with largest bend (32°) is shown in **yellow**. The two immunoglobulins differ markedly in tertiary organization due mainly to differences in the linker domain between the A and B subunits of each protein.

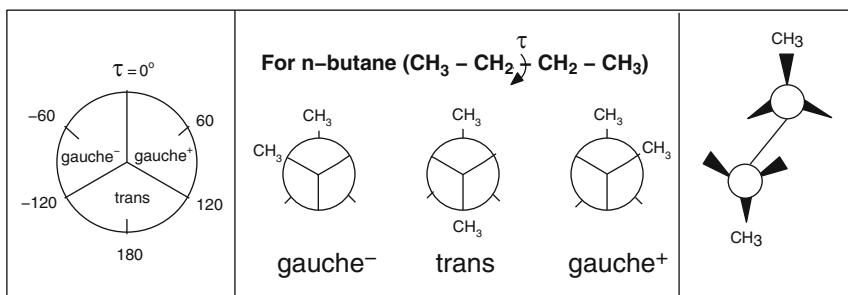


Figure 3.13. *Gauche* (g) and *trans* (t) dihedral-angle orientations for *n*-butane: (left) classification wheel; (middle) simple Newman projections that illustrate the three favored orientations of the two end methyl groups about the central C–C bond (perpendicular to the plane of the paper); and (right) the *trans* conformation, which has the least steric clashes.

## 3.4 Protein Conformation Framework

### 3.4.1 The Flexible $\phi$ and $\psi$ and Rigid $\omega$ Dihedral Angles

Polypeptides can have a wide variety of *conformations*, i.e., 3D structures differing only in rotational orientations about covalent bonds.<sup>2</sup> This type of rotational flexibility is characterized by a *dihedral angle*, which measures the relative orientation of four linked atoms in a molecule,  $i-j-k-l$ . A dihedral angle for a 4-atom sequence that is not necessarily covalently bonded can also be used for special terms in the potential energy function; see Chapter 9 for examples. See Box 3.4, Figure 3.14, and the equations in Appendix D, under the Addendum to Assignment 8 for a definition of a dihedral angle.

---

#### Box 3.4: Dihedral Angle

The dihedral angle  $\tau_{ijkl}$  defined for a sequence of linked atoms  $i-j-k-l$  (Figure 3.14) is the angle between the normal to the plane of atoms  $i-j-k$  and the normal to the plane of atoms  $j-k-l$ . The sign of  $\tau_{ijkl}$  is determined by the triple product  $(a \times b) \cdot c$ , where  $a$ ,  $b$  and  $c$  are the interatomic distance vectors for atoms  $i \rightarrow j$ ,  $j \rightarrow k$  and  $k \rightarrow l$ , respectively.

Strictly defined, the related *torsion angle*,  $\hat{\tau}$  is the angle between the two planes defined by  $i-j-k$  and  $j-k-l$ . Thus,  $\tau + \hat{\tau} = \pi$  ( $180^\circ$ ). However, the terms torsion and dihedral angle are often used interchangeably. We will often use *dihedral angle* to refer to the numerical value of the angle, and *torsion angle* or *torsional potential* when we discuss general properties of these rotations.

---

<sup>2</sup>see Chapter 8, Subsection 8.4.1, for the related definition of *configuration*.

When the dihedral angle is  $0^\circ$ , the four atoms  $i-j-k-l$  are coplanar and atoms  $i$  and  $l$  coincide in their projections onto the plane normal to the  $j-k$  bond; this orientation is defined as *cis* or *syn*. When the dihedral angle is  $180^\circ$ , the atoms are coplanar but atoms  $i$  and  $l$  lie opposite one another in the projection onto the plane normal to the  $j-k$  bond; such an orientation is defined as *trans* or *anti*. More generally, angular regions convenient to describe protein and nucleic acid conformations are the following: *cis* ( $\approx 0^\circ$ ), *trans* ( $\approx 180^\circ$ ) and  $\pm$  *gauche* ( $\approx \pm 60^\circ$ ). Another common terminology is: *syn* ( $\approx 0^\circ$ ), *anti* ( $\approx 180^\circ$ ),  $\pm$  *synclinal* ( $\approx \pm 60^\circ$ ), and  $\pm$  *anticlinal* ( $\approx \pm 120^\circ$ ). See Figure 3.13 for a simple illustration for *n*-butane.

While the peptide group (Figure 3.3) is relatively rigid — it has 40% double-bond character — there is a great deal of flexibility about each of the single bonds along the backbone,  $\{N-C^\alpha\}$  and  $\{C^\alpha-C\}=O$ . The two dihedral angles  $\phi$  and  $\psi$  are used to define rotations about the bond between the nitrogen and  $C^\alpha$  of the mainchain and between  $C^\alpha$  and the carbonyl carbon, respectively (Figure 3.15).

The dihedral angle  $\omega$  defines the rotation about the peptide bond, namely for the atomic sequence  $C_1^\alpha-\{C-N\}-C_2^\alpha$ , where  $C_1$  and  $C_2$  are the  $\alpha$ -carbons of two adjacent amino acids. Because of the partial double-bond character of the peptide bond and the steric interactions between adjacent sidechains,  $\omega$  is typically in the *trans* configuration:  $\omega = 180^\circ$ .<sup>3</sup> In this orientation, all four atoms lie in the same plane, with the distance between  $C_1^\alpha$  and  $C_2^\alpha$  as large as possible (see Figure 3.13 for a definition of various dihedral angle orientations).

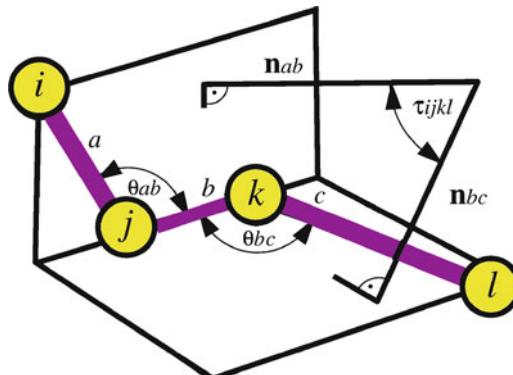


Figure 3.14. Definition of a dihedral angle  $\tau_{ijkl} = \cos^{-1}(\mathbf{n}_{ab} \cdot \mathbf{n}_{bc})$ , the angle between the two normals spanned by atoms  $i, j, k$  and  $j, k, l$ .

<sup>3</sup>Non-trivial deviations from planar peptide bonds can be shown by theory and experiment (e.g., as reviewed in [348]). A statistical survey of peptide and protein databases verified that the distribution of rotation angles (or energies associated with peptide bond rotations) follows Boltzmann statistics [799].

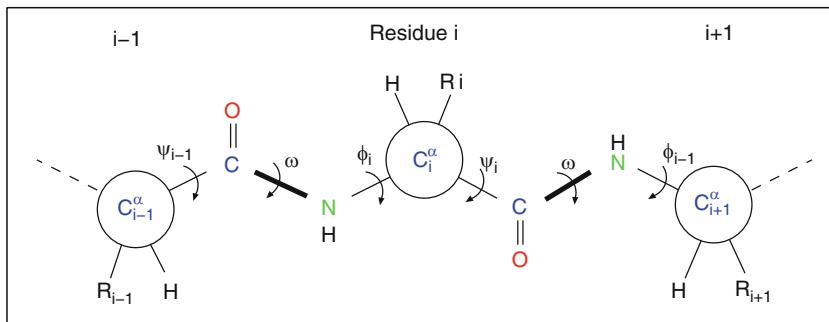


Figure 3.15. Rotational flexibility in polypeptides: definition of the  $\phi$ ,  $\psi$  and  $\omega$  dihedral angles.

### 3.4.2 Rotameric Structures

Besides the  $\{\phi, \psi\}$  flexibility associated with the two backbone bonds involving  $C^\alpha$ , multiple conformations are possible for 18 of the 20 amino acids when the sidechain geometries differ (excluded are glycine and alanine). *Rotameric structures* of amino acids (and hence proteins) are those that have the same  $\{\phi, \psi\}$  angles but differ in the sidechain conformations. The dihedral angles used to define sidechain rotations are denoted by  $\chi$ , with subscripts used as needed (see Figure 3.17).

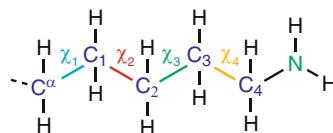


Figure 3.16. Lysine's four rotamers defined by torsional variables  $\chi_1$  through  $\chi_4$ .

For example, in lysine, whose sidechain has four carbons (see Figure 3.16), dihedral angles  $\chi_1$  through  $\chi_4$  denote the rotations about bonds  $C^\alpha-C_1$ ,  $C_1-C_2$ ,  $C_2-C_3$ , and  $C_3-C_4$ , respectively (see also Figure 3.17 for other amino acids). Rotameric structures for polypeptides and proteins depend on the environment of the polymer and on the secondary and tertiary structures.

### 3.4.3 Ramachandran Plots

The feasible combinations of the  $\phi$  and  $\psi$  angles are limited due to steric hindrance. That is, only certain combinations are typically observed, with some dependence on residue size and shape. Glycine is unique in its flexibility — it is therefore a good agent for turns in polypeptides and proteins — but other residues exhibit a highly limited range of sterically-permissible  $\phi$  and  $\psi$  combinations. In fact, only roughly one tenth the area of the  $\{\phi, \psi\}$  space is generally observed for polypeptides and proteins. Among the first to note this limitation were

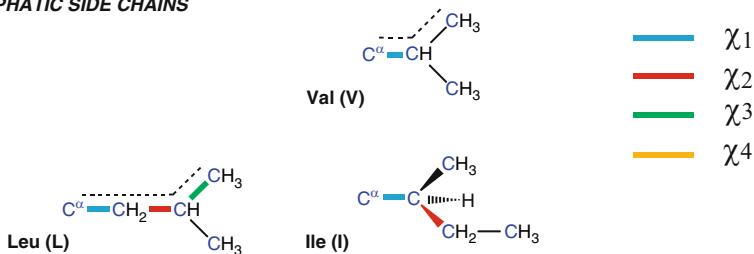
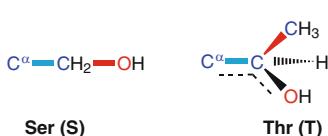
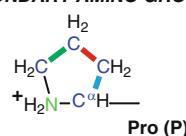
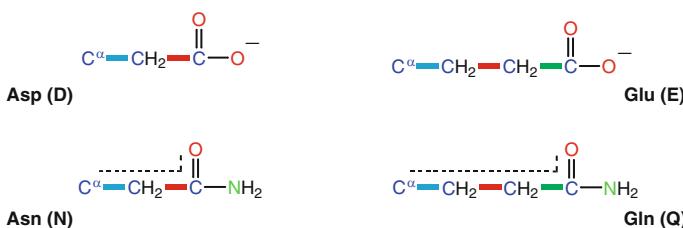
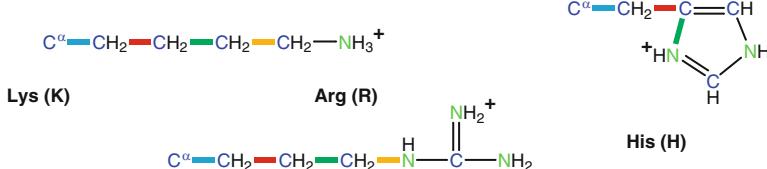
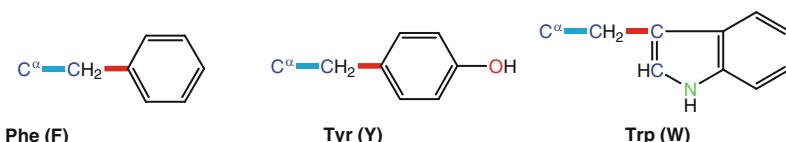
**ALIPHATIC SIDE CHAINS****ALIPHATIC HYDROXYL SIDE CHAINS****SECONDARY AMINO GROUP****ACIDIC SIDE CHAINS AND THEIR AMIDE DERIVITIVES****SULFUR-CONTAINING SIDE CHAINS****BASIC SIDE CHAINS****AROMATIC SIDE CHAINS**

Figure 3.17. Rotameric notation used for 18 of the 20 amino acids is illustrated using different colors for  $\chi_1$ – $\chi_4$ , as shown in the top right key.

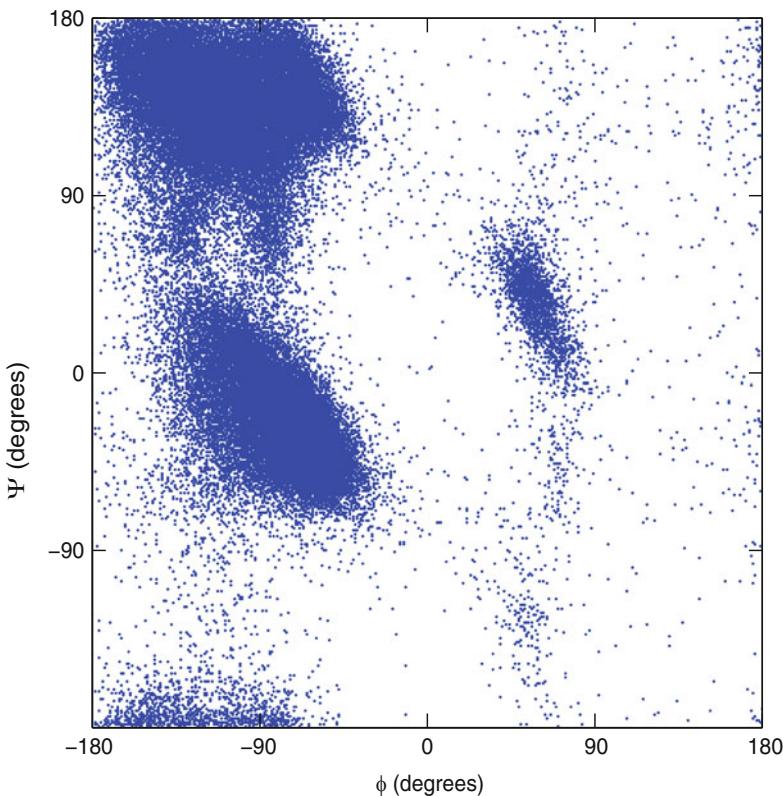


Figure 3.18. Ramachandran plots, obtained from a subset of the PDB40 dataset [959], corresponding to X-ray protein structures with resolution of 2.5 Å or better (470 proteins, 95778 total residues plotted, with proline and glycine excluded).

G.N. Ramachandran<sup>4</sup> and coworkers in 1963, after which *Ramachandran plots* are called. Around the same time, John Schellman and coworkers were working independently along the same lines of mapping the energetically favorable and excluded regions for protein conformations [1098].

These diagrams in the  $\{\phi, \psi\}$  space, as shown in Figure 3.18, are used to describe this  $\{\phi, \psi\}$  flexibility (actually inflexibility) in polypeptides and proteins. See also Figure 3.19 for a comparative view of Ramachandran plots derived from the moderate-resolution X-ray structures shown in Figure 3.18 versus high-resolution X-ray as well as NMR-derived structures.

Often, Ramachandran diagrams are presented by plotting the backbone dihedral angles of all nonterminal residues in a protein for a large group of known protein

---

<sup>4</sup>For lovers of scientific history, the 1998 biography of this renowned Indian molecular biophysicist (1922–2001) is recommended [1091]. His peppered poetry is an added bonus. (My favorite poem is number 9, on Superhelical Twisting and Replication of D N A [1091, p. 159].)

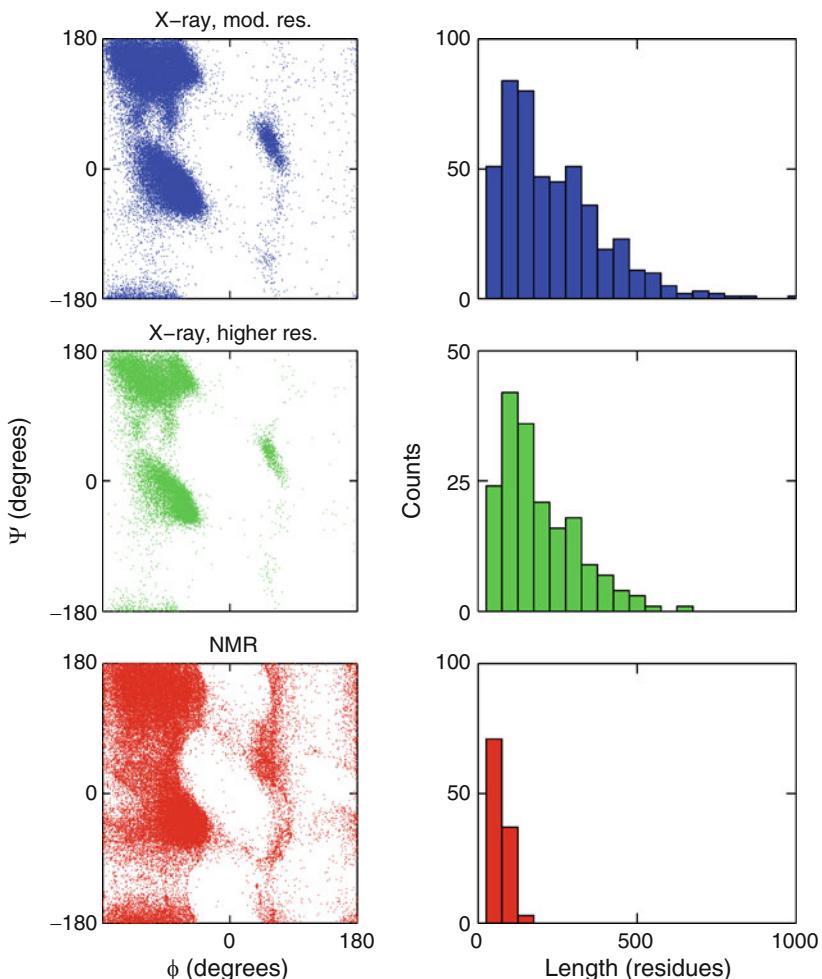


Figure 3.19. Three sets of Ramachandran plots based on the PDB40 dataset [959], corresponding to: (top) X-ray protein structures with resolution of 2.5 Å or better (470 proteins, 95,778 total residues plotted, with proline and glycine excluded); (middle) X-ray protein structures with resolution of 1.8 Å or better (183 proteins, 29,758 total residues plotted, proline and glycine excluded); and (bottom) NMR-derived structures (113 proteins, 84,719 total residues plotted, proline and glycine excluded). For each subset of structures, the length distribution is also shown.

structures. This superimposed view, averaged over many residues, approximates protein conformational tendencies. The favorable regions correspond to common secondary-structure elements, such as helices and sheets, with finer motifs also noted.

In addition to favorable combinations of  $\phi$  and  $\psi$  in polypeptides, the side-chain dihedral angle  $\chi_1$  has been found to cluster around one of three conformers

known as *gauche*<sup>+</sup> (or  $g^+$ ,  $\chi_1 = +60^\circ$ ), *gauche*<sup>-</sup> (or  $g^-$ ,  $\chi_1 = -60^\circ$ ), and *trans* (or  $t$ ,  $\chi_1 = 180^\circ$ ). These are the favored orientations about tetrahedral atoms (Figure 3.13). Some dependence of  $\chi_1$  on the residue's  $\phi$  and  $\psi$  values has also been noted.

### 3.4.4 Conformational Hierarchy

Most natural proteins adopt specific 3D structures that are associated with their biological activity. Of course, proteins are dynamic, but typical thermal fluctuations and local configurational arrangements revolve around a specific globally-folded structure. The majority of proteins is believed to be unknotted in a topological sense, though the polypeptide chain is frequently covalently bonded via disulphide links and noncovalently held together by hydrogen bonds [1254].

One of the hallmarks of biomolecular structure is that the amino acid sequence determines the 3D structure of a protein. This was first shown by Christian B. Anfinsen and his colleagues in the early 1960s [51]. Anfinsen shared the Nobel Prize in Chemistry in 1972 for his work on ribonuclease — connecting the amino acid sequence to the biologically active conformation — with Stanford Moore and William H. Stein — who connected ribonuclease's chemical structure to its catalytic activity.<sup>5</sup> In Anfinsen's work, the protein ribonuclease was denatured by destroying its hydrogen bonding network as well as intrinsic disulfide bonds. The researchers observed that the protein spontaneously refolded into its native state in a short time, regaining all its enzymatic activity. Of course, we recognize now that accessory chaperone molecules may be necessary to assist in the folding of many large proteins *in vivo*, as discussed in Chapter 2.

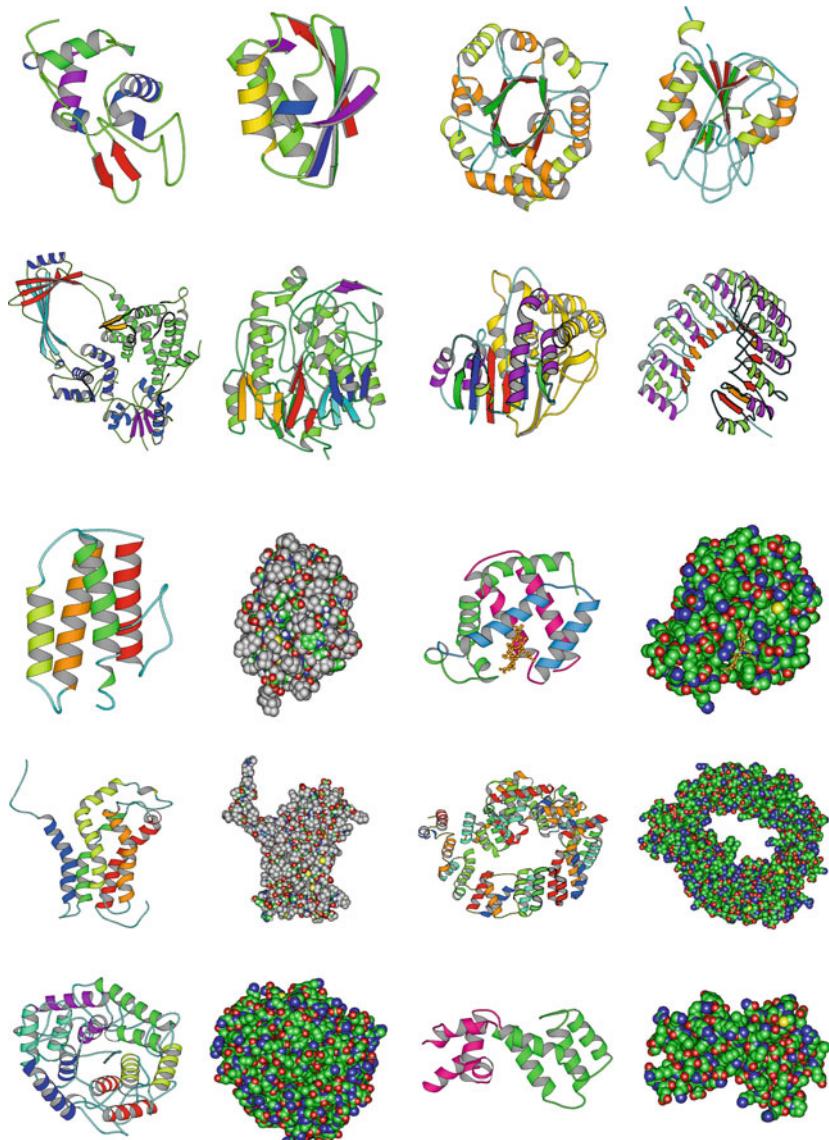
Four basic levels are used to describe protein structure:

- *primary structure* — the sequence of amino acids;
- *secondary structure* — regular local structural patterns such as  $\alpha$ -helices and  $\beta$ -sheets, or combination motifs thereof (*supersecondary structure*);
- *tertiary structure* — the 3D arrangement of all atoms in the polypeptide chain in space; and
- *quaternary structure* (used for large proteins with independent subunits) — the complete 3D interaction network among the different subunits.

The next chapter describes in turn the secondary and supersecondary, tertiary, and quaternary structure of proteins.

---

<sup>5</sup>Readers are invited to browse the electronic museum of Profiles in Science at [wwwprofiles.nlm.nih.gov/](http://wwwprofiles.nlm.nih.gov/) for a glimpse not only of Anfinsen's scientific activities but also of his other hobbies and interests.



# 4

## Protein Structure Hierarchy

Chapter 4 Notation

SYMBOL	DEFINITION
$\alpha_R$	classic right-handed $\alpha$ -helix
$\alpha/\beta$	protein class
$\alpha + \beta$	protein class
$\beta$ -sheet	aggregating amino-acid strands
$\beta_2$	hairpin motif
$\beta_4$	Greek key motif
$C^\alpha$	$\alpha$ -Carbon
$\pi$	helix form looser than $\alpha_R$
$\phi$	$\{N-C^\alpha\}$ rotation about peptide bond
$\psi$	$\{C^\alpha-C\}=O$ rotation about peptide bond
$3_{10}$	helix form tighter than $\alpha_R$

Try to learn something about everything and everything about something.

Thomas Henry Huxley (1825–1895).

## 4.1 Structure Hierarchy

The complexity of protein structures requires a description of their structural components. This chapter describes the elements of protein secondary structure — regular local structural patterns — such as helices, sheets, turns, and loops. Helices and sheets tend to fall into specific regions in the  $\{\phi, \psi\}$  space of the Ramachandran plot (see Figures 3.18 and 3.19). The corresponding width and shape of each region reflects the spread of that motif as found in proteins.

Following this description of each secondary structural element, we discuss the basic four *classes* of protein supersecondary or tertiary structure (the 3D spatial architecture of a protein), namely  $\alpha$ -proteins,  $\beta$ -proteins,  $\alpha/\beta$ -proteins, and  $\alpha + \beta$ -proteins. This is followed by a presentation of the *fold* motifs for each such class. Classes and folds are at the top of protein structure classification, as introduced in the last section. Describing these folds and structural motifs is far from an exact science, so variations in some of these aspects are common.

## 4.2 Helices: A Common Secondary Structural Element

### 4.2.1 Classic $\alpha$ -Helix

In the classic, right-handed  $\alpha$ -helix ( $\alpha_R$ ), a *hydrogen bonding* network connects each backbone carbonyl (C=O) oxygen of residue  $i$  to the backbone hydrogen of the NH group of residue  $i + 4$  (see Figure 4.1). This hydrogen bonding provides substantial stabilization energy.

The regular spiral network of the  $\alpha$ -helix is ubiquitous in proteins. It is associated with a  $\{\phi, \psi\}$  pair of about  $\{-60^\circ, -50^\circ\}$ . The resulting helix has 3.6 residues per turn, and each residue occupies approximately 1.5 Å in length. The helix may be curved or kinked depending on the amino acid sequence, as well as on solvation and overall packing effects. Such distortions are reflected by the  $\{\phi, \psi\}$  distribution around the  $\alpha_R$  region in typical Ramachandran plots. **Hemoglobin**, **myoglobin**, **bacteriorhodopsin**, **human lysozyme**, **T4 lysozyme**, **Trp repressor**, and **repressor-of-primer (Rop)** are all examples of proteins that are virtually entirely  $\alpha$ -helical. See Figures 4.2 and 4.3 for illustrations of such  $\alpha$ -proteins (see below) and Figure 3.10 for Rop.

An  $\alpha$ -helix is associated with a dipole moment: the amino terminus of the helix has a positive charge and the carboxyl end has a negative charge clustered about it. Thus, residues that are negatively charged on the amino end and positively-charged on the carboxyl end stabilize the helix; residues with the opposite charge allocation destabilize the helix.

Experimental and theoretical work has shown that both intrinsic and extrinsic (inter-residue interactions) factors are important for helix formation in proteins. Residues with restricted sidechain conformations, due to long or bulky groups, are poorer  $\alpha$ -helix participants than other residues. Glutamine, methionine,

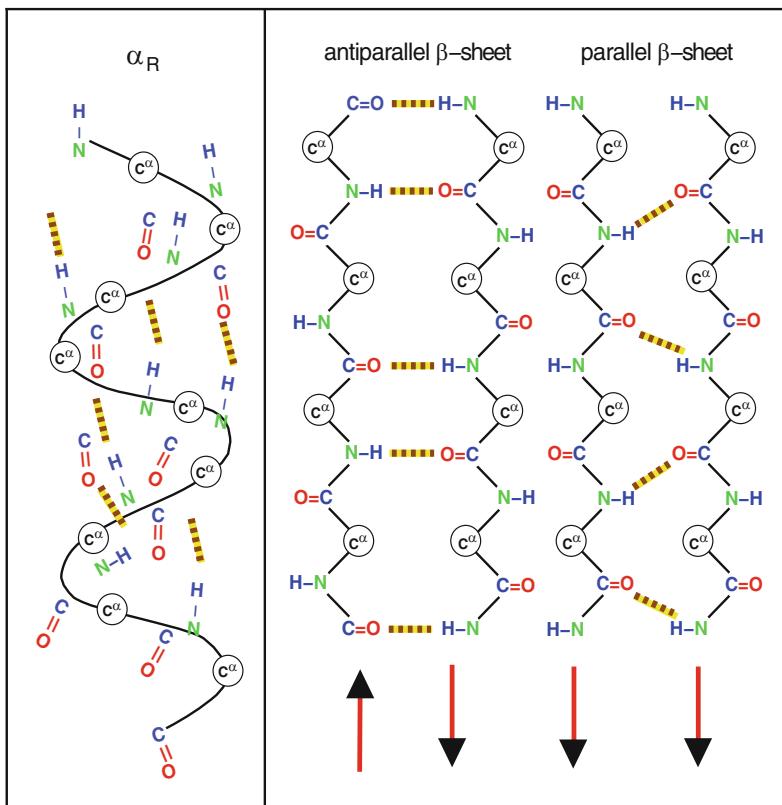


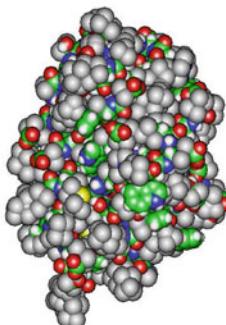
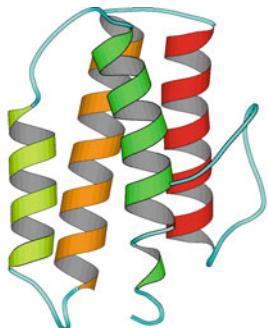
Figure 4.1. Hydrogen bonding patterns in the classic  $\alpha$ -helix ( $\alpha_R$ ), with the ribbon tracing the  $\alpha$ -carbons (left), anti-parallel  $\beta$ -sheet (middle), and parallel  $\beta$ -sheet (right).

and leucine favor  $\alpha$ -helix formation, while valine, serine, aspartic acid, and asparagine tend to destabilize  $\alpha$ -helices (e.g., due to steric and electrostatic considerations).

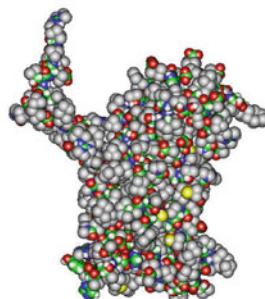
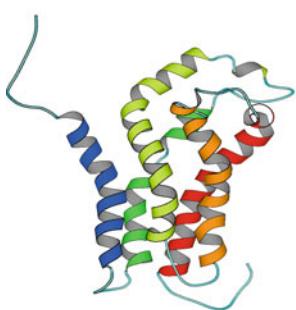
#### 4.2.2 $3_{10}$ and $\pi$ Helices

There are more common variants of the  $\alpha$ -helix motif that are typically not stable in solution but can play a part in overall protein structure. These include the tighter  $3_{10}$  and looser  $\pi$  helices, with  $\{\phi, \psi\}$  angles around  $\{-50^\circ, -25^\circ\}$  and  $\{-60^\circ, -70^\circ\}$ , respectively.

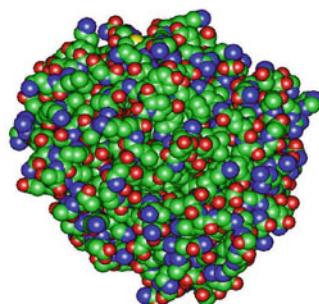
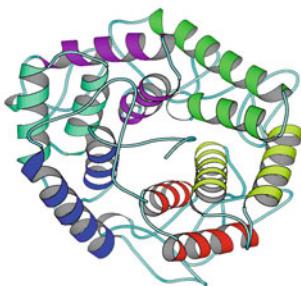
The tighter  $3_{10}$  helix of three residues per turn (instead of 3.6 in the classic  $\alpha$ -helix) involves hydrogen bonds between residues  $i$  and  $i+3$  instead of  $i$  and  $i+4$  as in  $\alpha_R$ . There are 10 atoms within the hydrogen bond; hence the nomenclature  $3_{10}$ . The more loosely coiled  $\pi$  helix has hydrogen bonds between residues  $i$  and  $i+5$  of the polypeptide.



Myohemerythrin (2MHR, 118 residues, four-helix bundle)



Pix (1BY1, 209 residues, five-helix bundle)



Cellulase Cela (1CEM, 363 residues, six-alpha hairpins)

Figure 4.2. Examples of  $\alpha$ -proteins: **myohemerythrin**, **pix**, and **cellulase cela**.

Because of their close packing,  $3_{10}$  helices generally form for a few residues only, often at the C-terminus end of classic  $\alpha$ -helices where the helix tends to tighten. Similarly, the  $\pi$  helix occurs rarely since the backbone atoms are so loosely packed that they leave a hole.

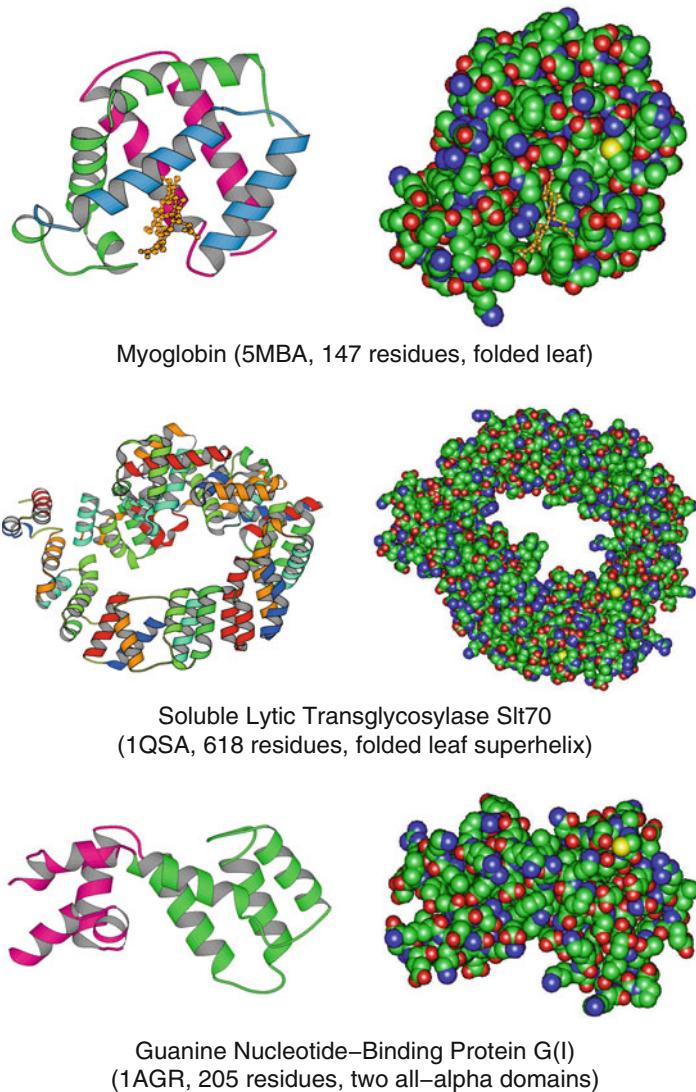


Figure 4.3. Examples of  $\alpha$ -proteins: **myoglobin**; **soluble lytic transglycosylate** protein of bacterial muramidase, in the N-terminal region of the enzyme muramidase in bacterial cell walls; and **guanine nucleotide-binding protein**, an irregular  $\alpha$ -helical protein with a fold containing a 4-helix bundle with left-handed twist.

#### 4.2.3 Left-Handed $\alpha$ -Helix

A left-handed  $\alpha$ -helix is theoretically possible, with  $\{\phi, \psi\} = \{+60^\circ, +60^\circ\}$ . However, this motif is generally unstable. The chirality preference for  $\alpha$ -helices follows the chirality of L-amino acids.

#### 4.2.4 Collagen Helix

The triple-stranded **collagen** helix is often considered a specific secondary element. It is associated with  $\{\phi, \psi\} = \{-60^\circ, +125^\circ\}$ . A large body of structural data has suggested that extensive hydration networks in the collagen triple helix (among the protein residues and with water molecules) are responsible for collagen stability and assembly (see [115, 680] and references cited therein). A recent hypothesis — that inductive effects by electron-withdrawing residue moieties might play a key factor in collagen's stability [562] — remains to be proven.

### 4.3 $\beta$ -Sheets: A Common Secondary Structural Element

Another common motif is a  $\beta$ -sheet. These sheet regions form by aggregating amino-acid strands, termed  $\beta$ -strands, via hydrogen bonds. Typical lengths of  $\beta$ -strands are 5–10 residues. The aggregation can occur in a parallel or anti-parallel orientation of the strands, as shown in Figure 4.1, each with a distinct hydrogen bonding pattern. Each such  $\beta$ -strand has two residues per turn and can be considered a special type of helix. The hydrogen bond crosslinking between strands — alternating  $\text{C}=\text{O} \cdots \text{H}-\text{N}$  and  $\text{N}-\text{H} \cdots \text{O}=\text{C}$  — is such that the sheet has a pleated appearance. Thus, in comparison to  $\alpha$ -helices,  $\beta$ -sheets require connectivity interactions that are much longer in range.

For parallel  $\beta$ -sheets,  $\phi \approx -120^\circ$  and  $\psi \approx +115^\circ$ . For anti-parallel  $\beta$ -sheets,  $\phi \approx -140^\circ$  and  $\psi \approx +135^\circ$ . As for  $\alpha$ -helices, the ring of proline does not adapt well into  $\beta$ -sheets since it cannot participate in the hydrogen bond network between strands. Valine, isoleucine, and phenylalanine have been found to enhance  $\beta$ -sheet formation.

Often, at the edges of  $\beta$ -sheets, an additional residue that cannot be included in the normal hydrogen bonding pattern produces a  $\beta$ -bulge of the extra residue. Figures 4.4 and 4.5 show the structures of proteins that are mostly  $\beta$ -sheets.

### 4.4 Turns and Loops

Other common structural motifs in proteins are turns and loops.

Turns (also called  $\beta$ -turns or reverse turns) occur in regions of sharp reversal of orientation, such as the junction of two anti-parallel  $\beta$ -strands. Such motifs are classified as turns based on distance criteria (e.g., the  $\text{C}^\alpha$  atoms of residues  $i$  and  $i + 3$  are less than 7 Å distant).

Loops occur often in short (five residues or less) regions connecting various motifs. Loop regions that connect two adjacent anti-parallel  $\beta$ -strands are known as hairpin loops. Short hairpin loops are found at protein surfaces.

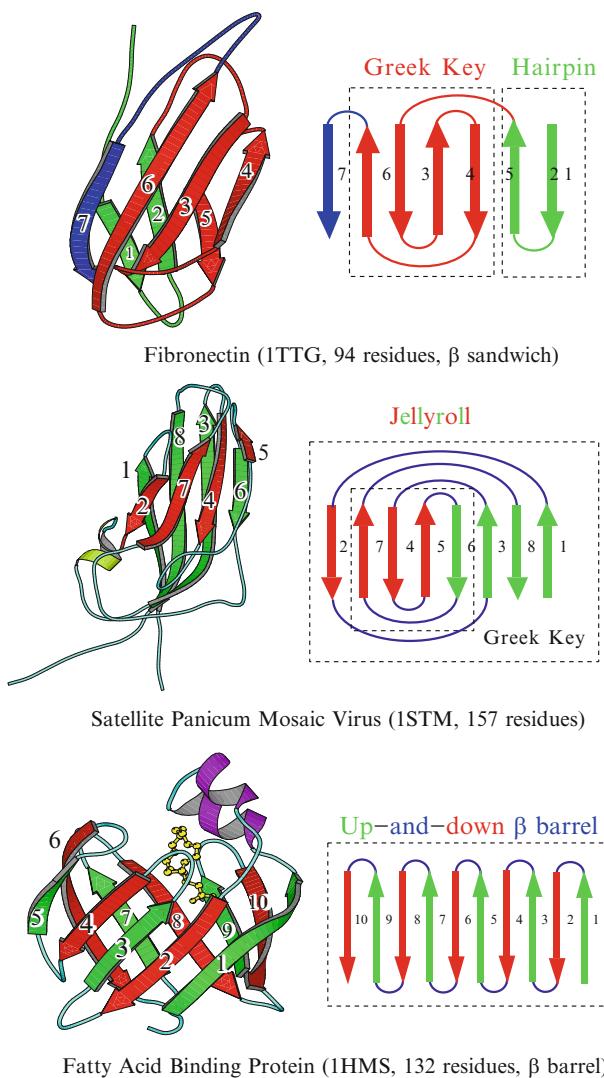


Figure 4.4. Examples of  $\beta$ -proteins and common motifs: **fibronectin**,  $\beta$ -sandwich illustrating hairpin and Greek key motifs; coat protein of **satellite panicum mosaic virus**; and **fatty acid binding protein**, up-and-down  $\beta$ -barrel.

The majority of turns and loops lies on the protein surface because of solvation considerations. They are important elements that allow, and possibly drive, protein compaction. Most loops interact with solvent and are highly hydrophilic (water soluble). Since protein core regions are more stable than short

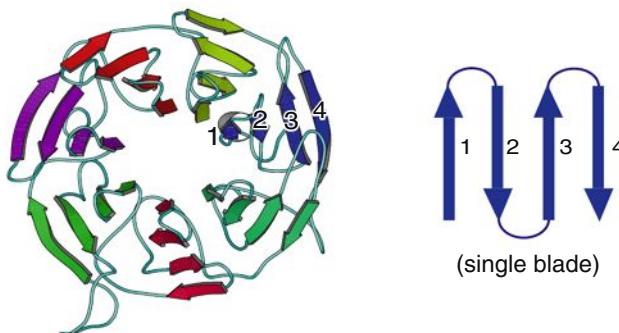
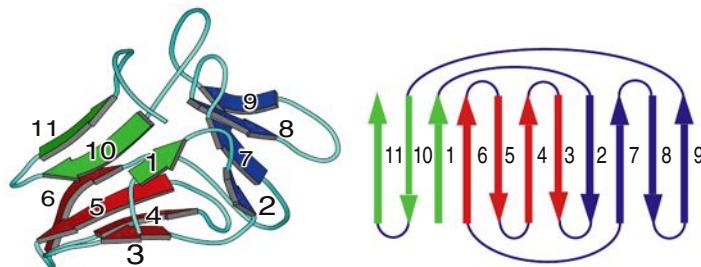
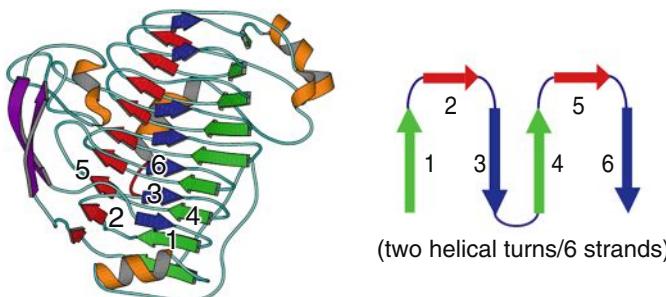
Galactose Oxidase (1GOF, 639 residues, 7-bladed  $\beta$ -propeller)Agglutinin (1BWU, 106 residues,  $\beta$ -prism)Pectin Lyase A (1IDK, 359 residues, right-handed  $\beta$ -helix)

Figure 4.5. Examples of  $\beta$ -proteins and common motifs: **galactose oxidase**, **agglutinin**, and **pectin lyase A**.

connective elements of helices and strands, evolutionary differences among homologous sequences are often localized to loop and turn regions. Non-coding regions (introns) are similarly found in genes that correspond to loops and turns in protein structures.

## 4.5 Formation of Supersecondary and Tertiary Structure

### 4.5.1 Complex 3D Networks

The secondary structural elements described above often combine into simple motifs that occur frequently in protein structures. Such motifs (or folds) are also called *supersecondary structure*. Examples are  $\beta$  hairpin ( $\beta$ -loop- $\beta$  units), Greek key, and  $\beta$ - $\alpha$ - $\beta$  units (see below).

Supersecondary and tertiary structures of proteins can be described by the specific topological arrangement of the secondary or supersecondary structural motifs. Although the 3D architecture of a protein can be a complex composite of various secondary and supersecondary structural motifs, the majority of the residues — roughly 90% — are found to be involved in secondary structural elements. In fact, on average 30% of the residues are found as helices, 30% as sheets, and 30% as loops and turns. Proteins can be monomeric or multimeric, with subunits that fold in a dependent or independent manner with respect to other domains.

The different polypeptide domains can be connected by disulfide bonds, hydrogen bonds, or the weaker van der Waals interactions. Tertiary structure is also affected by the environment. Hydrogen bonding with solvent water molecules can stabilize the native conformation, and the salt concentration can affect the compact arrangement of the folded chain.

Molecular graphics packages often display the secondary structural motifs clearly by using ribbon diagrams in which helices are depicted as coils and sheets are shown as twisting planes with arrows (see Figures 4.2, 4.3, 4.4, and 4.5, for example).

### 4.5.2 Classes in Protein Architecture

Based on the known protein structures at atomic resolution, four major *classes* can be used to describe the arrangement in space of the various secondary structural elements (or domains) of polypeptides:

- $\alpha$ -proteins — proteins which form compact aggregates by packing mainly  $\alpha$ -helices, often in a symmetric arrangement around a central hydrophobic core;
- $\beta$ -proteins — proteins which pack together mainly  $\beta$ -sheets, with adjacent strands linked by turns and loops and various hydrogen bonding networks formed among the individual strands, often resulting in layered or barrel structures;
- $\alpha/\beta$ -proteins — proteins that are folded with alternating  $\alpha$ -helices and  $\beta$ -strands, often forming layered or barrel-like structures; and
- $\alpha + \beta$ -proteins — proteins that combine largely-separated (i.e., non-alternating) helical and strand regions, often by hairpins.

Figures 4.2–4.7 illustrate members of each such class.

Recent statistics for PDB protein structures reveal that approximately 24% belong to the all- $\alpha$  class, 15% to all- $\beta$ , 12% to  $\alpha/\beta$ , and 32% to  $\alpha+\beta$ . The remaining 17% includes multidomain proteins, membrane and cell-surface proteins, and peptides, and small proteins (see Figures 4.8–4.10). For updated statistical information, check [scop.mrc-lmb.cam.ac.uk/scop/](http://scop.mrc-lmb.cam.ac.uk/scop/), click on ‘Statistics here’. (See last section of this chapter for SCOP description).

Other classes are defined for proteins found on membrane and cell surfaces, small and/or irregular proteins with multiple disulfide bridges, proteins with multiple domains or with bound ligands, and more. Included, for example, are small proteins like **rubredoxin** (PDB entry 1rb9), various zinc-finger and metal-binding proteins like the cysteine-rich domain of **protein kinase** (PDB entry 1ptq), disulphide-rich proteins like **sea anemone toxin k** (PDB entry 1roo), and **proteinase inhibitor PMP-C** (PDB entry 1pmc).

#### 4.5.3 Classes are Further Divided into Folds

The protein *classes* are further divided into observed *folds* for protein structures. Folds describe the arrangement of secondary structural elements and/or chain topology. Each protein class has common folds, as described in turn in the next three sections.

### 4.6 $\alpha$ -Class Folds

In the  $\alpha$ -class of proteins (Figures 4.2 and 4.3), bundles, folded leafs, and hairpin arrays are major fold groups.

#### 4.6.1 Bundles

Bundles occur when  $\alpha$ -helices pack together to produce a hydrophobic core. Typically, an array of  $\alpha$ -helices is roughly aligned around a central axis. The bundle can be right or left-handed depending on the twist that each helix makes with respect to this axis. A *coiled coil* (two intertwined helices) can be a building block of these bundles. A simple example of a coiled coil is seen in the **DNA-binding leucine zipper protein** shown in Figure 6.5 of Chapter 6.

Among the  $\alpha$ -protein bundles, the four-helix bundle motif (often written as  $\alpha_4$ ) is common, as in **myohemerythrin**, Figure 4.2, and **Rop** (a small RNA-binding protein involved in replication), Figure 3.10. Other  $\alpha_4$  proteins are **ferritin** (a storage molecule for iron in eukaryotes), **cytochrome c'** (heme-containing electron carrier), the **coat protein of tobacco mosaic virus**, and **human growth hormone**.

Multi-helical bundles are also observed in  $\alpha$ -proteins; 3–6 and 8-helix aggregates are more frequent than others. Figure 4.2 shows a 5-helix bundle for the transport protein **pix**.

### 4.6.2 Folded Leafs

Complex packing patterns involving layered arrangements are often features of long  $\alpha$ -proteins. For example, in folded leaf folds, a layer of  $\alpha$ -helices wraps around a central hydrophobic core. Like bundles, such multihelical assemblies (usually 3 or more) pack together as well as form layers. The longest helices are usually in the center, and often the arrangement contains internal pseudosymmetry.

The globin fold of **myoglobin** (Figure 4.3) shows such a compact arrangement of a folded leaf arrangement formed by 8 helices, leaving a pocket for heme binding. **Cytochrome C6** in Figure 3.12 also displays a folded leaf.

A more complex layered topology is the two-layered ring structure of one  $\alpha$ -helical domain in the N-terminal region of the enzyme **muramidase** in bacterial cell walls, **soluble lytic transglycosylate** (Figure 4.3). It is built from 27  $\alpha$ -helices, arranged in a two-layered superhelix, leaving a large central hole, thought to be important in its catalytic activity.

### 4.6.3 Hairpin Arrays

Other  $\alpha$ -helix assemblies that cannot be described by bundle or folded leaf motifs are often described as hairpin arrays (arrays of  $\alpha$ -helix /loop / $\alpha$ -helix motifs). The calcium binding protein **calmodulin**, for example, has a helix/loop/helix motif where the loop region between two helices binds calcium (see Figure 3.11). Figure 4.2 also shows **cellulase cela**, a toroid-like circular array composed from 6 hairpins.

An irregular  $\alpha$ -protein from an all- $\alpha$  subdomain of the regulator of G-protein signaling 4, namely **guanine nucleotide-binding protein**, is also shown in Figure 4.3. This protein's motif contains a 4-helical bundle with left-handed twist and up-and-down topology.

## 4.7 $\beta$ -Class Folds

Proteins in the  $\beta$ -class display a flexible and rich array of folds, as seen in Figures 4.4 and 4.5. Various connectivity topologies can exist within networks of *parallel*, *anti-parallel*, or *mixed*  $\beta$ -sheets that twist, coil and bend in various ways. Indeed, note the much wider regions of the Ramachandran plot associated with  $\beta$ -sheets than with  $\alpha$ -helices (Figs. 3.18 and 3.19).

### 4.7.1 Anti-Parallel $\beta$ Domains

To describe these intriguing folds, it is simpler to begin with folds associated with the large subclass of  $\beta$ -proteins made exclusively of *anti-parallel*  $\beta$  domains. Such proteins tend to form distorted barrel structures. They can be described in terms of building blocks of two-strand, four-strand, eight-strand units, etc., as follows.

#### Two-Strand Units

The basic two-strand unit, the hairpin (denoted  $\beta_2$ ), involves a  $\beta/\text{loop}/\beta$  motif. It has adjacent anti-parallel  $\beta$ -strands linked head-to-tail by a turn or loop; see the  $\beta$ -strands connected as  $1 \rightarrow 2$  or  $4 \rightarrow 5$  for the head-to-tail direction in **fibronectin** in Figure 4.4.

#### Four-Strand Units

Proceeding to connections of four  $\beta$ -strands, there are 24 ways to combine two  $\beta$ -hairpin units to form a 4-stranded anti-parallel  $\beta$ -sheet unit. The most common topology is the Greek key (or  $\beta_4$ ). The four strands of a Greek key produce a  $\beta$ -sandwich through the head-to-tail connectivity of  $3 \rightarrow 4 \rightarrow 5 \rightarrow 6$ , as shown in the diagrams for **fibronectin** and **satellite panicum mosaic virus** in Figure 4.4. The  $\beta$ -strands in these illustrations are labeled according to their connectivity in the protein.

#### Eight-Strand Units

Correspondingly, there are many more ways to combine a larger number of  $\beta$ -strands from motifs of smaller systems. The two most common folds for 8 anti-parallel  $\beta$ -strands are jellyrolls ( $\beta_8$ ) and up-and-down  $\beta$ -sheet.

- The appetizing jellyroll is illustrated in Figure 4.4 for the  $\beta$ -sandwich coat protein of **satellite panicum mosaic virus**. It is a network of 8 anti-parallel  $\beta$ -sheets with the connectivity  $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5 \rightarrow 6 \rightarrow 7 \rightarrow 8$ , where strands are *shuffled* when viewed in the diagram left to right. Note the Greek key submotif in the  $4 \rightarrow 5 \rightarrow 6 \rightarrow 7$  subunit of the jellyroll.
- In the up-and-down  $\beta$ -sheet, each  $\beta$ -strand is connected to the next by a short loop. It has the simpler connectivity  $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5 \rightarrow 6 \rightarrow 7 \rightarrow 8$ , where strands 1 through 8 are written left to right (no shuffling required). Figure 4.4 shows this fold for **fatty acid binding protein** ( $1 \rightarrow 2 \rightarrow \dots \rightarrow 9 \rightarrow 10$ ).

### 4.7.2 Parallel and Antiparallel Combinations

More generally,  $\beta$ -protein topologies made of composites of parallel and anti-parallel strands usually form layered or barrel structures. The sandwich, barrel, and  $\beta$ -propeller are three general reference fold groups.

### Sandwiches and Barrels

In sandwiches,  $\beta$ -sheets twist and pack with aligned strands, whereas in barrels the sheets twist and coil so that often the first strand is hydrogen bonded to the last strand to produce closed structures. See the sandwich protein **fibronectin** and barrel in **fatty acid binding protein** in Figure 4.4. The immunoglobulins in Figure 3.12(d) are also  $\beta$ -sandwiches where seven strands form two sheets.

### Propellers

In  $\beta$ -propeller folds, 6 to 8  $\beta$ -sheets, each with 4 anti-parallel and twisted strands, arrange radially to resemble a propeller. The 7-bladed propeller of **galactose oxidase** is shown in Figure 4.5.

### Other $\beta$ -Folds

Other  $\beta$ -folds include  $\beta$ -prisms (3 sheets that pack around an approximate 3-fold axis), barrel/sandwich hybrids (2  $\beta$ -sheets, each shaped as a half barrel and packing like a sandwich), and  $\beta$ -clips (3 two-stranded  $\beta$ -sheets, forming a long hairpin folded upon itself in two locations). **Agglutinin** in Figure 4.5 shows a  $\beta$ -prism fold.

Recently,  $\beta$ -helix structures have been identified [238]. The polypeptides contain up to 16 helical turns, each of which contains 2 or 3  $\beta$ -sheet strands. Unlike the  $\beta$ -sandwiches, the  $\beta$ -sheet strands of a  $\beta$ -helix have little or no twist. Most such  $\beta$ -helix folds known to date are right-handed, as seen in **pectin lyase A** in Figure 4.5. The  $\beta$ -helix motif has been suggested to occur in the infectious scrapie prion protein [1371].

## 4.8 $\alpha/\beta$ and $\alpha + \beta$ -Class Folds

Even more diverse fold patterns are known for the  $\alpha/\beta$  and  $\alpha+\beta$ -classes of proteins depending on the sheet types (parallel, anti-parallel, or mixed network) and the location of the helices (exterior, interior, or on both faces) with respect to the sheet assembly.

We can broadly classify three fold motifs in this class (see Figure 4.6): barrels — closely packed  $\beta$ -strands (usually 8) with  $\alpha$ -helices on the exterior, open structures made of twisted  $\beta$ -sheets (parallel or mixed) surrounded by  $\alpha$ -helices on both the exterior and interior, and leucine-rich motifs of curved  $\beta$ -sheets with exterior  $\alpha$ -helices in leucine-rich regions.

### 4.8.1 $\alpha/\beta$ Barrels

A classic example of a barrel core is the barrel structure of **triosephosphate isomerase** (TIM), an  $(\alpha/\beta)_8$  topology (see Figure 4.6). The TIM barrel is one of the most common polypeptide-chain folds known today. TIM's 8 parallel

$\beta$ -strands coil to form a central core, and its 8  $\alpha$ -helices pack along the exterior. The central barrel ‘mouth’ is the active site of the protein.

#### 4.8.2 Open Twisted $\alpha/\beta$ Folds

An example from the highly-variable class of open twisted  $\alpha/\beta$  structures is **flavodoxin** (Figure 4.6). Note that its helices lie on opposite sides of the  $\beta$ -sheet. Typically, the active sites of proteins in this fold class are near the loop regions that connect  $\beta$ -strands to  $\alpha$ -helices. Another member of this class is **maltate dehydrogenase**, characterized by the *Rossmann* fold (named after its discoverer Michael Rossmann). This  $(\beta\alpha\beta\alpha\beta)_2$  topology has a central, parallel twisted  $\beta$ -sheet surrounded by  $\alpha$ -helices and/or loops. It is an important motif in proteins that bind to nucleic acids.

#### 4.8.3 Leucine-Rich $\alpha/\beta$ Folds

**Ribonuclease inhibitor** is an example in the leucine-rich class of  $\alpha/\beta$  folds. Its *horseshoe* structure is formed by homologous repeats of right-handed  $\beta$ -loop- $\alpha$  units (see Figure 4.6). The 17 parallel  $\beta$ -strands lie on the inside of this horseshoe, with the 16  $\alpha$ -helices clustering on the outside. The leucine residues present in all three segments of the repeating unit — the  $\beta$ -strand, the loop, and the  $\alpha$ -helix — pack snuggly together to form a hydrophobic core between the  $\beta$ -strand and  $\alpha$ -helix regions.

#### 4.8.4 $\alpha+\beta$ Folds

Yet more complex fold patterns have been observed for the  $\alpha+\beta$ -class of proteins (see Figure 4.7). This diversity reflects the various topologies of the subdomains (or layers) as well as the richness of connectivity patterns among them.

#### 4.8.5 Other Folds

Examples of multi-domain proteins, membrane and cell surface proteins, and small proteins are shown in Figures 4.8, 4.9, and 4.10.

### 4.9 Number of Folds

It has been postulated that the number of folding motifs is finite and that the entire catalog of folds will eventually be known with the rapidly-increasing number of solved globular proteins [157, 237, 560]. Such postulates come from stereochemical considerations — for example, there is a small number of ways to link compactly  $\alpha$ -helices and  $\beta$ -strands — database analyses, and statistical sampling approaches.

### 4.9.1 Finite Number?

The exact number of folds has not been determined. Some studies estimate this number to be several thousand [266, 780], while others yield only several hundred [1340, 1434] (around 10,000 or 3000 total folds in the former group and 850 total folds in the latter works), so a minimal estimate of around 1000 [1259] and the range of 1000–10,000 seem reasonable [168]. Only time will tell how many folds Nature has produced.

Since many computational folding-prediction schemes use known folds for closely-related sequences or closely-related functions of proteins, a finite number of folds suggests that *eventually* we will be able to describe 3D structures from sequence quite successfully!

Zhang and DeLisi estimated in 1998 [1434], however, that with the technology available at that time, 95% of the folds will only be determined only in 90 years. They argued that, aside from technological improvements, we should carefully select new sequences for structure determination so as to maximize new fold detection and thereby reduce that time substantially. This is important since the annual number of new folds discovered during 1995–2000 has only averaged around 10%, with even less during 2000–2002. Certainly, careful selection of targets is even more critical if the number of folds is actually larger (e.g., of order 10,000) and associated with single sequence families [266]. The structural genomics initiatives (see beginning of Chapter 2) are certainly accelerating the discovery of new folds (see, for example, [48, 214]), but the effect of these projects will take time to assess (see, for example, differing opinions in [87, 993]). For updated fold information, search **PDB holdings**.

## 4.10 Quaternary Structure

Quaternary structures describe complex interactions for multiple polypeptide chains, each independently folded, with possibly other molecules (nucleic acids, lipids, ions, etc.). The interactions are stabilized by hydrogen bonds, salt bridges, and various other complex intermolecular and intramolecular associations in space. The classic example for a quaternary structure is that of the protein **hemoglobin**, which consists of four polypeptide chains. The four subunits, each of which contains an oxygen-binding heme group, are arranged symmetrically. Other examples of quaternary structure are DNA polymerases (with catalytic and regulatory components) and ion channels, and protein/nucleic acid complexes with complex structures involving many subunits like viruses, nucleosomes, and microtubules.

### 4.10.1 Viruses

Virus coats are often comprised of many protein molecules and have intriguing quaternary structures. These protein coats envelope the inner domain which

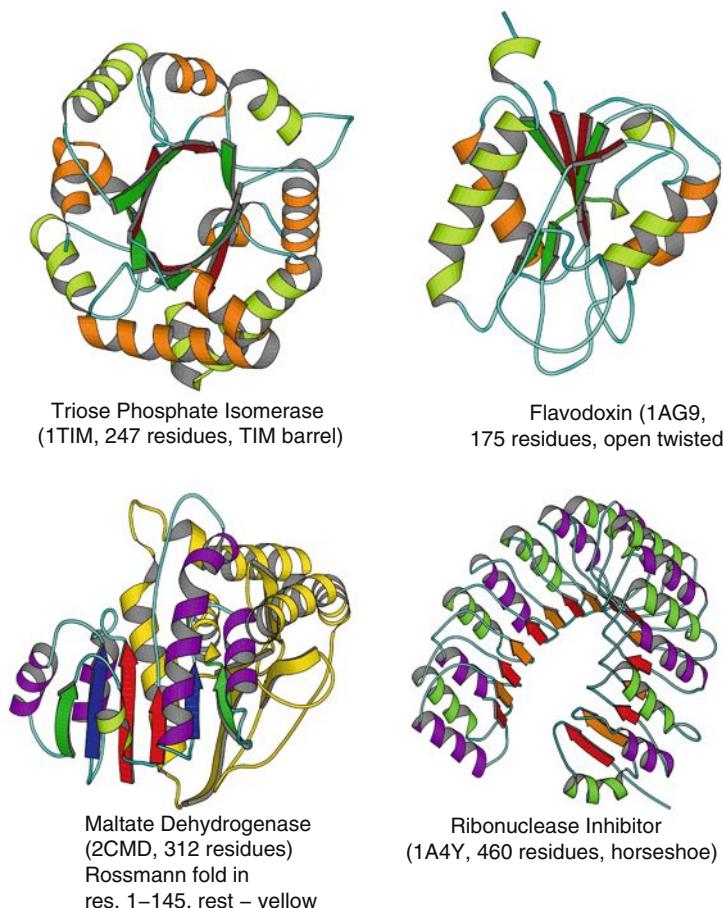


Figure 4.6. Examples of  $\alpha/\beta$ -proteins. **TIM** (triosephosphate isomerase) displays an architecture of 8 twisted parallel  $\beta$ -strands which form a barrel surrounded by  $\alpha$ -helices. **Flavodoxin**, an electron transport protein that binds to a flavin mononucleotide prosthetic group, displays an open twisted  $\alpha/\beta$  fold made of three layers (2 helices at left, 5  $\beta$ -strands in the middle, and 2 helices at right). **Maltate dehydrogenase** contains the  $(\beta\alpha\beta\alpha\beta)_2$  **Rossmann fold** in the subunit shown. **Ribonuclease inhibitor**, in the leucine-rich class of  $\alpha/\beta$  folds, displays a horseshoe structure.

consists of infectious nucleic acids. For example, the **poliovirus** — a spherical complex of 310 Å in diameter — has a shell of 60 copies of each of four proteins. The coat of **tobacco mosaic virus** combines 2130 identical protein units, each of 158 residues, arranged in a helix around a coiled RNA structure of 6400 nucleotides. This results in a rod-shaped complex 3000 Å long and 18 Å in diameter.

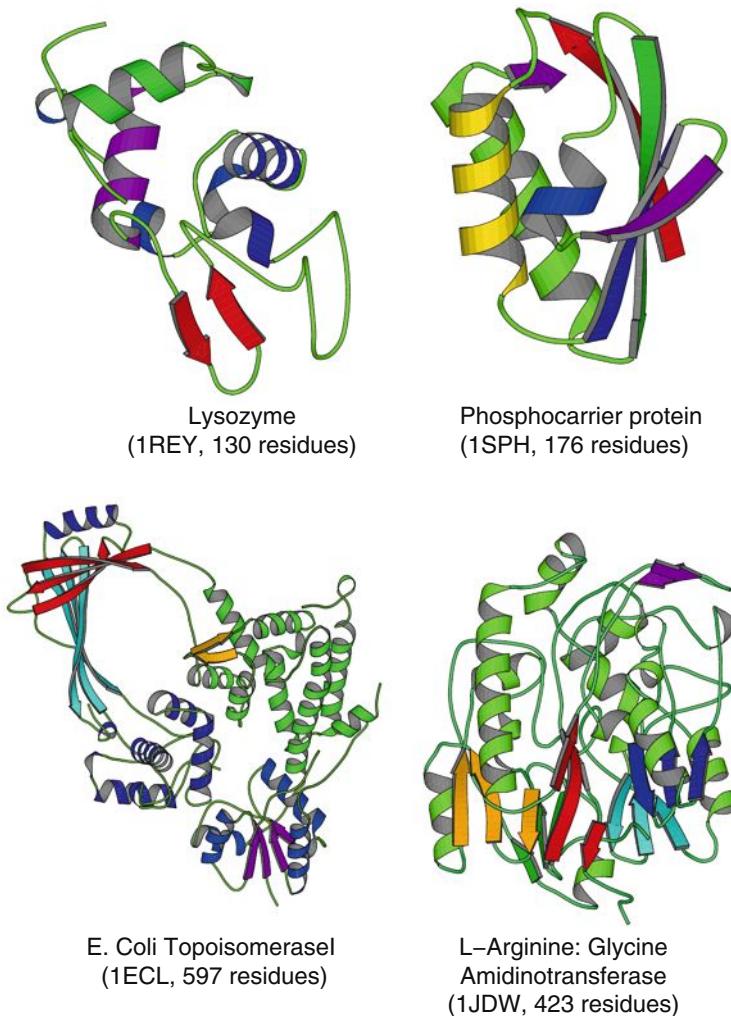


Figure 4.7. Examples of  $\alpha + \beta$ -proteins: **lysozyme**, **phosphocarrier protein**, **DNA topoisomerase I**, and **glycine amidinotransferase**.

Figure 4.11 illustrates the structure of the 180-chain **tomato bushy stunt virus** that infects many plants, including tomatoes and cherry trees. Interestingly, virus coats are assemblies of similar proteins rather than one huge protein or combinations of different proteins, because the relatively small amount of viral nucleic acids must encode this protein coat; at the same time, the nucleic acids must be covered completely. Hence a large protein shell consisting of repetitive motifs satisfies both of these criteria.

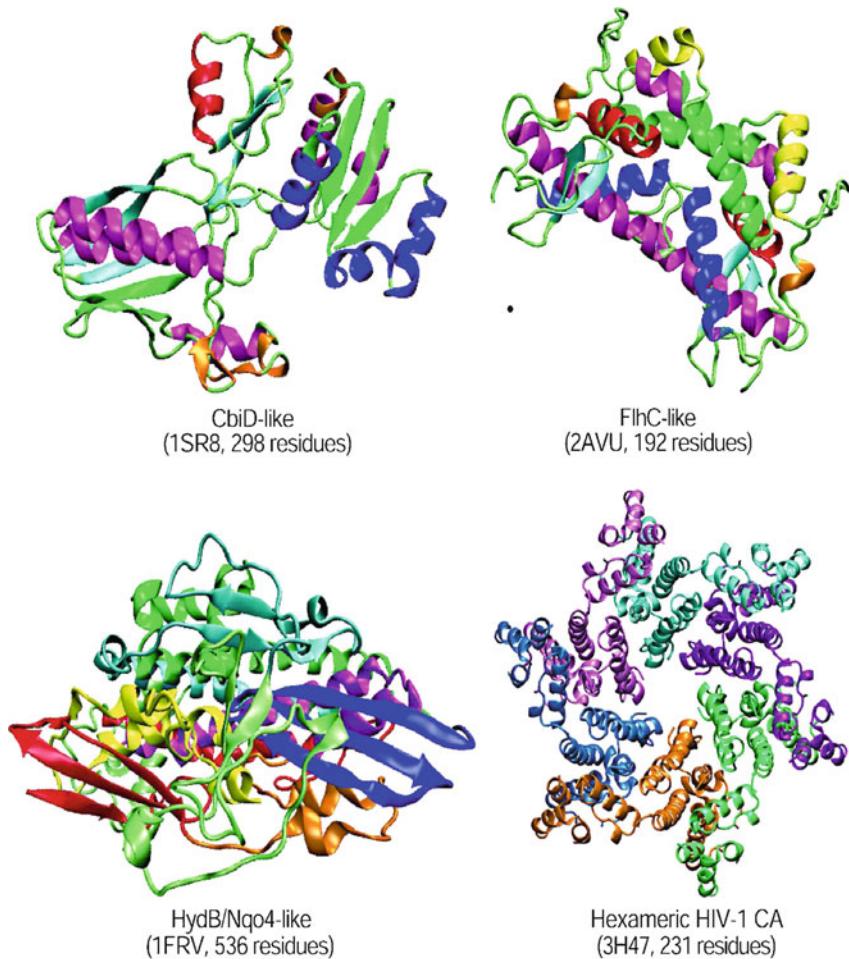


Figure 4.8. Examples of multidomain proteins: **CbiD-like** protein with two domains; **FlhC-like** protein with three domains; **HydB/Nqo4-like** with four domains; and **Hexameric HIV-1 CA** with two domains.

Among the larger molecular structures determined by X-ray crystallography at moderate resolution (i.e., approaching 3.5 Å) is the core particle of **bluetongue virus**, an agent of disease in both plants and mammals. Its transcriptionally active compartment measures 700 Å in diameter and is composed of two principal structural proteins that assemble in two layers, a core and a subcore, together encapsulating the RNA genome (10 segments of doubled-stranded RNA, ~19,000 base pairs total). The crystal structure revealed how these approximately 1000 protein components self-assemble through a complex mixture of packing mechanisms involved in each of the two layers, using triangulation and geometrical quasi-equivalence packing motifs [484].

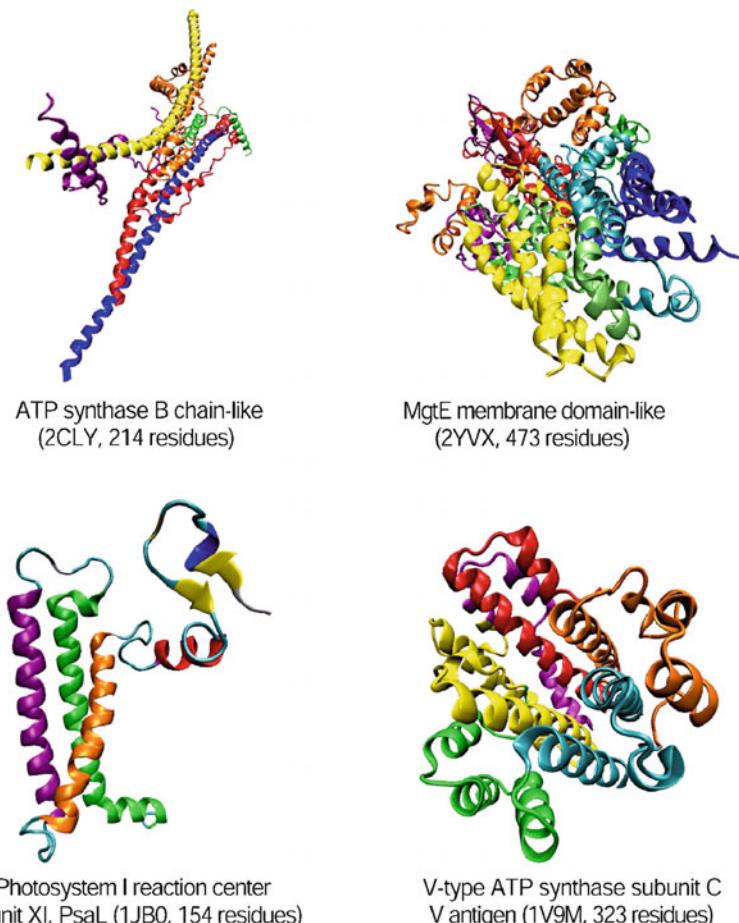


Figure 4.9. Examples of membrane and cell surface proteins and peptides: **ATP synthase B chain-like** protein, with a long helix; **MgtE membrane domain-like** protein, with five transmembrane helices; **Photosystem I reaction center subunit** protein, with three transmembrane helices; and **V-type ATP synthase subunit C** protein, with nine transmembrane helices.

#### 4.10.2 From Ribosomes to Dynamic Networks

Other examples of quaternary structure are noted for the ribosome, muscle-fiber complexes, bacterial flagellar filaments, and photosynthetic assemblies of membrane proteins.

The *E. Coli* **ribosome** is a ribonucleoprotein complex with a diameter of about 200 Å constructed from 3 RNA molecules and 55 protein chains [419]. The Nobel Prize in Chemistry was awarded in 2009 to three scientists who independently obtained atomic-level crystallographic views of this magnificent

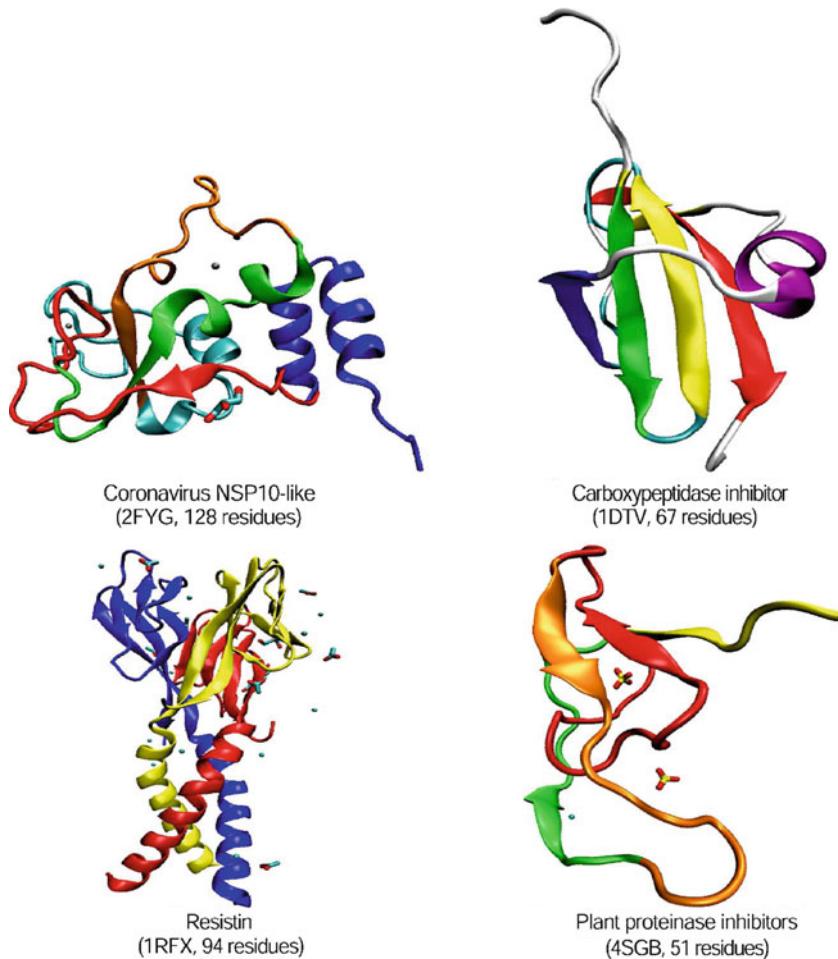


Figure 4.10. Examples of small proteins: **Coronavirus NSP10-like**, binds two zinc ion per subunit; **Carboxypeptidase inhibitor**, disulfide-rich,  $\alpha+\beta$ ; **Resistin**, disulfide-rich six-stranded  $\beta$ -sandwich; and **Plant proteinase inhibitor** complexed with calcium and  $\text{SO}_4$ .

RNA/protein machine: Ada Yonath, Venkatraman Ramakrishnan, and Thomas Steitz. For example, the Yonath lab solved the large ribosomal subunit from *Deinococcus radiodurans* [516] and the small ribosomal subunit from *Thermus thermophilus* [1135] (see Fig. 1.1). The Steitz lab reported the structure of the large ribosomal subunit from *Haloarcula marismortui* (2833 of the subunit's 3045 nucleotides and 27 of its 31 proteins) [85], and Ramakrishnan's group reported the structure of the small subunit of *T. thermophilus* [1379]. These eagerly awaited structures of the bacterial ribosome were aided by cryo-electron microscopy reconstructions — first reported in 1995 for the ribosome from *E. Coli* (see recent

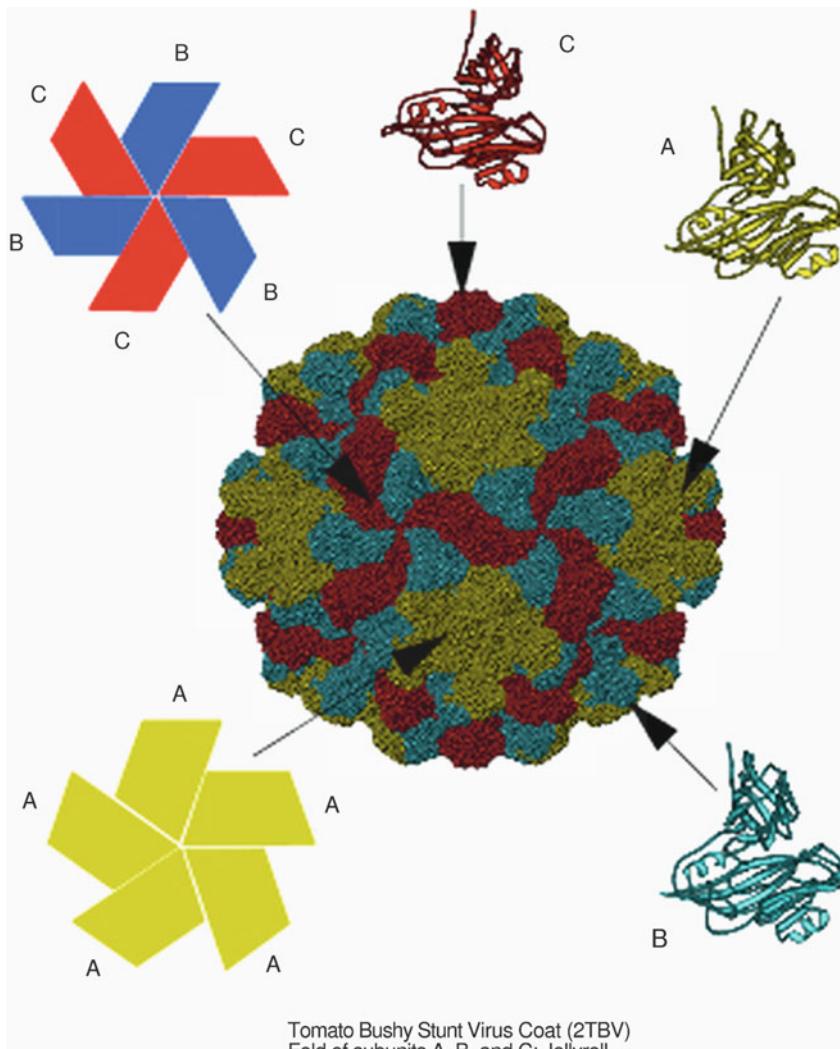


Figure 4.11. The structure of **tomato bushy stunt virus**, a spherical arrangement of 180 polypeptide chains, each of 387 amino acids, with every 3 chains making up an asymmetric unit (the subunits are colored blue, green, and red).

views in Figure 1.2 [710]) — which helped crystallographers estimate the initial phasing of their X-ray data (see [171] for a perspective). The combined structural characterizations of the ribosome provided clear evidence that the ribosome is a ribozyme — that is, that the ribosome RNA's component likely catalyzes peptide bond formation (see Chapter 7, RNA sections).

Muscle cells contain parallel myofibrils composed of two kinds of filaments, each with the following proteins: **myosin** (thick filament), and **actin**,

**tropomyosin**, and **troponin** (thin filament); around these filaments, **titin** — itself two extremely long proteins — plus nebulin form a flexible mesh. Muscle contraction is produced by the interaction of actin and myosin.

The bacterial flagellar motor of the protein **flagellin** [1085] represents another challenging motor complex solved recently. Filaments of flagellin are formed by an arrangement of stacked flagellin proteins ('protofilaments') lined up side by side; an arrangement like loosely rolled sheets of paper results. The remarkable cooperativity among the different filaments leads to conversions between a macroscopic left-handed form — used for swimming — and a right-handed form — used for reorientation of motion. The high-resolution flagellin crystal suggests how this possible structural switch (between left and right-handed supercoiled forms) might occur to direct function.

Insights into the solar energy converters in the membranes of bacteria and plants were provided by the crystal structure of photosystem I, a large photosynthetic assembly of membrane proteins and other cofactors from the thermophilic cyanobacterium *S. elongatus* [616]. The detailed atomic picture (at 2.5 Å resolution) of the network of 12 protein subunits and 127 cofactors (chlorophylls, lipids, ions, waters, others) shows the beautiful coordination of all components for efficient absorption and conversion of solar energy into chemical energy.

## 4.11 Protein Structure Classification

Many groups worldwide are working on classifying known protein structures; see [47, 48, 952, 1259] for a perspective of protein structure and function evolution. Several classification schemes and associated software products exist. A popular program is SCOP: "Structural Classification of Proteins" [887]. (See [scop.mrc-lmb.cam.ac.uk/scop/](http://scop.mrc-lmb.cam.ac.uk/scop/) or connect to SCOP through links available in many mirror sites such as PDB) [262]. These classifications are currently assigned manually, by visual inspection, but some automated tools are being used for assistance.

Also noteworthy is the PROSITE ([www.expasy.ch/prosite/](http://www.expasy.ch/prosite/)) database of protein families and domains intended to help researchers associate new sequences with known protein families. Other databases of patterns and sequences of protein families are PFAM and PRODOM; see [881] for a comprehensive list.

The SCOP levels (top-to-bottom) are: class, fold, superfamily, family, and domain. The sequence, or reference PDB structure, can be considered at the very bottom of this tree.

The top level of the SCOP hierarchy is the *class* (all- $\alpha$ , all- $\beta$ ,  $\alpha/\beta$ ,  $\alpha + \beta$ , multi-domain, membrane and cell-surface, and small proteins). Each *class* denotes common, global topologies of secondary structure.

Next comes the *fold*, which clusters proteins that have the same global structure, that is, similar packing and connectivity schemes for the secondary structural elements. Folds are often also called *supersecondary structure*. From 50 to several

hundred folds are currently known for each class, with the repertoire increasing steadily. An example mentioned above, the  $\alpha/\beta$  barrel fold, groups **TIM** with other proteins like **RuBisCo(C)**, **Trp biosynthesis**, and **glycosyltransferase** into a *superfamily*, the next level of the classification hierarchy.

The *superfamily* groups proteins with low sequence identity but likely evolutionary similarity, as judged by similar overall folds and/or related functions. Members of the same superfamily are thus thought to evolve from a common ancestor. Another superfamily, for example, contains **actin**, the ATPase domain of the **heat shock protein**, and **hexokinase**. Superfamilies often pose the greatest challenge in the task of protein classification.

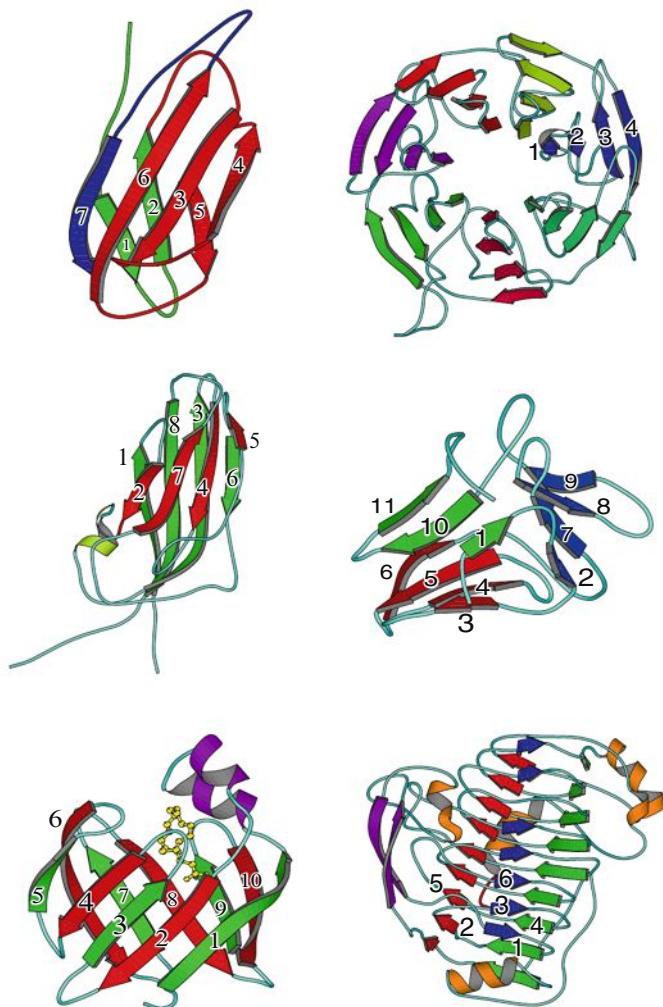
Superfamilies are further divided into *families*, which cluster proteins with substantial sequence, structure, and function similarity. Generally, this requirement implies a sequence identity of at least 30%, but there are instances of low sequence identity (e.g., 15%) but definitive structural and functional similarities, as in the case of globin proteins. For example, families of glycosyltransferase include  $\beta$ -galactosidases,  $\beta$ -glucanase,  $\alpha$ -amylase, and  $\beta$ -amylase.

Finally, at the bottom of the tree of the SCOP classification lies the *domain* category, to distinguish further structurally-independent regions that may be found in larger proteins.

For updated information on the number of identified folds, superfamilies, and domains, check [scop.mrc-lmb.cam.ac.uk/scop/count.html](http://scop.mrc-lmb.cam.ac.uk/scop/count.html).

As our knowledge of protein structure increases, our classification schemes and software tools will evolve quickly. Automation of the classification is important for rapid structural analysis and ultimately for relating the sequence and structure to biological function.

*The reader is encouraged to re-read at this point the sections in Chapter 2 on protein folding/misfolding (Sections 2.2 and 2.3).*



# 5

## Nucleic Acids Structure Minitutorial

### Chapter 5 Notation

SYMBOL	DEFINITION
<b>Vectors</b>	
$\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$	local base-pair coordinate system
<b>Scalars &amp; Terms</b>	
endo/exo	sugar labeling (e.g., C3'-endo)
$dx, dy$	$x$ and $y$ -displacements (helical parameters)
$h$	helical rise
$q$	wave amplitude (pseudorotation description)
$n$	number of DNA base pairs
$n_b$	number of base pairs per turn
$z_0-z_4$	perpendicular displacements of sugar ring atoms
$Dx, Dy, Dz$	shift, slide, and rise translations (between successive base pairs)
$P$	phase angle of sugar pseudorotation
$P_h$	helix pitch
$\alpha, \beta, \gamma, \delta, \epsilon, \zeta$	polynucleotide backbone torsion angles
$\eta$	inclination angle (helical parameter)
$\theta$	tip angle (helical parameter)
$\rho$	roll angle (base-pair step parameter)
$\tau$	tilt angle (base-pair step parameter)
$\tau_0-\tau_4$	internal sugar ring torsion angles
$\tau_{\max}$	puckering amplitude of sugar pseudorotation
$\phi$	phase-shift angle (pseudorotation description)
$\chi$	glycosyl (sugar/base) torsion angle
$\omega$	propeller twist (base pair parameter)
$\Omega$	twist angle

DNA plays a role in life rather like that played by the telephone directory in the social life of London: you can't do anything much without it, but, having it, you need a lot of other things — telephones, wires, and so on — as well.

Review of *The Double Helix* in *The Sunday Times, London* (1968).

## 5.1 DNA, Life's Blueprint

With the master molecule of heredity, deoxyribonucleic acid (or DNA), so frequently mentioned in the media — in connections ranging from polymerase chain reaction (PCR) applications (e.g., criminology, medical diagnoses) to cloning and art [1112] — it is difficult to imagine today that it was only a few decades ago when Watson and Crick reported their description of the DNA double helix [1354–1356]! Based on analysis of DNA fiber diffraction patterns and Chargaff's rules, they described a spiral image of an orderly helix — two intertwined polynucleotide chains, with a sugar/phosphate backbone on the exterior and pairs of hydrogen-bonded nitrogenous bases in the center. See [622] for a historical perspective of this discovery, including the contribution of all key players (a capsule of which is given in Chapter 1), and anniversary issues of DNA, for example issued in 2003 in many journals (e.g., *Nature* Vol. 421 and *Science* Vol. 300) at the occasion of DNA's golden anniversary.

Though the model was imperfect in several respects (e.g., sugar conformations and base positioning with respect to the helical axis were inexact), Francis Crick, James Watson, and Maurice Wilkins received the 1962 Nobel Prize in Chemistry for this monumental discovery. Indeed, soon after the description of the DNA double helix came a burst of scientific excitement and discovery regarding the mechanism of heredity. The field of *molecular biology* gained extraordinary momentum and, soon after, its tentacles extended to the disciplines of cellular biology, molecular genetics, genomics, and proteomics, all buttressed by impressive advances in technology and computing.

### 5.1.1 *The Kindled Field of Molecular Biology*

The description of DNA's three-dimensional (3D) structure provides a classic example of the structure/function relationship. The two key functions of DNA — replication and transcription — were suggested and later proven based on the ordered spatial arrangement of the DNA. *Replication*, or the accurate transmittal of genetic information from parent to progeny, was explained by base pairing specificity. *Transcription*, the transformation of DNA's genetic information into a form that directs protein synthesis, was later understood with the deciphering of the

genetic code relating triplet nucleotide codons (see Section 5.2 and Figure 5.2 for a description of the basic building blocks, including the pyrimidine bases T and C and purine bases A and G) to amino acids.

This genetic code (Table 5.1) reveals interesting patterns:

1. **It is highly redundant:** as 61 of the 64 triplet codons represent one of the 20 amino acids, and 3 signal ‘stop’ messages for protein synthesis (i.e., translation). In particular, Arg, Leu, and Ser are specified by 6 codons; Met and Trp are each represented by only one; and the rest are specified by 2–4 triplets.
2. **Pyrimidine-end commonalities exist:** Codons  $B_1B_2T$  and  $B_1B_2C$  (where  $B_1$  and  $B_2$  denote any of the 4 bases) always code for the same amino acid.
3. **Purine-end commonalities exist:** Codons  $B_1B_2A$  and  $B_1B_2G$  code for the same amino acid except in the cases of AT (ATA = Ile, ATG = Met) and TG (TGA = ‘STOP’, TGG = Trp).

Though the genetic code was initially thought to be ‘universal’, DNA sequencing in the early 1980s revealed code variations for certain organisms. Now it is clear that the genetic codes for mammalian and plant mitochondria and certain protozoa, for instance, exhibit some variations from the code described in Table 5.1. In organisms where nonstandard amino acids exist (e.g., certain Archaea and eubacteria), such a variation in the code may consist of a STOP codon that encodes a nonstandard amino acid rather than instruct to halt translation [68, 511, 1217].

Recent advances in experimental techniques that allow detailed study of the components of the translation apparatus are also making possible the study of how the modern code has evolved from a simpler form [662].

Thus the model for the DNA double helix led the way to establishment of the first iteration of the central dogma of biology: DNA is self-replicating, DNA makes messenger ribonucleic acid (mRNA) through transcription,<sup>1</sup> and mRNA makes protein through translation (DNA → RNA → protein).<sup>2</sup> Although other arrows in the dogma have been suggested or verified (e.g., RNA → RNA in certain viruses, RNA → DNA in retroviruses), the main components appear known.

---

<sup>1</sup>In transcription, an RNA polymerase glides along one strand of the DNA double helix and builds an RNA complement by coupling ribonucleotides through dehydration synthesis. The faithful replicate of one DNA strand then functions as the messenger RNA.

<sup>2</sup>In translation, the genetic code of the messenger RNA is read by transfer RNA molecules on cellular structures called ribosomes. Every transfer RNA molecule carries a specific sequence of three nucleotides on one end of an L-shaped structure and the corresponding amino acid on the other. The transfer RNA’s main task is to deposit its amino acids on the ribosomes in proper sequence. In this process of matching each messenger-RNA codon with the complementary transfer RNA molecule, the polypeptide chain is assembled. As the amino acids link to one another on the ribosomes, polypeptide folding is thought to begin.

Table 5.1. The genetic code in terms of the parental DNA; the mRNA transcript has uracil (U) instead of thymine (T).

Amino Acid (or Instruction)	Encoding Triplets in Parental DNA <sup>a</sup>
Arginine (Arg)	CG(*), AG(A,G)
Leucine (Leu)	CT(*), TT(A,G),
Serine (Ser)	TC(*), AG(T,C)
Alanine (Ala)	GC(*)
Glycine (Gly)	GG(*)
Proline (Pro)	CC(*)
Threonine (Thr)	AC(*)
Valine (Val)	GT(*) <sup>b</sup>
Isoleucine (Ile)	AT(T,C,A)
Asparagine (Asp)	GA(T,C)
Aspartic acid (Asn)	AA(T,C)
Cysteine (Cys)	TG(T,C)
Histidine (His)	CA(T,C)
Phenylalanine (Phe)	TT(T,C)
Tyrosine (Tyr)	TA(T,C)
Glutamine (Gln)	CA(A,G)
Glutamic acid (Glu)	GA(A,G)
Lysine (Lys)	AA(A,G)
Methionine (Met)	ATG <sup>b</sup>
Tryptophan (Trp)	TGG
STOP	TA(A,G), TGA

<sup>a</sup>Short-hand notation is used for the third base of the codon when the first and second bases are the same: GA(T,C) specifies both GAT and GAC, while GC(\*) denotes all four possibilities, namely GCT, GCC, GCA, and GCG.

<sup>b</sup>The codons ATG and (less frequently) GTG form part of the initiation signal in addition to their coding for Met and Val, respectively.

Much attention has now expanded to new areas of molecular biology, such as structural molecular biology, computational molecular biology, and offspring of genomics, such as structural genomics and functional genomics. The functional relationships among genes and the interaction of genes within the context of the complete organism are also of great interest.

### 5.1.2 Fundamental DNA Processes

As the genetic material, DNA carries structural information in the primary sequence that not only controls faithful duplication but also regulates expression

of the hereditary information [1021]. Genetic variability can result from errors (or *mutations*), such as insertions, deletions, or substitutions in the daughters compared to the parental DNA, during the template-copying process; these changes can lead to altered triplet codes and hence different polypeptide composition.

Since many basic biological processes rely on protein/nucleic-acid interactions, the base sequence of polynucleotides also affects profoundly the characteristic 3D structure of DNA and RNA and hence the nature of fundamental biological processes. On a higher level of structure (hundreds and thousands of base pairs), compact forms of long DNA — such as supercoils, knots, and chromosomes — are central to fundamental mechanisms for replication, transcription, and recombination [195, 1384]. DNA supercoiling, namely the coiling in space of the double helix axis itself, can condense the DNA by several orders of magnitude and readily store energy for various activities. This supercoiled state is essential for the storage of DNA in eukaryotic cells, where DNA is wrapped as a left-handed supercoil around cores of proteins to form the chromatin fiber. The wide range of characteristic timescales associated with the configurational rearrangements and hydrodynamic properties of supercoiled (or superhelical) DNA represents another area of intense study.

### 5.1.3 Challenges in Nucleic Acid Structure

As described above, there are several levels of nucleic acid structure, from the base-pair level to the cellular level of organization in the chromosomes (see end of next chapter and [1119] for example). This study of DNA's rich configurational levels is particularly challenging to modelers.

Much focus has been placed on *protein structure* (including protein/DNA complexes) and the *protein folding problem* — predicting the 3D architecture of a system from the primary sequence. Yet an analogous folding problem is also relevant to DNA and RNA polymers. In fact, the nucleic-acid analogue of the 'protein folding problem' might be viewed as more challenging than protein structure prediction in the sense that it extends over much larger spatial scales (thousands of Ångstroms) as well as temporal scales (picoseconds to minutes and longer) than typically associated with proteins. Biologically, elucidating the folding of DNA in the cell — supercoiling and chromosome condensation, for example — is important for understanding the regulatory role of DNA in fundamental biological processes.

For small segments of nucleic acids, X-ray crystallography and NMR data have been invaluable for providing detailed, atomic-resolution data on single, double, triple and other forms of nucleic acid oligonucleotides, as well as their complexes with proteins, other biomolecules, and ligands [126]. The Nucleic Acid Database (NDB, [ndbserver.rutgers.edu/](http://ndbserver.rutgers.edu/)) [127] has been beautifully cataloging these structures and offers many utilities for their analysis.

All-atom molecular dynamics (MD) simulations of nucleic acids have also shed important insights on DNA sequence/structure relationships, the nature of the hydration geometries surrounding nucleic acids, nucleic acid flexibility, and

nucleic-acid protein structures [126, 217, 808, 953, 1026, 1417, 1419, 1421, for example]. Indeed, simulation improvements and the availability of ultra-high resolution nucleic-acid crystal structures [234, 641, 1182] have made possible the study of fully solvated nucleic acid MD trajectories, with representative ionic atmospheres [219, 239, 283, 1412, 1416], and even millisecond simulations [987].

The theoretical advances resulted from improvements in long-range electrostatic modeling, force field parameters, representation of the ionic atmosphere, and novel conformational sampling approaches [217, 953, 988, 1116, 1117, 1319], as well as increases in computer memory and speed. The experimental advances reflect improved methods for crystallization and phase determination, the increased availability of very strong X-ray sources, improvements in algorithms for model refinement, and innovative approaches such as single-force microscopy which allows studies of DNA and RNA energetics and dynamics of folding and unfolding (e.g., [187, 759]) and improved ultra-structural visualization tools for DNA's higher levels of structural organization (e.g., chromatin, see next chapter). Significantly, such modeling and experimental advances have made it possible to simulate solvated RNA at atomic resolution [86, 502, 525, 526, 1446]. Advances on both the computational and experimental fronts are ongoing.

Computer scientists also find interest in DNA with the emerging possibility of using the strands of DNA *in vitro* for practical applications, such as to produce electronic devices like nanowires, or as parallel computers to solve very difficult, combinatorial optimization problems that have non-polynomial complexity [40, 117, 162, 825, 1156]. Very recently, a 'DNA Sudoku' approach for parallel DNA sequencing by combinatorial pooling strategies reminiscent of solving sudoku puzzles has also been reported for analyzing short genome segments associated with disease markers [366].

Undoubtedly, progress is expected in the bridging between all-atom and macroscopic representations of nucleic acids and between experiment and theory. This unity will enhance our understanding of DNA structure and DNA/protein interactions and, in turn, will likely have important biomedical applications, for example, in the design of new anti-viral drugs, antibiotics, and anti-cancer agents, some of which are being designed with DNA or RNA agents such as DNA plasmids and silencing RNAs.

#### 5.1.4 Chapter Overview

In this chapter, the basic elements of nucleic acids and DNA structure on the base-pair and helical level will be introduced: the fundamental building blocks and how they are linked to form polynucleotides, aspects of nucleic acid conformational flexibility (sugar pseudorotation, torsion angle preferences, and global and local base-pair parameters), and the three canonical DNA helices (A, B, and Z-DNA). The next chapter presents further topics regarding the structural diversity of DNA and RNA, and DNA folding on a higher level, namely supercoiling and chromatin organization.

While RNA is introduced in the next chapter, after the variety of base pairing arrangements is presented, this in no way means RNA is less fundamentally important than DNA. In fact, recent discoveries on RNA's prevalent regulatory roles in the cell place RNA in the forefront of biology and chemistry.

For basic books on DNA structure, see [100, 139, 195, 269, 893, 895, 1080, 1191, 1305]. Lively, less technical introductions can be enjoyed from [272, 420, 1353]. The reader is also well advised to examine some of the rich information on nucleic acid structure available in the public domain through the NDB [127], [ndbserver.rutgers.edu/](http://ndbserver.rutgers.edu/).

Throughout this chapter and the next, we abbreviate a *base pair* as bp and *base pairs* as bps. Note that though we mention the three canonical helices before introducing them formally (using all notation we systematically develop), novices should be able to follow the material. Students are encouraged to re-read the chapter after they are more familiar with all the presented material.

## 5.2 The Basic Building Blocks of Nucleic Acids

In the classic DNA double helix described by Watson and Crick, a flexible ladder-like structure is formed with the polymer wrapped around an imaginary central axis (Figure 5.1). The two rails of the ladder consist of alternating sugar (deoxyribose) and phosphate units; the rungs of the ladder consist of nitrogenous bps held together by hydrogen bonds (see Box 3.2 of Chapter 3 for a definition).

### 5.2.1 Nitrogenous Bases

Four nitrogenous bases can be found in DNA: the pyrimidines cytosine (C) and thymine (T) which are 6-membered rings, and the purines guanine (G) and adenine (A), each of which is a fused system with a 5 and 6-membered ring. In the single-stranded RNA polynucleotide, ribose replaces deoxyribose and uracil (U) replaces thymine (Figure 5.2). Some DNAs contain bases that are chemically modified (e.g., substitutions of a hydrogen by a methyl group).<sup>3</sup> Alternative but rare *tautomeric* forms due to proton shifts in aromatic molecules (as introduced in Chapter 1, subsection on the discovery of DNA structure),<sup>4</sup> are possible but may not be biologically significant.

---

<sup>3</sup>For example, *N6-methyl-dA* is a modified adenine base with the N6H<sub>2</sub> (attachment to the ring carbon C6) moiety replaced by N6CH<sub>3</sub>; *5-methyl-dC* is a modified cytosine where the C5H becomes C5CH<sub>3</sub>.

<sup>4</sup>Two classic tautomerization reactions are *keto/enol* and *amino/imino*; the keto and amino forms are typically favored and are shown in Figure 5.2. Keto/enol tautomerization involves alteration of the carbonyl group ( $=\text{C}=\text{O}$ ) to a hydroxyl group ( $=\text{C}-\text{O}-\text{H}$ ), shifting the double bond from the carbonyl group to the nitrogen-carbon bond in the ring (e.g., C6=O of G becomes C6-OH, accompanied by the change of H-N1-C6 to N1=C6). Similarly, an amino/imino tautomerization involves a change in an amino nitrogen  $-\text{NH}_2$  to an imino form,  $=\text{NH}$  (e.g., C6-NH<sub>2</sub> of A becomes C6=NH, accompanied by the change of N1=C6 to H-N1-C6).

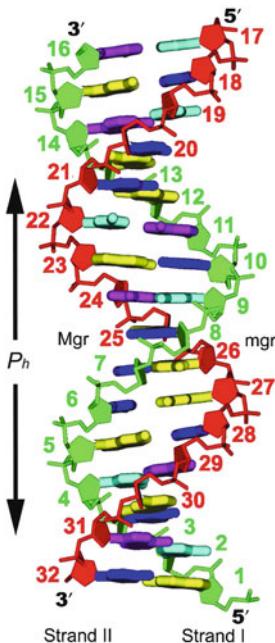


Figure 5.1. The two strands of the classic DNA double helix: alternating subunits of phosphates and deoxyriboses held together by hydrogen-bonded bps, normally with adenine (A) pairing with thymine (T) and guanine (G) pairing with cytosine (C). The two chains of the DNA double helix run in opposite directions. The chain ending in the lower right terminates with a phosphate group linked to the sugar carbon denoted as C5', while the chain ending in the lower left terminates with a free hydroxyl group linked to the sugar carbon denoted as C3'. Individual bases are numbered 1–16 (or 1 to  $n$  where  $n$  is the number of bps) for the sequence strand (Strand I) and 17–32 ( $n + 1$  to  $2n$  in general) for Strand II, both in the  $5' \rightarrow 3'$  direction; the  $n$  base pairs 1–16 are numbered so that they coincide with the bases of Strand I. The *pitch* of the helix  $P_h$  is the length along the helix axis for one complete turn, and  $n_b$  is the number of bps per turn (around 10.5 for DNA in solution). The unit *twist* is defined as  $\Omega = 360^\circ/n_b$ , and the helical rise is  $h = P_h/n_b$ . Mgr and mgr denote the major and minor grooves, respectively.

### 5.2.2 Hydrogen Bonds

In the Watson-Crick base pairing scheme (termed WC herewith), cytosine pairs with guanine by forming three hydrogen bonds, and thymine pairs with adenine by forming two hydrogen bonds (Figure 5.3). This arrangement produces CG and TA bps whose widths are nearly identical. The approximately uniform width is significant because any pyrimidine/purine or purine/pyrimidine sequence can be accumulated on one strand, with a pyrimidine opposing a purine, without much alteration in overall structure. The discovery of this orderly 3D structure of DNA

Sugars	Pyrimidine Bases	Purine Bases
<p><math>\beta</math>-D-2-Deoxyribose</p> <p>Ribose</p>	<p><i>Thymine</i></p> <p><i>Cytosine</i></p> <p><i>Uracil</i></p>	<p><i>Adenine</i></p> <p><i>Guanine</i></p>

Figure 5.2. Chemical structures and atomic labels for nucleic acid sugars (deoxyribose in DNA and ribose in RNA) and nitrogenous bases: T, C, A, G in DNA and U, C, A, G in RNA. The broken-line pattern in each compound indicates the bond where a link to another building block is made (see Figure 5.4).

helices explained for the first time how a regular polymer made of repeating phosphate/sugar/base units (*nucleotides*) could replicate and encode hereditary information.

### 5.2.3 Nucleotides

The basic building block of nucleic acid polymers, the *nucleotide*, consists of a 5-membered sugar ring — deoxyribose in DNA and ribose in RNA — a phosphate, and a purine or pyrimidine base (see Figure 5.4). The unit containing just the sugar and base is called the *nucleoside*. Nucleotides are linked together through the phosphate group to form the polynucleotide chain.

### 5.2.4 Polynucleotides

When nucleotides are polymerized into nucleic acid chains by the chemical removal of water molecules, a sugar/phosphate backbone is formed. The

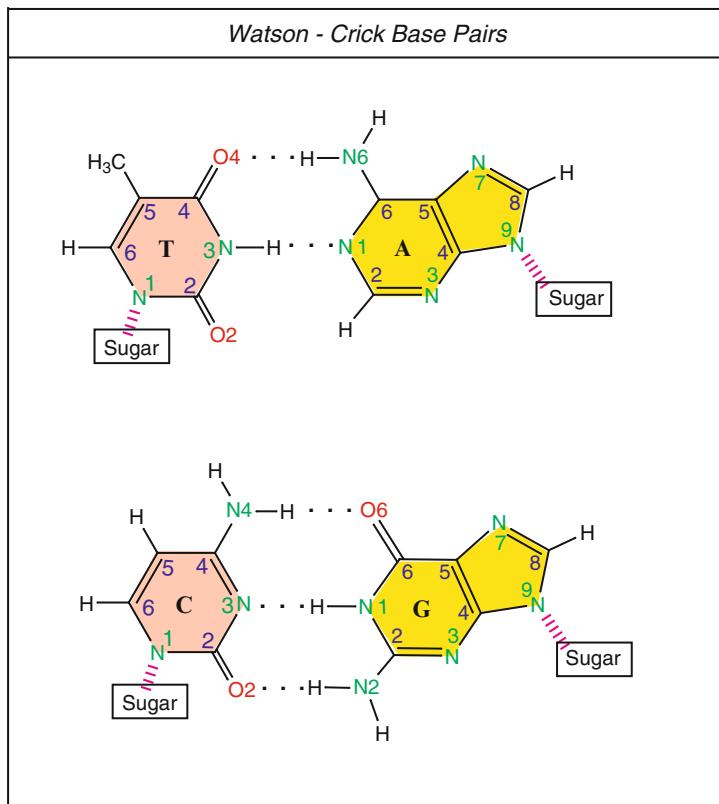


Figure 5.3. In the classic Watson-Crick base pairing scheme of DNA double helices, thymine (T) pairs with adenine (A) by forming two hydrogen bonds, and cytosine (C) pairs with guanine (G) by forming three hydrogen bonds. The hydrogen bonds are often represented by dots ( $\cdots$ ). The structural fit (e.g., as measured by  $C1'-C1'$  distances) is such that the difference in width between the two base pairs is less than 3%, allowing formation of a double helix with a nearly constant diameter.

C $3'$ -hydroxyl group of the  $n$ th nucleotide sugar is joined to the C $5'$ -hydroxyl group of the  $(n+1)$ th nucleotide by a *phosphodiester* bridge.<sup>5</sup> This negative phosphate group, located on the exterior of the helix, is readily available for physical and chemical interactions with solvent water molecules and metal ions present in the cell (see next chapter). Hydration effects are of great importance for DNA molecules in solution, since water plays a central role in stabilizing a particular helical conformation.

<sup>5</sup>An ester is an –OR group where R represents an organic chemical group.

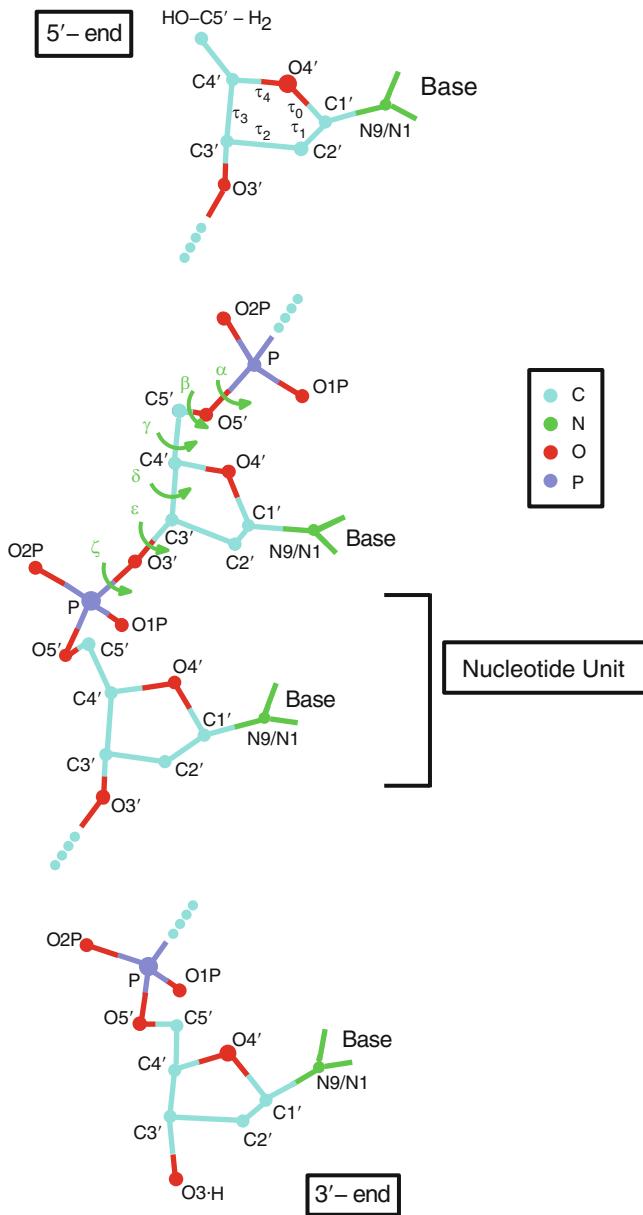


Figure 5.4. The polynucleotide chain, with standard atom labeling shown, runs from the 5' end (of the sugar atom C5') to the 3' end (of the sugar atom C3'). Nucleotide linkage is via the 3' to 5' phosphodiester bonds (i.e., C3'-hydroxyl group of one nucleotide sugar to C5'-hydroxyl group of another). Nucleic acid torsion angles are labeled as  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ,  $\epsilon$  and  $\zeta$  along the polynucleotide chain,  $\tau_0$  through  $\tau_4$  about the endocyclic bonds of the sugar, and  $\chi$  about the glycosyl bond, connecting the sugar and base. See Table 5.2 for full quadruplet sequences of these torsion angles.

### 5.2.5 Stabilizing Polynucleotide Interactions

Both *hydrogen bonding* and *base stacking* are considered to be intrinsic factors for helix stability. The estimates given for intrinsic energies of hydrogen-bonding and base-pair stacking in Chapter 7 (Subsection 7.2.1) suggest that the latter is at least as important for overall helix stability.

The bases in DNA double helices are normally linked by *hydrogen bonds* in the WC bps as depicted in Figure 5.3. Recall that in this weak noncovalent electrostatic interaction a hydrogen atom is shared between negatively-charged donor and acceptor atoms, typically nitrogens or oxygens in biological polymers. Besides WC, other hydrogen-bonding patterns involve variations in the positioning of the two bases relative to one another and in the hydrogen-bond defining atoms.

*Base-pair stacking* refers to favorable interactions between hydrogen-bonded neighboring bps (imagine weak interactions between the stacked ladder rungs). These arise from favorable van der Waals and hydrophobic contacts, which can optimize the water-insoluble areas of contact. Another factor in stability is electrostatics, which leads to preferred distributions of the partial electric charges spread over the aromatic rings. In addition, specific hydrogen bonds, other than classic WC bps, can form between adjacent bps if the bps are properly oriented. One example is *propeller-twisting* of base pairs (introduced below).

In fact, we are growing to appreciate the many hydrogen-bonding and base-pair stacking arrangements that one or more polynucleotide strands can form, including residues with modified bases, bases with bound adducts, and peptide nucleic acid structures [737, 907, 1191, 1216].

As will be further discussed, the notion of a perfectly ordered double-helical molecules is thus just an ideal reference. Many of the biological functions of DNA and DNA/protein complexes require appreciation of the sequence-specific, atomic-level variations.

### 5.2.6 Chain Notation

A DNA chain has *polarity*. Each strand of the chain has a terminal C5'-OH group on one end and a terminal C3'-OH group on the other. Thus, the two intertwined strands run in anti-parallel directions. Conventionally, the base sequence in a polynucleotide is specified for the 5' → 3' direction; the complementary strand is automatic for ideal WC bps.

To number the  $2n$  bases and  $n$  base pairs in a DNA duplex, Strand I — the sequence strand — and Strand II — its antiparallel companion — are specified, as shown in Figure 5.1. Bases along Strand I are numbered 1 to  $n$  in the 5' → 3' direction, and bases along Strand II are numbered from  $n + 1$  to  $2n$ , also in the 5' → 3' direction. Base pairs 1 to  $n$  are numbered so that they coincide with the bases of Strand I. Thus, bp 1 involves base 1 of Strand I with base  $2n$  of Strand II; bp 2 is the hydrogen-bonded pair of Strand I's base 2 with Strand II's base  $2n - 1$ ; and so on.

Nucleosides are abbreviated by a pair of letters: a lower-case Roman letter denotes the sugar type (“r” or “d”), and an upper-case Roman letter represents the base type (C, G, A, T, U). For example, rA is *adenosine* — ribose plus adenine, and dC is *deoxycytidine* — deoxyribose plus cytosine. The base prefixes are often dropped when the type of sugar is obvious. Nucleotides are abbreviated by adding a lower-case “p” (for phosphate) to the nucleoside symbol.

Thus, the sequence dGpdCpdApdC is a *tetramer* in a DNA double helix where G is at the C5'-OH end and C is at the C3'-OH end. For brevity, these lower-case “p”s are often omitted when a specific nucleic-acid sequence is specified. In the nucleic acid literature, duplex units of DNA for 2–12 base pairs are commonly termed as dimer, trimer, tetramer, pentamer, hexamer, heptamer, octamer, nonamer, decamer, undecamer, and dodecamer, respectively.

### 5.2.7 Atomic Labeling

Standard atomic numbering schemes have been recommended for nucleic acids [1080], as follows (see Figures 5.3 and 5.4):

- Sugar atoms are distinguished from base atoms by a prime suffix, and within the sugar the numbering sequence is counted clockwise from the ring oxygen in the direction of the carbon attached to the base nitrogen: O4' → C1' → C2' → C3' → C4' → O4'.
- In the polynucleotide backbone, the counting direction for torsion angles ( $\alpha, \beta, \gamma, \delta, \epsilon$  and  $\zeta$ ) is specified by the sequence: P → O5' → C5' → C4' → C3' → O3' → P.
- Base atoms are numbered systematically as shown in Figure 5.3. The procedure for labeling base atoms is different than that used for the sugar: Whereas the sugar atoms are numbered by atom *types* (O4', C1' through C4'), the base atoms are numbered according to *position* in the ring (in cytosine, for example, N1 and N3 are the two ring nitrogens, and C2, C4, C5 and C6 are the four ring carbons).

One nitrogen of the base is always connected to the C1' of the sugar by the *glycosyl* bond (N to C1'). Thus, according to the base numbering scheme, a pyrimidine base attached the glycosyl nitrogen has N1–C1', and a purine base has N9–C1'.

- When more than one oxygen or hydrogen are attached to the same atom, the two substituents are generally distinguished by the numbers 1 and 2. This applies to the two C5' hydrogens, H1 and H2, and the two oxygens at P, O1P and O2P. (Note that O2 is defined also for atoms of the pyrimidine bases). The order of these substituents is such that when we look along the counting direction of the main chain or the counting direction of the sugar ring, a clockwise direction gives substituent 1 and then 2. For example,

when we look along the  $O5' \rightarrow C5'$  bond, a clockwise counting gives substituents  $C4'$ ,  $H1$ , and  $H2$  of atom  $C5'$  (the hydrogens are generally labeled  $H5'1$  and  $H5'2$ ).

Table 5.2. Nucleic acid torsion angle definitions.

Angle	Sequence	Angle	Sequence
$\alpha$	$O3'-P-O5'-C5'$	$\tau_0$	$C4'-O4'-C1'-C2'$
$\beta$	$P-O5'-C5'-C4'$	$\tau_1$	$O4'-C1'-C2'-C3'$
$\gamma$	$O5'-C5'-C4'-C3'$	$\tau_2$	$C1'-C2'-C3'-C4'$
$\delta$	$C5'-C4'-C3'-O3'$	$\tau_3$	$C2'-C3'-C4'-O4'$
$\epsilon$	$C4'-C3'-O3'-P$	$\tau_4$	$C3'-C4'-O4'-C1'$
$\zeta$	$C3'-O3'-P-O5'$	$\chi$	$O4'-C1'-N1-C2$ (pyrimidine) $O4'-C1'-N9-C4$ (purine)

### 5.2.8 Torsion Angle Labeling

Recall that a torsion angle describes the relative orientation of a bonded 4-atom sequence  $i-j-k-l$  as the angle between the two planes defined by  $i-j-k$  and  $j-k-l$  (see Box 3.4 and Figure 3.14 in Chapter 3). As with polypeptides, the torsion angles in polynucleotide chains are denoted by a systematic set of Greek letters, as shown in Figure 5.4 and defined in Table 5.2. For the backbone rotations, the letters  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ,  $\epsilon$  and  $\zeta$  are used, in the order along the chain backbone direction.

For the sugar molecule, the endocyclic (i.e., ring) torsion angles are designated by  $\tau_0$  through  $\tau_4$  along the ring direction noted above (clockwise) so that  $\tau_j$  is the torsion angle for the sequence  $C_{j'-1}-C_{j'}-C_{j'+1}-C_{j'+2}$  where  $C_0$  is  $O4'$  and the integer  $j'$  is equal to  $j$  modulo 5 (hence  $C_5$  is  $O4'$  and  $C_{-1}$  is  $C4'$ , for example; see Table 5.2) [35].

For describing the orientation of the nitrogenous base to the sugar, the torsion angle  $\chi$  is defined about the glycosyl bond;  $\chi$  is the torsion angle of the  $O4'-C1'-N1-C2$  sequence for a pyrimidine base and the torsion angle of  $O4'-C1'-N9-C4$  for a purine base. Rotational sequences for the nitrogenous bases are not given special symbols, because the bases are approximately planar.

## 5.3 Nucleic Acid Conformational Flexibility

As in polypeptides, where conformational flexibility is limited due to steric hindrance (as described, for example, in the Ramachandran plots introduced in Chapter 3), the observed set of conformations in nucleic acids is limited by energetic and chemical considerations. A description of these conformations is more complex than those for the backbone  $\phi$  and  $\psi$  pairs and secondary-structure

elements as used for polypeptides, since a greater conformational variability exists in nucleic acid residues, with additional variation in helical parameters. Sequence and environmental factors (e.g., cations, bound ligands) affect sensitively the local structural fluctuations.

### 5.3.1 The Furanose Ring

The 5-membered furanose sugar ring is generally nonplanar in nucleic acids. One or two atoms may *pucker* out of the plane defined by the remaining skeletal ring atoms. When four ring atoms are planar, the pucker form is called *envelope*; when two atoms pucker at opposite sides of the plane defined by the remaining three ring atoms, the conformation is known as a *twist* (see Figure 5.5).

The sugar pucker type is described by the atom or atoms that deviate from that three or four-atom ring plane. Atoms displaced on the same side of C5' are designated as *endo*, and atoms displaced on the opposite side of C5' are called *exo*.

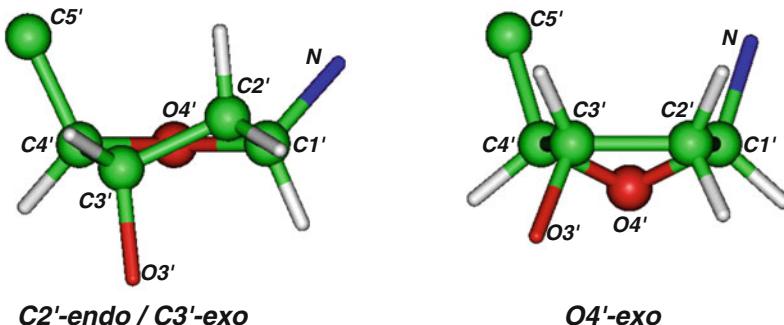


Figure 5.5. *Envelope* and *twist* puckering forms for a 5-membered ring: C2'-endo/C3'-exo symmetric twist (left) and O4'-exo envelope (right); for clarity, hydrogens attached to exocyclic atoms are not shown.

To describe nucleic acid sugar conformations, it is convenient to use a *pseudorotation path* developed on the basis of the 5-membered carbon ring cyclopentane [35, 271]. In cyclopentane, a wavelike motion between conformations of equal energy can be imagined with respect to a mean plane. This mean ring plane can be defined (in various ways) by positions of the five skeletal ring atoms. In both the Altona/Sundaralingam (AS) [35] and Cremer/Pople (CP) [271] descriptions, this sinusoidal motion is described by a wave amplitude and phase shift:  $(q, \phi)$  and  $(\tau_{\max}, P)$ , respectively, as follows.

In the AS description, the five endocyclic torsion angles are restricted to the values:

$$\tau_j = \tau_{\max} \cos [P + 4\pi/5(j - 2)], \quad j = 0, 1, 2, 3, 4. \quad (5.1)$$

In CP, the five perpendicular displacements of the ring carbons from a mean plane are described by the cosine series:

$$z_j = (2/5)^{1/2} q \cos [\phi + 4\pi/5(j - 2)], \quad j = 0, 1, 2, 3, 4. \quad (5.2)$$

The two formalisms are only equivalent under the assumption of infinitesimal displacements from a regular planar pentagon. A more general relation based on a simple analysis of model riboses was derived in [522].

The wave-like pseudorotation path described by eq. (5.1) defines 10 *envelope* conformations for  $P = 18^\circ, 54^\circ, 90^\circ, \dots, 342^\circ$  and 10 *twist* conformations for  $P = 0^\circ, 36^\circ, 72^\circ, \dots, 324^\circ$ . Figure 5.6 shows the 10 envelopes and two of the 10 twists. The quadrant terminology of N, S, E, and W is often used to describe North, South, East, and West sugar-pucker regions.

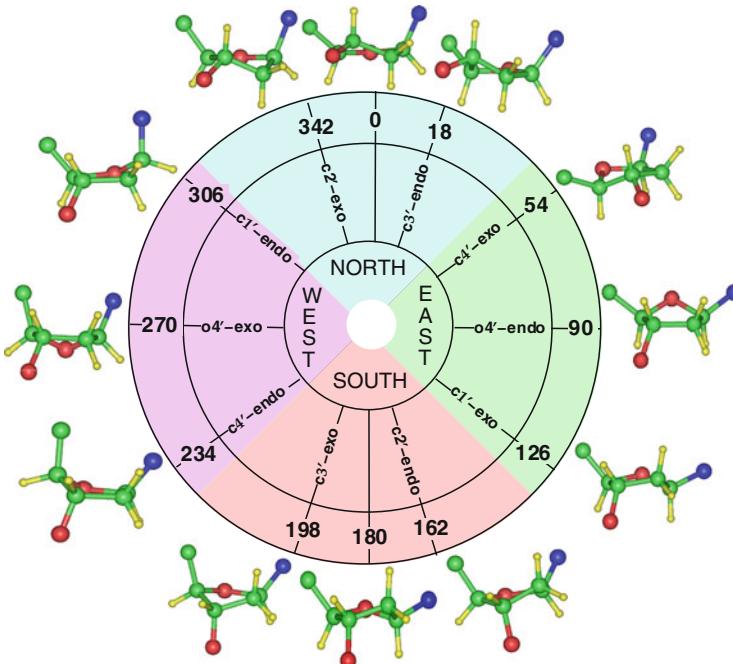


Figure 5.6. Sugar pseudorotation cycle.

Using this description for nucleic-acid sugars, we see from Figure 5.5 that O4'-exo, for example, is an *envelope* conformation with O4' puckering out of the C4'-C3'-C2'-C1' plane on the opposite side of C5'. Similarly, C2'-endo/C3'-exo is a *twist* conformation with C2' puckering out of the C4'-O4'-C1' plane on the same side of C5' and C3' puckering out of the same plane on the opposite side of C5'.

Two major types of puckering modes are observed in nucleic acid sugars. They cluster around the North C3'-endo ( $0^\circ < P < 36^\circ$ ) and South C2'-endo

( $144^\circ < P < 188^\circ$ ) puckers; see Figure 5.7. The puckering amplitude  $\tau_{\max}$  averages around  $40^\circ \pm 5^\circ$ . Still, significant deviations of sugar pseudorotation parameters are commonly observed [1080], especially as new forms of DNA are being discovered.

Figure 5.8 shows sugar puckering patterns of crystallographically determined nucleosides and nucleotides. Such analyses are possible using the above (Fourier) formalism extrapolated from cyclopentane motions. The numerical values for the two Fourier parameters are estimated by the Cartesian coordinates of atomic-resolution models.

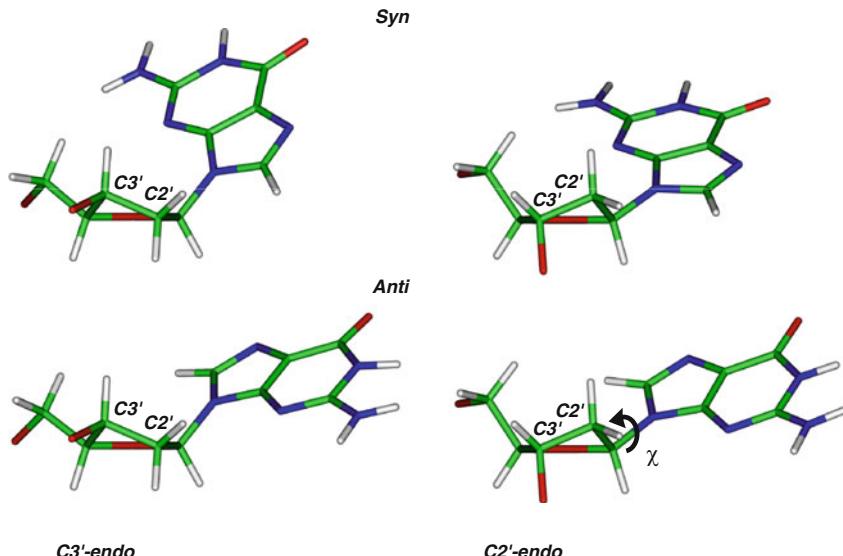


Figure 5.7. The common C3'-endo (left) and C2'-endo (right) sugar puckers for a deoxyguanosine in combination with the glycosyl torsion angle in the *syn* (top) and *anti* (bottom) conformations. Note that the 6-membered ring of G lies over the sugar in *syn* and away in *anti*.

The study of nucleic acid sugar conformations and conformational interchanges has been an active area of research, beginning with the pioneering work of Levitt and Warshel [750] (see works cited in [59]), since the sugar conformation strongly affects the overall helical structure.

### 5.3.2 Backbone Torsional Flexibility

In addition to the five internal sugar torsions described above, the six phosphodiester backbone torsion angles  $\alpha, \beta, \gamma, \delta, \epsilon$ , and  $\zeta$  are flexible but restricted to sterically allowable regions. Different values are also characteristic of various helical structures, as shown in Table 5.4 (discussed later).

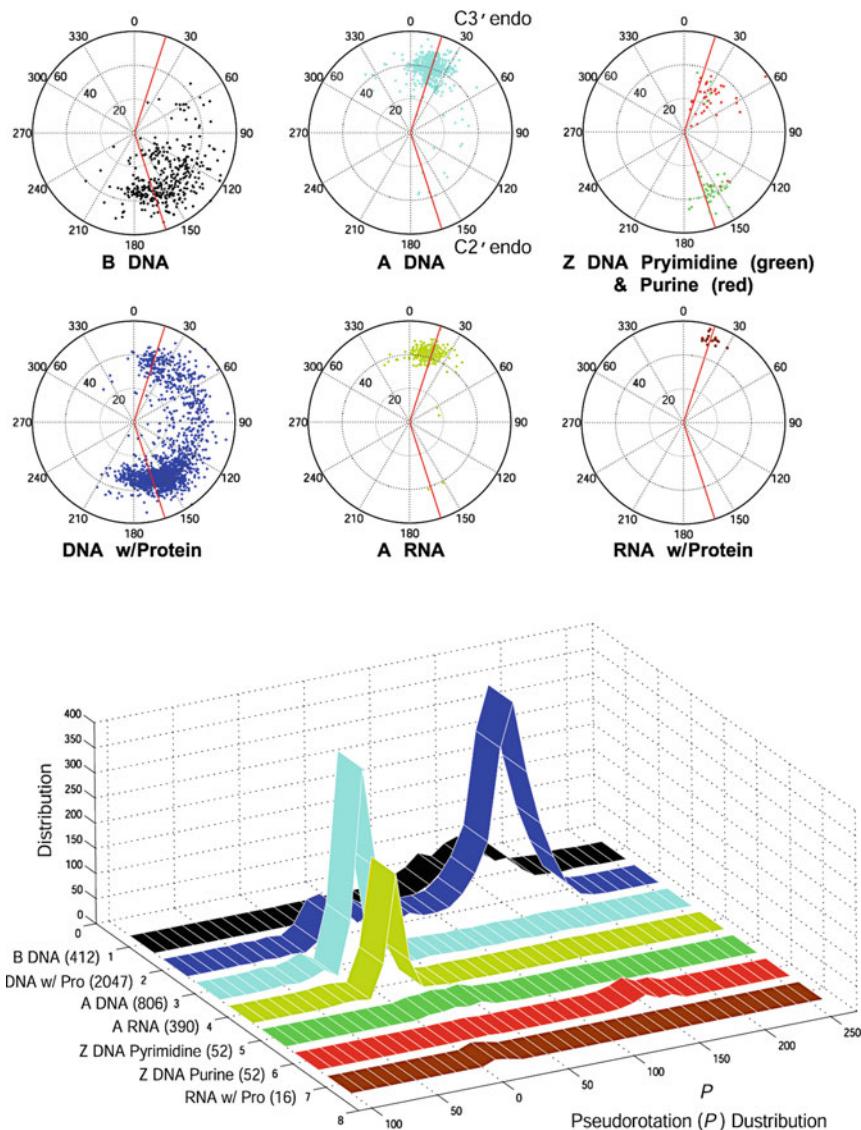


Figure 5.8. Sugar pucker conformations analyzed from the NDB for higher-resolution structures ( $<2 \text{ \AA}$ ) released as of July 31, 2009 and analyzed from the website of the 3DNA program (<http://w3dna.rutgers.edu/>). In the averaging, two nucleotides from each end are excluded. The total number of residues analyzed to obtain sugar parameters for A-DNA, B-DNA, Z-DNA pyrimidines, Z-DNA purines, A-RNA, DNA/Protein complexes, and RNA/Protein complexes is indicated in the bottom graph (in parentheses). In the sugar wheels, the radial coordinate is the phase amplitude  $\tau_{\max}$  and the angle is the phase angle of pseudorotation ( $P$ ). All the radial plots are colored monochromatically except for Z-DNA, where two colors are used for pyrimidines and purines.

Recall that a torsion angle range can be described by the three exhaustive regions of the conformational space — *gauche*<sup>+</sup> ( $\text{g}^+$ ), *gauche*<sup>-</sup> ( $\text{g}^-$ ), and *trans* ( $\text{t}$ ) (see also Figure 3.13 in Chapter 3). In terms of these classifications, both  $\alpha$  (about P–O5') and  $\gamma$  (about C5'–C4') exhibit relatively large flexibility, with the  $\text{g}^+$ ,  $\text{g}^-$ , and  $\text{t}$  positions observed. The angles  $\beta$  (about O5'–C5') and  $\epsilon$  (about C3'–O3') tend to cluster around the *trans* state, with  $\epsilon$  occupying a broad  $\text{t/g}^-$  range. The backbone torsion angle  $\delta$  (about C4'–C3') is correlated to the sugar pseudorotation pucker state, since the sugar torsion  $\tau_3$  is defined about the same C–C bond. This correlation can be described roughly by  $\delta = \tau_3 + 120^\circ$ . Finally,  $\zeta$  (about O3'–P) is rather flexible, with all three ranges observed.

See Figure 5.9 for distributions of these backbone torsion angles, as analyzed for the same high-resolution crystal structures used to generate Figure 5.8.

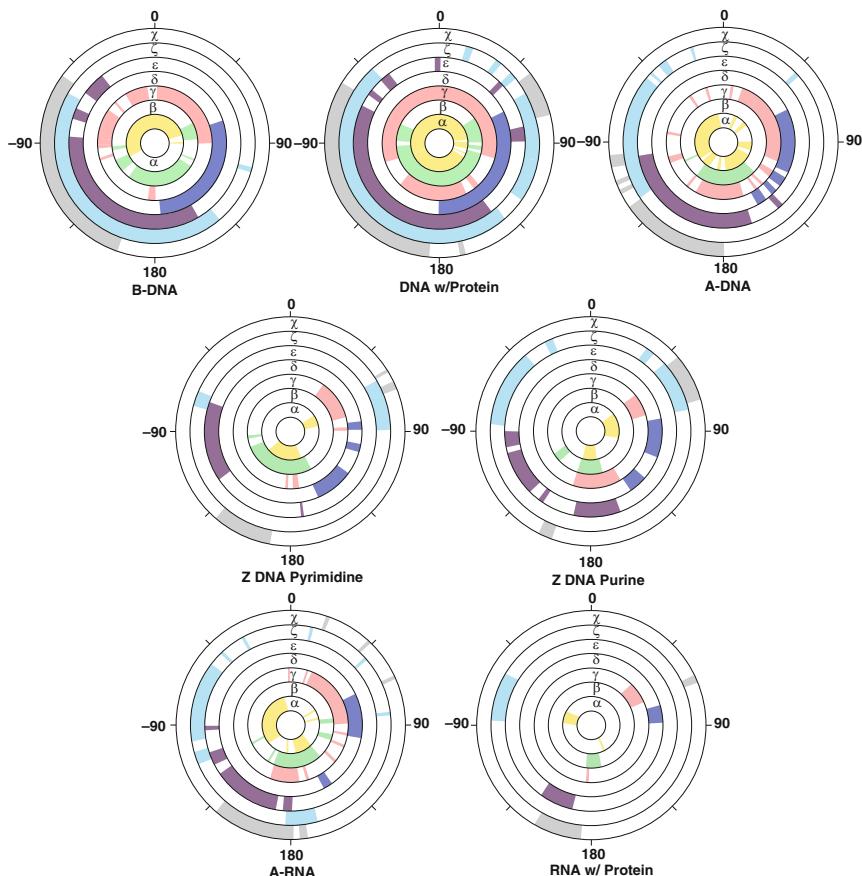


Figure 5.9. Observed ranges of nucleic-acid torsion angles for higher-resolution structures ( $<2 \text{ \AA}$ ) for the same residues analyzed for Figure 5.8.

### 5.3.3 The Glycosyl Rotation

Relative to the sugar moiety, the base can assume two major orientations about the glycosyl C1'-N bond: *syn* and *anti* (torsion angles of 0 and 180°, respectively) [1080]. Roughly speaking, four major conformations are favored. They correspond to the combinations of C3'-endo and C2'-endo sugar pockers with *syn* and *anti* values for  $\chi$ . Favored combinations of { $P$ ,  $\chi$ } pairs vary for the different nucleosides or nucleotides. They depend on the chemical structure of the sugar, the size of the base, and the nature of the nucleoside substituents (chemical derivatives). For example, deoxyribose nucleosides and nucleotides prefer the C2'-endo conformation over C3'-endo [944]. In pyrimidine nucleotides, the *anti* orientation of  $\chi$  about the sugar ring is finely tuned by the sugar pucker [1303].

In Figure 5.7 the two orientations of *syn* and *anti* bases are illustrated for deoxyguanosine in combination with the two common sugar pockers. The {C3'-endo, *syn*} combination of this figure (top left) is that observed in the purine of Z-DNA helices while the {C2'-endo, *anti*} combination (bottom right) is typically observed in the B-DNA varieties (and in Z-DNA pyrimidines). Figure 5.9 also shows the distributions of  $\chi$  in various nucleic-acid structures.

### 5.3.4 Sugar/Glycosyl Combinations

To further illustrate conformational tendencies in polynucleotides, we generated *adiabatic maps*<sup>6</sup> for two models of deoxyadenosine in the { $P$ ,  $\chi$ } space (Figure 5.10) based on the CHARMM force field [415, 803, 804]. The first model approximates solvation simply with a distance-dependent dielectric function. The latter uses explicit representation of water molecules.<sup>7</sup>

The adiabatic maps were calculated by dividing the { $P$ ,  $\chi$ } grid to 3600 points, using 6° intervals for each angle. For each selected  $P$ , the values of the 5 endocyclic sugar torsion angles were determined from eq. (5.1) using  $\tau_{\max} = 40^\circ$ . Starting structures were then generated from the set of variables { $\tau_0, \tau_1, \tau_2, \tau_3, \tau_4, \chi$ } and minimized over the remaining degrees of freedom. Of course, such a map only provides a reference for conformational flexibility, since constraining (or freezing) the angles does not allow complete energy relaxation over all the available degrees of freedom.

We note from Figure 5.10 essentially the same trends for the solvated (top) and more approximate (bottom) models of four minima corresponding to

---

<sup>6</sup>An adiabatic map is a simple way to examine molecular motion by characteristic low-energy paths along a prescribed reaction coordinate (i.e., variations in specific conformational variables). For each combination of these conformational coordinates, the entire potential energy of the system is minimized to approximate behavior for the motion under study. Though simple in principle, specification of the reaction coordinate is difficult in general, and the neglect of other degrees of freedom in the process is clearly an approximation whose validity depends on the motion in question.

<sup>7</sup>Namely, each nucleoside is enveloped in a water sphere of radius 11 Å, and the nonbonded interactions are truncated at 12 Å using a 2 Å buffer region, a potential shift function for the electrostatic terms and a potential switch function for the van der Waals terms; see Chapter 10 for details of such procedures.

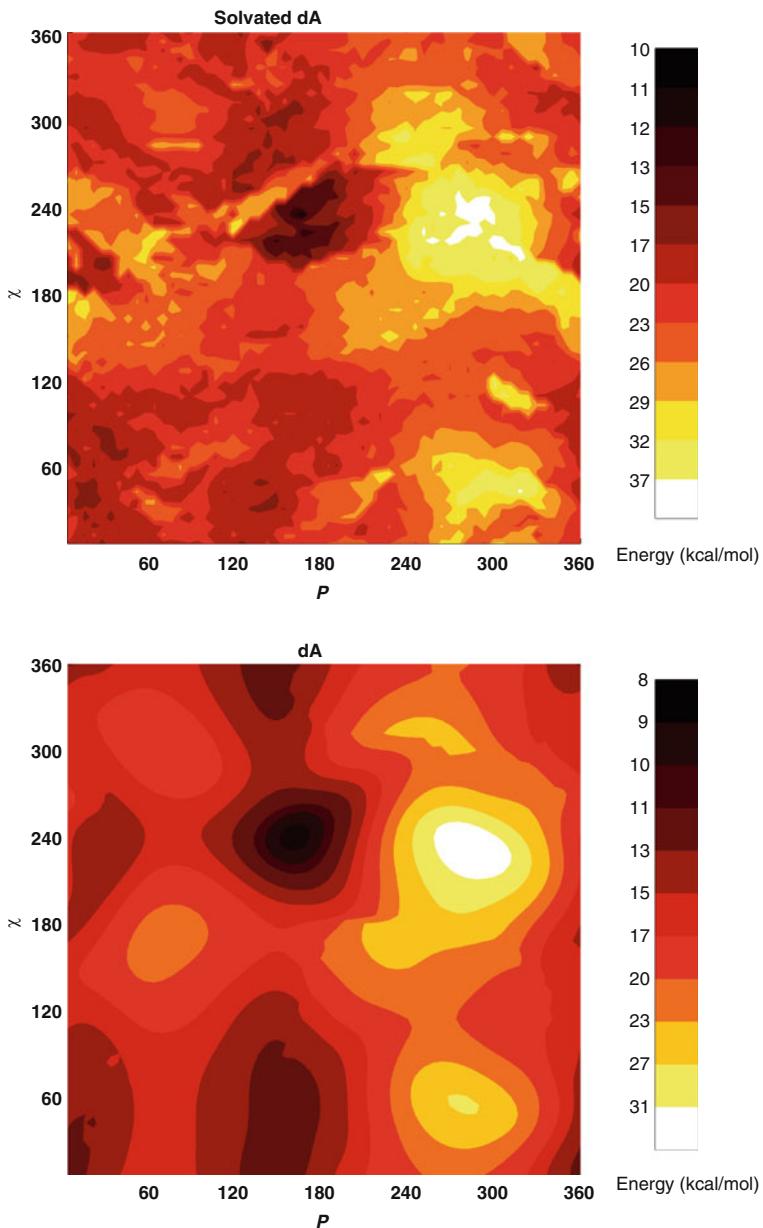


Figure 5.10. Adiabatic maps for two models of deoxyadenosine in the  $\{P, \chi\}$  space, the sugar pseudorotation parameter and the sugar-to-base torsion angle, respectively, as calculated with the all-atom CHARMM 27 force field for nucleic acids [803, 809]. Hydration is modeled with approximately 140 water molecules to produce the top figure, and approximated with a distance-dependent dielectric function to generate the bottom figure, as described in the text. Only the internal energy of the dA system (i.e., excluding water/dA and water/water energies) is used for plot generation in both cases.

the  $\{P, \chi\}$  combinations of C2'-endo/anti, C3'-endo/anti, C2'-endo/syn, and C3'-endo/ syn; two maxima are also apparent. See also [414] for a recent quantum-mechanical study of the glycosyl torsion energetics in DNA.

### 5.3.5 Basic Helical Descriptors

Besides the sugar, backbone, and glycosyl conformational variables introduced above, additional helical parameters are defined to describe the global arrangement of a base pair (bp) in a double helix [307, 1080]. See Dickerson [306] and the text [139] for complete definitions of all translational and rotational variables.

Names, symbols, and sign conventions were decided at an international workshop [307]. Before we introduce some of these (as well as other) conformational variables, we define basic features of helix descriptors that are relevant for model helices, whose geometries can be described by characteristic values, as shown in Tables 5.3 and 5.4.

- *Helix sense* refers to the handedness of the double helix.<sup>8</sup>
- *Helix pitch* (per turn),  $P_h$ , measures the distance along the helix axis for one complete turn (see Figure 5.1).
- *Number of residues per turn*,  $n_b$ , is the number of bps for every complete helix turn.
- *Axial rise*,  $h$ , is the characteristic vertical distance along the double helix axis between adjacent bps.
- *Unit twist or rotation per residue* for a repetitious helix is  $\Omega = 360^\circ/n_b$  and describes the characteristic rotation about the global helix axis between two neighboring base pairs.
- *Helix diameter* refers to the geometric diameter of the helical cylinder (around 20 Å for DNA).
- *Major and minor grooves* (Mgr and mgr in Figure 5.1) refer to the spaces generated by the asymmetry of the DNA. The two different-sized grooves are most apparent from a side view of the double helix (Fig. 5.13). The *minor groove side* is defined as the space generated along the edge closer to the two glycosyl linkages of a bp, and the *major groove side* is generated by the other edge, *farther* from those links (see also Figure 5.11).

In the classic DNA described by Watson and Crick, the minor groove is narrower (around 6 Å wide) compared to the major groove (which is doubly wide) and slightly deeper (8.5 vs. 7.5 Å). For the A-model of DNA

---

<sup>8</sup>In a right-handed form, a right hand held with the thumb pointing upward in the direction of the helix axis will wrap right (counterclockwise) and around the axis to follow the chain; a left hand with an upward-pointing thumb will wrap to the left (clockwise) to follow the chain direction of a left-handed helix.

discussed below, the ‘minor’ groove is as large, or larger, and also more shallow, than the ‘major’ groove according to the above definition.

Characterizing helical grooves is important for describing interactions of nucleic acids with solvent molecules and with proteins. A larger accessible area of a groove can facilitate nonspecific, as well as sequence-directed, protein recognition and binding. The edges of the bps, which form the bottom of the grooves, contain nitrogen and oxygen atoms available for contacting protein side chains via hydrogen bonds. The hydrogen bonds form the basis of sequence-specific recognition of DNA by proteins.

Table 5.3. Mean properties of representative DNA forms.

Property	A-DNA	B-DNA	Z-DNA
Handedness	Right	Right	Left
Representative Structures	GGCCGGCC CGTATACC	CGCGAATTGCG	CGCGCG
Bps/turn	11	10	12 (6 dimers)
Rise/base pair	2.6 Å	3.4 Å	3.8 Å (ave.)
Helix diameter	≈26 Å	≈20 Å	≈18 Å
Helix pitch	≈28 Å	≈34 Å	≈45 Å
Twist/residue	33°	36°	–60°/dinuc.
Bp inclination	20°	0°	–7°
Sugar pucker	C3'-endo	C2'-endo	C2'-endo (C)/ C3'-endo (G)
Glycosyl rotation	<i>anti</i>	<i>anti</i> (higher)	<i>anti</i> (C)/ <i>syn</i> (G)
Major groove	narrow & deep	wide & deep	convex
Minor groove	very wide & shallow	narrow & deep	very narrow & deep

### 5.3.6 Base-Pair Parameters

The geometric variables above and others are necessary to describe the *global* and *local* arrangements of base pairs in nucleic acid helices. The global parameters describe the overall arrangement of the bps in double-stranded helices, while the local variables specify the orientation between successive bps. The global helical parameters are thus measured for a particular bp with respect to the overall (global) helix axis, while the local variables are defined in the local framework of two successive bps. *The global and local helical parameters can be entirely different quantities. See [678] for related transformations between global and local variables.*

Table 5.4. Selected parameters for constructing model DNA from nucleotide geometric variables (dinucleotide for Z-DNA) according to Figure 5.13, as developed in [1102] and used to generate the structures in Figure 5.13. See also Figure 5.11 for definitions of the translational variables  $dx$  and  $dy$  and the rotational variables tip ( $\theta$ ), inclination ( $\eta$ ), and twist ( $\Omega$ ). The parameter  $h$  is the helical rise.

Helix	$\alpha$	$\beta$	$\gamma$	$\delta$	$P/\tau_{\max}$	$\chi$
A-DNA	-62	173	52	88	3/38	-160
B-DNA	-63	171	54	123	131/36	-117
Z-DNA (dG)	47	179	-165	9	-1/23	68
Z-DNA (dC)	-137	-139	56	138	152/35	-159

Helix	$dx$	$dy$	$h$	$\theta$	$\eta$	$\Omega$
A-DNA	4.0	0	2.87	0	13.5	32.2
B-DNA	0	0	3.33	0	0	37.3
Z-DNA (dG)	-3.0	2.5	-3.72	0	-7	52 (G→C)
Z-DNA (dC)	-3.0	-2.5	-3.72	0	-7	8 (C→G)

In addition to these two groups of conformational variables, other parameters describe the orientation of the *two bases* in a hydrogen-bonded bp.

Below, we define some of the global parameters for bp orientations (like *tip* and *inclination*), local variables (associated with successive bps) like *roll*, *tilt*, and *twist*, and a variable that describes orientations within a bp (*propeller twist*). See Dickerson [306] for complete definitions of all rotational and translational variables.

The description of these quantities requires definition of a reference coordinate frame, which we introduce next.

### Reference Frame

The commonly used reference frame shown in Figure 5.11 [307] defines a coordinate system with unit vectors  $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$  so that  $\mathbf{e}_1$  represents the short bp axis,  $\mathbf{e}_2$  represents the long bp axis, and  $\mathbf{e}_3$  is the normal to  $\mathbf{e}_1$  and  $\mathbf{e}_2$  which completes the right-handed coordinate system (i.e., defined as the cross product  $\mathbf{e}_3 = \mathbf{e}_1 \times \mathbf{e}_2$ ). The direction of the long-bp axis can be defined by connecting the two C1' atoms of the pyrimidine and purine atoms. The short-bp axis (also called the *dyad*) can be constructed by passing a perpendicular vector from the midpoint of this C1' (purine) to C1' (pyrimidine) line. The intersection of the dyad with the C8 (purine) to C6 (pyrimidine) line is considered the origin of this plane.

An alternative standard reference frame describing nucleic-acid bp geometry was proposed at an international workshop held in 1999 in Tsukuba, Japan [941].

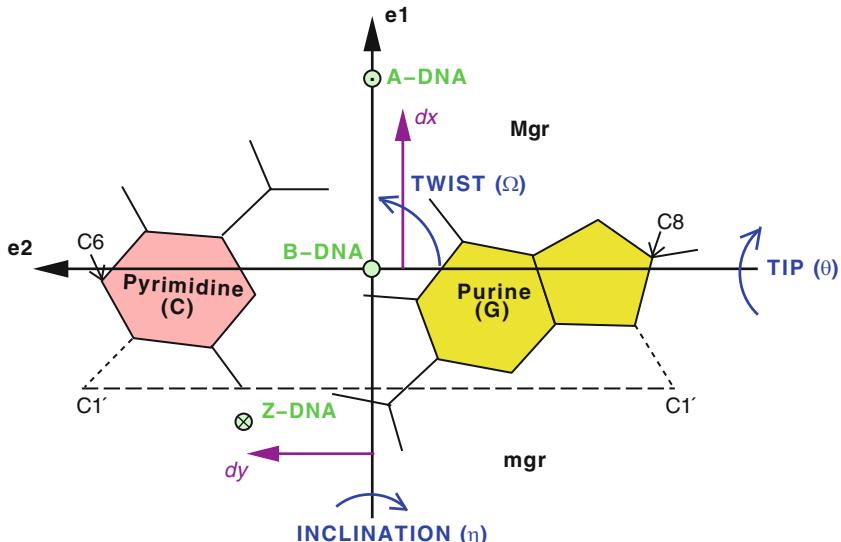


Figure 5.11. The local base-pair coordinate system  $\{e_1, e_2, e_3\}$  representing the short base-pair axis, the long base-pair axis, and the normal to both which completes a right-handed coordinate system. Associated translational and rotational parameters are indicated as detailed in the text. The symbols Mgr and mgr define major and minor groove sides of the base pair, respectively. The positions of A-DNA and Z-DNA helix centers are only illustrated for perspective relative to B-DNA. The global helix direction for Z-DNA would point in the opposite direction (down from the paper plane) according to standard definitions [307].

#### Global Variables (Base Pair Orientations With Respect to Helical Axis)

The reference frame defined above can be used to define deviations of the bp as a whole with respect to the overall helical axis. These include the rotational variables *tip* and *inclination*, and translational variables like *dx* and *dy*.

- Tip ( $\theta$  in standard conventions) measures the rotational deformation of the bp as a whole about the long bp axis  $e_2$ . It is considered positive when the rotation is clockwise as shown in Figure 5.11, moving the far side of the bases, or the major groove side, as viewed along  $e_1$ , below the paper plane.
- Inclination ( $\eta$  in standard conventions) measures the rotational deformation of the bp as a whole about the short bp axis  $e_1$ . This angle is considered positive when it is clockwise as shown, moving the far base when viewed along  $e_2$  (G in Figure 5.11) down below the paper plane.
- The displacement parameters *dx* and *dy* denote the translations of the mean bp plane from the global helical axis, along  $e_1$  and  $e_2$ , respectively (see Figure 5.11). They indicate the shift of the bp origin (the point through which  $e_3$  passes for a particular helical model). A positive *dx* indicates

translation towards the major groove direction, and a positive  $dy$  denotes displacement toward the first nucleic acid strand of the duplex (see Strand I in Figure 5.12).

For A and B-DNA, the helix axis lies approximately on the dyad, but for Z-DNA the helix axis is displaced from the dyad toward a pyrimidine atom and points down rather than up.

The A-DNA double helix lies on the major groove side ( $dx > 0$  as shown in Figure 5.11 with respect to the B-DNA helical axis), while the Z-DNA helix lies on the minor groove side ( $dx < 0$ , as shown in Figure 5.11). Note that the signs of  $dx$  and  $dy$  should be reversed when meaning the displacements of the mean A-DNA and Z-DNA bp planes from their respective global helical axes.

### Local Variables (Base-Pair Step Orientations)

The *roll*, *tilt*, and *twist* angles (see Figure 5.12) define the rotational deformations that relate the local coordinate frames of two successive bps. The three local translational variables are *slide*, *shift*, and *rise*.

- Roll ( $\rho$ ) defines the deformation along the long axis of the bp plane and describes DNA groove bending: a positive roll angle opens up a bp-step towards the minor groove, while a negative roll angle opens up a bp-step towards the major groove.
- Tilt ( $\tau$ ) is the deformation defined with respect to the short axis of the bp plane. A positive tilt angle opens the bps on the side of the sequence strand (Strand I in Figure 5.12).
- Twist ( $\Omega$ ) is the helical rotation between successive bps, as measured by the change in orientation of the C1'-C1' vectors between two successive bps, projected down the helix axis, as shown in Figure 5.12.
- The translational *slide* ( $Dy$ ) motion describes the relative displacement of successive bps along their long axes as measured between the midpoint of each pyrimidine-purine long-bp axis. It is considered positive when the direction is toward the first nucleic acid strand (i.e., positive  $dy$  direction), as shown in Figure 5.12. The other local translational variables are the shift ( $Dx$ ) and rise ( $Dz$ ).

### Deviations Within a Base Pair

- The propeller twist ( $\omega$  in standard conventions) measures the angle between the normal vectors associated with the planes of the two bases in a hydrogen-bonded pair (from the torsion angle between the individual base planes). Imagine the motion of a helicopter propeller where the two bases twist in opposite directions about the long bp axis **e2** (one up and one down), as shown in Figure 5.12.

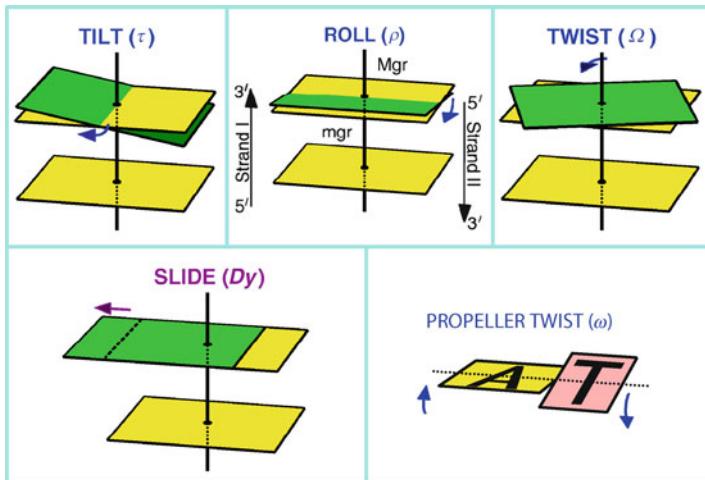


Figure 5.12. Base-pair step (tilt  $\tau$ , roll  $\rho$ , twist  $\Omega$ , slide  $Dy$ ) and base pair (propeller twist  $\omega$ ) orientation parameters, with positive displacements shown; see text for details. Mgr and mgr denote the major and minor-groove sides of the bps, respectively. The sequence strand (I) and the complementary strand (II) are indicated in the roll illustration.

According to nucleic-acid conventions [307], when the angle is viewed from the minor groove edge of a bp, it is positive when the base on the left moves its minor groove edge up while the base on the right moves down. Alternatively, we define a positive angle when each base rotates in a counterclockwise manner when viewed from its attached sugar/phosphate backbone. The propeller twist has a *negative* sign under normal conditions.

- The two other rotational variables that describe deformations within a hydrogen-bonded bp are *buckle* ( $\kappa$ ), about the short bp axis **e1**, and *opening* ( $\sigma$ ), about the **e3** [307].

## 5.4 Canonical DNA Forms

Small changes in the global and local parameters introduced above can lead to large overall changes in helix geometries. The double helix described by Watson and Crick in 1953 — now known as B-DNA — was deduced by adjusting wire models so as to fit the X-ray diffraction patterns recorded in the 1950s from calf thymus DNA fibers, first manually and later by various model-building and refinement analyses (summarized in [215]). The fiber diffraction patterns provide an excellent reference for describing features of canonical B-DNA forms since they are generally devoid of the end effects evident in analysis of crystal structures of DNA oligomers. They also represent average structures over all sequences in the fiber.

The right-handed B-form is the dominant form under physiological conditions. One possibility for its prevalence is that the B-DNA helix can be smoothly bent about itself to form a (left-handed) superhelical (*plectoneme* or toroid-like form; see next chapter) with minimal changes in the local structure. This was first suggested by Levitt by early molecular simulations [745]. This deformability property facilitates the packaging of long stretches of the hereditary material in the cell (especially of circular and topologically-constrained DNA) by promoting volume condensation as well as protein wrapping.

Yet we now recognize numerous variations in polynucleotide structures — both helical and nonhelical forms — that depend profoundly on the nucleotide sequence composition and the environment (counterions, relative humidity, and bound ligands or other biomolecules).

The canonical B-DNA was deduced from X-ray diffraction analyses of the sodium salt of DNA fibers at 92% relative humidity. Another form of DNA — now termed A-DNA — emerged from early X-ray diffraction studies of various forms of nucleic acid fibers at the much lower value of 75% relative humidity. This alternative helical geometry is prevalent in double-helical RNA structures and in duplex DNA under extreme solvation conditions in certain sequences (such as runs of guanines).

Though both these early diffraction-based models were inherently low in resolution and contained several incorrect structural details, later analyses of single DNA crystals concurred with these basic fiber diffraction findings. The DNA fiber structure analyses also served as a reference by which to analyze the sequence-dependent trends that emerged from oligomer crystallography [215].

Both the A and B-DNA forms are right handed. A rather surprising finding, first discovered by single crystal X-ray diffraction and rediscovered 25 years after Watson and Crick's description of DNA, was a peculiar left-handed helix with a zigzag pattern. Andrew Wang, Alexander Rich, and their collaborators observed this form in crystals of cytosine/guanine polymers (dCGCGCG) at high salt concentrations and dubbed it Z-DNA (for its zigzag pattern) [1327]. This high ionic environment stabilizes Z-DNA relative to B-DNA by shielding the closer phosphate groups on opposite strands and hence minimizing the otherwise increased repulsive interactions.

The biological function of Z-DNA remains in the forefront of research, but recent evidence suggests that the conversion of helical segments from B to Z-like acts as a genetic regulator.

Below, these three families of DNA helices are detailed; see Figures 5.13, 5.14, and 5.15 for comparative illustrations.

### 5.4.1 B-DNA

B-DNA can be distinguished by the following characteristics:

- The helix axis runs through the center of each bp ( $dx, dy \approx 0$ ).

- The bps stack nearly perpendicular to the helix axis (small inclination of the bps and very small roll and tilt values). This implies that bases in adjacent steps of the same strand overlap vertically (*stack*), and bases on the opposite strands do not stack.
- The mean helical twist ( $\Omega$ ) is about 34–36°.
- There are about 10–10.5 bps per turn.
- Deoxyriboses favor a C2'-endo sugar conformations (S region of pseudorotation cycle); see Figure 5.8.
- The glycosyl bond orientation is typically higher *anti*; see Figure 5.9.
- The major groove is wider (12 Å) and deeper (8.5 Å) than the minor groove, which is 6 Å wide and only slightly less deep.

Overall, these features produce a model helix of the form shown in Figures 5.13, 5.14, and 5.15. Note that the top view in the space-filling stereo figure (5.15) reveals no hole in the helix cylinder since the global helix axis intersects the bps. Ordered water spines in DNA structures have been reported along the minor groove and around the phosphate groups (see next chapter) [125].

#### 5.4.2 A-DNA

The A-DNA helix is very different in overall appearance than B-DNA. Specifically:

- The bp center is shifted from the global helix axis ( $dx \approx 4$  Å,  $dy \approx 0$  in Figure 5.11).
- A prominent inclination is noted for the bp planes, as large as 20° on average. This implies a combination of both intrastrand and interstrand base stacking for most sequences (the exception being pyrimidine/purine steps, which overlap in an interstrand manner only).
- The mean helical twist is less than for B-DNA, about 33°.
- There are thus 11 bps per turn, producing a shorter helix length than B-DNA for the same number of bps.
- The sugars favor a C3'-endo pucker (N region of the pseudorotation cycle) rather than C2'-endo as in B-DNA; see Figure 5.8.
- The glycosyl bond orientation is typically *anti*, as in B-DNA; see Figure 5.9.
- The minor groove is not as deep as in B-DNA, but the major groove is narrower and deeper than the minor groove.

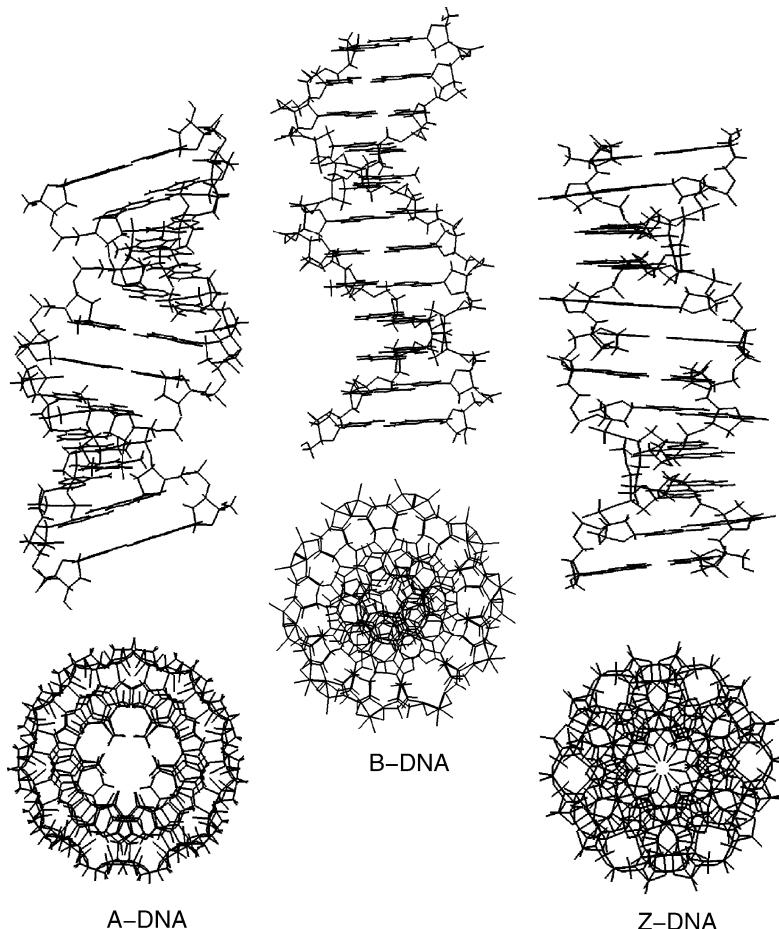


Figure 5.13. Model A, B, and Z-DNA (top and side views) as generated by the polynucleotide building program developed in [1102] based on parameters listed in Table 5.4.

From Figure 5.13, note the dramatic inclination of the bps and the hollow top view of the helix, a consequence of the bps being pulled closer to the sugar/phosphate backbone.

A-DNA regions might exist within a generally B-DNA helix (e.g., in runs of poly(dG)-poly(dC)) under extremes conditions only. Certain RNA molecules that adopt partially double-helical forms, such as tRNA, rRNA, and parts of mRNA, tend to be A-like in the duplex regions, with characteristic C3'-endo sugar puckers, because the B-DNA conformation leads to steric clashes between the two sugar hydroxyl groups.

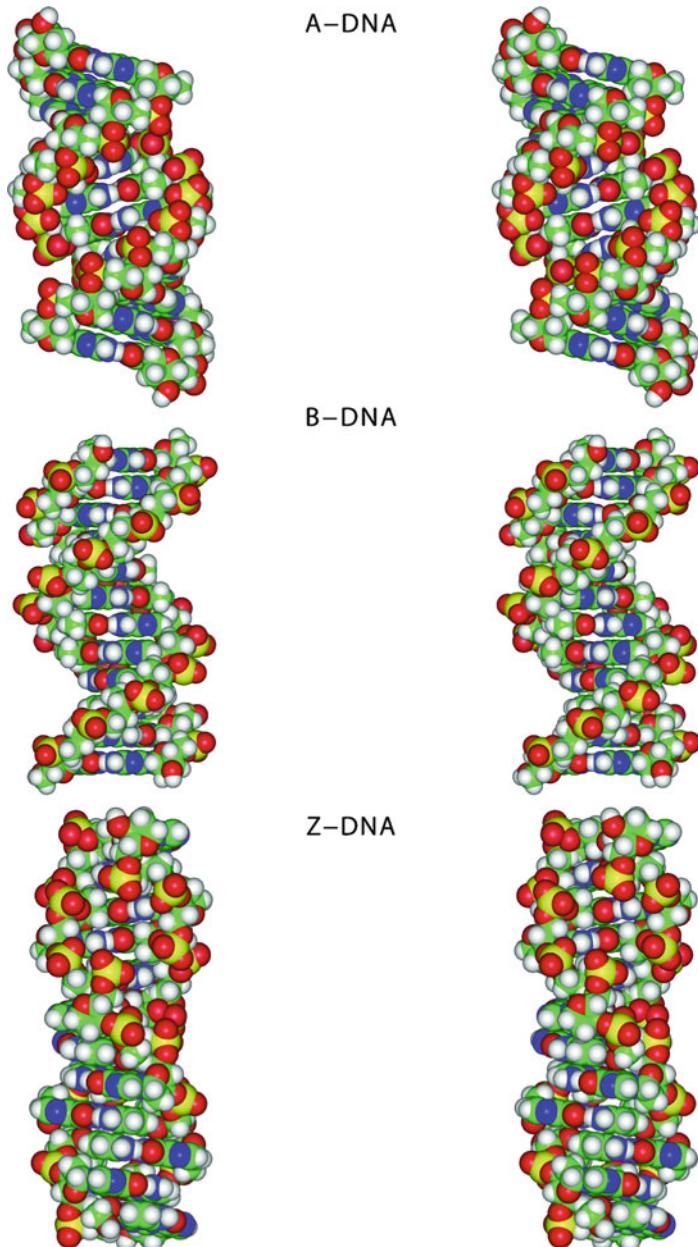


Figure 5.14. Space-filling stereo figures of model A, B, and Z-DNA, as generated by the polynucleotide building program developed in [1102] based on parameters listed in Table 5.4. The stereo side-view images are rotated about 8° relative to one another, with the center of images separated by about 6.5 cm, the average distance between two human eyes. Images are prepared for cross-eyed viewing from about 46–51 cm. See Figure 5.15 for corresponding top views.

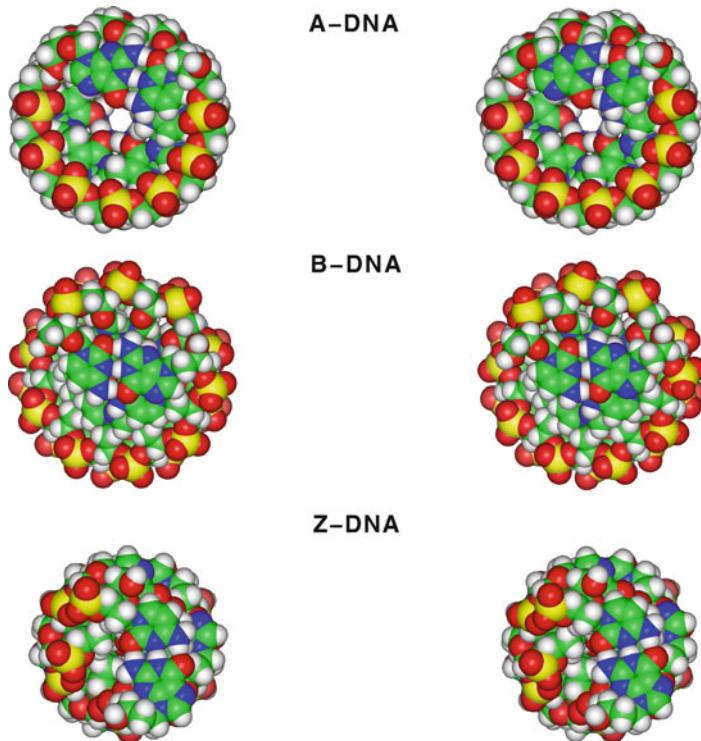


Figure 5.15. Space-filling stereo figures of model A, B, and Z-DNA, as generated by the polynucleotide building program developed in [1102] based on parameters listed in Table 5.4. The stereo top-view images are rotated about  $8^\circ$  relative to one another, with the center of images separated by about 6.5 cm, the average distance between two human eyes. Images are prepared for cross-eyed viewing from about 46–51 cm. See also Figure 5.14 for corresponding side views.

### 5.4.3 Z-DNA

In contrast to the A and B-DNA models, specific sequences are required for Z-DNA (e.g., alternating GC). The repeating unit for Z-DNA is a dinucleotide rather than a mononucleotide. The regular composition of alternating pyrimidines and purines allows alternating features in geometry, for example in the sugar and glycosyl orientations. Besides its distinguishing left-handedness, the following are properties of the slender Z-DNA helix (modeled for GC polymers):

- The bp center is shifted slightly from the helix axis along **e1** and **e2** ( $dx \approx -3 \text{ \AA}$ ,  $dy \approx \pm 2.5 \text{ \AA}$  in Figure 5.11).
- The bps are inclined slightly ( $\eta \approx -7^\circ$ ).
- The mean helical twist is about  $60^\circ$  *per dimer*, with about  $52^\circ$  for the G → C step and  $6^\circ$  for the C → G step [1102].
- There are 12 bps per turn.

- The sugars adopt C2'-endo conformers at C but C3'-endo forms at G; see Figure 5.8.
- The glycosyl link is *anti* for C but *syn* for G in alternating CG sequences; see Figure 5.9.
- The major groove bulges out and the minor groove is narrow and deep.

Like the A-form, a stretch of Z-DNA might occur within B-DNA, but direct experimental evidence is lacking. The negative supercoiling of naturally occurring DNA (see next chapter) may also promote Z-DNA, or other left-handed, helix formation.

### Biological Significance

The biological role of Z-DNA forms continues to be an active area of research. It has been recently found that when DNA adopts the Z conformation, an editor RNA molecule — the double-stranded RNA enzyme adenosine deaminase (ADAR1) — can bind to left-handed DNA and alter the base in a codon (for example from adenine to guanine) and thereby produce alternative forms of the translated proteins. The Z-DNA binding domain of ADAR1, namely Z $\alpha$ , was recently co-crystallized with a 6-bp DNA fragment [1147]. This newly-identified role for Z-DNA — a regulator of the nucleotide template which directs protein synthesis — may have practical benefits as another element of biological control.

#### 5.4.4 Comparative Features

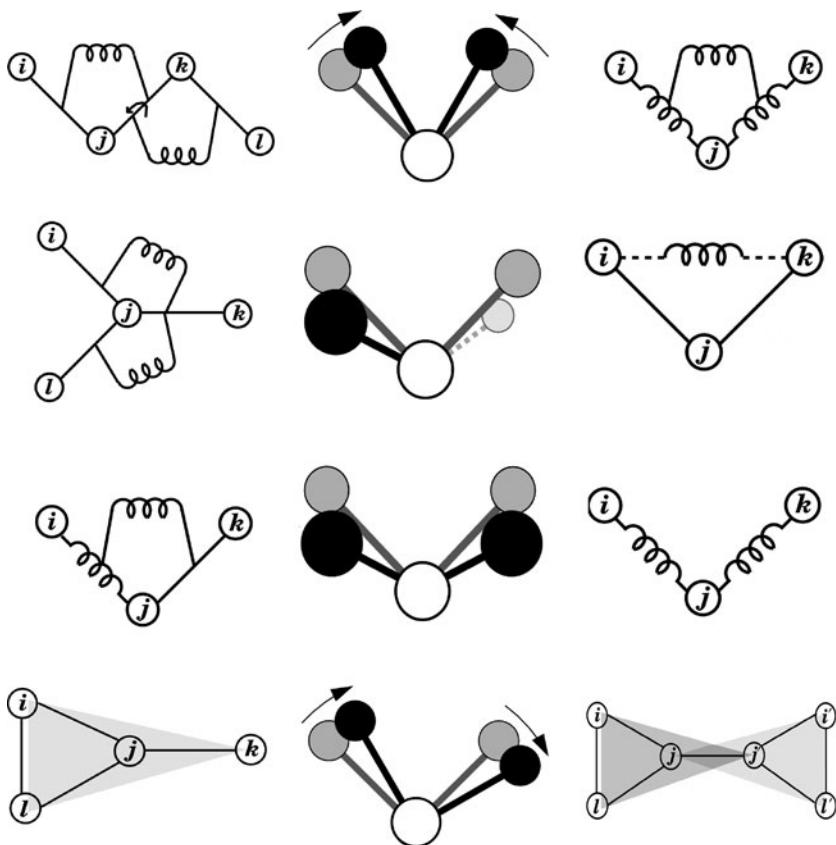
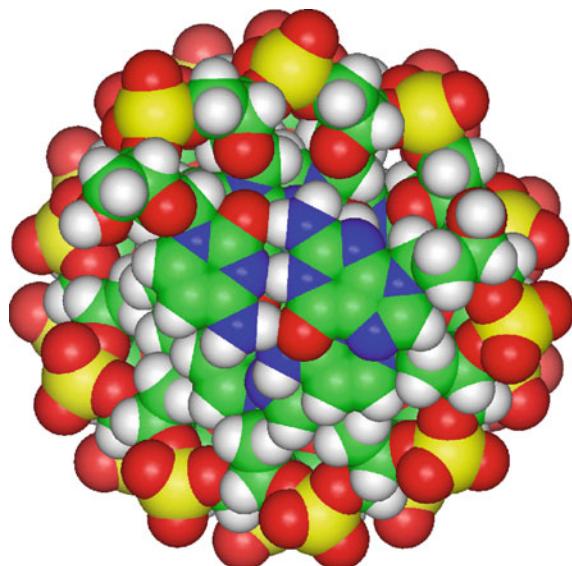
Table 5.3 summarizes basic features of model A, B, and Z-DNA forms as compiled and reconciled from several textbooks. See also Figure 5.8 for sugar analysis and Figure 5.9 for torsion-angle analysis.

Such model helices can also be constructed from building blocks of nucleotides (for A and B-DNA) and dinucleotides (for Z-DNA), for example using the geometric variables shown in Table 5.4 [1102].

In practice, departures from model-helix variables occur frequently and the observed ranges in some of these quantities are quite large and differ markedly from one residue to the next.

For example, analyses of high-resolution DNA structures collected since the early 1980s indicate a mean twist angle ( $\Omega$ ) for the right-handed B-DNA of around 34° (rather than 36°), though a broad sequence-dependent range (roughly 20–50°) is observed. The helical twist of 34° corresponds to about  $n_b = 10.5$  bps per turn (rather than 10), 36 Å for the helical pitch, and  $h = 3.4$  Å for the axial rise (or rise per residue).

A large field of research is devoted to unraveling the sequence-dependent features of DNA. These sensitive patterns are important for numerous biological functions involving DNA, such as protein binding and transcription regulation. The next chapter introduces basic elements of DNA sequence effects, DNA hydration and ion interactions, DNA/protein interactions, RNA structure, cellular aspects of DNA organization, and DNA supercoiling.



# 6

## Topics in Nucleic Acids Structure: DNA Interactions and Folding

### Chapter 6 Notation

SYMBOL	DEFINITION
$h$	helical rise, or base-pair step separation (e.g., 3.4 Å)
$k_B$	Boltzmann's constant
$n$	number of DNA base pairs
$n_b$	number of base pairs per DNA turn
$p_b$	DNA bending persistence length
$p_{tw}$	DNA twisting persistence length
$r$	elastic ratio of bending to twisting elastic constants, $A/C$
$s$	arclength
$A$	DNA bending rigidity
$C$	DNA torsional-rigidity constant
$Dy$	slide (between successive base pairs)
$\mathcal{L}$	DNA contour length
$Lk$	linking number (topological invariant)
$Lk_0$	reference linking number
$\langle R^2 \rangle$	mean square displacement
$T$	temperature
$\langle T_c \rangle$	site juxtaposition time
$Tw$	total twist (geometric variable)
$Wr$	writhing number (geometric variable)
$\theta_b$	bend angle (for isotropic polymer models)
$\theta_R$	net roll angle
$\theta_T$	net tilt angle
$\kappa$	curvature (also Debye screening parameter)
$\rho$	roll angle (between successive base pairs)
$\sigma$	$\Delta Lk/Lk_0$ (superhelical density)

Chapter 6 Notation Table (continued)

SYMBOL	DEFINITION
$\sigma_e$	Poisson's ratio (for an elastic material)
$\tau$	tilt angle (between successive base pairs)
$\chi$	glycosyl (sugar/base) torsion angle
$\omega$	DNA twist rate; also propeller twist (within a base pair)
$\omega_0$	intrinsic twist rate of DNA (e.g., $2\pi/10.5$ radians)
$\Delta Lk$	linking number difference with respect to $Lk_0$
$\Omega$	twist angle

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

*Molecular Structure of Nucleic Acids*, J.D. Watson and F.H.C. Crick (1953).

## 6.1 Introduction

This chapter introduces further topics in nucleic acid structure, building upon the minitutorial of the previous chapter. These topics include DNA sequence effects, DNA hydration, DNA/protein interactions, and the cellular organization of DNA, including supercoiling and chromatin structure. The next chapter expands upon the related topics of alternative hydrogen bonding schemes, non-canonical helical and hybrid structures, DNA mimics, overstretched and understretched DNA, and RNA structure and folding.

If I dare write something on any of these topics — given the excellent books and book sections [100, 137, 139, 163, 195, 893, 895, 1080, 1191, 1305] and numerous articles already written on these subjects — it is only because these important areas in nucleic acid structure cannot be omitted.

The sequence-induced variations of DNA, for example, profoundly affect the local structure of DNA and that of DNA complexed with other molecules (water, ions, ligands, and proteins). Such local variations translate into global structural effects when considered on the molecular and cellular levels. The associated energetic and dynamic effects, in turn, affect key functional processes, from the action of regulatory proteins (replication, repair, transcription, etc.) to genome packaging and processing. This functional influence of DNA sequence on biological activity has been shown by the works of Olson, Zhurkin, Dickerson, and many others (e.g., [308, 942]).

Until recently, nucleic acids might have been considered ‘the silent partner’ in regulatory complex formation with proteins. This view is beginning to change as our appreciation grows for the subtle, sequence-dependent effects of nucleic acids, including the recently discovered “second genetic code”, the nucleosome positioning code [1157, 1288, 1425].

Though presenting these topics in this text contrasts with the very brief and basic protein chapters, I hope to encourage mathematical and physical scientists who may not read structural biology texts to pursue research in the area of nucleic acids. Aspects of DNA and RNA structure that may interest mathematicians, computer scientists, or engineers, include DNA supercoiling; RNA folding prediction; application of elasticity theory, topology, and knot theory to DNA structure prediction and functional mechanisms; and using DNA as a parallel computer (e.g., [40, 117, 162, 825, 1155, 1156]). The presentations here might introduce readers to some of the exciting areas of research on DNA.

Besides the excellent texts [139, 195, 1080, 1191], the reader is invited to explore the wealth of related structural data on the nucleic acid database NDB and the protein data bank/research collaboratory for structural bioinformatics resource [128, 1365].

As in the previous chapter, we abbreviate *base pair* and *base pairs* as bp and bps, respectively.

## 6.2 DNA Sequence Effects

### 6.2.1 Local Deformations

What makes DNA such a rich resource of structural and functional information? The subtle geometric, electrostatic, and mechanical differences inherent in each nucleotide (e.g., C), bp (e.g., G–C), or bp step (e.g., AT in the 5' GCGATCGC 3' octamer) combine to affect patterns in local hydration and local helical parameters, such as *twist* ( $\Omega$ ), *tilt* ( $\tau$ ), *roll* ( $\rho$ ), and *slide* ( $Dy$ ) between two successive bps (one bp-step), and propeller twist ( $\omega$ ) within each bp, as introduced in the previous chapter and Figure 5.12.

Recall that *twist* is the relative rotation between two successive bps about the helical axis as measured by the change in orientation between the corresponding C1'–C1' vectors, projected down the helical axis.

*Roll* and *tilt* define the deformations between two bp planes along the long and short bp axes, respectively, perpendicular to the helical axis. The sign conventions are such that a positive roll angle opens up the minor groove side, while a positive tilt angle opens up bps in the direction of the phosphate backbone of the first strand.

*Slide* describes the translational displacement of neighboring bps along their long axes, as measured between the midpoint of each long-bp axis. It is considered positive when the direction is toward the first nucleic acid strand (i.e., positive **e2** direction).

The *propeller twist* defines the angle between the planes of the two *bases* in a hydrogen-bonded pair. A positive value corresponds to rotation of each base in a counterclockwise manner when viewed from its attached sugar/phosphate backbone [307].

Normal propeller twists are negative in DNA. This is a consequence of base-pair stacking interactions along polynucleotide strands, as first discussed in [309, 745]. (Note: some of the earlier sign conventions were opposite than the current standard).

Propeller twisting can improve bp stacking interactions through enhanced van der Waals, electrostatic, and hydrophobic contacts. It can also promote stability through additional hydrogen-bonding contacts. Such contacts involve diagonally-distant bases in two adjacent residues. Stability is enhanced since the distance that would result from a perfectly stacked arrangement (perpendicular to the global helix axis) is decreased due to propeller twisting. Regions of repeating AT bps exhibit greater propensity than GC for propeller twisting [195, 527].

The *bending anisotropy* in DNA was established over twenty years ago [1096, 1454]. Namely, DNA prefers to bend more easily into the minor and major grooves than other directions, and the preference to major or minor-groove depends on the nucleotide sequence.

Such bending preferences already emerge for dinucleotide steps, as described below. However, because sequence context is important, recent work has focused on the effect of flanking sequences as well (see Subsection 6.2.3).

### 6.2.2 Orientation Preferences in Dinucleotide Steps

There are 10 unique dimer (dinucleotide) steps in DNA, conventionally specified on the  $5' \rightarrow 3'$  strand. Thus, associated with each dimer step is the complementary bp step on the opposite strand (e.g.,  $5' [AC] 3'$  with  $5' [GT] 3'$ ); when the usual orientation conventions (for tilt, roll, etc.) are applied to the complementary bp step, the same deformations defined for the main strand result.

Local geometric and energetic trends can be associated with each bp step. Trends can also be clustered into three broad classes: pyrimidine/purine (Pyr/Pur), purine/purine (Pur/Pur), and purine/pyrimidine (Pur/Pyr) steps. (Often pyrimidines and purines are also denoted by Y and R for short). This is because common properties such as residue size and electrostatics produce similar geometric and energetic characteristics with these classes of dinucleotide steps.

The three *Pyr/Pur* steps are:

TA (TA)    CA (TG)    CG (CG);

the four *Pur/Pur* (Pyr/Pyr) steps are:

AA (TT)    AG (CT)    GA (TC)    GG (CC);

and the three *Pur/Pyr* steps are:

AT (AT)    GT (AC)    GC (GC).

Table 6.1. Base-pair step parameters for roll ( $\rho$ ), tilt ( $\tau$ ), and twist ( $\Omega$ ) angles (in degrees) for free [473] and protein-bound DNA [942] (denoted as DNA and DNA<sup>+</sup>, respectively), as analyzed from crystal structures of DNA and DNA complexes from the NDB. The overline and underline symbols indicate the largest and lowest values, respectively, associated with each angular value in each class (free and complexed DNA). They help distinguish features among the three dinucleotide classes as well as highlight similarities in patterns between free and bound DNA of the same bp step. Note a tie for the minimal value of 0.7° roll for AA and GT steps.

Nuc.		<i>Pur A</i>		<i>Pur G</i>		<i>Pyr T</i>		<i>Pyr C</i>	
		DNA	DNA <sup>+</sup>						
<i>Pyr T</i>	$\rho$	2.6	3.3						
	$\tau$	0.0	0.0						
	$\Omega$	40.0	37.8						
<i>Pyr C</i>	$\rho$	1.1	4.7	6.6	5.4				
	$\tau$	0.6	0.5	0.0	0.0				
	$\Omega$	36.9	37.3	31.1	36.1				
<i>Pur A</i>	$\rho$	0.5	0.7	2.9	4.5	-0.6	1.1		
	$\tau$	-0.4	-1.4	-2.0	-1.7	0.0	0.0		
	$\Omega$	35.8	35.1	30.5	31.9	33.4	29.3		
<i>Pur G</i>	$\rho$	-0.1	1.9	6.5	3.6	0.4	0.7	-7.0	0.3
	$\tau$	-0.4	-1.5	-1.1	-0.1	-0.9	-0.1	0.0	0.0
	$\Omega$	39.3	36.3	33.4	32.9	35.8	31.5	38.3	33.6

Let us first examine in Table 6.1 the published average roll ( $\rho$ ), tilt ( $\tau$ ), and twist ( $\Omega$ ) values for these 10 dinucleotide steps as collected over a set of crystallographically-determined for free (unbound) DNA systems [473] (first column in each dinucleotide entry) and protein-bound DNA systems [942] (second column). For updated values, consult newer works by the authors.

This  $4 \times 4$  table is symmetric as presented, so the upper triangle is not filled. As ordered, the three top matrix entries (upper triangle cluster: TA, CA, CG) correspond to Pyr/Pur bp steps; the four lower square entries (for AA, AG, GA, GG) correspond to the Pur/Pur steps; and the three entries in the lower triangle cluster (AT, GT, GC) correspond to Pur/Pyr steps. Note that the six maxima (overline symbols) associated with roll, tilt, and twist values for both free and bound DNA occur in the Pyr/Pur cluster; four extremes (lower bounds) are noted in the Pur/Pur cluster, and three in the Pur/Pyr group. (The minimal value of 0.7° average roll displacement occurs for both AA and GT steps).

This observation is related to the decreasing overall flexibility associated with these dinucleotide steps, with Pyr/Pur steps most easily deformed. Note also the similarity in trends for the free and protein-bound DNA entries.

The observed trends in each dinucleotide step can be explained in part by the tendency of DNA to adjust local geometry so as to improve bp *stacking* (overlap) interactions. Since the bp-plane areas are water insoluble, stacking

energies can be strengthened by reducing the area between successive bps via local variations, and/or enhancing longer-range interactions through formation of bifurcated hydrogen bonds (between successive bps on opposite strands).

See Box 6.1 for a discussion of the inherent flexibility of Pyr/Pur and Pur/Pur steps, including the contribution of some examples of biological systems where flexibility in Pyr/Pur steps has been noted.

### Box 6.1: Inherent Flexibility of Pyr/Pur and Pur/Pur Steps

**Pyr/Pur Favorable Configurations.** Pyr/Pur steps (i.e., TA, CA, and CG) can produce favorable stacking in one of two general configurations: a combination of *large positive roll and small negative slide*, or *near-zero roll and small positive slide*, both in association with *propeller twisting* in the two bps [195] (see Figure 5.12 of Chapter 5).

In the former arrangement (large positive roll and small negative slide), the purine bases slide toward each other to improve the cross-chain stacking between them. The concomitant large positive roll inclines the smaller pyrimidine partners to maintain the favored propeller twist in both bps. In the latter (near-zero roll and small positive slide), the large purine bases slide away from each other to avoid a steric clash between them since the propeller twisting of the two bps brings them closer together. The roll is zero since, in the same strand, stacked pyrimidines and purines remain parallel to one another in this propeller-twisted arrangement. (See illustrations in [195].)

Given at least these two options for favorable stacking, the Pyr/Pur class of dinucleotide steps generally displays a wide range of roll values when many structures are analyzed.

**CG Example in E2.** The high positive roll associated with a CG dinucleotide step has been used to explain the importance of the central CG step in the DNA-binding sequence of the bovine papillomavirus E2 protein [1074], whose sequence is ACCGACGTCGGT. This step contributes to the needed deformation of the DNA-binding region — a large overall bending toward the protein.

**TG Example in CAP.** The characteristic flexibility of another Pyr/Pur step, TG (equivalent to the CA step discussed above), has also been used to explain the importance of the central TG dinucleotide step in the DNA-binding sequence of the catabolite gene-activating protein (CAP) in *E. coli*, a highly conserved dinucleotide segment in both monomers of this dimer protein in different CAP-binding sites. The binding of this complex requires substantial distortion of the DNA, largely modulated by a 40° kink at this central TG step [163, 1143].

**TA-Rich Regions in Functional Sites.** Furthermore, the combination of an observed small energetic barrier to unwinding in Pyr/Pur steps (though not fully understood) and the intrinsic curvature associated with consecutive AT bps (equivalently, AA or TT dinucleotide steps) has been used to explain the prevalence of TA-rich regions in key functional sites. Examples of such sites are TATATATA, also known as the adenovirus E4 promoter, and TATAAAAG, in the adenovirus major late promoter, both of which binds to transcription proteins. These proteins (like the TATA-binding protein TBP in eukaryotes) must

unwind DNA, and hence the relatively low barrier to twisting is well exploited. Note also that the standard, Watson-Crick (WC) AT bp has one fewer hydrogen bond than a WC GC bp, which has three, and thus the AT bp should be easier to deform.

**Pur/Pur Tendencies.** It has also long been noted that the Pur/Pur AA dinucleotide step prefers an orientation where the two successive AT bps are propeller-twisted (i.e., they deviate from planar bp orientations) with near zero roll values (as bases remain parallel to their neighbors on the same strand) [195]. This out-of-plane bending shortens cross-residue distances between the adenines on one strand and the thymines on the opposite strand and allows offset, cross-strand hydrogen bonds to form between an oxygen of thymine and a nitrogen of adenine on the major groove side. This interaction also implies close contact between an oxygen of thymine and carbon of adenine on the minor groove side.

The large positive roll characteristic of CG dinucleotide steps and low roll associated with AA (or TT) steps are particularly consistent trends noted through various structural analyses [274].

---

### 6.2.3 *Orientation Preferences in Dinucleotide Steps With Flanking Sequence Context: Tetranucleotide Studies*

The importance of bp flexibility in the structural and functional properties of DNA as well as of DNA/protein, DNA/RNA and other complexes combined with the significance of flanking sequence effects has led to extensions in recent sequence-effect studies from dinucleotide to tetranucleotide steps. The latter represent the minimal structural unit that can reveal near-neighbor bp structural patterns.

There are 136 unique tetranucleotide steps, and unraveling these patterns has become an international initiative called the **Ascona B-DNA Consortium** (ABC). See [133, 321, 705] for emerging information from MD simulations of many solvated DNA systems containing all unique tetramer sequences. Such studies reveal fine patterns for sequence context, and they are being used to compile a database of geometric parameters for all dinucleotide steps to address all possible flanking bp combinations.

For example, these modeling and simulation studies have revealed that while Pyr/Pur steps are intrinsically flexible, they are least affected by the neighboring units. Their flexibility is especially pronounced when these Pyr/Pur steps occur near the intrinsically rigid Pur/Pur or Pur/Pyr steps due to their significant impact. ABC-driven dynamics simulations are also uncovering a large degree of backbone fluctuations and persistent canonical and non-canonical B-DNA sub-states [321].

### 6.2.4 *Intrinsic DNA Bending in A-Tracts*

Now consider all these sequence effects taken together. For heterogeneous, or mixed sequence DNA, all these differing trends will very roughly average out on a global scale for free DNA. In other words, these trends will accumulate slowly

in the sense of causing large global distortions. However, this would not be the case if the sequence composition was very regular, as in poly(AT), or contains a motif (such as  $A_n$ ) that recurs every about 10 bps (i.e., within the repeating unit of the DNA helix); such ‘phased’ sequences may lead to substantial global effects as a sum of local variations.

A dramatic sequence-dependent pattern of this type from the sum of small net curvature was noted in the early 1980s in Paul Englund’s laboratory [829] in segments of kinetoplast DNA, which displayed anomalously slow gel migration rates.<sup>1</sup> Besides kinetoplast DNA, A-tracts are associated with other structures and functions *in vivo*, such as core regulatory elements (found in DNA origins of replication and recombination), intergenic regions in prokaryotes, tips of supercoiled DNA, DNA loops, and nucleosomal DNA in chromatin.

Studies of this intriguing phenomena continued in Don Crothers’ laboratory [276, 741]. Upon close analysis, this behavior was explained by the AT-bp rich content of these DNA sequences. Specifically, kinetoplast DNA contains stretches of consecutive groups of adenines residues (4 to 6), separated by other sequences, *phased* with the helical repeat. This composition of blocks of adenine repeated every  $\approx 10$  bps is known as *phased A-tracts*. For example, a representative part of the sequence is:

CCCCAAAATG|TCAAAAAAATA|GGCAAAAAAT|GCCAAAAATC|,

where the vertical bars mark each 10 residues. Thus, the 5 or 6-membered A-tracts in this sequence occupy positions 3–8, 4–8, or 4–9 within each 10 residues. The significant curvature observed in association with each A-tract (roughly  $15^\circ$ ) [276, 514] thus translates into a highly-curved helix when viewed over hundreds of residues, as shown in Figure 6.1 (right).

The models in Figure 6.1 were built from dodecamer curvature as deduced from MD simulations of two solvated systems [1230]: the A-tract system CGCGAAAAAAACG solved by crystallography (1D89) [312] and a control sequence CGCGAATTCTCGCG (known as the Dickerson/Drew dodecamer) [334, 1380]. The control sequence — though containing two adenines and thus exhibiting limited aspects of intrinsic bending — has a small overall helix bend. The A-tract oligomer, in contrast, exhibits an overall helix bend of about  $14^\circ$  per dodecamer. This curvature has profound implications on the plasmid level, as seen in the figure, and explains the anomalous migration rates observed in kinetoplast DNA.

Though we now recognize the bent helical axes in DNAs containing phased A-tracts, a clear understanding of the origin and details of intrinsic bending has not yet been achieved despite numerous experimental and theoretical studies [310, 492, 514, 551, 800, 946, for example]. Part of the dilemma in explaining bending involves reconciling experimental data obtained by crystallography [310] versus

---

<sup>1</sup>The anomaly is a deviation from the usual linear relationship between DNA length and the logarithm of the distance migrated on the gel.

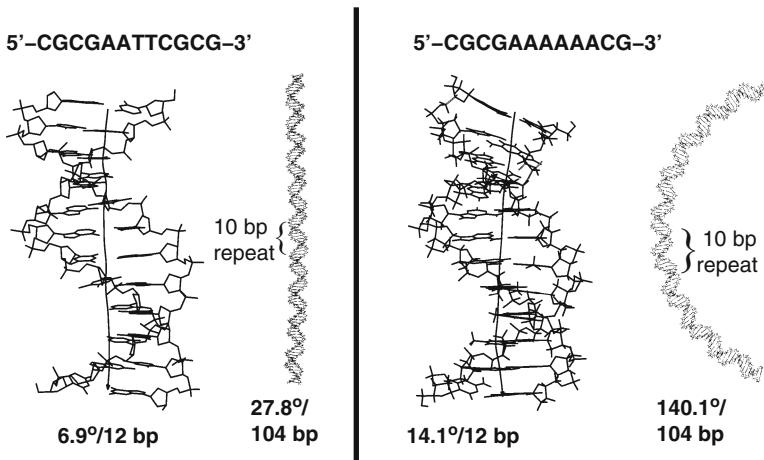


Figure 6.1. Bending on the dodecamer and longer-DNA levels: models of 110 bp DNA were constructed from dodecamer systems with differing bending trends: a gentle degree of helix bending ( $7^\circ$  per dodecamer) with no preferred direction of bending, and more substantial bending ( $14^\circ$  per dodecamer) with preferred bending realized as minor groove compression, for an A-tract system (1D89 in Figure 6.2). Data are based on molecular dynamics simulations [1230]. Scales are different for the long and short DNAs.

solution studies since environmental effects can be critical; another component is the consideration of static versus dynamic structures; finally, there is a technical difficulty in quantifying large helical bending in DNA; see Box 6.2.

Indeed, various bending models have been proposed for adenine-rich sequences [1191]. At the two extremes, the *junction-type model* (developed from models of junctions between two types of helices [1159]) suggests that a localized bending between two bps at regions separating A-tract regions from others largely explains the large degree of bending; the *wedge-type model* proposes that bending is smooth, a cumulative effect from the small bending associated with each AA dinucleotide step. Clearly, the realistic model involves a combination of sharp and smooth bending models and depends on the residues that flank both the 5' and 3' sides of the A-tract [1230]. A growing body of simulations and experiments are suggesting that the A-tract moieties themselves are relatively straight and that the overall curvature results from substantial rolls at the A-tract junctions, between A-tracks and external sequences, the magnitude of which is sequence modulated [800, 1230, for example].

Moreover, the factors that induce and stabilize bending include the tendency of AT bps to be propeller-twisted and exhibit systematic differences in sugar conformations between the adenine and thymine sugars ( $\sim 15^\circ$  pucker angle) [1171, 1212, 1230, 1417, 1422]. These factors combine to compress the DNA minor groove and stabilize it by bifurcated hydrogen bonds (between thymines on one strand and adenines on the opposite strand at successive steps), as illustrated in Figure 6.3. This geometry also leads to a stabilizing, ordered spine of hydration

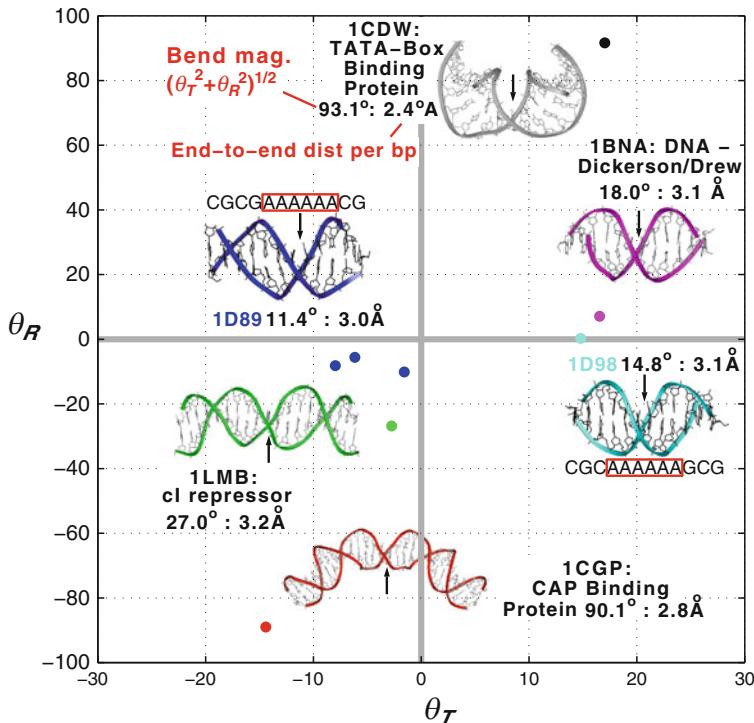


Figure 6.2. Macromolecular examples of *net tilt* ( $\theta_T$ ) and *net roll* ( $\theta_R$ ) combinations, as calculated by eqs. (6.1) and (6.2), for each system in the reference plane indicated by arrows. The bending angle magnitude ( $\sqrt{\theta_T^2 + \theta_R^2}$ ) and the end-to-end distance (normalized per bp) are also shown. Regions of positive and negative  $\theta_R$  are often termed “*major-groove compression*” or “*minor-groove compression*”, respectively. Co-crystals of dimeric major-groove binding proteins (like CAP and cI repressor) tend to wrap the DNA around the protein with the minor groove at the center of curvature ( $\theta_R < 0$ ). Structures with bends toward the major groove, such as the Dickerson/Drew dodecamer or co-crystals of minor-groove binding proteins like the TATA-box binding protein (TBP), have  $\theta_R > 0$ . The net bend angles for the two A-tract crystal structures (1D89 and 1D98) correspond to near-zero  $\theta_R$  (and opposite magnitudes of  $\theta_T$ ) rather than bending toward the minor groove ( $\theta_R < 0$ ), as suggested by simulations and solution models. For 1D89, three crystal forms have been solved (three filled blue circles).

in the minor groove of the A-tracts, with counterion coordination playing a role [1182, 1417]. In addition, many factors such as temperature, organic solvent, and monovalent and divalent ions influence A-tract structure.

In their recent extensive review of the entire field of A-tract structures and DNA bending, Haran and Mohanty postulate based on the accumulating evidence that A-tracts appear to serve as a multi-tasking DNA element involved in various *in vivo* functions and that it is the unique *structure* of A-tracts — rather than their intrinsic bending *per se* — that explains their features and effects [515]. They

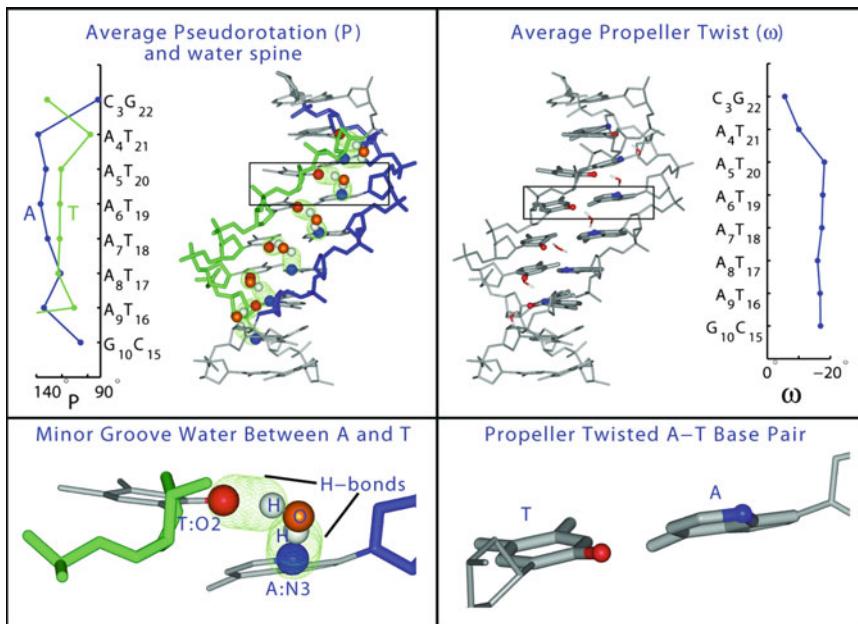


Figure 6.3. Bend-stabilizing interactions in A-tract DNA as revealed by MD simulations [1230] of two A-tract DNA dodecamers, 1D89 and 1D98, also shown in Figure 6.2: sugar puckering differences between adenines and thymines on opposite strands (left plot), ordered minor-groove water spine (left dodecamer), and propeller twisting in the A-tract region (right dodecamer and right plot).

rationalize their conjecture by the fact that since sequence mutations can easily modify curvature and thus affect function drastically, this would not be advantageous if A-tract's role was primarily functional. In contrast, protein binding is a much more effective vehicle for affecting global DNA structure and hence function, especially in eukaryotes where the DNA is packaged in the chromatin fiber around nucleosomes.

Research continues on this important front.

### 6.2.5 Sequence Deformability Analysis Continues

While Crothers' sentiment [274] is now well-accepted: “*the days are over when one dinucleotide step was as good as another in determining the structure and mechanical properties of DNA*”, our understanding of sequence-dependent deformability in DNA and DNA complexes is far from complete. While clearly much effort has focused on finding a ‘set of rules’ relating nucleic-acid sequence to preferred structure, it is probably fair to say that we have not yet been illuminated by a useful set of predictive rules.

Consideration of longer bp-step regions, such as trinucleotide and tetranucleotide steps, as described above, is one obvious route for establishing clearer

predictive rules (and exceptions to the rules); the increasing resolution of DNA crystal structures will undoubtedly facilitate such analyses. Still, even if a set of rules cannot be identified, the local tendencies of DNA are important, as they are known to work along with the deformations induced by proteins.

### 6.3 DNA Hydration and Ion Interactions

The binding of water molecules and ions to DNA atoms in the cellular environment is not only important for explaining DNA's conformational stability and energetic properties; it can also help interpret the dependence of DNA conformation on the environment. For example, differences in hydration patterns (specifically with regard to accessible surface area in the grooves) are believed to be a factor in DNA's preference for the canonical A-form at low humidity, higher salt concentrations, and larger amounts of organic solvents, and that of Z-DNA at increased ionic strengths for certain sequences. Both bases and phosphate groups have organized hydration shells [1141].

Protein/DNA and drug/DNA interactions are also dependent on water contacts. Proteins can displace water molecules that surround certain DNA regions and replace them with interactions involving amino acid side chains. Thus, hydration sites represent preferential, energetically favored binding sites for hydrophilic DNA binders, like proteins and drugs [1382].

Positive counterions are fundamentally important in shielding the negatively charged DNA backbone and allowing DNA segments to come closer together. Dramatic effects are manifested in long supercoiled DNA molecules, as observed experimentally and studied by simulation (e.g., [1125, 1309]).

#### Box 6.2: Net DNA Bending and A-Tract Analysis

A quantitative description of helical DNA bending is far from trivial. Indeed, depending on the choice of the analysis program used for computing DNA parameters, differences in measurements frequently result when the DNA is significantly curved. Part of these differences can be explained by the various protocols used to define reference frames for each base and base pair [74]. This can be resolved by using standard definitions [74]. Still, helical bending is often measured as the angle between two *extremal* vectors connecting two separated bp steps [292, 405, 406, 1284, 1285, 1455].

As an alternative, introduced in [1230] are two net-bend angles termed *net tilt* and *net roll* ( $\theta_T$ ,  $\theta_R$ ). These angles are calculated by considering the cumulative projection of tilt and roll at each dinucleotide step, adjusted for twist, into the plane of a reference 'average' bp (typically at the oligomer center). From the tilt, roll, and twist at each bp-step  $j$ , namely  $\tau_j$ ,  $\rho_j$ , and  $\Omega_j$ , we compute the net tilt and roll angles as:

$$\theta_T = \sum_j \left[ \tau_j \cos \left( \sum_{i=N_c}^j \Omega_i \right) + \rho_j \sin \left( \sum_{i=N_c}^j \Omega_i \right) \right], \quad (6.1)$$

$$\theta_R = \sum_j \left[ -\tau_j \sin \left( \sum_{i=N_c}^j \Omega_i \right) + \rho_j \cos \left( \sum_{i=N_c}^j \Omega_i \right) \right], \quad (6.2)$$

where  $\sum_{i=N_c}^j \Omega_i$  is the cumulative twist over  $i$  bp-steps from the plane of the reference bp,  $N_c$ , to bp-step  $j$  [1230]. Bends in the helical axis defined by  $\theta_R < 0$  are equivalent to ‘minor groove compression’, a property associated with A-tracts [276] or with the center oligonucleotide of the CAP/DNA complex [163, 1143]; bends defined by  $\theta_R > 0$  are equivalent to ‘major groove compression’, as noted for the TATA-box binding protein TBP [163, 971] (see Figure 6.2).

The crystal versus solution data of A-tract systems differ in their characterization of the degree of bending. Crystallographic studies have indicated various bend directions for solved A-tract systems: negative [312, for example] versus positive [896, for example] net tilt, as shown in Figure 6.2. Solution models of A-tract oligomers, in contrast, have suggested a unique bend direction along negative net roll. Such data come from gel electrophoresis (mobility data), NMR (minor-groove water life times), and hydroxyl radical footprinting (compressed minor groove, as inferred from the reduced accessible surface area in the A-tract minor groove to hydroxyl radical  $[\text{OH}^-]$  cleavage of the DNA backbone). See [186, 276, 310–312, 324, 439, 520, 1017, 1213] and other citations collected in [1230]. Simulations have shown, however, that small changes from those initially dissimilar bends (from crystallographic models) can yield similar bend directions, equivalent to minor groove compression, in accord with solution data [1230]. The crystal models may reflect the inherent conformational disorder (multiple conformations) in A-tract systems, as well as effects by the organic solvent needed for crystallization (2-methyl-2,4-pentanediol or MPD). See also discussion in [946] emphasizing the importance of considering the role of thermal fluctuations and the local chemical environment on DNA curvature and its interpretation.

### 6.3.1 Resolution Difficulties

Though fundamentally important, it has been a challenge in the past to establish precise hydration and ion patterns for biomolecules (see [352], for example, concerning the latter).

First, the number of water molecules, captured in crystal structures of DNA and DNA complexes, is highly dependent on structure resolution, though the high-resolution (e.g., better than 2 Å) structures now appearing should alleviate this problem (e.g., [641]). Counterions, which in the past could not be located in the electron density maps, can also be detected in ultra-high resolution DNA structures [234, 641, 1182]. See [352], for example, for a perspective, and recent overviews [216, 675, 1076].

Second, the crystal environment of the lattice — periodicity or organic solvent — can also introduce artifacts, so interpretations must be cautious (see also [946]).

Third, NMR spectroscopy measurements for locating protons and other techniques are complicated by the relatively fast, reorganization component of water dynamics (0.1 ps and longer) that can lead to overlap of spectra signals; such experiments can yield more detailed information on the less-transient, and more structured, water molecules, such as the minor groove hydration patterns in DNA A-tracts.

Fourth, though computer simulations are vulnerable to the usual approximations (force fields, protocols, etc.), they are nonetheless becoming important in deducing biomolecular hydration and ion patterns [216, 380, 675, 1412], especially in identifying counterion distribution patterns for DNA [239, 1076, 1419, 1421] and RNA [69, 70, 72, 73, 172, 225, 240, 557]. Continuum solvent models are also beginning to provide information on the relative stabilities of different sequences and different helical forms (e.g., [203, 1273]), as are innovative methods such as integral methods using a 3D reference interaction-site model [1412], modified Poisson-Boltzmann implementation [239], and grand canonical Monte Carlo [675].

### 6.3.2 Basic Patterns

From all available techniques, the following facts have now been established regarding DNA hydration and ion patterns:

1. **Multiple Layers.** Hydration patterns are inhomogeneous and multi-layered, from the first layer closest to the DNA (including the nucleotide-specific hydration patterns and minor-groove ‘spine of hydration’ [125, 126, 1140]) to the outermost layer of the highly transient bulk water molecules. (Intermediate layers are characterized by fast exchange of water molecules and ions with bulk solvent). Water interactions around nucleic acids are important for stabilizing secondary and tertiary structure.
2. **Local Patterns.** Hydration patterns are mostly local, i.e., short range and largely dependent on the neighboring atoms [1138]. For example, the water patterns around guanines and adenines are very similar and there are clear differences in the distributions of hydration sites around guanines and cytosines that canonical A, B, and Z-DNA helices share [1140]. (Conclusions for adenine and thymine bps are not available due to their strong preference for B-form DNA). The strong local patterns generated near individual nucleotides permit canonical helices to be reconstructed, complete with preferential hydration sites [1140].

To analyze hydration patterns and ions around nucleic acids, thermodynamic, spectroscopic, and theoretical calculations have been used. A useful concept for quantifying hydration patterns is the solvent-accessible surface — introduced to describe the proportion of buried to accessible atomic groups [723]. The three-dimensional quantification of hydration is typically computed as a volume-dependent probability [653, 1138, 1140].

3. **Structured First Layer.** The first hydration shell is generally more structured. It contains roughly 15–25 water molecules per nucleotide, with varying associated life times. The accessible electronegative atoms of DNA are water bound in this model, including the backbone phosphate region (around phosphate oxygens), sugar oxygen atoms (especially O4'), and most suitable base atoms. In particular, hydrogen bonds form in roughly one quarter of these molecules, bridging base and backbone atoms or two base sites. Interestingly, the accessible, hydrophilic O3' and guanine N2 groups are statistically un-hydrated [1140, 1141].
4. **Cation Interactions and Counterion Condensation Theory.** Debye-Hückel theory, along with the counterion condensation theory (see Box 6.3), are invaluable frameworks for understanding how counterions distribute in electrolyte solutions.

Counterions tend to cluster around a central ion of opposite charge. The *Debye length*  $\kappa^{-1}$  (or Debye radius) is the salt-dependent distance at which the probability of finding a counterion from a central ion is greatest. Debye-Hückel theory predicts the *screening factor*  $\exp(-\kappa r)$  (see Chapter 10, Subsection 10.6.4 on continuum electrostatics) by which the ordinary Coulomb electrostatic potential is reduced (see Figure 10.10 in Chapter 10). For DNA, the negatively charged (and thus mutually repulsive) phosphate residues on different strands — separated from each other by about 10 Å — are effectively screened by added salts, thereby stabilizing the helix and allowing the phosphates to come closer together. At physiological ionic concentrations, the Debye length for DNA is about 8 Å. Along with Debye-Hückel theory, counterion condensation theory can be used to analyze the ionic environment around polyelectrolyte DNA and its effect on thermodynamic processes [386, 822, 824, 1040].

As a result of screened Coulomb effects, the association of ions with DNA spreads over a broad range, from very loose associations to tight binding, and the timescale ranges are also broad. Experimental methods tend to capture only average effects, but some insights are emerging from simulation studies. For example, the counterion distribution around DNA that emerged from Monte Carlo simulations [1419] revealed two distinct layers of sodium ions, the innermost of which corresponds to the condensed counterion layer, and the outer layer more diffusive [1419, Figure 1].<sup>2</sup>

It is well appreciated that cations interact in a non-uniform and sequence-specific manner with the DNA grooves and the phosphate backbone [234, 1417]. Major groove cation binding appears to be a more general

---

<sup>2</sup>In nucleic acid simulations, typically one sodium ion per base is incorporated into the first hydration shell for charge neutralization; additional ions (positive and negative) are placed, with positions adjusted by minimization or Monte Carlo procedures so as to produce ions in the most probable regions compatible with the salt concentration of the solution (e.g., [1421]).

motif than the more sequence-specific minor-groove cation binding, as observed in A-tracts [578, 850, 1255]. Furthermore, divalent cations bind to duplex DNA in a more sequence-specific manner than univalent ions since the former tend to be fully hydrated and thus can donate or accept hydrogen bonds to base atoms through their water ligands [234].

Such details of cation binding to DNA have come from nucleic acid simulations and are beginning to emerge from ultra-high resolution crystal structures of DNA [234, 641, 1182]. The facilitation of DNA bending motions by proper cation shielding was demonstrated by classic studies of Mirzabekov, Rich, and Manning on the role of asymmetrically neutralized phosphates in promoting bending around nucleosomes [823, 865]. These findings were highlighted by Maher and co-workers, who examined bending promoted by neutral phosphate analogs [815], and by Williams and co-workers, who examined ion distribution and binding in crystallographic nucleic acid structures [850, 1182]. See [352] for a perspective.

5. **Second Hydration Layer.** The second water shell mainly stabilizes the first shell, allowing it to have a favorable tetrahedral structure (see box on water structure in Chapter 3). The outermost, bulk solvent hydration shell is generally more disordered, with its network of interactions rapidly changing.
6. **Hydrated Grooves.** The DNA major and minor grooves are well hydrated. Though the wider major groove can accommodate more water molecules, the minor groove hydration patterns tend to be more structured. The different groove shapes in canonical A, B, and Z-DNA lead to systematic differences in hydration patterns [1139, 1140, 1182]. For example, hydration sites are better defined and form an extensive network in the very deep major groove surface of A-DNA than those in B-DNA (less deep major groove) and Z-DNA (very flat major groove surface); hydration sites have higher densities in the narrower minor groove of B and Z-DNA than in corresponding A-DNA minor-groove sites.
7. **Sequence-Dependent Hydration.** Hydration pattern details depend on the sequence and helical context [1139, 1140]. The compressed minor groove waters of adenine-rich sequences (A-tracts) are particularly associated with long lifetimes. This results from well-ordered water interactions ('spine of hydration') stabilized by cross-strand, bifurcated hydrogen bonds (e.g., N3 of A at step  $n + 1$  with O2 of T on the opposite strand at step  $n$ ), where the adenine and thymines are on opposite strands and offset by one step), as shown in Figure 6.3. Such patterns have been extensively studied in association with the central AT-rich region of the Dickerson-Drew dodecamer (CGCGAATTCGCG) [334, 763, 1380]. An example of sequence-dependent local water is shown in Figure 6.4 for three DNA sequences that differ by one bp from one another.

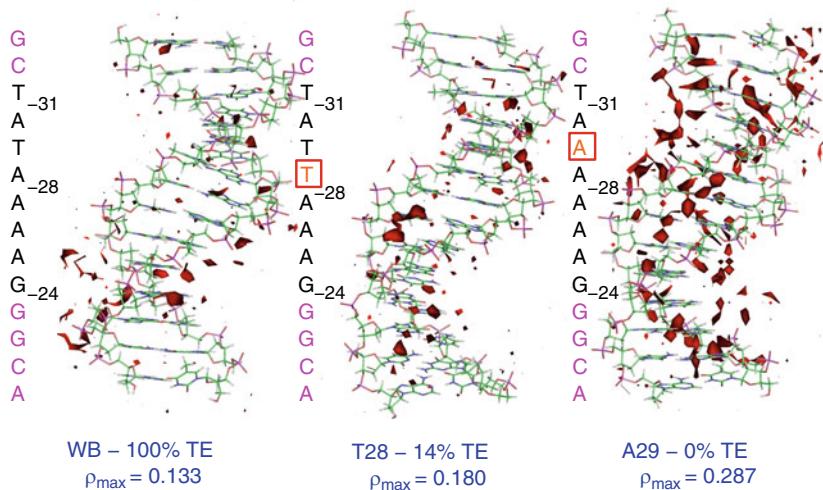


Figure 6.4. Local hydration patterns in three DNA systems (TATA elements differing by one bp) as deduced from molecular dynamics simulations [1026] based on crystallographic studies [971]. The sequences are variants of the promoter octamers called TATA elements (for their rich A and T content) [163]. The experimental studies [971] examined the structures of resulting DNA/TBP complexes (TBP is the eukaryotic transcription ‘TATA-box Binding Protein’) and their relative transcriptional activities (TE values given in figure). The simulations [1026] discerned sequence-dependent structural and flexibility patterns, like the maximal water density around the DNA, as shown here for three TATA sequences. The transcriptionally inactive variant shown at right, an A-tract ( $\text{GCTA}_6\text{GGGCA}$ ), has a much denser local water environment than the wildtype sequence (WB,  $\text{GCTATA}_4\text{GGGCA}$ , left). The maximal water densities (number of oxygens) were computed on cubic grids of 1 Å.

8. **Varying Lifetimes.** Water lifetimes vary, from residence times in the several-hundred picosecond range for major groove waters associated with thymine methyl groups to life times of order nanosecond in the minor groove of A-tracts [763, 995, 1230].
9. **Phosphate Localization.** Hydration patterns around the phosphate groups exhibit characteristic hydration ‘cones’ with three tetrahedrally-arranged water molecules in the first hydration shell of each phosphate-charged oxygen [1141]. Furthermore, the arrangements depend on the DNA helix type. For example, the shorter distances between successive phosphate groups in A-DNA (with respect to B-DNA) encourage water molecules to form bridges between them. In B-DNA, these waters favor forming hydrogen bonds with the anionic phosphate oxygens O1P and O2P; O5' is inaccessible and O3' is surprisingly under-hydrated [1382].

Given this multilayered organization of hydration shells in DNA and the large costs associated with fully-solvated simulations of nucleic acids, practitioners have attempted to economize on system size by modeling a limited number of water molecules to represent only the thin, ordered first solvation shell around the DNA [230, 843].

---

### Box 6.3: Counterion Condensation Theory

Counterion condensation theory essentially predicts formation of a condensed layer of mobile and hydrated counterions at a critical polyionic charge density in the vicinity of the DNA surface ( $\sim 7 \text{ \AA}$ ). For monovalent cations, the concentration of this cloud is relatively independent of the bulk cation concentration. This concentrated cloud of ions effectively neutralizes  $\sim 76\%$  of the DNA charge, thereby reducing the negative charge of each phosphate to about one quarter its magnitude (hence the charge scaling done in early nucleic-acid simulations [523, 1013, 1260]). Divalent and trivalent counterions reduce the residual charge further according to this model (to about one tenth of its magnitude). The resulting phosphate screening reduces the electrostatic stiffness of DNA and also explains the favorable entropy of binding by cation ligands to DNA, since they lead to release of counterions from the condensed ion layer to the bulk solvent.

Various experimental and theoretical models of DNA electrostatics support counterion condensation of polyion charge above a critical threshold of polyion charge density. Theories differ mainly in the structure of the condensed layer [1419]. Work on assessment of models in light of experimental observations continues; see [850, 1373], for example, for reviews.

---

## 6.4 DNA/Protein Interactions

Fundamental biological processes involve the interaction of nucleic acids and proteins. These complexes may involve various types of proteins: regulatory, packaging, replication, repair, recombination, and more. Many questions are now being addressed in this large area of research involving protein/DNA complexes, including:

- How do these proteins recognize DNA? That is, how do the protein and the DNA accommodate each other? What geometric, energetic mechanisms are involved?
- How are the mutual interactions stabilized? What are the specific and nonspecific interactions involved?
- How are nucleotide and/or amino acid substitutions tolerated in the target DNA (e.g., [971]) and bound protein regions (e.g., [734, 1100])? Numerous mutation studies (experimental and theoretical) are shedding insights into this question and highlighting the residues that are essential to structure

Table 6.2. Representative protein/DNA and drug/DNA complexes.

Complex	Binding Motif <sup>a</sup>	Binding Groove <sup>b</sup>	Details of Complex
$\lambda$ repressor	HTH	Mgr	Canonical HTH; homodimers; 2 helices of Cro dimer cradle Mgr, stabilized by direct H-bond and vdW contacts; little DNA distortion.
CAP repressor	HTH	Mgr	About 90° bend.
<i>trp</i> repressor	HTH	Mgr	Indirect, water-mediated base contacts.
Purine rep.	HTH	Mm	$\alpha$ -helices inserted in mgr.
Yeast MAT $\alpha$ 2	HTH	Mgr	Homeobox domains bind as monomers.
Zif268	Zn	Mgr	Zinc finger subfamily; each Zn finger recognizes 3 bps.
GATA-1	Zn	Mm	Transcription factors subfamily; single domain coordinated by 4 cysteines.
GAL4	Zn	Mgr	Metal binding subfamily; each of two Zn ions, coordinated by 6 cysteines, recognizes 3 bps.
GCN4	Leu/Zip	Mgr	Canonical; basic region/leucine zipper ( $\alpha$ helices) motif; slight DNA bending.
fos/jun	Leu/Zip	Mgr	$\alpha$ -helices resemble GCN4; unstructured basic region folds upon DNA binding.
fos/jun/NFAT	Leu/Zip	Mgr	$\alpha$ -helices bend to interact with NFAT.
MetJ	$\beta$ -ribbon	Mgr	Two anti-parallel $\beta$ -strands in Mgr; bends each DNA end by 25°.
papillomavirus E2 DNA target	$\beta$ -barrel	Mgr	Domed $\beta$ -sheets form an 8-strand $\beta$ -barrel dimer interface with 2 $\alpha$ -helices in Mgr; strong tailored fit for every base of the recognition element; bent DNA; compressed mgr; DNA target crystallized without protein.
TBP	$\beta$ -saddle	mgr	Ten- $\beta$ -strand saddle binds in Mgr; significant distortion, $\approx$ 90° bend.
<i>p53</i> tumor supp.	Loop/other	Mm	Binds to DNA via protruding loop and helix anchored to anti-parallel $\beta$ -barrel.
SRY	Loop/other	mgr	Isoleucine intercalated into mgr.
NFAT	Loop/other	Mm	Flexible binding loop stabilized by DNA.
histones	Loop/other	Mm	Nonspecific PO <sub>4</sub> interactions.
distamycin (drug)		mgr	Selective to AT bps; binds in mgr without distortion.

<sup>a</sup>HTH: helix/turn/helix; Zn: Zinc-binding;  $\beta$ :  $\beta$ -strand motif; Leu/Zip: Leucine zipper/bZIP; Loop/other: motifs with few representative members.

<sup>b</sup>Mgr: binding mainly to major groove; mgr: binding mainly to minor groove; Mm: binding to both grooves.

and function and those that are more variable, as evidenced by evolutionary trends. Enhancing interactions by design is also an exciting application of such studies.

- What is the relation between structural stability of the complex and biological function? For example, Burley and co-workers [971] showed that TBP can successfully bind to several single-bp mutation variants of the wild type

TATA-box octamer element of adenovirus major late promoter (AdMLP), 5' TATAAAAG 3', but that the biological transcriptional activity can be sensitively compromised by these mutations.

These are complex questions, but we are addressing them steadily as our collection of complexes between proteins and DNA and between DNA and other molecules (drugs, various polymers) is rapidly growing. See [808], for example, for a review of protein/DNA and protein/RNA molecular dynamics simulations and [1313] for a comprehensive review of DNA/protein interactions.

There are now hundreds of solved protein/DNA complexes known, a representation of which is collected in Table 6.2 and illustrated in Figure 6.5. What has certainly been established is that the observed protein/DNA interactions span the whole gamut of possibilities — regarding binding specificity for sequence or grooves, stabilizing motifs, protein topology, overall deformation, tolerance to mutations, and more.

## 6.5 Cellular Organization of DNA

Thus far we have discussed the structure of DNA at the atomic and molecular levels. Understanding the organization of DNA in the cell is important for appreciating DNA's role as the hereditary material and its versatility in structure and function.

### 6.5.1 *Compaction of Genomic DNA*

DNA's cellular organization is critical because of the enormous content of genomic DNA. The genome size — in terms of nucleotide bps per chromosome haploid (eukaryotic chromosomes are each made of two haploids) — varies from organism to organism. Though the number of bps that specifies our makeup basically increases with the number of different cell types present in each organism, it ranges greatly within an organism. Some organisms also have more genomic content than mammals. For example, bacterial genomes have about  $10^6$ – $10^7$  bps per haploid; for algae, the range is  $10^8$  to  $10^{11}$ ; for mammals it is around  $10^9$ ; yet salamanders reach the large number of  $10^{10}$ – $10^{11}$  residues.

Table 6.3 shows representative examples of the genomic content of different organisms, along with the corresponding total length of DNA (assuming the DNA is fully stretched). We see that this total DNA length isolated for bacterial chromosomes is only of order 1 mm, but for human DNA the comparative length is 3 orders of magnitude greater. In fact, if the genomic content in each human chromosome would be stretched, 4 cm of DNA would result; stretching out all

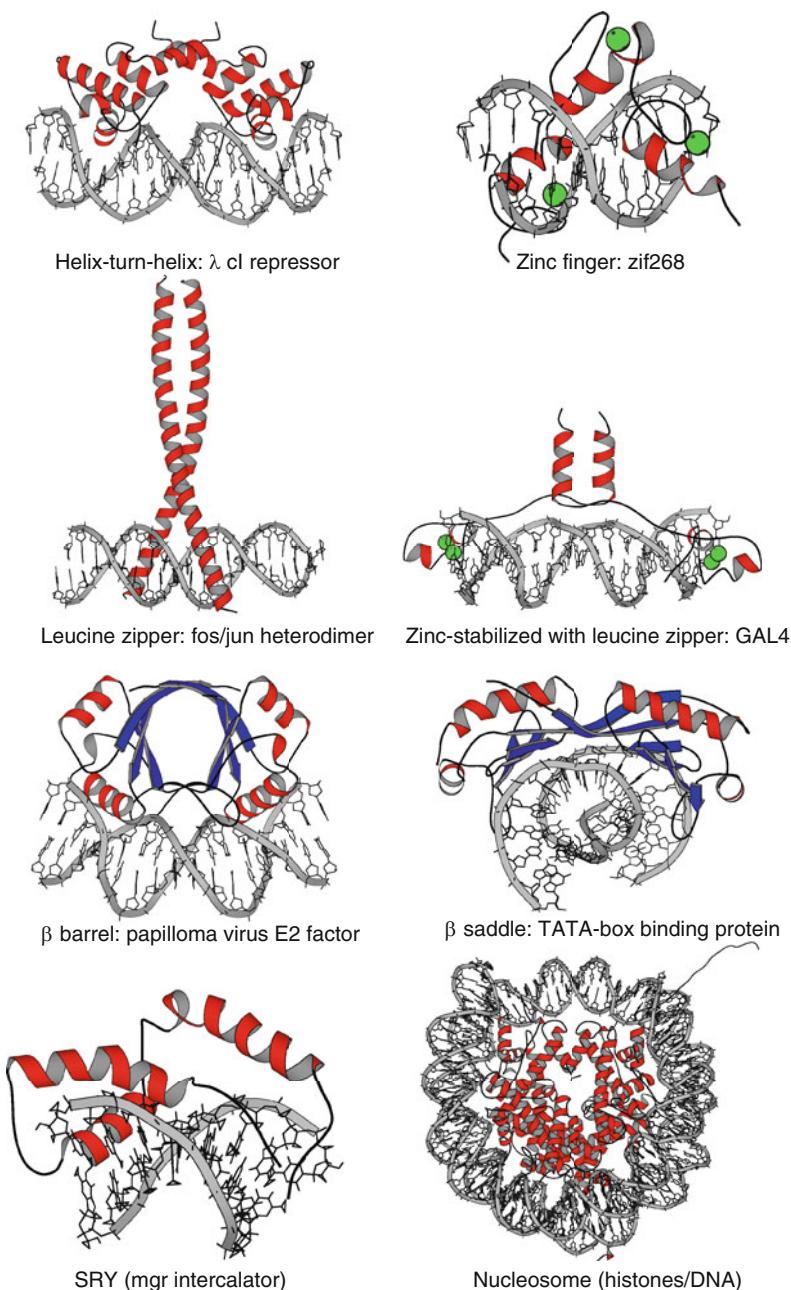


Figure 6.5. DNA/protein binding motifs for various complexes, with secondary structural elements such as  $\alpha$ -helices (red),  $\beta$ -strands (blue), metal ions (green), and DNA (grey) shown.

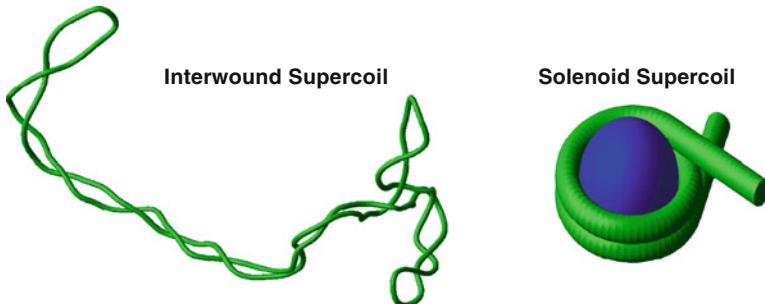


Figure 6.6. Interwound and toroidal supercoiling. Interwound supercoiling with writhing number  $Wr \approx -13$  (left), and a solenoidal supercoil, drawn wrapped around a protein core to mimic the nucleosome (right).

the DNA (two haploids for each of the 23 diploid human chromosomes) would produce 2 *meters of DNA!* Yet, the eukaryotic nucleus size (or the cell size in prokaryotes) is much smaller, around 5  $\mu\text{m}$ , thus more than 5 orders of magnitude smaller. *This necessitates extreme condensation of the DNA.*

This condensation is largely achieved by *supercoiling*, the coiling imposed on top of the double-helical coiling, involving the twisting and bending of the DNA about the global double helix axis itself. The right-handed B-DNA form is most suitable for this coiling, as first shown via modeling by Levitt [745]: B-DNA can be bent smoothly about itself to form a (left-handed) superhelical structure with minimal changes in the local structure. This property facilitates the cellular packaging of DNA, not only by reducing the overall volume DNA occupies but also by promoting protein wrapping in the chromatin fiber complex.

Different types of DNA compaction can be distinguished for prokaryotes and eukaryotes. The former have supercoiled closed circular genomes (leading to the solenoids introduced below), while the latter have linear DNA duplexes and nucleosomes (leading to the toroids discussed below). Multivalent cations, such as polyamines, are known to be important in the spontaneous condensation of DNA to form compact, orderly toroids [138] and are likely to have a significant role in chromosomal condensation.

### 6.5.2 Coiling of the DNA Helix Itself

Supercoiling is a property of both free and protein-bound DNA. When this wrapping of the global DNA helix itself involves self-interaction, like a braid, a *plectoneme* or *interwound* configuration results. If instead, the winding occurs around the imaginary axis of a torus, a *toroidal* or *solenoidal* supercoil (or superhelix) is formed (see Figure 6.6).

Interwound supercoiling is common for circular DNA, such as found in the genomes of many viruses and bacteria, as well as for topologically constrained DNA in higher organisms. In addition, eukaryotic DNA is circular in certain

Table 6.3. DNA content of representative genomes.

Organism <sup>a</sup>	kb <sup>b</sup>	Total DNA (# haploids) <sup>c</sup>
Bacteriophage $\lambda$ virus	49	17 $\mu\text{m}$
Haemophilus influenzae bacterium	1800	0.6 mm
<i>E. coli</i> bacterium	5000	1.6 mm
Yeast ( <i>S. cerevisiae</i> )	13,000	4.6 mm (16)
Roundworm ( <i>C. elegans</i> )	100,000	3.4 cm (6)
Mustard plant ( <i>Arabidopsis thaliana</i> )	135,000	4.6 cm (5)
Fruitfly ( <i>Drosophila</i> )	137,000	4.7 cm (4)
Mouse ( <i>M. musculus</i> )	3,100,000	1.1 m (21)
Human ( <i>H. sapiens</i> )	3,300,000	1.1 m (23)
Salamander ( <i>A. tigrinum</i> , axolotl)	42,000,000	14.3 m (14)

<sup>a</sup>All listed are eukaryotes except the virus and bacteria; a bacteriophage is a virus that invades bacteria.

<sup>b</sup>One kb = 1000 bps.

<sup>c</sup>Contour lengths of stretched DNA are calculated based on 3.4 Å per bp. The number of haploids reflected in this total DNA length is given in parentheses.

energy-producing organelles, the mitochondria. Toroidal-type supercoiling is characteristic of the packaged form of chromosomal DNA in the chromatin fiber (see below).

The orderly packaging of the DNA in the cell has two major roles. It contributes to the flexibility of the DNA fiber and to the accessibility of DNA for performing vital biological processes — replication, transcription, and translation — all of which require the DNA to unwind. The packaging also compacts the chromosomal material by orders of magnitude, as required.

### 6.5.3 Chromosomal Packaging of Coiled DNA

Cellular DNA is organized in the chromosomes. Each chromosome in eukaryotes (two haploids) is made up of a fiber called the *chromatin* that contains DNA wound around proteins. Specifically, in this packaged form of the DNA, DNA wraps around many large *histone* protein aggregates like a long piece of yarn around many spools (see Figure 6.7). The diameter of this DNA/protein fiber in its compact form is around 300 Å, an order of magnitude greater than the diameter of the double helix [1384].

Several levels of folding are recognized for the chromatin fiber, but only the basic units of the chromatin fiber and the associated low-level packaging are well characterized; see [383, 567] for an overview of the basic components of chromatin and levels of structure. This view has been deduced from various experimental studies of the individual globular histone proteins (class types H1, H2A, H2B, H3, H4) as well as of the chromatin fiber. Techniques used for the fiber

analysis include electron microscopy, X-ray diffraction, neutron diffraction, nucleic acid digestion combined with gel electrophoresis,<sup>3</sup> and chromatin reconstitution *in vitro*.<sup>4</sup>

### The Nucleosome: DNA + Histones

A key fact established in 1974 by Roger Kornberg and Jean Thomas was that the repeating unit of the chromatin is the *nucleosome* [674]. This unit consists of about 200 bps of DNA, most of which is wound around the outside core of histones; the remainder, *linker DNA*, joins adjacent nucleosomes (Figures 6.6 and 6.7).<sup>5</sup>

The histone proteins have a large proportion of the positively-charged residues Arg and Lys. Both residues make up between 20 and 30% of all residues: the percentages of Lys/Arg residues for H1, H2A, H2B, H3, and H4 are 29/1, 11/9, 16/6, 10/13, and 11/14, respectively. (These proteins have 215, 129, 125, 135, and 102 amino acids, respectively). Therefore, electrostatic interactions between the negatively charged DNA backbone and the positively-charged histone side chains are thought to stabilize this protein/DNA complex.

In addition, as mentioned in connection to our heightening appreciation for the subtle, sequence-dependent effects in nucleic acids, the DNA wrapped around nucleosomes has specific regions that are more favorable to bind to nucleosomes. Such recent discoveries concerning a “second genetic code”, the nucleosome positioning code [1157, 1288, 1425], are actively being pursued.

### Nucleosome Structure

The earlier works, combined with recent chemical, enzymatic, and structural studies (e.g., [55, 790, 1020]) suggest detailed organization of the nucleosome units and reveal some aspects of stabilizing electrostatic interactions. For example, based on a nucleosome crystal structure without the wound DNA [55], Moudrianakis and collaborators have shown that the nucleosome has a tripartite organization — assembly of two dimers of H2A–H2B, one on each side of a centrally located H3–H4 tetramer [55]. The nucleosome was later shown to be surrounded by a positive ion cloud with an average local density exceeding the bulk ion concentration significantly.

In 1997, the 11-nanometer nucleosome core particle, including the wrapped DNA, was solved by X-ray crystallography at 1.9 Å resolution [790] (see Figure 6.8 for a rendering of the nucleosome refined later [287] with the histone tails), revealing further details. Namely, 146 bps of core DNA are wound on

<sup>3</sup>In this procedure, the phosphodiester bond of DNA in solution is cleaved. This leaves chromatin protected and therefore reveals overall chromatin organization when analyzed by gel electrophoresis.

<sup>4</sup>Reconstitution can involve the construction of a chromatin-like fiber by adding histones to specific sequences of DNA.

<sup>5</sup>The lengths associated with the total nucleosomal DNA and with the linker component vary from organism to organism and tissue to tissue. Specifically, the total length ranges from around 160 to 260 bps; the length of linker DNA ranges broadly from about 10 to 110 bps, though it is usually around 55 bps.

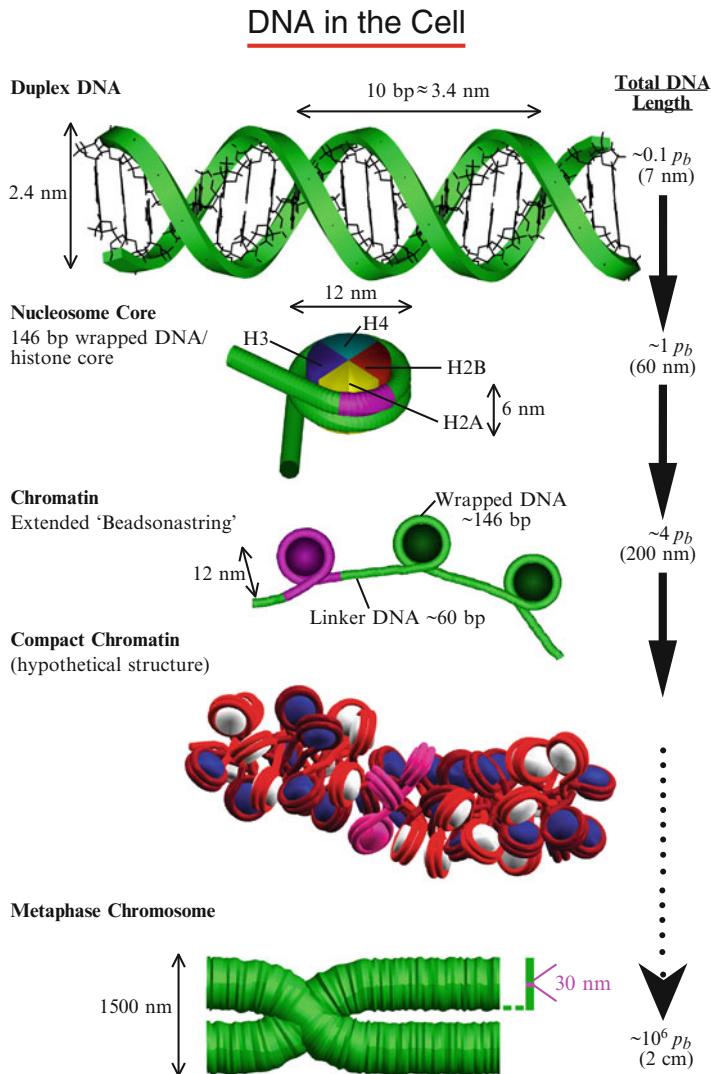


Figure 6.7. Schematic view of DNA's many levels of folding. On length scales much smaller than the persistence length,  $p_b$ , DNA can be considered straight. In eukaryotic cells, DNA wraps around a core of histone proteins (see also Figure 6.8) to form the chromatin fiber. The fiber is shown in both the extended view and a hypothetical compact zigzag view (the '30-nm fiber') deduced from a modeling study [483]; the compact structure of the chromatin fiber is unknown. Chromosomes are made up of a dense chromatin fiber, shown here in the metaphase stage. For reference, we highlight with pink in all the DNA/protein views the hierarchical organizational unit preceding it. The length scale at right indicates the level of compaction involved.

the outside of an octamer core of the histone proteins (two dimers each of proteins H2A, H2B, H3, and H4) to form 1.75 turns of a left-handed *supercoil*; the linker histone component H1 is thought to be a key player in regulation through binding to the *outside* of this core particle and contacting the linker DNA. The wrapped DNA has a structure very different from free oligonucleotides as well as DNA in other protein/DNA complexes [1050]. Distinct features include excess curvature, bending into the minor groove, twist alterations, and DNA stretching. Five years later, a high-resolution nucleosome structure with tails resolved was published [287].

Figure 6.8 illustrates the electrostatic view that emerges from Poisson-Boltzmann calculations (see discussion of theory and methodology at the end of Chapter 10) for the high-resolution nucleosome core particle with resolved tails [287]. Note the positively charged (blue) H3 and H2B tails and histone regions inside the complex, as well as the negatively charged (red) wrapped DNA.

In 2005, a tetranucleosome was solved crystallographically [1094], and several other nucleosome structures have since become available, with modified tails [786], modified DNA sequences [88], as shown in Figure 6.9, and bound to drugs like cisplatin [1392]. The nucleosome core structure shows that each histone contains unstructured end regions which make important points of contact between the protein and the DNA. Specifically, an underwinding (10.2 vs. 10.25 bp/turn) of the nucleosome-bound DNA superhelix lines up neighboring DNA grooves to form a channel through which the histone ends can pass. These tails likely play key roles in regulating biological processes, such as transcription, that require a conformational change of the complex for initiation. Ongoing investigations of the transition between the more open and more compact nucleosome structure will help us understand better transcriptional regulation and DNA packaging.

### Polynucleosome Assembly

The nucleosomal packing described above — superhelical DNA around a histone core — represents only a tenfold compression of the DNA. This is because  $\approx 166$  bps of DNA (or  $\approx 560$  Å of contour length if stretched) are organized as a core cylindrical particle of dimensions  $110 \times 110 \times 55$  Å, where 110 Å is the diameter and 55 Å represents the height of this unit. Chromatin fibers within interphase chromosomes possess a hierarchical structure. At low salt concentrations, a *beads-on-a-string* model is produced which compacts the DNA further by a factor of about 40. At physiological ionic strengths, it is believed that a next level of chromosomal organization is the so called “30 nm” (300 Å) chromatin filament observed by electron microscopy.

Several models for this polynucleosomal level of folding have been suggested based on X-ray crystallography and electron microscopy imaging. Yet, a consensus has not been reached despite decades of intense research, as reviewed recently [1271, 1297]. The two broad classes of structures for the chromatin fiber [330] include *two-start zigzag* structures, where the nucleosomes criss-cross one another around the helical axis with straight linker DNA and dominant interactions between next-nearest neighbors ( $i \pm 2$ ), and *one-start solenoid* struc-

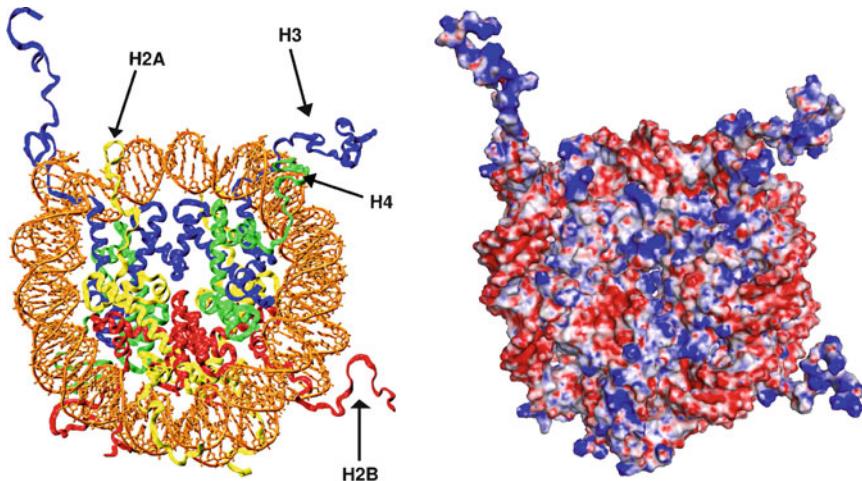


Figure 6.8. The nucleosome core particle (left) [287] and its corresponding electrostatic surface potential (right) as computed from the Poisson Boltzmann (PB) equation with the PBEQ module [586] in the CHARMM program [175] (see Chapter 10 for theoretical framework and algorithms). The grid resolution for solving the linear PB equation was 1.5 Å and 1.0 Å after focusing; the salt concentration was set to 0.15 M, and partial charges from the CHARMM program were used as input. As customarily done for macromolecules [903], the unitless (i.e., relative) dielectric constant  $\epsilon_r$  is set to the numerical value of 2 inside and 80 outside the macromolecule (i.e., 2 and 80 times the vacuum dielectric constant  $\epsilon_0$ ).

tures, where nucleosomes arrange helically around the fiber axis with bent linker DNA and dominant interactions between nearest neighbors of nucleosomes ( $i \pm 1$ ) (see Figure 6.10).

The solenoid was proposed in 1977 by John Finch, Aaron Klug, and collaborators [401] based on electron micrographs that suggested a helical form with 6 nucleosomes per turn stabilized by contacting linker histone molecules. The zigzag form [330] is evident in the recent crystallographic structure of the tetranucleosome [1094], for example, with short linker DNAs between nucleosome and without linker histones.

Indeed, this unsettled puzzle for the fiber's organization is evident in views presented over 22 years in various editions of the classic Molecular Biology textbook [17]. The 1986 (second) edition showed chromatin as a zigzag, while the 1994 (third) edition presented a solenoid; eight years later, the solenoid remained in the fourth edition, but the 2008 (fifth) edition illustrates both zigzag and solenoid models.

Only recently, have researchers begun to dissect the influence of key internal and external factors such as length of the connecting linker DNA segments between nucleosomes (which can vary from 10 to 70 bp), the binding of linker histones, and the presence of various concentrations of monovalent and divalent ions on chromatin structure. For example, electron microscopy of reconstituted

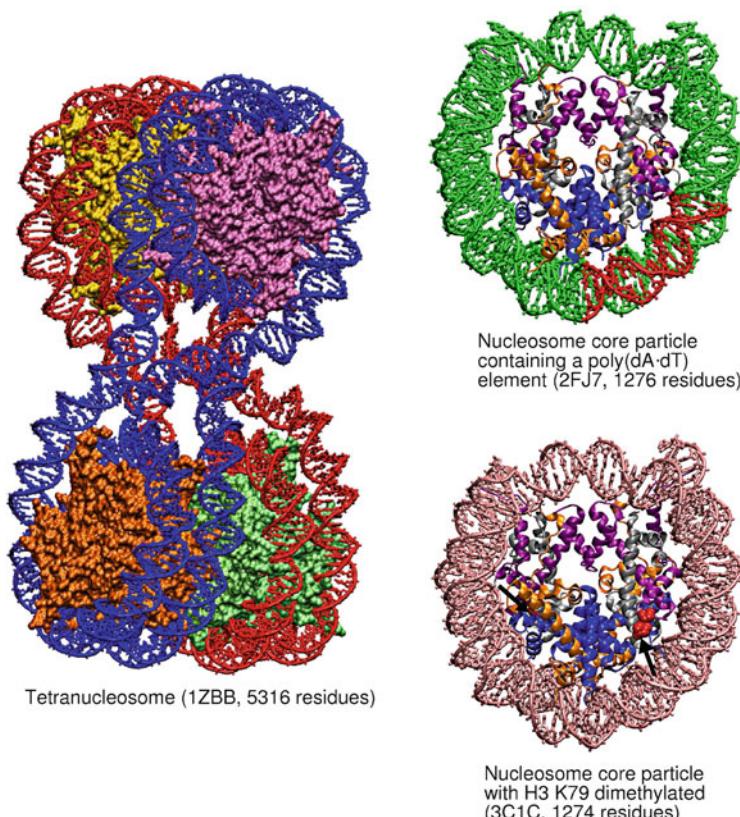


Figure 6.9. Solved crystal structures for the tetranucleosome and several nucleosomes: **tetranucleosome complex** [1094]; **nucleosome core particle containing a poly(dA•dT) tract** [88] (red); **nucleosome core particle with H3 K79 dimethylated residues** [786]. In single nucleosome particles, the histone proteins are colored by type (i.e., H3, purple; H4, silver; H2A, orange; H2B, blue). Arrows point to dimethylated H3 K79 residues (red).

fibers with long linker DNAs and with linker histone and divalent ions produced strong evidence for solenoid model [1055]. Studies also suggest that, depending on the linker DNA length and presence of linker histone, different fiber dimensions are produced; in particular, short linker DNAs cannot produce compact fibers [1055].

Modeling by Wong et al. [1388] also showed the dependence of fiber width on the linker DNA length and the orientation of linker histones. Modeling of simplified coarse-grained nucleosome models by the Rippe group reinforced the large effect of the linker length and nucleosome twist angles on the extent of fiber compaction [1219]. Our mesoscale modeling (Fig. 6.11) combined with experimental studies of EM and nucleosome interaction measurements [483] (see Fig. 6.12 and Box 6.4) suggest a compact zigzag organization for the chromatin fiber at typical linker DNA lengths with linker histones and a more heteromorphic architecture

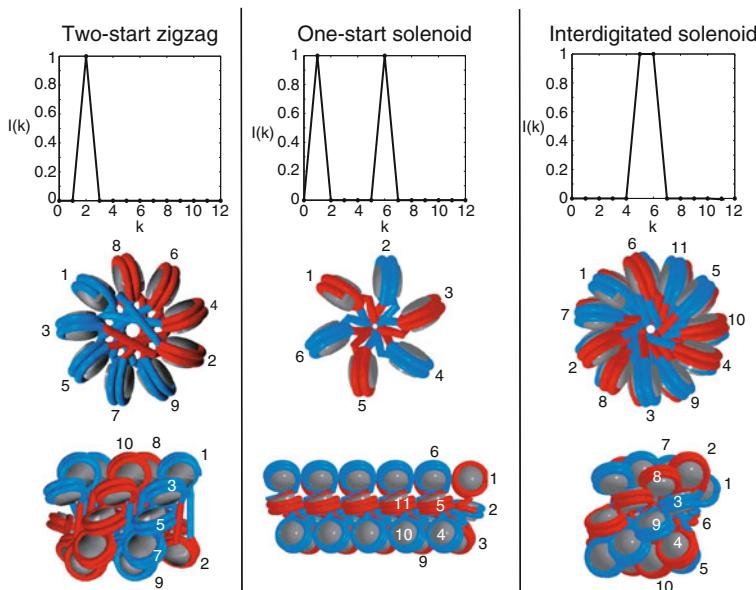


Figure 6.10. Various hypothetical models for the structure of the chromatin fiber. Three hypothesized chromatin arrays are shown from the top and side with black nucleosome cores and alternating colors for the wound DNA in consecutive nucleosomes. In the classic two-start zigzag model [1094], nucleosomes  $i$  interact with  $i \pm 2$ . In the classic one-start solenoid model [851], nucleosomes  $i$  interact with  $i \pm 1$  and  $i \pm 6$ . In the interdigitated solenoid model [1055], nucleosomes  $i$  interact on the flat sides with  $i \pm 5$ ,  $i \pm 6$ , and on the narrow edges at  $i \pm 11$ .

of zigzag forms with straight linker DNA interspersed with bent DNA linkers at divalent ion environments. Single-molecule force microscope studies [682] subjecting 25-nucleosome arrays with two linker DNA lengths (167 and 197 bp for the wrapped plus linker DNA) to forces up to 4 pN suggested a fundamental solenoid organization, stiffer fibers with short linkers, and only a stabilizing but not structure-determining effect of the linker histone. While studies are ongoing and no consensus has been reached regarding the structure of the 30-nm fiber, both the zigzag and solenoid models appear to be viable and relevant at various external and internal conditions (e.g., presence of linker histones, length of linker DNA, monovalent and divalent salt concentrations). Indeed, chromatin structure is likely heteromorphic [483, 1388].

Still, the condensation at this polynucleosomal level does not achieve the 5 orders of spatial compaction realized by the chromatin fiber near the end of the cell cycle (*metaphase* chromosomes). Various looping, scaffolding, wrapping, and specific contacts with other proteins and possibly RNA have been suggested for this higher folding to occur. Beyond the 30-nm fiber, a higher degree of coiling is thought to lead to fibers of dimension about 130 nm in diameter, and these forms in turn are condensed to chromatids of dimension 200–400 nm ([844] and refs. quoted therein).

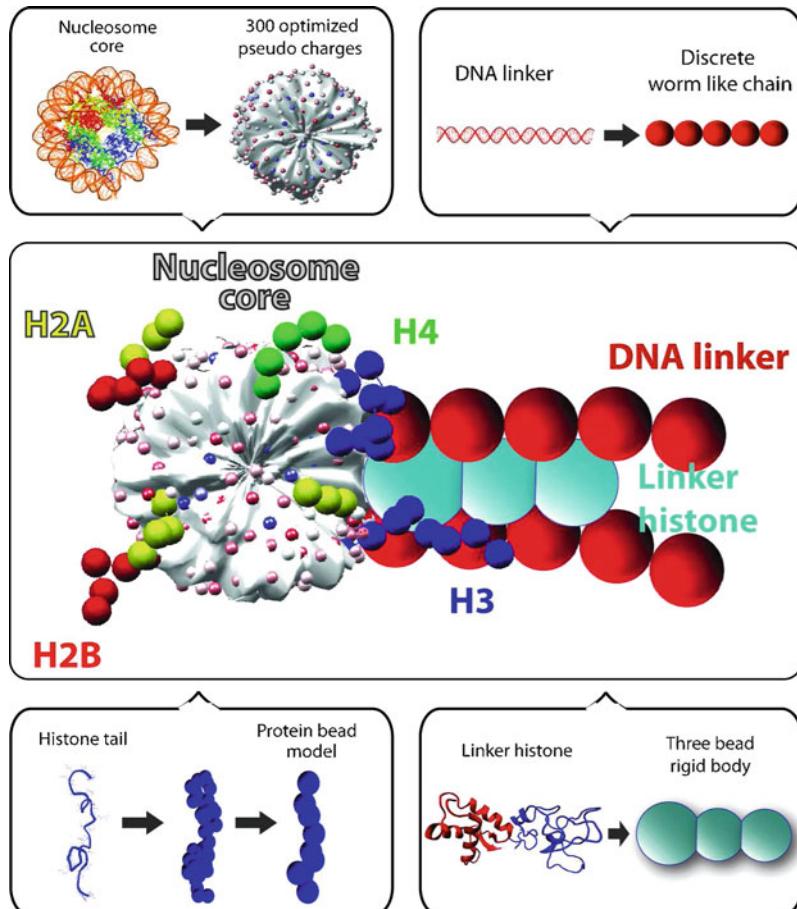


Figure 6.11. The mesoscale oligonucleosome model in which the nucleosome core with wrapped DNA (excluding histone tails) is treated as an electrostatically charged object with Debye-Hückel charges optimized to approximate the electric field using the Discrete Surface Charge Optimization (DiSCO) algorithm [108, 109, 1240, 1441]; linker DNA are treated as beads using the wormlike chain model for supercoiled DNA; and the histone tails and linker histones are coarse grained to approximate atomistic models [64, 65, 483].

Undoubtedly, in the next decade, we will begin to understand better how this nucleoprotein chromosomal material is stabilized and packed in the cells, and how packaging works with biological processes, such as replication and transcription, that require full access to the DNA. Much work is in progress on understanding the biochemical mechanisms by which the electrostatic charge density of polynucleosomes is modulated (e.g., by acetylation and phosphorylation mechanisms) to regulate transcription, and the sequence-dependent “code” in the DNA that binds to nucleosomes.

---

**Box 6.4: Simulations of Polynucleosome Folding**

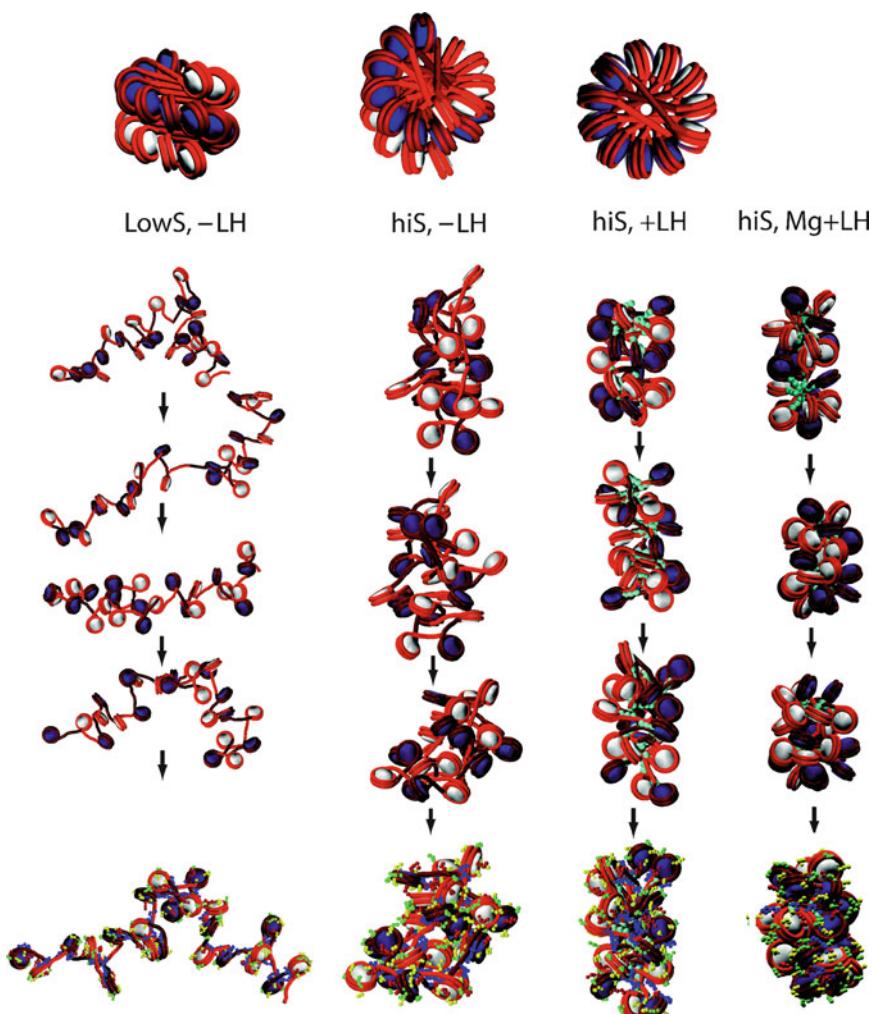
As mentioned in the text, little is known about the detailed organization of the chromatin fiber — the DNA/protein complex in eukaryotes made of DNA wound around histone proteins (see Figure 6.7) — both in terms of the linker DNA geometry and the nucleosome packing arrangement in the fiber. However, the crystallographic triumphs that produced an atomic-level view of the basic building block of the chromatin — the nucleosome core [55, 287, 790, 1094] — provide firm foundations for complementary modeling work.

Over the past several years, our group has developed a mesoscale model of chromatin (see recent overview [1114]) in which the nucleosome, excluding the histone tails, with wrapped DNA is treated as an electrostatically charged object with Debye-Hückel charges to approximate the atomistic electric field computed by the Poisson-Boltzmann equation using the DiSCO (Discrete Surface Charge Optimization) algorithm [108, 109, 1240, 1441]. In this model, each nucleosome unit is represented by several hundred charges, as shown in the Figure 6.11, optimized so that the effective Debye-Hückel electrostatic field matches that predicted by the nonlinear Poisson-Boltzmann equation [108]. This electrostatic representation is important, since properties of chromatin are sensitively dependent on the internal charges as well as on the ionic concentration of the medium.

Connecting these charged bodies of nucleosomes is the linker DNA, which we treat as beads in the wormlike chain model used for supercoiled DNA. The histone tails [62, 65] and linker histones [64] are coarse grained by beads from united-atom protein model (Figure 6.11). Following detailed model validation against available experimental data (e.g., translational-diffusion constants, radii of gyration, sedimentation coefficients, etc.), as recently summarized [64], such a model simulated using Monte Carlo [63] and Brownian dynamics permits detailed analysis of many structural, energetic, and dynamic questions, such as the folding and unfolding of dinucleosomes and trinucleosomes as a function of salt [109], the dynamics and energetics of nucleosome arrays as a function of salt [1240], the roles of histone tails in stabilizing fiber architecture [64, 65], the influence of linker histones and divalent ions in compacting the fiber structure [483], and the influence of linker DNA length on fiber organization [989].

Recent studies show that without linker histones, the fiber structure has a loose zigzag conformation [483]. Linker histones further compact the fiber by forming stems in the linker DNA where they closely interact with the linker histone, thereby promoting an ordered zigzag fiber organization. When divalent ions are introduced, bending in some linker DNAs results to minimize steric clashes along the fiber axis; this produces a mostly zigzag fiber accented with some bent linker DNA, resembling the solenoid form. These studies thus lend support for the both solenoid and zigzag models; moreover, they underscore the heterogeneous and polymorphic nature of the chromatin fiber. These dynamic fiber structures and their dependence on the ionic strength and linker histones can be seen from Figure 6.12, which shows snapshots at low and high salt conditions (including divalent ions), with and without linker histones. A condensation trend is apparent as the ionic concentration increases and as linker histones provide additional screening of the negatively charged linker DNA.

---



## 6.6 Mathematical Characterization of DNA Supercoiling

Now that we have covered many local features of nucleic acids as well as basic global characteristics, I will introduce quantitative tools to describe how DNA is condensed in the cell.

### 6.6.1 DNA Topology and Geometry

Topological methods are important for analyzing some reactions of supercoiled DNA. For example, successful collaborations between biologists and mathematicians have led to techniques that establish mechanisms for various reactions that produce knots and catenanes (linked DNA rings) from supercoiled DNA substrates by recombination [1239].

#### Basic Topological Identity

The topological method, as well as the many computational methods used to study DNA supercoiling, rely on the fundamental identity attributed to Grigore Calugareanu, Jim White, and F. Brock Fuller [279, 431, 1369]. This equation relates the topological invariant  $Lk$ , or *linking number*, to the geometric quantities *twist*,  $Tw$ , and *writhe*,  $Wr$  as:

$$Lk = Tw + Wr. \quad (6.3)$$

#### Linking Number

Essentially,  $Lk$  characterizes the order of linkage of two closed-oriented curves in space (Figure 6.13). It is unchanged by continuous deformations but altered when



Figure 6.12. Configurational snapshots of 24-core oligonucleosomes simulated at various conditions [989]. The top shows three different orientations of the zigzag starting configuration used for all series. The four simulated conditions from left to right involve: low (0.01 M) monovalent salt without linker histone, physiological monovalent salt (0.15 M) without linker histone, physiological monovalent salt (0.15 M) with linker histone, and physiological monovalent salt (0.15 M) with moderate levels of divalent ions (as modeled to a first order approximation [483]) and linker histones. The configurations are simulated for 10 million steps using Monte Carlo simulations [63]. The nucleosomes are colored white and navy in alternating order, and linker DNA segments are rendered red. The linker histones (each represented by 3 beads) are colored in turquoise. The final configurations at bottom also show the other tail beads colored as: H2A—yellow, H2B—red, H3—blue, and H4—green (see mesoscale model in Fig. 6.11). Overall, we see the increased compaction trend from left to right. At low salt, the negatively charged linker DNAs (red segments) repel one another and expand the array, producing the beads-on-a-string form shown also in Figure 6.7. At higher salt concentrations, the chromatin fiber condenses. Note the compaction as enhanced by linker histone as well as divalent ions. With divalent ions, the mostly irregular zigzag conformation is accented by some linker DNA bending [483].

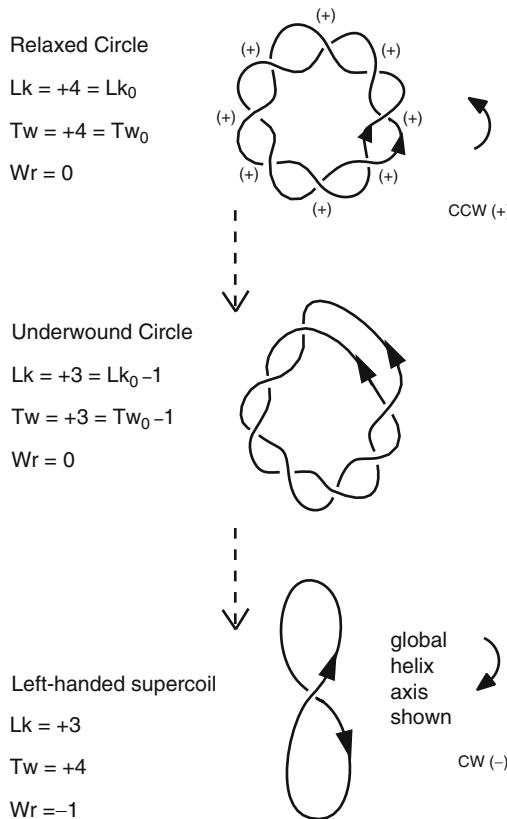


Figure 6.13. Descriptors of supercoiling topology and geometry. For the relaxed and underwound circles of DNA, each line represents one polynucleotide strand of the double helix. For the supercoil, the curve represents the *double helical axis* of the DNA.

strand breaks occur. By convention, for right-handed DNA, the linking number of the relaxed state,  $Lk_0$ , is the number of primary turns:  $Lk_0 = n/n_b$  where  $n$  is the number of bps and  $n_b$  is the number of bps/turn (e.g., 10.5).

$Lk$  can be rigorously computed from any planar projection of these curves onto the plane by summing all signed crossings (ignoring self interaction) and dividing the result by two, as shown in Figure 6.13. By convention, a crossover index is considered *positive* if the tangent vector of the top vector is rotated *councclockwise* to coincide with the tangent of the bottom curve; the sign is *negative* if this rotation is *clockwise*. For precise mathematical definitions, refer to [431, 1370].

### Twist

The variable  $Tw$  reflects the familiar concept of the helical twist angle introduced earlier, namely the winding angle per residue. For the case of a helix spinning about a global helical axis,  $Tw$  measures the number of times the he-

lix revolves about that axis. As conventional for  $Lk$ ,  $Tw$  is positive if the helix is right-handed and negative if it is left-handed (see Figure 6.13).

When the global helix axis is a closed circle and the DNA winds  $n$  times around this planar axis,  $Tw = Lk \equiv Lk_0 = n/n_b$ . From eq. (6.3), this implies that  $Wr = 0$ . This case is shown for the relaxed circle in Figure 6.13.

A nonzero writhing number is introduced by the freeing of the two ends of the closed ring and overwinding or underwinding the duplex before resealing the ends. Topoisomerase enzymes perform this magical act in the cell. The unwrithed (planar) state is unfavored because there is substantial local disruption to the natural twist of the DNA (Figure 6.13). The induced torsional stress can be alleviated through nonplanar writhing, or supercoiling, namely the twisting of the global helix axis itself. This is illustrated in Figure 6.13 where a figure-8 interwound forms the path of the double helical axis.

### Writhe

The writhing number describes the geometry of this global helix axis in space. Essentially, it is the average number of crossings in space of this curve. Like the linking number,  $Wr$  can be obtained for each planar projection and averaged over all these projections. The same sign convention used for  $Lk$  is used for  $Wr$  (Figure 6.13).

Rigorously, the writhe can be computed by the *Gauss double integral* [1370] (mathematically speaking, a map that associates each pair of curve points with the unit sphere). For example, the simplest interwound — a figure-8 form — has  $Wr = \pm 1$ ; the interwound with two self-crossings, separating the chain into three regions, has  $Wr = \pm 2$ . The figure-8 interwound in Figure 6.13 (bottom) has  $Wr = -1$ , and the interwound supercoil in Figure 6.6 (left) has  $Wr \approx -13$ . See [1127, for example] for computational determination of  $Wr$ .

Typically, DNA is subjected to a *linking number deficit*, or *underwinding*, with respect to its relaxed state:  $\Delta Lk = Lk - Lk_0 < 0$ . This state is often described by the normalized quantity  $\sigma = \Delta Lk / Lk_0$ . The value  $-0.06$  is typical of DNA *in vivo*. In the absence of loop-constraining agents like histone proteins, DNA tends to absorb most of this stress into writhing, by forming left-handed supercoils (negative  $Wr$ ).

## 6.7 Computational Treatments of DNA Supercoiling

While the mathematical concepts of  $Lk$ ,  $Tw$ , and  $Wr$  have been invaluable for interpreting supercoiling, the energetic and geometric details of the process remain unknown and require further computational treatments. The processes of interest involve fundamental reactions such as replication, recombination, and transcription, that often require, or are facilitated by, a supercoiled DNA substrate [631, 1307]. Experimental techniques such as gel electrophoresis and electron microscopy (atomic, scanning, and cryo) have been crucial for studying properties

associated with large supercoiled DNA molecules. The former allows separation of DNA molecules into topological isomers, and the latter offers views of overall shapes of supercoiled DNA. However, the experimental resolution is limited due to the large size (thousands of bps) of the DNA subjects and their extreme floppiness in solution. Hence, theoretical tools based on numerical analysis and computational modeling have been useful.

Computational treatments of the energetics and dynamics of supercoiled DNA have largely relied on the theories of polymer statistical mechanics and elasticity in the framework of elastic rod mechanics and dynamics. For realistic results, non-elastic contributions — electrostatics and hydrodynamics — must be included. These can be easily incorporated into computational models and investigated using various configurational sampling techniques (Metropolis/Monte Carlo) [1308] and dynamic techniques (by molecular, Langevin, and Brownian dynamics) [1105].

### 6.7.1 DNA as a Flexible Polymer

Very long DNA can be described by concepts from polymer statistical mechanics [410]. The DNA polymer is characterized by two key quantities: a contour length  $\mathcal{L}$ , and a bending rigidity  $A$ . This view of a random coil slithering through space is reasonable only for naturally occurring DNA of mixed sequences that are not intrinsically bent at lengths much greater than DNA's *persistence length*,  $p_b$ .

The persistence length is essentially the *length scale on which the polymer directionality is maintained*. It can be computed as the average projection angle between the end-to-end distance vector on the first bond vector of a polymer, in the infinite-length limit [410]. The main result from polymer statistical mechanics is that the length of this vector is proportional to the square root of the contour length; that is, the mean square displacement,  $\langle R^2 \rangle$ , satisfies:

$$\langle R^2 \rangle = 2p_b \mathcal{L} \quad (6.4)$$

with  $2p_b$  as the proportionality constant.

Hence, for lengths  $\ll p_b$ , the DNA can be considered straight, but for lengths  $\gg p_b$ , a better description is a bent random coil undergoing Brownian motion. This is evident from the views of DNA on different length scales, as shown in Figure 6.7. For DNA, the persistence length *in vivo* is around 500 Å or about 150 bps at physiological monovalent salt concentrations [491]. A salt dependence of the persistence length is recognized for DNA [104], but the magnitude of this effect is not currently understood.

The persistence length is also related to the bending force constant of a long flexible polymer of length  $\mathcal{L}$  and bending rigidity  $A$  via

$$A = p_b k_B T \quad (6.5)$$

( $k_B$  is Boltzmann's constant and T is the temperature). In this description, the floppy polymer writhes through space as a worm-like chain. The bending rigidity — which tries to keep the DNA straight — is balanced by thermal forces — which tend to bend it in all directions.

A decomposition of the persistence length into static and dynamic components has been developed [1097, 1272].

### 6.7.2 Elasticity Theory Framework

Since the pioneering applications of Fuller [431, 432], the elastic energy approximation has proven valuable for studying global (i.e., long range and time flexibility) features of superhelical DNA. In this approximation, the DNA polymer is idealized as a long, thin and naturally straight (i.e., no intrinsic curvature) elastic isotropic rod with a circular cross section. Homogeneous bending (i.e., equal flexibility in all directions) is often assumed as a first approximation, though current computational models follow twist locally, for example by using body-centered coordinate frames along the helical axis with associated Euler transformations.

The elastic deformation energy can then be written as a sum of bending and twisting potentials, with bending and torsional-rigidity constants ( $A$  and  $C$ , respectively) deduced from experimental measurements of DNA bending and twisting [491]. The bending term is proportional to the square of the curvature  $\kappa$ , and the twisting energy is proportional to the twist deformation:

$$E = E_B + E_T = \frac{A}{2} \oint \kappa^2(s) ds + \frac{C}{2} \oint (\omega - \omega_0)^2 ds. \quad (6.6)$$

In these equations,  $s$  denotes arclength, and the integrals are computed over the entire closed DNA curve of length  $\mathcal{L}_0$ . The intrinsic twist rate of the DNA is  $\omega_0$  (e.g.,  $2\pi/10.5$  radians between successive bps). The parameters  $A$  and  $C$ , denoting the bending rigidity and the torsional rigidity constants, respectively, can be estimated from various experimental measurements of DNA, such as the persistence length. These force constants are key characteristics of an elastic material.<sup>6</sup> See Box 6.5 for the relationship between the force constants  $A$  and  $C$  and the bending and twisting persistence lengths, and Box 6.6 for the relationship between  $A$  and measured variations in roll and tilt angles in solved DNA structures.

---

<sup>6</sup>For example, a rubber band, for which bending is facile, has a small ratio  $r = A/C$ , while a stiff material with strong bending resistance has large  $r$ . This bending to torsional-rigidity ratio is a key descriptor of an elastic material. It is related to both Poisson's ratio ( $\sigma_e$ ) — a characteristic of a homogeneous isotropic material — and the geometry of the rod cross section. Specifically, for a homogeneous elastic rod of circular cross section,  $r = 1 + \sigma_e$ . **Note:** the ratio  $A/C$  is usually designated by the symbol  $\rho$ , but here we reserve  $\rho$  for the roll angle.

---

**Box 6.5: Elastic Constants and Persistence Length**

As discussed in the text (eq. (6.5)), the elastic bending constant of DNA,  $A$ , can be linearly related to the bending persistence length of DNA,  $p_b$ , as:  $A = k_B T p_b$ , where  $k_B$  is Boltzmann's constant and  $T$  is the temperature. In addition,  $A$  can be related to the root-mean-square (RMS) bending angle  $\langle \theta_b^2 \rangle^{1/2}$  of a semiflexible chain with a preferred axis of bending (perpendicular to the helix axis), for which the angular fluctuations are independent from one another [694]:

$$A = (2k_B T h) / \langle \theta_b^2 \rangle. \quad (6.7)$$

Here  $h$  is the bp-separation distance in DNA (around 3.4 Å) [491, 1095]. (Box 6.6 describes how  $\theta_b$  is related to the roll and tilt variables introduced earlier).

Similarly, the torsional rigidity constant can be related to the persistence length of twisting  $p_{tw}$  and the RMS twist angle  $\Omega$  as [830]:

$$C = (p_{tw} k_B T) / 2 \quad (6.8)$$

and

$$C = (k_B T h) / \langle \Omega^2 \rangle. \quad (6.9)$$

For the bending to torsional rigidity constant ratio  $r = A/C$ , the relationship

$$\langle \Omega^2 \rangle^{1/2} = \langle \theta_b^2 \rangle^{1/2} \sqrt{\frac{r}{2}} \quad (6.10)$$

follows.

The typical values for DNA ( $A = 2.0 \times 10^{-19}$  erg cm and  $C = 3.0 \times 10^{-19}$  erg cm) correspond to persistence lengths for bending and twisting of  $p_b = 500$  Å [491],  $p_{tw} = 750$  Å, and  $r = 2/3$ . These values correspond to RMS bend and twist-angle values of  $\langle \theta_b^2 \rangle^{1/2} = 6.7^\circ$  and  $\langle \Omega^2 \rangle^{1/2} = 3.9^\circ$  at room temperature, in the context of isotropic bending. The bending constant  $A$  can also be related to measured local angular variations in solved DNA structures, as detailed in Box 6.6. The torsional rigidity constant has recently been estimated from pulling experiments on single molecules of DNA to be 40% higher than generally accepted values [182].

---

### 6.7.3 Simulations of DNA Supercoiling

Many groups are studying the geometric, thermodynamic, statistical, and dynamical properties associated with supercoiling using a variety of models both for representing the DNA and for simulating conformations. See, for example, [103, 107, 108, 145, 231–233, 512, 577, 656, 657, 697–699, 830, 939, 947, 1036, 1036, 1052, 1105, 1119, 1124, 1127–1129, 1144, 1175, 1211, 1249, 1250, 1256, 1274, 1306–1308, 1308, 1310, 1311, 1339, 1366, 1410].

---

**Box 6.6: Relationship between the DNA Bending Constant and Local Angular Measurements in Solved Nucleic Acid Structures**

To relate the elastic bending constant  $A$  and *measured* standard deviations of DNA bending angles, the two components of bending — namely roll ( $\rho$ ) and tilt ( $\tau$ ), each with associated stiffness constants  $A_\rho$  and  $A_\tau$  [943] — can be related to  $A$  via:

$$A = \frac{2k_B Th}{\langle \rho^2 \rangle + \langle \tau^2 \rangle}. \quad (6.11)$$

As above,  $h$  is the bp-separation distance in DNA (around 3.4 Å). By comparing the above to eq. (6.7), we see that the bending persistence length of  $p_b = 500$  Å corresponds to an isotropic model where  $\langle \theta_b^2 \rangle^{1/2} = 6.7^\circ$ , as well as to an anisotropic bending model where  $\langle \rho^2 \rangle^{1/2} = 5.7^\circ$  and  $\langle \tau^2 \rangle^{1/2} = 3.6^\circ$ . Values from analysis of B-form crystal structures yield somewhat smaller values for these roll and tilt fluctuations (e.g.,  $\langle \rho^2 \rangle^{1/2} = 5^\circ$  and  $\langle \tau^2 \rangle^{1/2} = 3^\circ$ ) and thus a larger effective bending persistence length and rigidity constant (by about 1.3, or  $A = 2.6 \times 10^{-19}$  erg cm and  $p_b = 500$  Å). However, the static fluctuations in the crystal structures do not directly correspond to the dynamic range of DNA flexibility.

For sufficiently small angular deflections, we can partition the total bending magnitude as

$$\langle \theta_b^2 \rangle = \langle \rho^2 \rangle + \langle \tau^2 \rangle = [k_B Th(A_\rho + A_\tau)]/(A_\rho A_\tau). \quad (6.12)$$

This first-order model does not account for the well-recognized preferential directions of bending of DNA into the major and minor grooves with respect to other directions [473, 940, 1096, 1285, 1454]. A non-uniform bending chain view is required for better representations, especially for sequences with intrinsic curvature, like A-tracts.

---

A few of the many topics that have been studied using such models are the behavior of supercoiled DNA as a function of salt [1125, 1309], solvent [1036], superhelical density [1129], and length [830, 1214, 1250]; extended theoretical treatments based on elastic rod mechanics and dynamics [305, 651, 1175]; the response of supercoiled DNA to constraints imposed by proteins [945]; theory of elastic rods with applications to DNA [255, 256]; the geometry of DNA in small minichromosome systems [834]; and the *site juxtaposition* time (the bringing together in space of linearly-distant DNA segments due to supercoiling) [607, 1211, 1310]. See Boxes 6.4 and 6.7 for examples of nucleosome folding and site juxtaposition studies, respectively, as well as corresponding Figures 6.12 and 6.14–6.15 and related experimental studies of EM and nucleosome interaction measurements [483].

Many models for studying nucleosomes and oligonucleosome arrays have also been developed (e.g., [64, 108, 353, 883, 1219, 1240, 1312, 1441]) and should offer further insights into the nature of the complex folding and dynamics of the chromatin fiber.

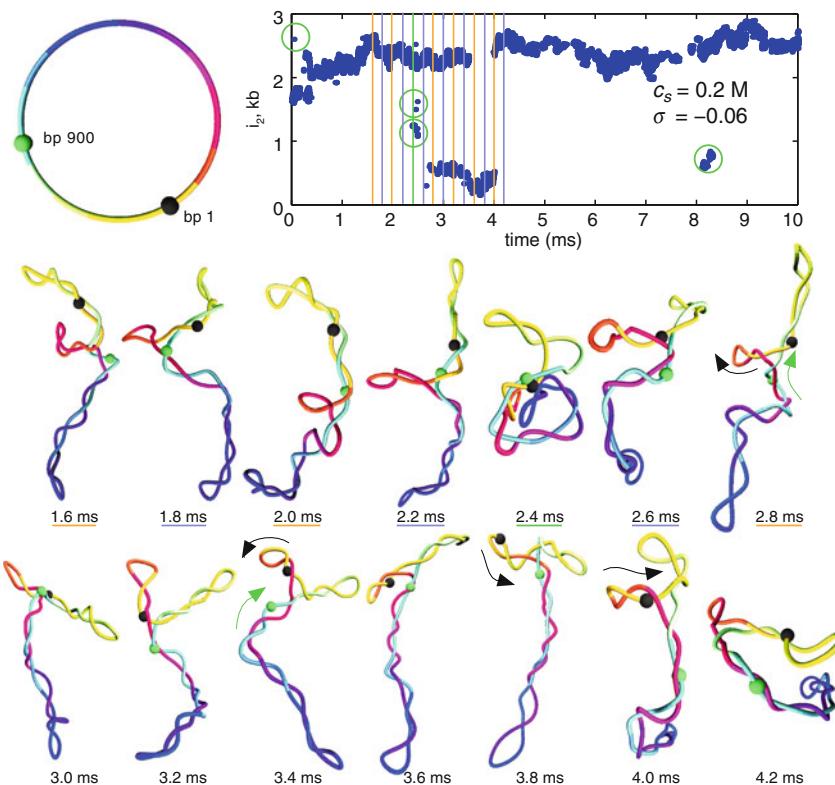


Figure 6.14. Juxtaposition kinetics, as analyzed for two beads separated by 900 bps in circular DNA of length 3000 bps and  $\sigma = -0.06$  by Brownian dynamics simulations at high salt ( $c_s = 0.2$  M sodium ions) [577]. (Top) Juxtaposition event plot, showing bps  $i_2$  that juxtapose with bp 1; the circled values correspond to more random juxtaposition events, as illustrated for the snapshot at 2.4 ms. (Bottom) BD snapshots, showing ordered motions along the DNA contour and rarer random collision events.

---

#### Box 6.7: Simulations of DNA Site Juxtaposition

Illustrative results from computer simulations of supercoiled DNA are shown in Figure 6.14. These snapshots are taken from a Brownian dynamics simulation of a 3000-bp DNA system at the superhelical density of  $\sigma = -0.06$  at high salt using a homogeneous bead/wormlike model [577]. Simulations have tabulated the times for distant sites to *juxtapose* due to the ambient floppiness of DNA as a function of the superhelical density, DNA length, site separation, and the salt concentration [607].

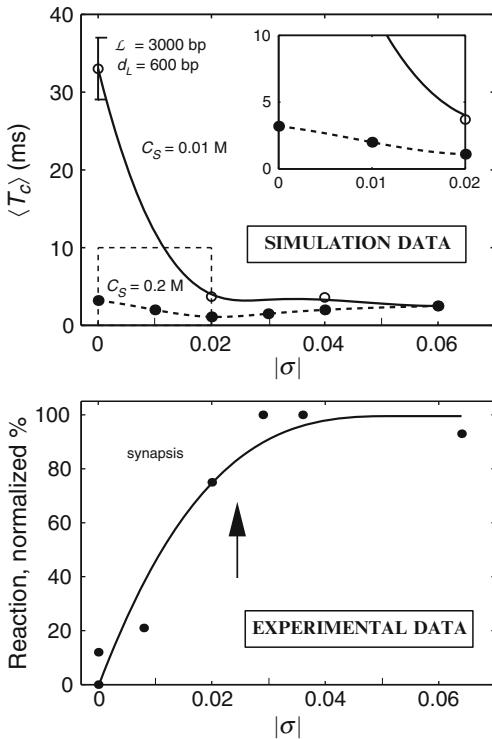
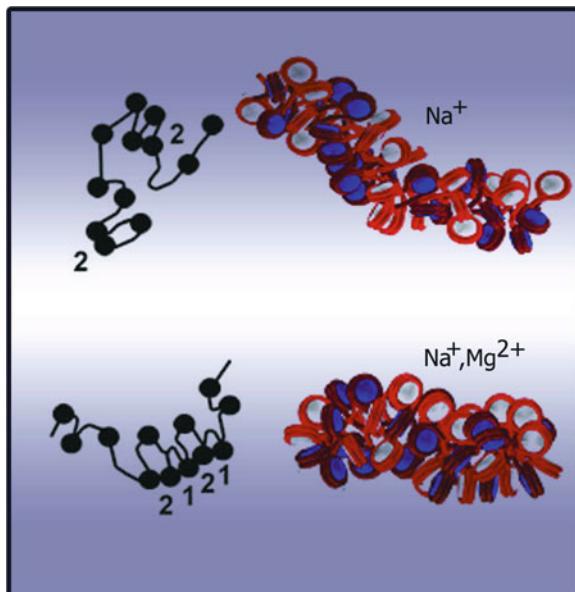


Figure 6.15. (Top): Site juxtaposition time measurements,  $\langle T_c \rangle$ , defined as the mean time for two sites  $d_L$  bps apart to come within 100 Å, as averaged over all such separated pairs, as a function of  $|\sigma|$  for  $d_L = 600$  bps for a 3000 bp system and two salt concentrations [577]. The inset focuses on the large salt effect at low  $|\sigma|$ . Note also the plateauing effect of  $\langle T_c \rangle$  around  $|\sigma| = 0.03$  for both salt concentrations. This is also the experimental observation (bottom) [119] for the synapsis-time dependence on the superhelical density.

In the study of juxtaposition, two linearly-distant sites are considered *juxtaposed* when the distance between them is 100 Å or less. Interestingly, we find (Figure 6.15) that at low salt [608] juxtaposition times are accelerated by a factor of 10 or more due to supercoiling (i.e., juxtaposition times for relaxed DNA are much slower), but that supercoiling beyond  $|\sigma| = 0.03$  does not accelerate such site synapsis times [608]. Such critical behavior at moderate  $|\sigma|$  agrees with experimental findings regarding supercoiling-dependent synapsis rates for the resolvase recombination system [119], as also displayed in Figure 6.15. The explanation for the fact that further supercoiling does not enhance juxtaposition times significantly emerges from our simulations: the balance between flexibility and tight supercoiling is optimized at this mean value of superhelicity.

At high salt, the effect of superhelicity is far less dramatic. Further analysis of the mechanism for juxtaposition shows that large conformational rearrangements due to Brownian motion operate at low superhelical density and low salt, whereas systematic *slithering* motions (reptation, or bidirectional conveyer-like motion) along the polymer contour [118] are more prominent at high salt, as shown in Figure 6.14. At high salt, site juxtaposition

can be described by a combination of slithering, branch rearrangements, and interbranch collisions, as also shown in the figure. The figure, prepared for studying site juxtaposition kinetics [577], shows at top a plot in time of the bp  $i_2$  that juxtaposes with the first bp (i.e., is within 100 Å of it). The continuous pattern in  $i_2$  in time indicates ordered slithering; at low salt, the patterns are much more random.



# 7

## Topics in Nucleic Acids Structure: Noncanonical Helices and RNA Structure

### Chapter 7 Notation

SYMBOL	DEFINITION
$\alpha$	polynucleotide backbone torsion angle
$\gamma$	polynucleotide backbone torsion angle
$\chi$	glycosyl (sugar/base) torsion angle

Anything that DNA can do, RNA can do better.

*Susan Gottesman, Cold Spring Harbor symposium, 2006.*

### 7.1 Introduction

This chapter builds upon nucleic acid concepts introduced in the prior two chapters to include a description of alternative hydrogen bonding schemes in nucleic acids, non-canonical helical and hybrid structures, DNA mimics, over-stretched and under-stretched DNA, and RNA structure and folding, including secondary and tertiary-structure RNA modeling.

Besides the excellent texts [139, 195, 1080, 1191], the reader is invited to explore the wealth of related structural data on the nucleic acid database NDB

([ndbserver.rutgers.edu/](http://ndbserver.rutgers.edu/)) and the protein data bank/research collaborative for structural bioinformatics (PDB/RCS) resource [128, 1365], as well as information concerning RNA on the **RNA Ontology Consortium** (ROC) [735].

As in the previous chapter, we abbreviate *base pair* and *base pairs* as bp and bps, respectively.

## 7.2 Variations on a Theme

As evident from the previous chapter, nucleic acids are extremely versatile. The classic right-handed double helical structure described by Watson and Crick is an excellent textbook model. In Nature, however, polynucleotide helices have developed an enormous repertoire of sequence-dependent structures, helical arrangements, and folding motifs on both the bp and higher level of folding.

### 7.2.1 Hydrogen Bonding Patterns in Polynucleotides

Hydrogen-bonding variations are important for DNA's adaptability to base modifications (e.g., methylation, drug or carcinogen binding), bp mutations and mismatches, and various interactions between DNA and proteins and among polynucleotide helices.

#### Classic Watson-Crick (WC)

The *classic* WC hydrogen-bonding arrangement as shown in Figure 5.3 of Chapter 5 is a particularly beautiful arrangement. It can be appreciated by noting that the C1' (pyrimidine) to C1' (purine) distance is around 10.6 Å in both the AT and GC pairs, in good preparation for helix stacking. The intrinsic energy of hydrogen bonds — that is, relative to vacuum — is generally in the 3–7 kcal/mol range but only up to 3 kcal/mol in nucleic acids due to geometric constraints.<sup>1</sup> Though much weaker than covalent bonds (80–100 kcal/mol) [1191, p. 12–14], the cumulative effect in the double helix is substantial due to the collective impact of hydrogen-bonding interactions resulting from cooperative effects [1018]. Based on theoretical studies, the strength of the hydrogen-bond energy depends on base composition but is very similar for the two WC bps. The energy of bp stacking, 4–15 kcal/mol, substantially contributes to overall helix stability and depends on the nucleotide sequence [169].

Yet, hydrogen-bonding patterns in polynucleotides are extremely versatile. There are several nitrogens and oxygens on the bases, allowing various donor and acceptor combinations involving different interface portions of the aromatic rings.

---

<sup>1</sup>Estimates of base-pairing energetics in solvated nucleic acids come mainly from theoretical calculations [1225]. This is because the resolution of macromolecular thermodynamic measurements into subcontributions (hydrogen bonding, stacking, and electrostatic forces) remains a challenge, despite numerous thermodynamic studies on nucleic acid systems [169, 449, 1007].

To appreciate this versatility, we first examine the classic WC pattern in detail.

For an AT pair, two hydrogen bonds form between the C4–N3 face of T and the C6=N1 face of A (see Figure 5.3 of Chapter 5 and Figure 7.1 here). One hydrogen bond involves the sequence O4 (T) · · · H–N6 (A), i.e., between the thymine carbonyl oxygen O4 attached to ring atom C4 and the adenine N6 atom, which is attached to the ring C6 atom. The other hydrogen bond is N3–H (T) · · · N1 (A), between the thymine ring nitrogen N3 and the adenine N1.

For the GC pair, three hydrogen bonds form between the C4=N3–C2 face of C and the C6–N1–C2 face of G: N4–H (C) · · · O6 (G), N3 (C) · · · H–N1 (G), and O2 (C) · · · H–N2 (G). All hydrogen-bond lengths in both the AT and GC bps are very similar, with individual lengths ranging from about 2.85 to 2.95 Å between the heavy atoms.

### Reverse WC

*Reverse WC* base-pairing schemes can result when one nucleotide rotates 180° with respect to its partner nucleotide, as shown in Figure 7.1. Thus, one hydrogen bond from the original WC scheme might be retained, but another can involve a different atom combination. This flip of one bp changes the position of the glycosyl linkage. Hence, an AT pair can more easily accommodate this flip due to near symmetry about the N3–C6 axis of T. That is, since the ring face C4–N3–C2 of T has carbonyl attachments at both C4 and C2, the hydrogen bonding face can be changed from C4–N3 to C2–N3 (see Figure 7.1). This type of reverse WC base pairing exists in parallel DNA [1080].

### Hoogsteen

*Hoogsteen* bps utilize a different part of the aromatic ring for hydrogen bonding. Specifically, the N7–C5–C6 face of A and G, where N7 is in the 5-membered ring and C6 is in the 6-membered ring of the purine, is used in the Hoogsteen arrangement (see Figure 7.1). This is instead of the C6=N1 face of A and G — both atoms of which are in the 6-membered ring of the fused-ring compound — in combination with the C4–N3 face of T and C — as in WC interactions.

The Hoogsteen arrangement also implies a change in the glycosyl torsion angle  $\chi$ , from *anti* to *syn*, and a shortening of the C1'–C1' distance, from about 10.5 Å in WC to about 8.5 Å in Hoogsteen base pairs (e.g., [12]).

Hoogsteen base pairing helps stabilize structures of highly bent DNA (e.g., TATA elements [971]), tRNAs, and DNA triplexes where two strands are held by WC hydrogen bonds and where two strands are stabilized by Hoogsteen-type pairing (see Figure 7.1). They also appear occasionally in complexes of DNA with anti-cancer drugs. Recently, an unusual Hoogsteen A–T bp was discovered in a high-resolution complex of unbent DNA bound to 4 MAT $\alpha$ 2 homeodomains (see Figure 7.2) [12, 13].

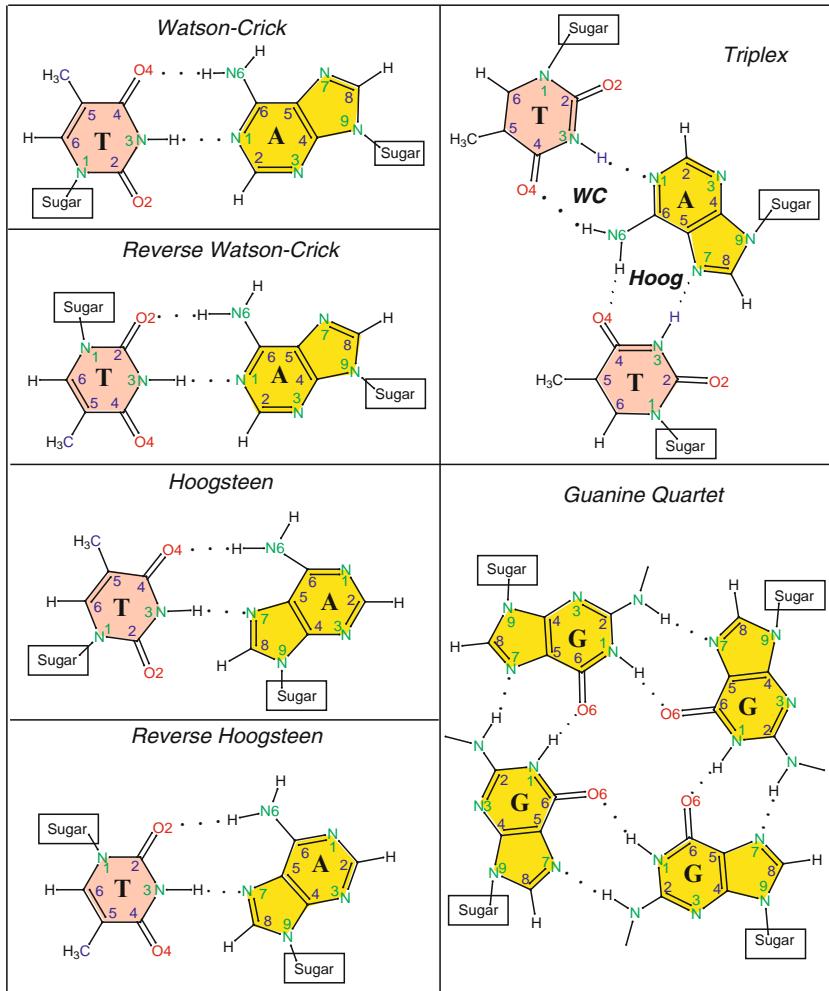


Figure 7.1. Various hydrogen-bonding schemes: (left) classic Watson-Crick, reverse Watson-Crick, Hoogsteen, and reverse Hoogsteen in base pairs; (right) WC/Hoogsteen in a triplex (from an H-DNA structure, which forms in sequences that have stretches of purine followed by stretches of pyrimidines, PDB code 1B4Y) [1290], and patterns in a guanine-quartet quadruplex (or guanine tetraplex, PDB code 1JB7) [569].

### Reverse Hoogsteen

Similarly, a *reverse Hoogsteen* arrangement involves the flipping of one nucleotide by 180° with respect to its partner, by analogy to reverse WC, as shown in Figure 7.1.

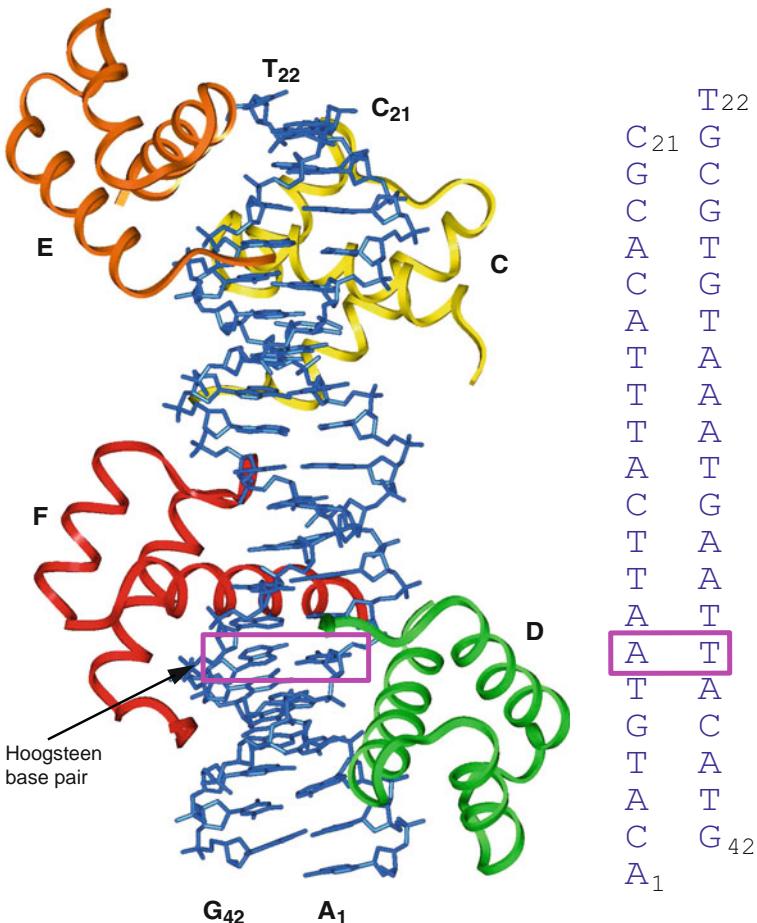


Figure 7.2. Unusual DNA/protein complex in which a Hoogsteen A–T base pair is found in unbent DNA. The complex is a high-resolution crystal of four MAT $\alpha$ 2 homeodomains bound to DNA [12, 13].

### Mismatches and Wobbles

Other bp arrangements involve *anti/syn* bps between two purines (a *mismatch*), in which one base adopts an atypical *syn* glycosyl configuration. This type of pairing can occur due to protonation of one base (favored at acidic conditions) or keto-enol tautomerizations. (See introduction to the term in Chapter 1, subsection on the discovery of DNA structure, and Chapter 5, subsection on nitrogenous bases).

Such ionized bases reverse the polarity of the hydrogen-bonding sites and hence lead to various hydrogen-bonding schemes, such as between  $G^+ \cdot C$  (Hoogsteen) or the ionized mispair  $A^+ \cdot C$  (*wobble*). Wobble pairs are formed when one base

is shifted in the bp plane relative to its partner due to steric misalignment of the bases. Experimental evidence for bp wobbling has come from tRNA structural analysis.

Thus, various non-WC base-pairing schemes can occur when bases are chemically modified, found in the rare enol and imino tautomeric forms, or mismatched, for example. The alternative arrangements for mispaired bases, in particular, can in turn lead to mutations upon DNA replication if not corrected by the elaborate repair machinery of the cell.

### Other Patterns

For further details regarding these and many other base-pairing schemes in nucleic acids, see [1191] and [737] for an early encyclopedic compendium of the “bewildering” variety of observed canonical and non-canonical nucleic acid arrangements, as compiled from RNA base-pairing interactions. This variety led to many further works and a proposal for new nomenclature for RNAs [738] which stimulated a large international initiative termed the **RNA Ontology Consortium** (ROC) [735]. ROC aims to establish a standard vocabulary for studying RNA by creating a common system to describe RNA sequences, 2D, and 3D structures and by developing software tools that merge sequence and structural databases for RNA for use by the general scientific community. See ROC website (<http://roc.bgsu.edu/>) for details.

In general, non WC bps are generally lower in stability because they typically incur a greater distortion in the sugar/phosphate backbone (assuming that the rest of the DNA is in a canonical A, B, Z-type conformation). Though they are more difficult to accommodate in general, many local alterations in sequence composition, environmental factors, and conformational patterns favor these alternative base-pairing arrangements.

#### 7.2.2 Hybrid Helical/Nonhelical Forms

##### Alternative Helical Geometries

Numerous variations are noted in the helical structures of polynucleotides (see [139, 893, 1080], for example). For example, C-DNA, D-DNA, and T-DNA helices have been defined with 9.3, 8.5, and 8.0 bps/turn, respectively. The C-DNA motif forms in fibers of low humidity (57–66%), D-DNA is observed in poly(dA)·poly(dT) strands, and T-DNA has been noted in bacteriophages that contain modified C residues and glucose derivatives for the sugar attachments.

Within polynucleotide structures, local distortions can also produce *hairpins* (DNA and RNA strands that fold back on themselves), *cruciforms* (intrastrand hydrogen bonds between complementary bases, often leaving a few unpaired bases at the hairpin tip), including *bulges* and *loops* (unpaired extrusions) (Figure 7.4). Such motifs are especially important in stabilizing the folding of single-stranded RNA.

### DNA Triplexes and Quadruplexes

Though Linus Pauling was ridiculed for his early suggestion of a triple helix structure for DNA [975], we now recognize the existence or the possibility for formation of various DNA triplexes (involving WC and Hoogsteen bps) [421, 1008]. Other existing or possible forms are quadruplexes (stabilized by hydrogen bonding and cations as in guanine-quartets; see Figure 7.1) [961, 968, 998, 1160, 1183, 1319, 1320, 1337]; parallel DNA (requiring reverse WC base pairing); DNA analogues; and hybrids of RNA, DNA, and other polymers such as oligonucleotide mimics (see below) [712].

Triplex forming oligonucleotides are promising pharmaceutical targets since they modulate gene activity *in vivo* through recognition of duplex DNA [905]. Triplexes and other unusual oligonucleotides are good subjects for simulation techniques since the factors that govern sequence-dependent properties can be explored [784].

Triple-helical DNA gained wide attention in the mid-1980s when various groups demonstrated that homopyrimidine (polynucleotide chains containing pyrimidines, such as poly(dT) or poly(dCT)) and some purine-rich oligomers (such as poly(dA) or poly(dAG)) can form stable and sequence-specific triple-stranded complexes with corresponding sites on duplex DNA; the third strand associates with the duplex via Hoogsteen pairing (see Figure 7.1) [421]. Homopurine/homopyrimidine stretches in supercoiled DNA were also found to adopt an unusual structure which includes a triplex called H-DNA.

Four-stranded DNA structures called G-quadruplexes are believed to play important functional roles in genome maintenance through their stabilization of chromosome ends or *telomeres*. Likened to plastic caps firming shoe-lace ends (by Stu Borman, C&EN writer), telomeres are protein/DNA assemblies that confine chromosome ends. Such protection from the possible fusing of chromosomes into one another helps maintain the integrity of the genome in living cells.

Telomere structure is of great biomedical importance since aging and cancer can be associated with malfunctioning telomeres. Thus, drugs targeted to telomeres can affect telomere function in DNA synthesis and transcription.

Telomeric DNA contains short guanine-rich repeats. At the very end of telomeres, G-quadruplexes are found, elaborately folded stacks of G-quartets (also known as GGGG tetrads) — the planar arrays of hydrogen-bonded guanines as shown in Figure 7.1.

Direct evidence that quadruplexes exist in human cells was provided by Hurley and coworkers [1183], who reported a chair-shaped quadruplex in a purine-rich strand of DNA in a promoter region of the oncogene *c-Myc* activated in cancer cells. Their work also highlights the potential role of quadruplex-targeted compounds as agents to combat disease, for example through control of oncogene expression to induce tumor shrinkage [601].

The architecture of G-quartets in G-quadruplexes remains an area of intense study. The planar G-quartets, as shown in Figure 7.1, can stack vertically with monovalent ions ( $\text{Na}^+$  or  $\text{K}^+$ ) sandwiched in between the layers. Wang and

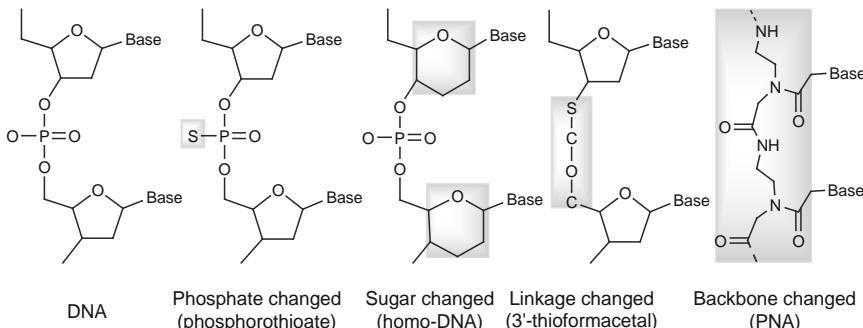


Figure 7.3. Various oligonucleotide analogues that involve modifications of the DNA phosphate unit, sugar, entire phosphodiester link, or the entire sugar/phosphate backbone.

Patel [1337] revealed in 1993 an NMR structure of a G-quadruplex in *sodium* solution. The sequence studied, with three repeats of TTAGGG, has a topology of G-quartets connected by loops (the TTA sequences that join each GGG series to the next) at the top and bottom of the quadruplex rather than the sides (see original article for figures). An exciting 2002 crystal structure by the Neidle group [961] of the same single-stranded sequence motif (specifically, TTAGGG<sub>2</sub> and TTAGGG<sub>4</sub>) in the presence of *potassium* ions reveals a radically different arrangement, with propeller-like connections through the sides of the stack (double-chain reversal topology [968]) rather than top and bottom (or one diagonal and two edgewise loops [968]).

Since mammalian cells are likely to have quadruplexes coordinated by potassium ions (since K<sup>+</sup> concentration exceeds that of Na<sup>+</sup> by an order of magnitude), this propeller-form quadruplex structure is of great importance and has clinical implications as well [961]. The topologies of quadruplexes and their interactions with other DNA, RNA, and proteins continue to define an active research frontier with significant potential for biomedical applications, for example in cancer therapy. See [1319, 1320] for recent computational studies.

### DNA Mimics

A growing number of DNA analogues (or ‘DNA mimics’) is also appearing (see Figure 7.3) [905]. Analogues can be sought in several ways (see Figure 7.3 and Box 7.1):

- by modifying the *phosphate backbone* segment (e.g., one P–O link changed to P–S or P–CH<sub>3</sub>);
- by modifying the *sugar* (e.g., 6-membered instead of 5-membered ring);
- by modifying the entire *phosphodiester linkage* of 4 atoms, O<sub>3'</sub> through C<sub>5'</sub> (to increase DNA hybrid stability through charge neutrality); or
- by replacing the entire *sugar/phosphate backbone*, retaining just the bases as in PNA, peptide (or polyamide) nucleic acids.

Such close analogues of oligonucleotides are designed to bind selectively to DNA bps through triple strands or duplexes. Various duplex DNA/RNA hybrids are key elements in transcription and replication and thus are potential therapeutic agents in anti-sense techniques. Triplex structures are also potential agents for sequence-specific cutting of DNA and drugs (by attacking various disease targets at the genetic level). Non-complementary or ‘anti-sense’ binding to a sequence of complementary messenger RNA or single-stranded DNA can also be exploited in antisense technology to suppress gene expression.

To be suitable, DNA analogues must also be biologically stable (i.e., resistant to nuclease digestion) and have favorable cellular absorption properties. All issues of geometry, flexibility, hydrophobicity, and triplex stability must be weighed in designing DNA mimics with practical utility.

### **Box 7.1: PNA (Peptide Nucleic Acid)**

PNA is an especially interesting oligonucleotide analogue that forms very stable complexes with double-stranded DNA [228, 350, 351, 905, 907, 1161]. In studies of evolution, PNA has been proposed as a candidate for the backbone of the first genetic material that predates the “RNA world”. While RNA is unstable and difficult to synthesize, the polymeric constituent of PNA, N-(2-aminoethyl)glycine (or AEG), may spontaneously polymerize and is accessible in prebiotic syntheses. No firm evidence, however, exists regarding this hypothesis [662].

The sugar/phosphate backbone of oligonucleotides is replaced in PNA by a pseudopeptide chain of AEG units with the N-acetic acids of the bases (N9 for purines and N1 for pyrimidines) linked via amide bonds (Figure 7.3). This substitution of the DNA backbone makes the PNA backbone electrically neutral rather than negatively charged; though the same number of chemical bonds as found in DNA and RNA exist in PNA (albeit different types), the replacement of the sugar ring by a linear bond sequence introduces additional torsional flexibility.

Stable hybrids between PNA and complementary DNA or RNA oligonucleotides form in a sequence-selective manner. Binding of PNA to double-stranded DNA leads not only to triple helices but also to P-loops strand-displacement complexes [905]. Model building, molecular mechanics and dynamics calculations, as well as experimental studies have shown how the replacement of DNA by PNA disrupts the canonical DNA triple helix, by displacing the hydrogen-bonded bases away from the global helical axis with respect to their positions in B-DNA helices, and changes the groove structure substantially [1216]. This makes PNA/DNA hybrids (involving both single and double helical DNA) A-like rather than an intermediate between A and B-DNA forms, as is typical for canonical DNA triplets. Helix stabilization in PNA/DNA hybrids stems from favorable base pairing, stacking, and (reduced) electrostatic interactions.

PNA molecules have already found practical applications as probes — an alternative to conventional DNA probes — for the detection of bacteria like *Salmonella enterica* or *Staphylococcus aureus*; chemical advantages stem from PNA’s strong resistance to degradation by proteases and nucleases in the cell and its salt-independent hybridization kinetics.

Triplexes and DNA analogues like PNA also have potential as drugs in gene therapy [905]. In particular, PNA is quite attractive for drug development, given its high triplex stability. However, given the A-like form of PNA/DNA hybrids, the design of stable B-like hybrids requires a different chemical approach than that used to construct PNA.

---

### 7.2.3 Unusual Forms: Overstretched and Understretched DNA

#### Single-Molecule Manipulations

Recently, a new experimental achievement of single-molecule observation and manipulation experiments using laser tweezers or single glass fibers [187, 1196] has been applied to DNA [682, 1049], re-invigorating interest in DNA's structural versatility [120, 182, 188, 250, 990, 1203, 1233]. Such experiments can also be used to determine DNA force constants, such as torsional rigidity [182]. It has been found that DNA subjected to previously unattainable forces of 10–150 picoNewtons (pN) displays a highly cooperative, sharp, and reversible transition to a new DNA structure at around 70 pN. This new overstretched helical structure, termed the *S-DNA ladder*, has been suggested to have a rise per residue of around 5.8 Å (compare to values in Table 5.3 of Chapter 5) and a notable inclination (possibly as large as 70°) [671], but structural details remain controversial. The transition under a force of extension is also affected by changes in the ionic strength of the medium, sequence, and addition of intercalators. Interestingly, DNA can sustain extensions to roughly *twice* its original length without severe distortions to the base pairing.

Modeling studies soon followed these experiments to investigate global conformational features of overstretched DNA duplexes [671, 711] and the local morphological features associated with the deformation process [678] (see Box 7.2).

Single-force spectroscopy has also been applied to study the mechanical properties of nucleosome arrays of the chromatin fiber (see Section 6.5), to suggest chromatin organization and to examine the effect of linker histones on fiber compaction [682].

Similar experiments have also been applied to RNA [759, 771, 1452] and proteins [1453].

#### Biological Relevance and Other Applications

The extreme deformations of DNA are of great interest because flexible DNA molecules undergo a variety of distortions in biological environments. Notable examples occur during DNA recombination and cell division. It has been suggested, for example, that a helix to ladder transition in DNA near the chromosomal centromere region may occur during cell division and thus play a regulatory role [671]. Intriguingly, the overstretched DNA in recombination filaments of 5.1 Å per bp is close to that associated with the phase transition observed in single-molecular micro-manipulation experiments.

It has also been suggested that longitudinal deformations of DNA might be associated with many DNA/protein binding events, such as DNA binding to the TATA-box binding protein TBP (where DNA is locally compressed, strongly bent, and unwound) and DNA binding to nucleosomes [1050], where variable lengths of DNA associated with nucleosome wrapping might accommodate sequence and ionic variations [678].

More generally, single-molecule biochemistry experiments help investigate the energetics and dynamics of folding and unfolding, the stability of various 2D and 3D intermediates, conformational landscapes, and folding pathways.

### **Box 7.2: Modeling Overstretched DNA**

Results of modeling studies of overstretched DNA — by energy minimization or molecular dynamics — are protocol dependent. For example, results and interpretations depend on which end of the DNA is pulled, what minimization method is used, how minimization is implemented (e.g., how fine the helical rise increments are), what force field is employed, and how solvent is treated.

The minimization work of Lavery and coworkers [711] suggested that the stretched conformation may be a flat, unwound duplex or a narrow fiber with substantial bp inclination [711]. The molecular dynamics simulations of Konrad and Bolonick [671] reproduced the helix to ladder transition and analyzed the geometric and energetic properties stabilizing the *S-ladder*. The constrained minimization studies of Olson, Zhurkin and coworkers [678] produced a wealth of structural analyses for both compressed and stretched DNA duplexes (from 2 to 7 Å per bp) of poly(dA)·poly(dT) and poly(dG)·poly(dC) homopolymers under high and low salt conditions. It was found that DNA can stretch to about double, and compress to half, its length before the internal energy rises sharply. Energy profiles spanning four families of right-handed structures revealed that DNA extension/compression deformations can be related to concerted changes in rise, twist, roll, and slide parameters. The lowest energy configurations correspond to canonical A and B-DNA. These models may be relevant to nucleoprotein filaments between bacterial Rec-A-like proteins and overstretched, undertwisted DNA [349].

Such fascinating single-molecule biochemistry manipulations have also shown that under much smaller forces (<3 pN), a new DNA phase is achieved, with about 2.6 bps/turn and thus about 75% more extended than B-DNA [21]. This new DNA conformation, termed *P-DNA* (for Pauling, see below), occurs at moderate positive supercoiling for molecules that cannot relieve the torsional stress via writhing (nonplanar bending). (For negative supercoiling, DNA denatures in this force range). Intriguingly, some modeling suggests that such a structure is ‘inside-out’, with bases on the helical exterior and the sugar/phosphate backbone at the center, as Linus Pauling once suggested for the structure of the double helix based on a model with un-ionized bases [975]. Such a conformational transition arrives from major rotation of torsion angles (notably  $\alpha$  and  $\gamma$ , toward the *trans* conformation) and is compatible with both C2'-endo and C3'-endo sugar pucksers. The intrastrand P-P distances in P-DNA (around 7.5 Å) are larger than the corresponding values in canonical A and B-DNA (5.8 and 6.6 Å, respectively). Evidence shows that this P-DNA conformation is found in the packed DNA inside some virus complexes, where

the DNA is constrained by the helical coat protein [289, 775]. *Still, the interpretation of positively-supercoiled, overstretched DNA as an ‘inside-out’ model remains controversial.*

Polymer statistical-mechanics calculations suggest that the experimental data involved in these force-versus-extension measurements of DNA can be fit to the elastic theory of the entropic force required to extend a worm-like chain [190]. However, such studies of entropic force are only relevant to small fluctuations about the B-DNA equilibrium conformation.

---

## 7.3 RNA Structure and Function

### 7.3.1 *DNA’s Cousin Shines*

While proteins are household words and DNA is an icon, in science as well as art (for the latter, see [1112] for an overview), their biomolecular cousin, RNA, was largely left behind until recently. Indeed, RNA’s starring role in the cell has emerged with new discoveries concerning RNA’s vital regulatory roles, as introduced in Chapter 1. Namely, our appreciation for RNA has heightened with discoveries that RNA molecules are integral components of the cellular machinery for protein synthesis and transport, RNA editing, chromosome replication and regulation, catalysis, and many other functions (see Table 7.1 for some of RNA’s diverse roles).

At a 2006 symposium organized by NCI at Cold Spring Harbor, Susan Gottesman suggested on a presentation slide: “*Anything that DNA can do, RNA can do better*”. At the same symposium, Gary Ruvkun announced 23 challenges facing scientists concerning regulatory RNAs. James Watson simply summed the atmosphere: “*This is a revolution*”. Indeed, we are now discovering that biological and synthetic RNAs prepared in the laboratory perform a broad range of functions and have numerous applications. This comes with the discovery not only of regulatory RNAs but also of many synthetic RNAs developed from *in vitro* selection experiments that have significantly expanded our knowledge of RNA’s repertoire.

### 7.3.2 *RNA Chains Fold Upon Themselves*

Many of the base pairing variations described for DNA in the prior sections are common for RNA, the single-stranded polynucleotide chain, which can fold upon itself to form double-stranded segments interspersed with loops. Such patterns are governed by the primary sequence and can accommodate a variety of hydrogen bonding patterns. The double stranded regions (stems) can be imperfect with bulges, mismatched pairs, and unusual hydrogen bond schemes, as shown in Figure 7.4. The stem/loop structures themselves can fold further, seeking

Table 7.1. Some classes of non-coding RNA (ncRNA).

RNA	Function
transfer RNA (tRNA)	protein synthesis
ribosomal RNA (rRNA)	protein synthesis
Signal recognition particle (SRP)	protein recognition
small nucleolar RNA (snoRNA)	rRNA modification
micro RNA (miRNA)	translation regulation
transfer-messenger RNA (tmRNA)	protein stability in ribosome
telomerase RNA	replication
guide RNA (gRNA)	mRNA editing
spliced leader RNA (SL RNA)	mRNA trans-splicing
small nuclear RNA (snRNA)	RNA splicing
hairpin, hammerhead, and HDV ribozymes	self-cleavage
Group I intron	self-splicing
Group II intron	self-splicing
RNase P	pre-tRNA processing
23S rRNA	peptide bond formation
G, A, glmS, TPP and other riboswitches	gene regulation

favorable stacking, hydrogen bonding, and other interactions between distant partners. See [1113] for an introduction to RNA structure and associated biological problems.

The multitude of RNA secondary structures and tertiary interactions are stabilized by various double-stranded regions, interior and terminal loops, bulges, K-turns, U-turns, S-turns, A-platforms, tetraloops, and other motifs (e.g., [102, 545, 654, 736, 739, 740, 875]).

RNA *pseudoknot* motifs, which can be considered supersecondary structural elements, are produced from a special intertwining of secondary structural elements by Watson-Crick base pairing (see Figure 7.5). Namely, pseudoknots form when a consecutive single-stranded domain with segments **a**, **a'**, **b**, **b'**, **c**, **c'** and **d** (**a'**, **b'**, and **c'** are connectors) folds to form two hydrogen bonds: **a** with **c**, and **b** with **d**. This is a pseudoknot rather than a knot since the strands do not actually pass through one another [1367]. See Figure 7.6 for examples of RNAs with pseudoknots.

### 7.3.3 RNA's Diversity

The wonderful capacity of RNA to form complex, stable tertiary structures has been exploited by evolution. RNA molecules are integral components of the cellular machinery for protein synthesis and transport, RNA editing, chromosome

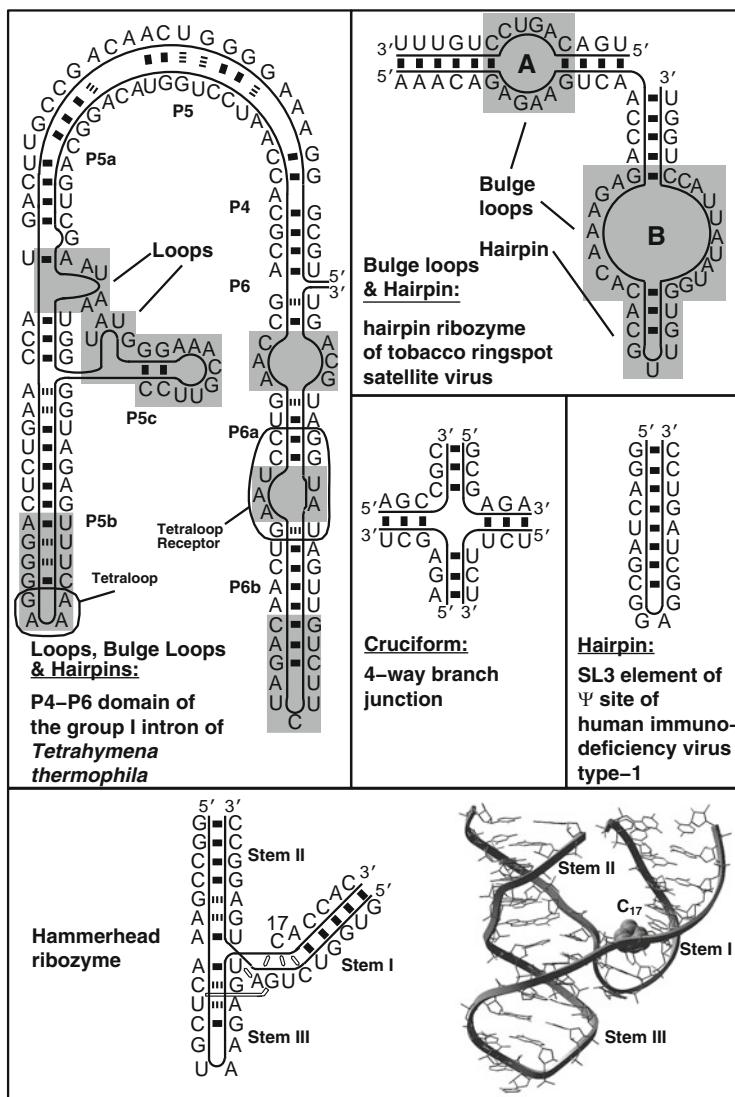


Figure 7.4. Various nucleotide-chain folding motifs from solved RNA systems: loops, bulge loops, hairpins, hairpin loops, cruciforms (four-way junctions, found in recombination intermediates and synthetic designs), and the complex tertiary contacts found in RNA (illustrated on the hammerhead ribozyme), which compact the secondary-structure elements. Broken connections are used for non-WC bps, including mismatches. The PDB files for the structures (clockwise from left) are as follows: tetrahymena intron, 1GID; tobacco ringspot ribozyme, 1B36; HIV-1  $\Psi$ -stem loop 3, 1A1T; and hammerhead ribozyme, 1MME. The tertiary interactions shown for the hammerhead drawing (ellipses or connecting line segments) are based on [1154]. The nucleotide C17 shown in space-filling form is the strand cleavage site of the ribozyme.

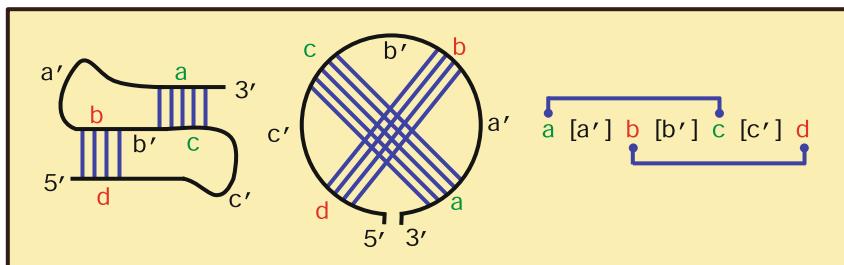


Figure 7.5. RNA pseudoknots have an intertwined form of base pairing, which can be evident from a circular representation of base pairing.

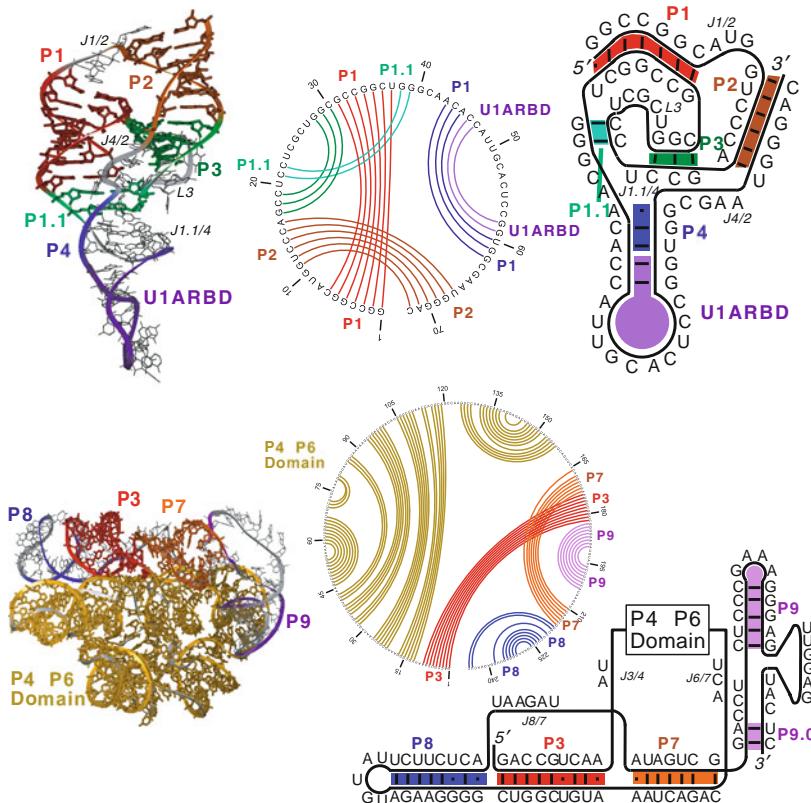
replication and regulation, catalysis and many other functions<sup>2</sup>. Indeed, some scientists hypothesize that life was based on RNA ('RNA World') before the evolution of the modern nucleoprotein universe [454, 906]. RNA/DNA hybrid double helices also occur in transcription of RNA onto DNA templates and in the initiation of DNA replication by short RNA segments.

As mentioned in Chapter 1, a highlight of RNA's functional diversity is its catalytic capability, as discovered in the 1980s and recognized in the 1989 Nobel Prize in Chemistry to Thomas Cech and Sidney Altman. Such catalysis is performed by RNA molecules termed *ribozymes*; hundreds of ribozyme types are now known in a diverse range of organisms (e.g., [1142, 1208, 1251, 1275], and some have been designed in the laboratory by *in vitro* selection (e.g., [1228, 1245]), as spare in composition as two base building-block units (rather than four) [1042].<sup>3</sup> Many ribozymes make or break phosphodiester bonds in nucleic acid backbones, but other biological and chemical functions are continuously being discovered.

RNA's conformational flexibility, modularity, and versatility [493] are important for recognition by, and interactions with, ligands (e.g., as in the RNA of the human immunodeficiency virus, HIV) and other molecules. These features are of practical importance for the design of therapeutic agents that exploit RNA's functional sites as potential drug targets (e.g., [76, 211, 539, 546, 725, 981, 1268]). The construction of ligands that bind RNA and interfere with protein synthesis, transcription, or viral replication have potential as new therapeutic agents, such as antibiotic/antiviral drugs [984]. These pharmaceuticals are urgent given the resurgence of previously known viral and bacterial diseases and the emergence of resistant mutants of common antibiotic and antiviral drugs.

<sup>2</sup>For example, tRNA molecules carry amino acids and deposit them in correct order, mRNAs translate hereditary information from DNA into protein, rRNA are involved in protein biosynthesis (within a complex of ribosomal RNA and numerous proteins), gRNAs edit RNA messages (by 'guide' sequences), cRNAs play roles in catalysis and autocatalysis, and snRNAs are splicing agents (by 'small nuclear' components of the mRNA splicing machinery known as the spliceosome).

<sup>3</sup>The 83-nucleotide ribozyme composed only of two different building blocks — uracil and 2,6-diaminopurine — was shown to catalyze the ligation of two RNA molecules with a rate 36,000 times faster than the uncatalyzed reaction [1042]. The fact that RNA's genetic code may be simpler than today's four bases lends further support to the "RNA world" hypothesis.



**Figure 7.6. Ribozymes with pseudoknots.** **Top:** Hepatitis Delta virus (HDV) ribozyme (PDB code 1DRZ) shows two pseudoknots, with pseudoknot basepairing pattern drawn in the circular diagram with basepair number along the perimeter: a main one formed by regions P1 (red) and P2 (orange-red), and a minor one formed by regions P1.1 (cyan) and P3 (green). A hairpin formed by the P4 helix (blue) and the U1A ribonucleoprotein binding domain (purple) is also present. **Bottom:** P3-P9 domain of the group I intron *Tetrahymena thermophila* ribozyme (PDB code 1GRZ) shows one pseudoknot formed by regions P3 (red) and P7 (orange-red). The P4-P6 domain (yellow) folds independently. See Box 7.3 for a discussion of some of those structures. The intertwined or pseudoknotted regions are clearly seen by the crossings in the circular diagrams.

These applications — together with design of novel RNAs and of RNA sequences called *aptamers* that bind specific molecular targets (e.g., [166, 199, 724]) — have important potential as regulators of gene expression, therapeutic agents, molecular switches, and molecular sensors<sup>4</sup>. See [546, 889, 1207, 1208, 1275, 1381], for some examples.

<sup>4</sup>RNA sensors are RNA molecules that are turned on or off when they contact a target, like small organic molecules, and can trigger a chemical reaction.

Designing new RNAs based on this idea of aptamers has led to the exciting experimental field of *in vitro* selection. This experimental technology involves generating and screening large random-sequence libraries of nucleic acid molecules for a specific function, such as binding or catalysis. Numerous nucleic acid molecules binding targets (aptamers) have been developed, as diverse as organic molecules, antibiotics, proteins, and whole viruses. In addition, new classes of RNA enzymes (ribozymes) have been discovered by this approach, with applications in biomolecular engineering, such as in the design of allosteric ribozymes and aptamer-based biosensors, aptamers capable of inhibiting protein function, and therapeutic aptamers (e.g., inhibiting the TAR RNA element of HIV-1). Other emerging applications of engineered RNAs include RNA nanotechnology, where RNAs are assembled into functional arrays, and RNA synthetic biology, where designed RNAs are used to control cellular functions (e.g., regulate gene expression). See Subsection 7.5.2 on analysis of random nucleotide pools to better understand the relationship between sequence and structure/function and development of a computational procedure for mimicking aspects of this selection approach *in silico*.

#### 7.3.4 Non-Coding and Micro-RNAs

Such newly found roles for RNAs, especially concerning tiny RNAs that do not encode proteins (ncRNA for non-coding RNAs) but can influence gene action won DNA's cousin the venerable trophy of "Breakthrough of The Year" by *Science* editors in 2002 (see the 20 December 2002 issue of *Science*, volume 298). Non-protein coding stretches of mRNAs range in size from only 20 nucleotides to over 10000 nucleotides [1228]; they are required to control the translation from the mRNA transcript into protein.

The 2002 award recognized a large group of papers that unraveled various fascinating features of small RNAs (affectionately termed nanoRNAs). Micro-RNAs (miRNAs), generally 21 to 25 nucleotides long, form a regulatory class of ncRNAs [200]. These RNAs control gene expression by repressing translation of target genes through, for example, binding to 3' untranslated regions of the messenger-RNA targets.

Such small RNAs in animals, plants, and fungi collectively became associated with the title RNAi (for RNA interference). The agents that initiate RNAi in a sequence-specific manner are double-stranded RNA segments that silence gene expression (siRNAs, for small interfering RNAs). For example, they may seek out the messenger RNA and destroy it, or they may bind to chromatin and/or modify chromatin structure [383, 1004, 1432]. Though initially regarded as anomalies, work is revealing that such siRNAs regulate gene expression in a variety of organisms.

Such interference mechanisms by RNA silencing can provide an organism a natural defense against invading viruses and transposons (DNA segments that migrate within and across organisms and are associated with bacterial pathogenicity) [11]. Consequently, this natural protection is being exploited by scientists using

siRNAs to target viral genes that can inhibit the replication of HIV-1, polio, or other viruses (e.g., [757]). Moreover, RNA interference mechanisms are being investigated by companies who apply them to discover the functions of genes by turning them off to determine the effect on the plant or the animal. For example, a landmark study on obesity employed RNA interference [626] to inactivate about 85% of the roundworm's predicted 19,757 genes that code for proteins in a single experiment [66]. More recent studies have used RNA interference to uncover genes involved in regulating the immune response to pathogenic bacterial infections [273] and genes regulating programmed cell death (or *apoptosis*) [229].

These fascinating discoveries regarding RNA's interference with gene activity are associated with many *epigenetic* phenomena — changes in gene expression (inheritance of features) that do not involve alterations in the genome and persist across at least one generation [31]. A different kind of epigenetic control was also discovered by Breaker, Nudler, and coworkers [863, 864, 889, 926, 1275, 1381] in some bacterial messenger RNAs containing sequences that sense small molecules directly to control translation of mRNA into protein. Namely, specific control regions of mRNA can bind directly to metabolites associated with vitamin B synthesis and import, and induce a conformational change in RNA's folding state; this metabolite-triggered conformational change acts as part of the signal transduction pathway that senses vitamin level and controls enzyme production. Such RNAs that control gene expression upon binding metabolites or other ligands have been termed *riboswitches*.

A riboswitch can undergo local and global conformational changes upon binding to its substrate. The secondary and tertiary structures of these natural aptamers can be very diverse, and several structures have been solved [873]. Examples include the type-I glmS ribozyme and purine and SAM-II riboswitches in which one binding region induces local conformational changes upon binding. TPP, SAM-I, and the M-box magnesium are examples of type II riboswitches because of their two regions which undergo conformational changes upon binding their metabolite, making possible global as well as local conformational rearrangements (see Figure 7.7).

Such a switch of RNA conformation between two states in a ligand-dependent manner (see Figure 7.7) also opens new avenues for thinking about RNA design in a variety of contexts (e.g., [166, 199, 242, 599, 1028, 1245]). Like the Paracelsus challenge for proteins (see Chapter 2), one can formulate a similar challenge for RNA design — describe minimal changes in the nucleotide sequence to trigger a conformational rearrangement in the folding of a given RNA molecule — that can be approached by a combination of computational and experimental wizardry. Indeed, Science editor Jennifer Couzin [267] writes: “Having exposed RNAs’ hidden talents, scientists now hope to put them to work”.

### 7.3.5 RNA at Atomic Resolution

Until fairly recently, less was known at atomic resolution on RNA structure in comparison to DNA, but this has changed as the ‘RNA era’ has begun.

Progress in RNA structure elucidation can be attributed to vast improvements in crystallization procedures (e.g., RNA structure determination through crystallization with a protein that would not interfere with the enzyme's activity [391]), as well as alternative approaches for studying RNAs such as high-resolution NMR, spectroscopy, crosslinking reactions, and phylogenetic data analysis. Our knowledge of RNA structures has also increased dramatically with recent solutions of ribosomes [85, 201, 206, 210, 910, 922, 1135, 1374, 1379, 1427], since ribosomes contain numerous tertiary motifs of RNAs and therefore provide a rich resource of information on RNA structural elements and organization.

The clover-leaf structure of the tRNA molecule has been known for decades, and for a long time was the only well-characterized major structure of an RNA molecule [1173]. Its structure whet our appetite for RNA appreciation by revealing the long-distance tertiary interactions. By now, RNA folds characterized by X-ray crystallography include the tRNA, hammerhead ribozyme, *Tetrahymena* group I intron, hepatitis delta virus ribozyme, Group I intron from Azoarcus and Twort, Group II intron, various ribozymes (e.g., hairpin, GlnS, Diels-Alder), various riboswitches (e.g., purine, M-box, and TPP), RNase P types A and B, signal recognition particle, and various fragments such as kissing hairpin and sarcin/ricin motifs. See some examples in Figure 7.7 and details of some solved catalytic RNAs in Box 7.3.

### Box 7.3: RNAs at Atomic Resolution

The *hammerhead* [1005, 1282] and *Hepatitis Delta Helper virus* (HDV) [391] ribozymes, in the family of self-cleaving catalytic RNA, were solved in 1994 and 1998, respectively. The hammerhead's Y, or wishbone-shaped, structure has three base-paired stems resembling the head and handle of a carpenter's hammer (see Figure 7.4). The RNA is unpaired in its U-turn core and stabilized by non-WC, non-wobble bps in the stems. Visualizing its structure led to further analysis of the mechanism of RNA self cleavage through trapping of intermediates in the ribozyme reaction pathway [331, 886]. While the hammerhead's active site is open, that of HDV is hidden (see Figure 7.6), resembling the catalytic sites of globular proteins, and contains two pseudoknots.

The crystal structure of one self-folding *Tetrahymena thermophila* Group I intron [488], was solved in 2004. Its secondary structure consists of nine regions (P1-P9), which fold into two major domains: P3-P9 and P4-P6. Both domains are stabilized through extensive inter-domain tertiary interactions. The P3-P9 domain is formed by a coaxial helix between P3 and P8, with a slight bend with respect to P7, while helix P9 is oriented perpendicular to P7 (see Figure 7.6). The P4-P6 domain has a helical *tetraloop* region connected to another helical segment, the *tetraloop receptor*, by a large bend ( $\approx 150^\circ$ ) (see Figure 7.4). Thus, its active site is hidden and exemplifies the complexity of RNA structure, involving complex intertwining of secondary structure elements, including pseudoknots. Indeed, both WC and non-canonical base-paired regions are interspersed with internal loops, and an *adenosine-rich bulge* region mediates the long-range tertiary contacts in the RNA, improving base stacking interactions. The *adenosine platform* motif emerges from this structure as an important architectural component of RNA that might have arisen early

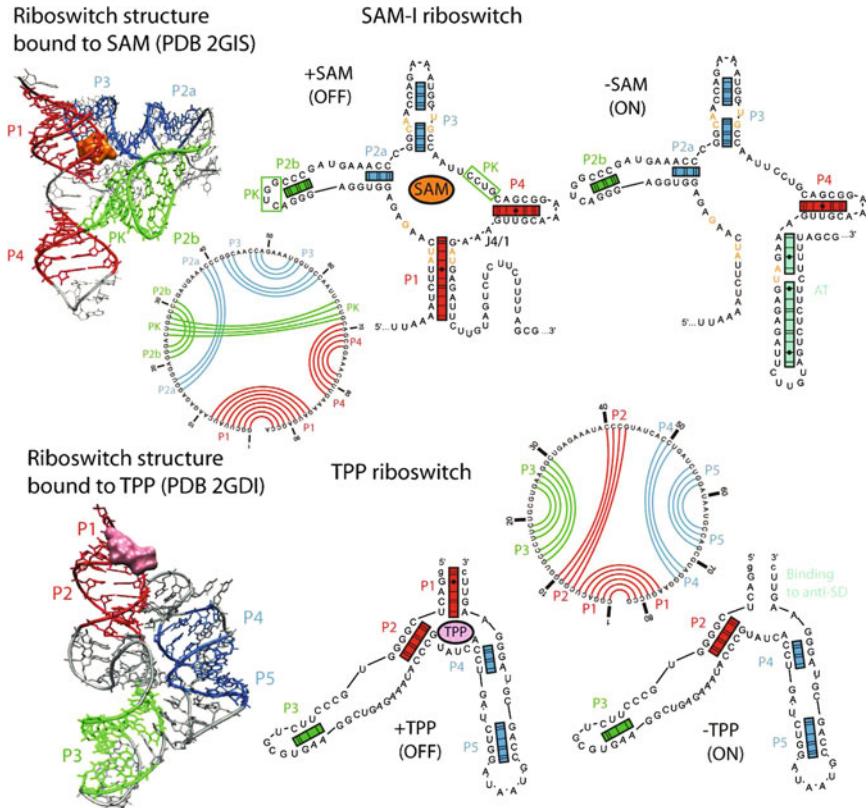


Figure 7.7. Examples of two riboswitches: SAM-I and TPP riboswitch. In the presence of its substrate, S-adenosylmethionine (SAM), the SAM riboswitch is composed of a four-way junction forming a pseudoknot and a pair of coaxial helices P1–P4 and P2a–P3. The junction is stable due to hydrogen bonding between nucleotides (orange) in helices P1 and P3 which bind to SAM. The bound state of the riboswitch was solved by crystallography [872]. In the absence of SAM, the 3'-side of helix P1 forms an anti-terminator element (AT), allowing the mRNA to be fully transcribed. The TPP riboswitch The thiamine pyrophosphate (TPP) riboswitch from *Escherichia coli* thiM mRNA also has two forms depending on the presence of TPP. The bound-state form has been solved by X-ray crystallography [1166]. Without TPP, the 3'-side of helix P1 binds to an anti-ShineDalgarno (anti-SD) region, allowing the ribosome to access the SD element so that translation can start as proposed by the Breaker group [1381].

in evolution. As found in other RNA structures, metal-phosphate coordination is also important in stabilizing tertiary contacts in this intron RNA domain. In addition to the Group I intron of the *Tetrahymena thermophila*, the Group I intron of the *Azoarcus* [3] and *Twort* [466] species are also available.

The complete *Group II intron* solved in 2008 [1266] is a self-splicing ribozyme that catalyze their own excision from precursor mRNAs and joins together flanking exons without the help of proteins. The secondary structure is characterized by six domains (I to VI) composed of step-loop structures connected by a common central core and stabilized by long-range tertiary interactions. Domain I is the largest element and contains recognition sequences responsible for the correct assemble of the intron in its active form. Domains II and III enhance the catalytic efficiency of the splicing, while Domain IV contains the open reading frame (ORF) for expression of a reverse-transcriptase enzyme. Domain V is the most phylogenetically conserved element and contains a hairpin loop involved in catalysis. Domain VI contains an adenosine that interacts with the splice site during splicing. The Group II intron and the spliceosome share common structural and functional features, supporting the hypothesis of a common ancestor [1266].

---

## 7.4 Current Challenges in RNA Modeling

### 7.4.1 RNA Folding

Deducing the functional structure of RNA molecules from the primary sequence has been called the *RNA folding problem* [172, 241, 771, 1261]. The challenge is to understand how the strong electrostatic repulsions between closely packed phosphates in RNA are alleviated. Indeed, the stability of compact RNA forms is strongly maintained through interactions with both monovalent and divalent cations and by pseudoknotting.

As mentioned in connection to the multitude of hydrogen bonding patterns for polynucleotides in Section 7.2.1, the rich variety of possibilities in RNA, as reviewed in [736, 739, 740], has led to proposal for new nomenclature for RNAs and the annotation of RNA structure through the international **RNA Ontology Consortium**, ROC (<http://roc.bgsu.edu/>) [735].

### 7.4.2 RNA Motifs

We now recognize at least seven major tertiary interaction motifs for RNA, as shown in Figure 7.8 — subdivided into three classes. These classes separate interactions between two double stranded helices, between a single strand and a helix, and between two single stranded regions. A recent annotation study of 54 representative high-resolution solved RNA structures showed the dominance of A-minor motifs (37%), coaxial helices (32%), and ribose zippers (20%), which together account for 89% of the total motifs (which number 613) [1403] (see Figure 7.9). Correlations among motifs, such as a pseudoknot or coaxial helix with A-minor, reveal patterns of higher organization and underscore RNA's hierarchical structure (Figure 7.9). Analyses of RNA junctions of orders 4 through 10 further suggested global patterns and subnetworks that organize RNA in complex

ways as well as a classification into distinct families (Figure 7.10) [690, 691]. Such studies can help in the goal of RNA structure prediction by defining major motifs in RNA and pinpointing specific sequence/structure relations.

### 7.4.3 RNA Structure Prediction

Predicting the secondary and tertiary folding of RNA is a difficult and ongoing enterprise [172, 225, 241, 277, 689, 1022, 1145, 1465]. Secondary structure elements are easier to identify through free energy minimization combined with comparative analysis (sequence alignment) using evolutionary and database relationships [838, 839, 1460, 1461]. The minimized objective function is empirically derived based on basic physiochemical laws. However, caution is warranted in the

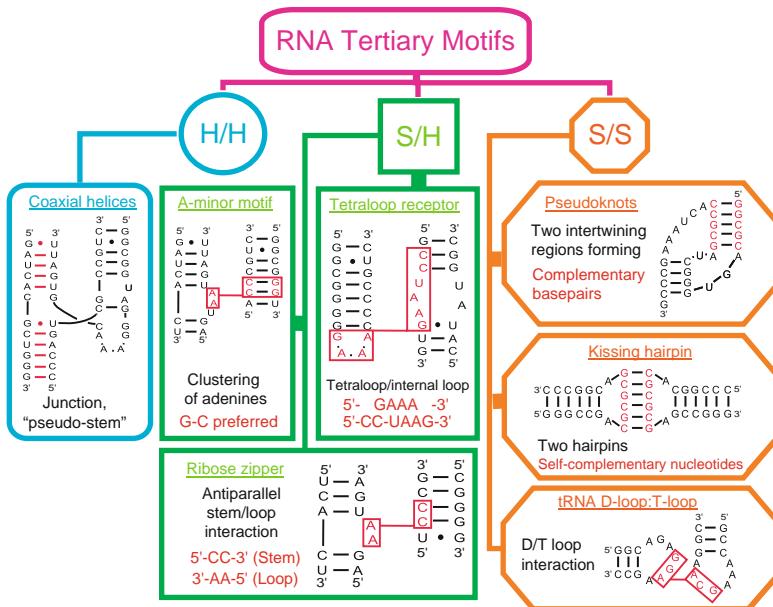


Figure 7.8. Seven major motifs of tertiary interactions in RNA, involving two double stranded helices (H/H), single strand and a helix (S/H), and two single stranded regions (S/S). *Coaxial helices* form when nucleotide bases from two separate helices stack and align axes to form a pseudo-continuous coaxial helix. *A-minor* motifs originate from clustering of adenosine interactions, often present within other motifs such as coaxial helices. *Tetraloop receptor* is an interaction between a tetraloop (GNRA, where N = A,G,C,U and R = A,G) and an internal loop plus two GC bps. *Ribose zippers* form when two consecutive residues from one chain segment interact in an antiparallel fashion with two consecutive residues from another chain segment distant in sequence but close in space. *Kissing hairpins* form by base pairing between single-stranded residues of two hairpins with complementary sequences, and *tRNA D-loop:T-loop* are interactions between two conserved hairpins in tRNA. See glossary of [1403] for more details and references to original motif definitions.

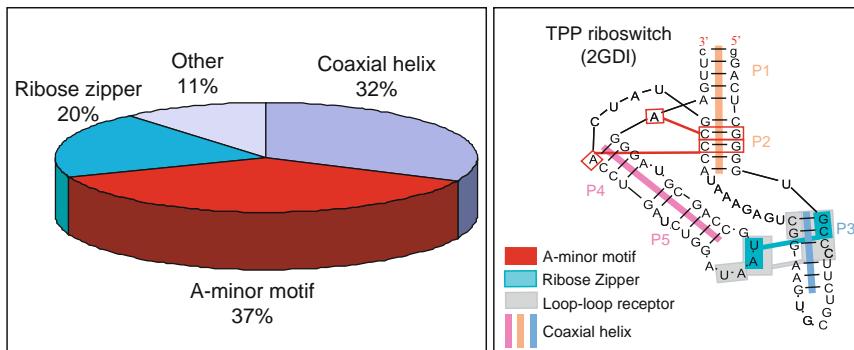


Figure 7.9. The distribution of RNA tertiary motifs in a non-redundant set of 54 high-resolution crystal structures and annotated diagram of the TPP riboswitch (PDB 2GDI) showing correlated motifs [1403].

interpretations of such 2D structures since predictions are imperfect, especially for more than 200 nucleotides. The new structures, however, provide opportunities for learning what works, as well as what fails, in structure prediction. Thus, discriminating among the possible tertiary interactions to obtain the final folded state remains a challenge [277, 1115, 1261]. Still, findings concerning the folding kinetics of *Tetrahymena* ribozyme [1149] have suggested that, as thermodynamic data on tertiary structure interactions become available [1261], the RNA folding problem might be easier to solve than protein folding [101, 1261]. Therefore, with advances in RNA synthesis and structure determination [558] and the availability of thermodynamic data on tertiary interactions [1261], it is likely that our understanding of RNA structure, RNA folding, and RNA's role in enzyme evolution will dramatically increase in the coming decade. These developments are propelling studies in RNA informatics (or *ribonomics*) [332] and RNA design (e.g., [166, 199, 241, 242, 359, 599, 600, 689, 1115, 1142, 1245, 1245, 1251, 1281]).

Emerging themes in RNA structure include the importance of metal ions and loops for structural stability, hidden active sites in some ribozymes like catalytic sites of proteins, various groove binding motifs (e.g., adenines interacting with minor-groove helices to stabilize tertiary contacts, as deduced from crystal structures [327, 911] and statistical analyses [1383]), architectural motifs tailored for intermolecular interactions [546], hierarchical folding, fast establishment of 2D elements, and extreme flexibility of the molecule as a whole [172, 277, 390, 493, 1261]. See recent review [1465].

Some key challenges concerning RNA include finding novel RNA genes, identifying the biological roles of these RNA genes, determining the structural repertoire of RNA, determining RNA tertiary folds from sequence, and designing novel RNAs [1112]. Bioinformatics tools hold great promise in addressing these challenges.

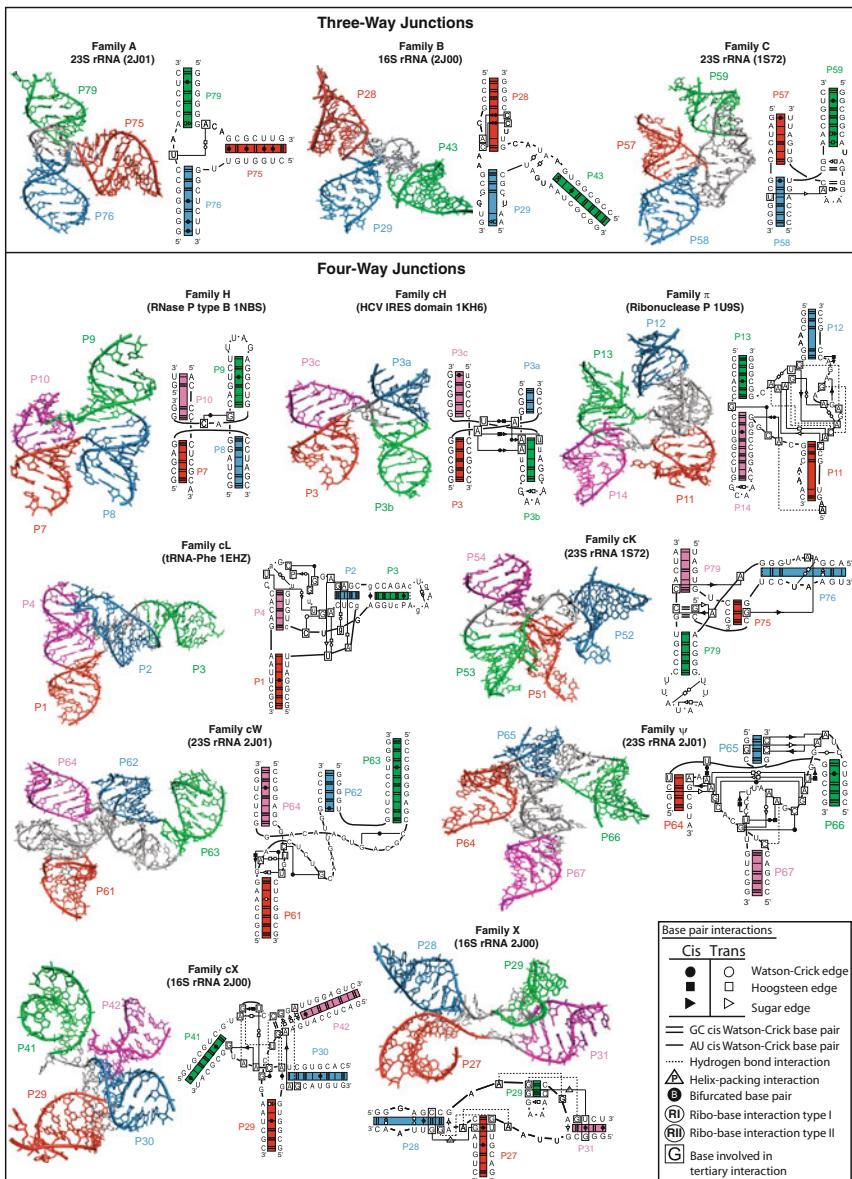
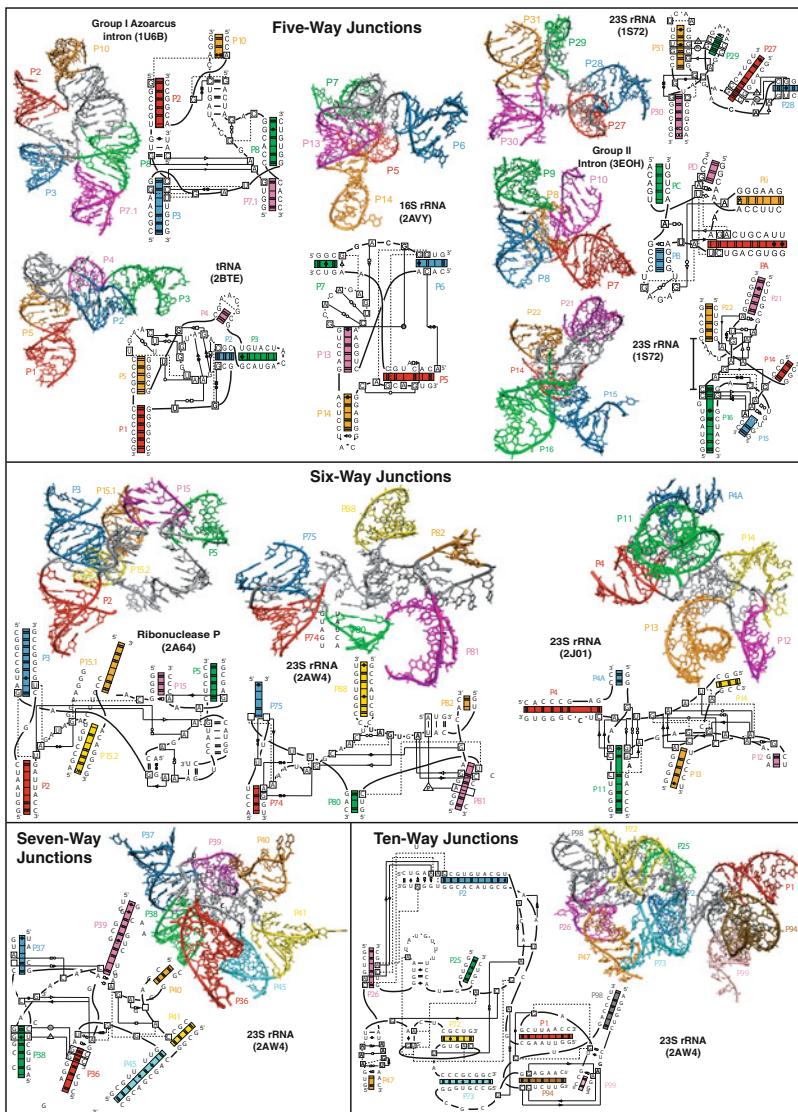


Figure 7.10. Classification of RNA junctions into families according to coaxial stacking properties, perpendicular helical configurations, and flexible helical arms. Lines inside the helices represent the canonical WC bps GC, AU, and the GU wobble bp. The network symbology follows the Leontis Westhof notation [738].



## Classification of RNA junctions (continued).

## 7.5 Application of Graph Theory to Studies of RNA Structure and Function

### 7.5.1 Graph Theory

The application of graph theory to describe secondary structure motifs of RNA, as pioneered by Waterman in 1978 and extended by others [116, 438, 708, 1350],

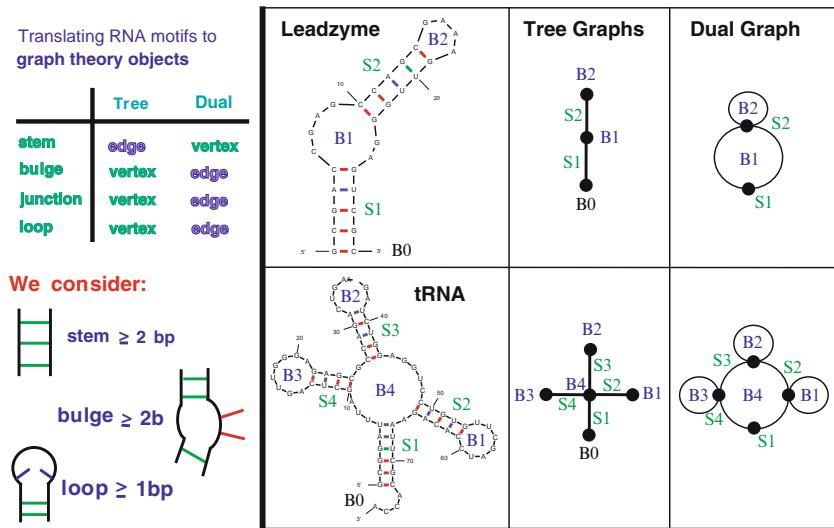


Figure 7.11. Graphic Representations of RNA Secondary Structures shown for two RNAs, using Tree and Dual Graphs, using the definitions at right [388,438].

holds promise in the area of RNA structure analysis. Graph theory is suitable for RNA because this field of mathematics is widely used for analyzing networks and enumerating structural possibilities, including chemical structures, genetic and biochemical networks, monetary transactions, terrorist organizations, social networks, and the Internet [33,900]; see also the 24 July 24 2009 issue of *Science*, volume 325, devoted to networks.

Graph theory can be used to represent RNA secondary structures as two types of graphical objects: *trees* and *dual* graphs according to specified rules for defining RNA stems, bulges, junctions, and loops (Fig. 7.11). All possible RNA 2D structure motifs, as simplified by this approach, can then be cataloged by the graph vertex number, as well as ranked by topological complexity according to the second-smallest eigenvalue of the Laplacian matrix of the graph [388,438].

### 7.5.2 RNA-As-Graphs (RAG) Resource

One such RNA topology resource, RAG (RNA-As-Graphs) (<http://monod.biomath.nyu.edu/rna>), is being used to classify/analyze topological characteristics of existing RNAs [438,645] (see Fig 7.12 for graph enumeration segments from RAG) and to design and predict novel RNA motifs [643–646]. See also some RNA reviews commenting on graph theory applications to RNA [542,736].

Specifically, RAG and the associated graph theory framework for RNA [388, 435] have been used to classify and catalog RNA motifs [438, 645], estimate the size of RNA's structural/functional repertoire [645], detect structural

and functional similarity among existing RNAs [965], identify RNA motifs of antibiotic-binding aptamers (found synthetically) in genomes [702–704], analyze the structural diversity of random pools used for *in vitro* selection of RNAs [450], simulate aspects of the process of *in vitro* selection *in silico* [643, 644, 646], and analyze RNA thermodynamics landscapes to better understand riboswitch mechanisms to ultimately enhance their design [1028].

Other applications of RAG in the community which include graph theory extensions involve classification and prediction of ncRNAs [503, 633, 801, 901, 902, 1181], various RNA structure analyses [50, 78, 148, 170, 533, 534, 538, 982, 1041], and various applications of graph theory [82, 416, 470, 471, 538, 761, 1324, 1431]. Examples of graph extensions include labeled dual graphs [633] and directed tree graphs [503]. Our application of spectral theory to catalog RNA graphs has also been extended to other biological and physical systems [82, 416, 470, 471, 761, 1324, 1431]. See Kim’s thesis for a more detailed descriptions on these applications [642].

### RNA Structure Enumeration

Cataloging based on graph theory enumeration suggests that the RNA structure universe is dominated (more than 90%) by pseudoknots, in agreement with available data [645], as also discussed in [78, 170, 982, 1041]. Significantly, the existing RNA classes represent only a small subset of possible 2D RNA motifs as enumerated by graph index; some of these motifs may be natural while others may be possible to generate in the laboratory. Still, others may not exist.

### RNA-Like Motifs

The usage of clustering techniques to separate graphs that are ‘RNA-like’ from those that do not resemble natural RNAs also led to predictions of many new RNA-like motifs, including ten specific examples of sequences that might lead to novel-like RNA topologies [645], as shown in Figure 7.13. Some of these motifs predicted in 2004 have since been solved: C1 in mammalian CPEB3 ribozyme [1084], C2 in a purine riboswitch [819], C3 in the tymovirus/Pomovirus tRNA-like 3' UTR element [840], C4 in the tombusvirus 3' UTR region IV [1463], and C7 in the flavivirus DB element [235]. Significantly, the predicted and actual sequences have between 45 to 51% homology.

Graph theory tools are also natural for comparing RNA structures to find existing RNA motifs within large RNAs based on graph isomorphisms [965]. This idea was applied to identify topological similarities among existing RNA classes and to define motifs of RNA within larger RNA topologies for major RNA classes (e.g., tRNA, tmRNA, hepatitis delta virus RNA, 5S, 16S, 23S rRNAs).

Furthermore, the representation of RNAs as graphs led to identification of RNA motifs in genomes. Since natural aptamers exist in many bacterial genomes and other organisms, it appeared likely that natural counterparts of synthetic motifs exist *in vivo*. This led to development of an efficient search tool for identifying small RNA motifs in genomes by exploiting many artificial motifs derived from

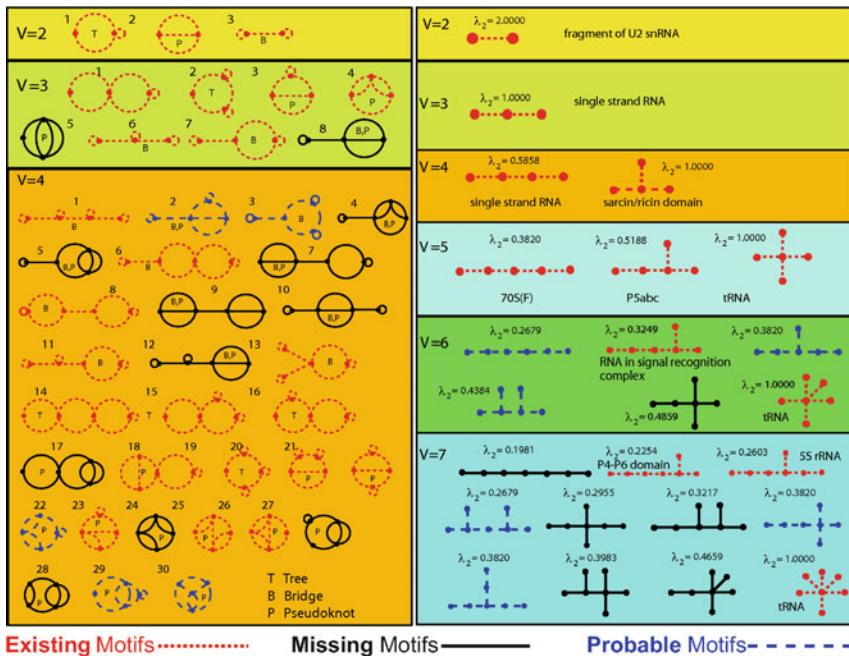


Figure 7.12. Segments from RAG showing some enumeration of graphs for tree and dual segments of low V numbers. The graphs are coded according to motifs found in nature (red), motifs not yet found that are RNA-like (blue), as determined by clustering analysis, and remaining motifs (black) [645]. See RAG website for details and updates.

RNA *in vitro* selection experiments [702–704]. The search for antibiotic-binding aptamers produced 37 candidate sequences from bacterial and archaeal genomes.

### RNA Design

Finally, graph theory tools can also advance the design of novel RNAs by mimicking the process of *in vitro* selection *in silico*. As a first step, understanding the structural diversity of random pools was important for improving *in vitro* technology. It is simple to generate random pools computationally, but graph theory allowed rapid analysis of the resulting 2D motifs using 2D folding algorithms [450]. By characterizing the distribution of secondary structure motifs in computer-generated random pools using sets of possible RNA tree structures from graph theory [450], we showed that random pools do not have a uniform distribution of possible topologies and instead favor simple topological motifs. This was expected from experimental observations that typically yield simple motifs. Further studies also found that the proportion of multiply branched structures increases with sequence length, in agreement with other studies.

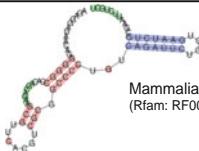
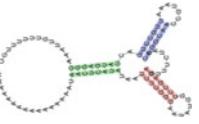
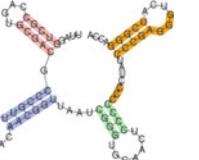
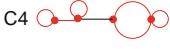
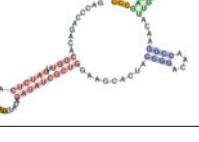
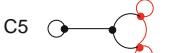
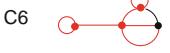
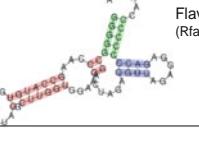
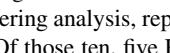
Graph Representation With Natural Submotif	RNA Secondary Structure With Natural Submotif	Candidates Discovered After 2004
C1	 single strand RNA (NDB:PR0055)	 Mammalian CPEB3 Ribozyme (Rfam: RF00622)
C2	 bulged hairpin (Rfam:CopA)	 Purine Riboswitch (Rfam: RF00167)
C3	 DsrA RNA (Rfam:DsrA)	 Tymovirus tRNA-like 3' UTR element (Rfam: RF00233)
C4	 bulged hairpin (Rfam:CopA) single strand RNA (NDB:PR0055)	 Tombivirus 3' UTR region IV (Rfam: RF00176)
C5	 bulged hairpin (Rfam:CopA)	
C6	 DsrA RNA (Rfam:DsrA)	
C7	 single strand RNA (NDB:PR0055)	 Flavivirus DB element (Rfam: RF00525)
C8	 single strand RNA (NDB:PR0037)	
C9	 DsrA RNA (Rfam:DsrA)	
C10		

Figure 7.13. Ten candidates predicted to have RNA-like topologies, as determined by clustering analysis, represented as dual graphs; the red submotif occurs in natural RNAs [645]. Of those ten, five RNAs were discovered since the published predictions, as shown in the third column.

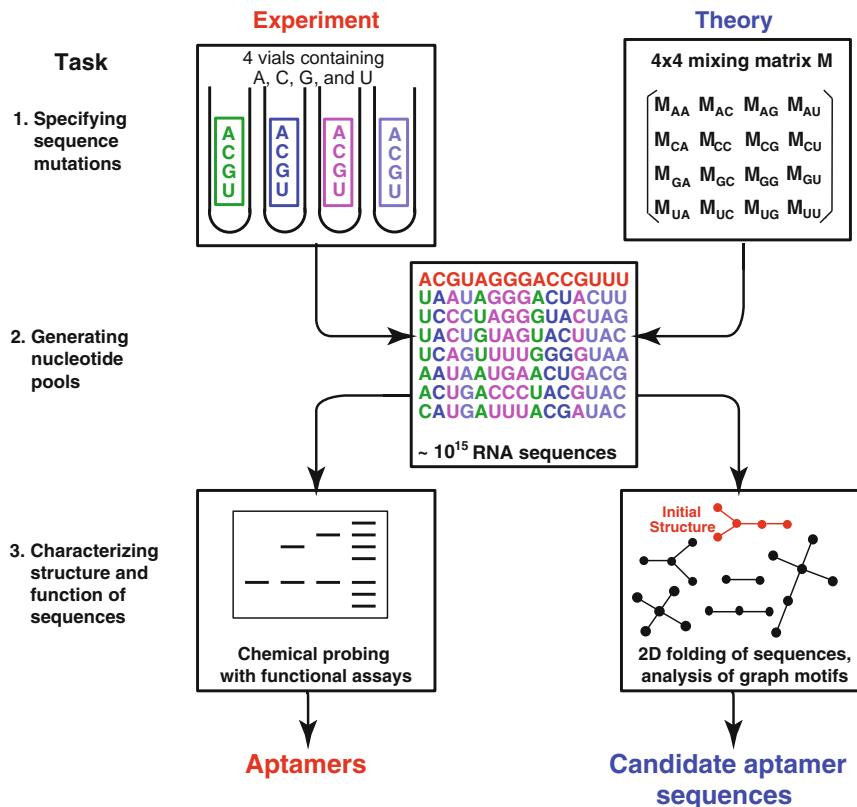


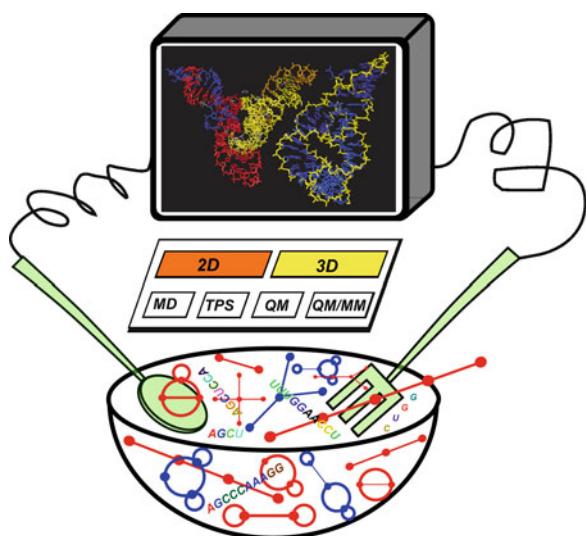
Figure 7.14. Experimental RNA pool synthesis and an *in silico* approach to mimic the process [643, 646]. Pool generation can be simulated using the mixing matrix, which specifies nucleotide mixtures in synthesis port or mutation rates for all nucleotide bases and can be defined to mimic specific biological situations [646]. The resulting sequences are “folded” into 2D structures using existing algorithms and analyzed further to screen and filter the candidates.

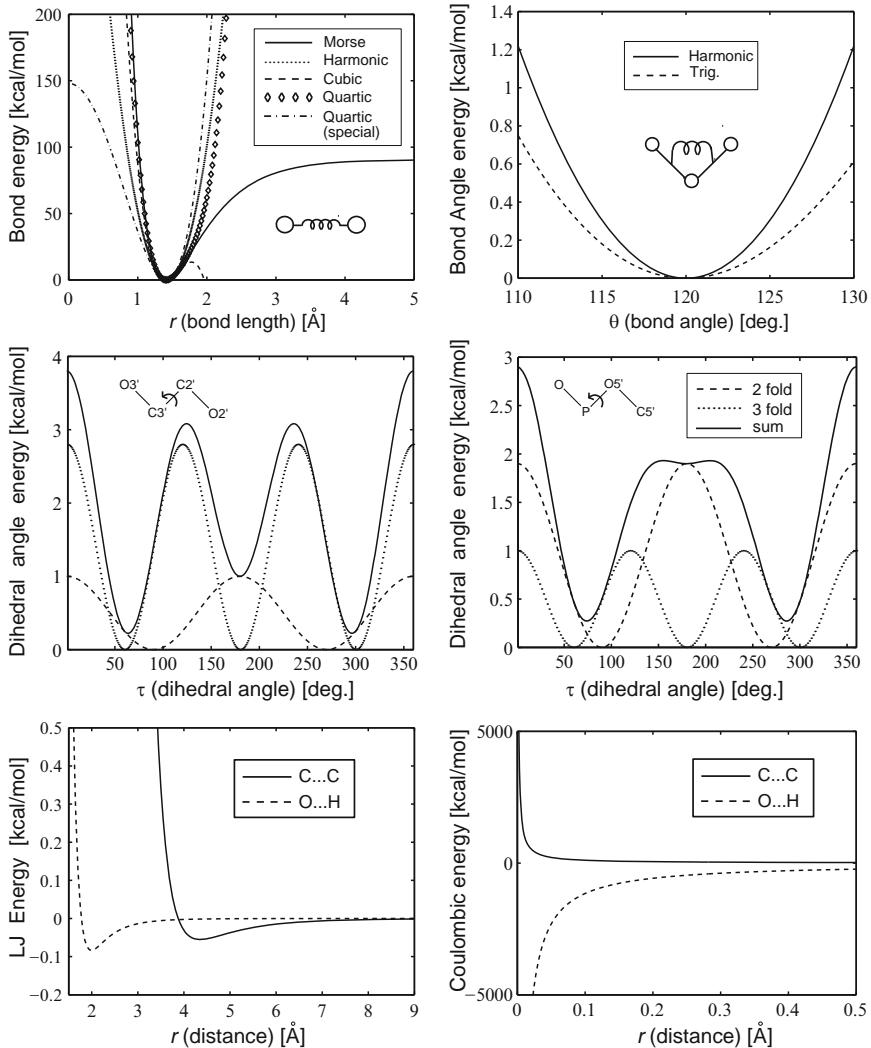
Mimicking the experimental process *in silico* can be done on the basis of a “mixing matrix” framework (see Fig. 7.14), which led to development of RAGPOOLS, a publicly-available web server for pool design [643]. Such mixing matrices specify the mixing ratios of nucleotides in synthesis ports (see Fig. 7.14). When applied to a starting sequence, each mixing matrix will generate a pool of specific sequences via mutations of the given sequence. Thus, different design strategies (based on covariance mutations like AU to CG or conversion of AU to CG base pairs) produce different sequence pools, and the main idea is to design these pools to yield the desired structure and/or function.

Using this new tool and other computational and experimental resources, very large pools of nucleotides (up to  $10^{14}$ ) can be generated, screened, and filtered according to various 2D-structure similarity and flanking sequence

analyses [644]. Such computational and theoretical yields agree for simple RNA motifs. For real aptamer targets, the *in silico* procedure overestimates the yields found experimentally, as expected, because experimental yields represent lower bounds and the screening does not yet involve 3D structural aspects.

Advances on all these fronts are ongoing.





# 8

## Theoretical and Computational Approaches to Biomolecular Structure

### Chapter 8 Notation

SYMBOL	DEFINITION
<b>Vectors</b>	
$\mathbf{a}, \mathbf{b}, \mathbf{c}$	interatomic distance vectors (in definition of $\tau$ )
$\mathbf{n}_{ab}, \mathbf{n}_{bc}$	unit normals
$\mathbf{r}_{ij}$	interatomic distance vector ( $\mathbf{x}_j - \mathbf{x}_i$ )
$\mathbf{x}_i$	position vector of vector $i$ , components $x_{i1}, x_{i2}, x_{i3}$
$\tilde{P}$	collective momentum (for nuclei and electrons)
$\tilde{X}$	collective position (for nuclei and electrons)
$X$	collective position (nuclei only), components $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathcal{R}^{3N}$
$V$	collective velocity (nuclei only)
$\psi_n$	eigenfunctions of the Hamiltonian operator
<b>Scalars &amp; Functions</b>	
$b_i$	bond length $i$
$q_i$	Coulomb partial charge for atom $i$
$r_{ij}$	interatomic distance (between atoms $i$ and $j$ )
$t_0$	initial time
$A_{ij}, B_{ij}$	Lennard-Jones coefficients
$\hat{H}$	Hamiltonian operator ( $E_k + E_p$ )
$E_{\text{bond}}$	bond length energy
$E_{\text{bang}}$	bond angle energy
$E_{\text{coul}}$	Coulomb energy
$E_k$	kinetic energy
$E_{\text{local}}$	local (short-range) energy
$E_n$	eigenvalues of the Hamiltonian operator (quantum states)
$E_{\text{nonlocal}}$	nonlocal (long-range) energy
$E_p$	potential energy (also $E$ )

Chapter 8 Notation Table (continued)

SYMBOL	DEFINITION
$E_{\text{tor}}$	torsional (dihedral) angle energy
$E_{\text{LJ}}$	Lennard-Jones energy
$K_{ijk}$	bond angle force constant
$M_F$	dimension of the Fock matrix
$N$	number of atoms
$N_e$	number of electrons
$\mathcal{R}^n$	Euclidean space of dimension $n$
$S_{ij}$	bond length force constant
$S_B$	set of bonds
$S_{BA}$	set of bond angles
$S_{DA}$	set of dihedral angles
$S_{NB}$	set of atom pairs computed in the nonbonded energy
T	temperature
$V_n$	torsional barrier height of periodicity $n$
$\epsilon$	dielectric function
$\theta_i$ or $\theta_{ijk}$	bond angle (e.g., $\theta_{dha}$ for donor–hydrogen···acceptor sequence in a hydrogen bond)
$\tau_i$ or $\tau_{ijkl}$	dihedral or torsion angle
$\Delta t$	timestep

Science says the first word on everything, and the last word on nothing.

Victor Hugo (1802–1885).

## 8.1 The Merging of Theory and Experiment

### 8.1.1 Exciting Times for Computationalists!

Computational techniques for exploring three-dimensional (3D) structures of nucleic acids and proteins are now well recognized as invaluable tools for revealing details of molecular conformation, motion, and associated biological function. Over a decade ago, the theoretical chemist Henry Schaefer declared: “*It is clear that theoretical chemistry has entered a new stage . . . with the goal of being no less than full partner with experiment*” [1092].

Commenting on the two 1998 Chemistry Nobel Prize awardees in quantum chemistry — Walter Kohn of the University of California, Santa Barbara, and John Pople of Northwestern University — a reporter in *The Economist* wrote: “*In the real world, this could eventually mean that most chemical experiments are*

*conducted inside the silicon of chips instead of in the glassware of laboratories. Turn off that Bunsen burner; it will not be wanted these ten years*" (17 October 1998)!

It remains to be seen how effective "in silico biology" will be, but clearly a new enthusiasm is stirring in the molecular biophysics community in the wake of many methodological and technological improvements. The following categories reflect improvements that lend computation a stronger basis than ever before:

1. **Improvements in instrumental and experimental techniques** [1043]. Rapid advances in sequencing and mapping genomes are being made, as discussed in Chapter 1, making available an enormous amount of biological data requiring analysis [83]. Many procedures in biomolecular NMR and electron spectroscopy [361, 908, 1184], including 4D NMR [1433] and X-ray crystallography [110, 994, 1224], are accelerating the rate and improving both the accuracy and scope of structure determination; much of this progress was stimulated by the structural genomics initiatives. And newer microscopic techniques such as scanning tunneling and cryo [61, 189, 336, 361, 685, 1077, 1286] are being applied. With such advances, structures that were considered unconquerable a decade ago are being resolved — RNAs, nucleosomes, ribosomes, ion channels, and membrane proteins, for example — although limitations still remain for macromolecular complexes. Theoreticians can use these structures with other experimental data and biological information as solid bases for simulations and computer analyses.
2. **New models and algorithms for molecular simulations.** Improved force fields are being developed, some simpler (e.g., coarse-grained [1117]) and others more complex (e.g., polarizable, more variables [223]) that are sophisticated versions of early coarse-grained (e.g., [749]) or polarizable (e.g., [1344]) protein folding models. Innovative models for protein and DNA folding and dynamics and for macromolecular complexes are shedding insights into important biological processes. Faster and more advanced dynamic simulations of complex systems with full account of long-range solvation and ionic effects are being used. Quantum-mechanical and classical/quantum hybrid studies of biomolecules are becoming routine. And many enhanced sampling methods are making possible simulation of a wide range of conformational states of biomolecules, including computation of reaction rates [1116, 1117].
3. **The increasing speed and availability of supercomputers, parallel processors, and distributed computing.** Faster, cheaper, and smaller computing platforms and graphics workstations are entering the laboratory at reduced costs, though computer memory requirements are exploding. Supercomputing centers are making available very fast computing platforms, and specialty hardware like Anton [1169] hold promise for advancing macromolecular modeling and simulation.

4. **Successful multidisciplinary collaborations.** Notable examples are collaborations between mathematicians and biologists in relating knot theory with DNA topology and geometry [246, 613, 1370]; between computer scientists/engineers and biologists regarding DNA computing [245, pp. 26–38], [40, 117, 162, 779, 929, 1083, 1155]; between biological and mathematical/physical scientists in propelling biology-information technology, or bioinformatics; and between material scientists and biomedical scientists regarding nanomaterials for biological applications, creating the field of nanotechnology/bionanotechnology [442, 787, 904, 1059, 1202].<sup>1</sup> See [635], for example, for a description of some mathematical challenges in genomics and molecular biology and [252, 530] for biological challenges in the 21st century.
5. **The wealth of readily available Internet and web resources, sequence and structure databases, and highly-automated analysis tools.<sup>2</sup>**

### 8.1.2 *The Future of Biocomputations*

*“One of these days,”* believes *Nature*’s former editor John Maddox, “*somebody will begin a paper by saying, in effect, ‘Here is the Hamiltonian of the DNA molecule’ and will then, after a little algebra, explain just why it is that the start and stop codons of the genetic code have their precise functions, or how polymerase molecules work in the process of transcription*” [812].

Some of us might be somewhat skeptical that “a little algebra” will suffice to explain DNA’s functional secrets, but many remain confident that *a large amount of computation* with carefully developed models will bring us closer to that goal in the not-too-distant future.

### 8.1.3 *Chapter Overview*

In this chapter, we introduce molecular mechanics from its quantum-mechanical roots via the Born-Oppenheimer approximation. Following a brief overview of current quantum mechanical approaches, we discuss the three underlying

---

<sup>1</sup>Nanotechnology is an emerging science of creating functional materials, devices, and systems on the basis of matter at the nanometer scale, including macromolecules, and the exploitation of novel properties and phenomena on this scale for biomedicine, technology, and more. See, for example, general principles for a National Nanotechnology Initiative on [www.nano.gov](http://www.nano.gov); the 24 November 2000 issue of *Science*, volume 290, highlighting issues in nanotechnology; the September 2001 special issue of *Scientific American* devoted to ‘The Science of the Small’; and recent reviews on the application of nanomaterials to biology and medicine [442] for cell imaging, cell tracking, and cancer treatment (diagnosis and therapy) [904], as well as innovative synthetic DNA-based enzymes and aptamers [787].

<sup>2</sup>Caution is certainly warranted regarding the quality of some unreviewed online information. As stated in the New York Academy of Sciences Newsletter of Oct./Nov. 1996, “There may be debate about whether the explosive growth in electronic communication has made life better or worse, but there’s no question that it has made life faster”.

principles of molecular mechanics: the thermodynamic hypothesis, additivity, and transferability. We then describe the choices that must be made in formulating the potential energy function: configuration space, functional form, and energy parameters. We end by mentioning some of the current limitations in force fields. The next chapter discusses details of the force field form and origin.

## 8.2 Quantum Mechanics (QM) Foundations of Molecular Mechanics (MM)

Quantum mechanical methods are based on the solution of the Schrödinger equation [742, 978]. This fundamental approach is attractive since 3D structures, molecular energies, and many associated properties can be calculated on the basis of fundamental physical principles, namely electronic and nuclear structures of atoms and molecules. Indeed, the quantum mechanics pioneer Paul Dirac is believed to have expressed the sentiment that the Schrödinger equation reduces theoretical chemistry to applied mathematics [765]!

Although historically quantum calculations were practical only for very small systems, exciting developments in both software and hardware (computer speed as well as memory) have made quantum-mechanical calculations feasible for larger systems, including biomolecules, with various approximations (see [465], for example; more below). See the Nobel Lecture address by John Pople for field perspectives [1011].

The presentation of this important area of modeling is very brief here. For comprehensive treatments, see textbooks by Warshel and Cramer [270, 1342] and excellent reviews [573, 628, 1163–1165, 1348, 1442].

### 8.2.1 The Schrödinger Wave Equation

The Schrödinger wave equation describes the motions of the electrons and nuclei in a molecular system from first principles. This equation can be written as

$$\hat{H}\psi_n = E_n\psi_n, \quad (8.1)$$

where the Hamiltonian operator  $\hat{H}$  is the sum of the kinetic ( $E_k$ ) and potential ( $E_p$ ) energy of the system:

$$\hat{H}(\tilde{P}, \tilde{X}) = E_k(\tilde{P}) + E_p(\tilde{X}), \quad (8.2)$$

where  $\tilde{P}$  and  $\tilde{X}$  denote the collective momentum and position vectors for all the nuclei and electrons in the molecule. The potential energy  $E_p(\tilde{X})$  originates from electrostatic interactions among all the variables.

According to this description, the quantum states  $E_n$  (eigenvalues) form a discrete set, corresponding to the eigenfunctions  $\psi_n$ , for the system of electrons and nuclei. The Schrödinger equation thus describes the spatial *probability distributions* corresponding to the energy states in a *stationary* quantum system.

Traditional electronic structure methods — important more from a historical perspective — calculate these eigenstates associated with the discrete energy levels by diagonalization of the Hamiltonian matrix, of order of the number of basis functions. This cubic scaling is prohibitive for large systems.

### 8.2.2 The Born-Oppenheimer Approximation

In the Born-Oppenheimer approximation to the Schrödinger equation, the motions of the molecule are separated into two levels: electrons and nuclei.

First, only the electrons are considered as independent variables of  $\hat{H}$ , while positions of the nuclei are assumed fixed. This is generally a good approximation because the nuclei — much heavier than the electrons — are typically fixed on the timescale of electronic vibration. The resulting eigenvalues from this analysis of electron motion represent electronic energy levels of a molecule as a function of atomic coordinates. These energy states are known as Born-Oppenheimer energy surfaces (BOES).

In the second level of the Born-Oppenheimer approximation, the BOES of the electronic ground state ( $E_p(X)$ , where  $X$  is the collective position vector of the nuclei in the system), are used as the *potential* energy of the Hamiltonian (eq. (8.2)) instead of the Coulombic potential. The quantum mechanical behavior of the nuclei is then investigated.

In theory, quantum mechanical treatments should be the tool of choice for reliable description of complex chemical processes, since no experimental information is needed as input. Yet, an accurate analytic solution of the Schrödinger equation is not feasible except for small molecules. Thus the equation must be generally solved through standard approximations, which naturally reduce the accuracy obtained. The range of application is further limited by the computational complexity required by these techniques, but advances are rapidly occurring on this front.

Two basic quantum-mechanical approaches are used in practice: *ab initio* and *semi-empirical*. Both rely on the Born-Oppenheimer approximation that the nuclei remain fixed on the timescale of electronic motion, but different types of approximations are used. The former is more rigorous than the latter and hence more computationally demanding. For example, in Hartree-Fock type calculations, the computational requirements, including computer time and memory, scale as  $M_F^2$  to  $M_F^4$  where  $M_F$  is the dimension of the Fock matrix. This dimension corresponds to the size of the computational basis set used to approximate the wave functions, related to the number of electron orbitals,  $N_e$ .

### 8.2.3 Ab Initio QM

The name *ab initio* implies *non-empirical* solution of the time-independent Schrödinger equation, or solutions based on genuine theory. However, besides the Born-Oppenheimer approximation, relativistic effects are ignored, and the concept of molecular orbitals (or wave functions) is introduced.

In *ab initio* methods, molecular orbitals are approximated by a linear combination of atomic orbitals. These are defined for a certain basis set, often Gaussian functions. The coefficients describing this linear combination are calculated by a variational principle, that is, by minimizing the electronic energy of the molecular system for a given set of chosen orbitals.

This energy (known as Hartree Fock energy) and the associated coefficients are calculated iteratively by the Self Consistent Field (SCF) procedure with positions of the nuclei fixed. This calculation is expensive for large systems — as computation of many integrals is involved — and makes *ab initio* methods computationally demanding.

In practice, the basis set for the molecular wave functions is represented in computer programs by stored sets of exponents and coefficients. The associated calculated integrals are then used to formulate the Hamiltonian matrix on the basis of interactions between the wave function of pairs of atoms (off-diagonal elements) and each atom with itself (diagonal elements) via some potential that varies from method to method. An initial guess of molecular orbitals is then obtained, and the Schrödinger equation is solved explicitly for a minimum state of electronic energy.

The quality of the molecular orbitals used, and hence the accuracy of the calculated molecular properties, depends on the number of atomic orbitals and quality of the basis set. The electronic energy often ignores correlation between the motion of the electrons, but inclusion of some correlation effects can improve the quality of the *ab initio* results.

### Density Functional Theory (DFT)

DFT can be formulated as a variant of *ab initio* methods where exchange/correlation functionals are used to represent electron correlation energy [962]. DFT methods are based on the use of the electron density function as a basic descriptor of the electronic system. In the DFT Kohn-Sham formulation, the electronic wave function is represented by a single ground-state wave function in which the electron density is represented as the sum of squares of orbital densities.

DFT schemes differ by their treatment of exchange/correlation energy, but in general this class of methods offers a good combination of accuracy and computational requirements, especially for large systems. DFT methods are computationally more efficient than conventional *ab initio* methods that correct for electron correlations, but have a similar scaling complexity due to the  $N_e^3$  diagonalization cost of the Hamiltonian matrix. With linear-algebra advances, however, this standard diagonalization expense can be circumvented with approaches that yield *linear scaling* by localization of the electronic degrees of freedom (i.e., electron density, density matrix, and orbital calculations) [282, 465, 1148, 1409, for example] (see below).

### 8.2.4 *Semi-Empirical QM*

In semi-empirical methods, the matrix elements associated with the wavefunction interactions are not explicitly calculated via integrals but are instead constructed from a set of predetermined parameters. These parameters define the forms and energies of the atomic orbitals so as to yield reasonable agreement with experimental data. Thus, most integrals are neglected in semi-empirical methods, and empirical parameters are used as compensation. Although good parameterization is a challenging task in these approaches, and parameters are not automatically transferable from system to system, semi-empirical methods retain the flavor of quantum approaches (solution of the Schrödinger equation) and are less memory intensive than *ab initio* methods. As in *ab initio* methods, the quality of the results depends critically on the quality of the approximations made.

### 8.2.5 *Recent Advances in Quantum Mechanics*

Traditional quantum calculations are dominated by the cost of solving the electronic wave function expressed in terms of the Hartree-Fock operator. Solution of the wave functions via inversion of the Fock matrix for a minimal basis set (roughly the number of electrons,  $N_e$ ) requires  $\mathcal{O}(M_F^4)$  work. While still much better than classical orbital calculations that include correlations, this scaling limits applications to large systems. It has been noted, however, that by exploiting the sparsity of the Fock matrix — zero elements due to the rapid decay with distance of orbital overlapping — algorithms developed in linear algebra for banded systems can reduce the cost to linear scaling, i.e.,  $\mathcal{O}(M_F)$ , or to the near-linear cost of  $\mathcal{O}(M_F \log M_F)$  when Ewald forces are included [726, 951, 1414, 1445]. See also [464] for a physics-community perspective of electronic structure calculations (rather than the chemistry community).

#### Linear Scaling

Linear scaling algorithms are also possible by various divide-and-conquer algorithms that localize electronic calculations (for both *ab initio* and semi-empirical methods) [465, 1409] and by reformulating the Hamiltonian diagonalization as a minimization problem. The localization is achieved by using different treatments for the strong intramolecular interactions and the weak intermolecular interactions [755]. The reformulation as an optimization problem entails construction of an energy functional, which is minimized with respect to some variational parameters that represent the electronic ground state [282].

The nonlinear conjugate gradient method (see Chapter 11) exploits the matrix sparsity pattern and is very modest in both computational and storage requirements. See Daniels and Scuseria [282] for a comparison among several linear-scaling approaches for semi-empirical calculations, including pseudo-diagonalization of the Fock matrix via orthogonal rotations of the molecular

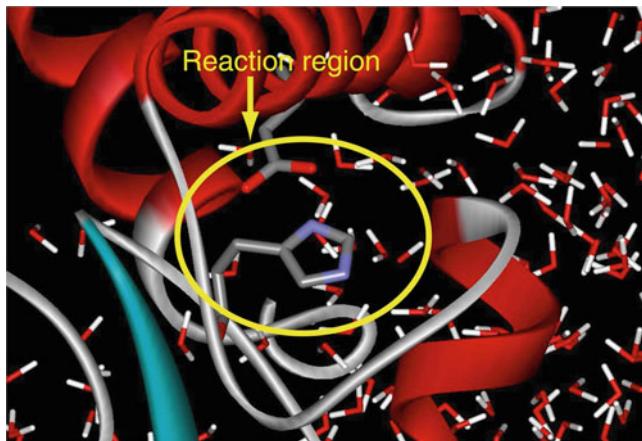


Figure 8.1. The concept of QM/MM methods, as introduced by Warshel & Levitt [1344], in which a limited part of the system (reaction region) is treated quantum mechanically, while the surrounding solvent and remaining biomolecular system is treated classically. Figure kindly provided by Arieh Warshel.

orbitals, conjugate-gradient density matrix search, and purification of the density matrix via transformations applied to a function expanded in terms of Chebyshev polynomials.

### Biomolecular Applications

Such advances in complexity reduction are making possible quantum calculations on biomolecules under various simplifications. Semi-empirical methods have been applied to protein, DNA, and RNA systems (e.g., [86, 638, 639, 754, 1415]), and *ab initio* approaches have been applied to smaller systems (e.g., DNA bases and base pairs [552, 755], cyclic peptides [756], and the reactive centers of enzymes [144, 1185]). Indeed, hybrid QM/MM (quantum mechanics/molecular mechanics) methods, as first introduced in 1976 [1344] (see Figure 8.1), are routinely used for studies of enzyme mechanisms and related aspects of protein dynamics [776, 1442]. See Figures 8.2 and 8.3 and Box 8.1 for examples.

As shown in Figure 8.1, QM/MM methods rely on dividing the biomolecular systems into two regions: a small region enclosing the active site that is treated quantum mechanically, and the remaining region (including solvent and biomolecule), which is modeled by classical molecular mechanics force fields. The key to such QM/MM methods is the coupling between the electric field from the surrounding and the QM Hamiltonian in the active-site region. This requires careful treatment of the boundary between the QM and MM regions, either by using hybrid orbitals for the connection or a linked atom approach.

Recent advances in the field have come from using higher-level *ab initio* QM Hamiltonians, as first proposed by Singh & Kollman [1194] or, alternatively, empirical versions, like the valence bond method (EVB) proposed by Warshel

and co-workers [54, 408, 1342]. However, using *ab initio* QM Hamiltonians has introduced new challenges of proper sampling and computing the associated free energy pathway. The calculation of free energies from QM/MM simulations can be performed by averaging over the system's configurations via perturbations from a reference surface; however, such sampling for accurate free energy evaluations as well as calculations of pK<sub>a</sub> values remain challenging and form an active area of research. See recent reviews in [441, 452, 572–574, 628, 1165, 1442].

---

### **Box 8.1: Illustrative QM and QM/MM Applications**

Semi-empirical quantum-mechanical applications have been used, for example, to study aqueous polarization effects on biological macromolecules, by comparing free energies in the solvated versus gas-phase states [1415]. Figure 8.2 from [638] illustrates the quantum-mechanically derived electrostatic potential of the polarized electron density of A, B, and Z-DNA relative to the gas phase density. The maps indicate how the electrostatic potential changes when the electron density of the DNA is polarized by the reaction field of the solvent. Another example of semi-empirical applications is the study of the active site of an enzyme (cytidine deaminase) to delineate a mechanism for ligand attack [754].

In *ab-initio* applications, electronic and vibrational properties have been determined from optimized geometries. *Ab initio* methodologies have also been applied to many nucleic-acid base systems and their complexes, for example, to study stacking interactions [20, 552, 1317], hydrogen-bonding [1318], and basepair planarity properties.

Combined *ab initio* QM/MM approaches can be used to study enzyme catalysis [54, 1344], for example to deduce the reaction paths and free energy barriers for the two steps of the reaction catalyzed by enolase [1443, 1444]. This two-step reaction involves abstraction of a proton from the substrate to produce an intermediate, followed by departure of a hydroxyl group with the assistance of a general acid.

The calculations have identified catalytically important residues and water molecules at the active site of the enzyme (see Figure 8.3) and have shown that the electrostatic interactions driven by metal cations at the active site strongly favor the first, but strongly disfavor the second, step of the reaction. This dilemma appears to be resolved by a tailored organization of polar and charged groups at the enolase active site, exploiting the two different orientations of charge redistribution involved in each of the two reaction steps. Thus, the enzyme environment might provide an essential platform for the reaction mechanism. This finding may ultimately be tested experimentally.

Enzyme catalysis associated with DNA repair and replication is another important example where QM/MM applications have been insightful. Understanding the fidelity of DNA repair and replication mechanisms requires tracking the conformational and energetic pathways associated with these processes, and simulations have the potential to suggest the transient intermediates that are beyond the capabilities of experiment. In DNA polymerases, the repair involves the chemical incorporation of a nucleotide in the DNA by phosphodiester bond formation [1029]. QM/MM studies of DNA pol  $\beta$ , for example, suggest a Grotthuss hopping mechanism of proton transfer between water molecules and three

conserved aspartates in the active site. When a correct unit (i.e., C opposite a template G) is incorporated, a lower activation energy is involved compared to the incorrect unit (A opposite G) [1033]. Similar water-assisted mechanisms were later identified in other polymerase systems such as Dpo4 [1334, 1338] and T7 DNA polymerase [1333]. Other QM and QM/MM studies [18, 49, 144, 408, 409, 769] also suggest alternative pathways, depending on the structure of the complex, the protonation states, and the environment (water and ions), underscoring the versatility of polymerase mechanisms (see Fig. 8.4).

---

### 8.2.6 From Quantum to Molecular Mechanics

An alternative approach is known as *molecular mechanics*, also referred to as the *force-field* or *potential energy* method [24, 175, 185, 764, 871, 898, 938, 1335, 1359]. The Born-Oppenheimer approximation to the potential energy with respect to the nuclei,  $E_p(X)$ , can be imagined as the target function in molecular mechanics ( $X$  represents the collective position vector for the nuclei). The electrons can be regarded as implicit variables of this potential. However, unlike quantum mechanics, this potential function must be evaluated empirically.

#### Mechanical Molecular Representation

An underlying principle in molecular mechanics is that *cumulative* physical forces can be used to describe molecular geometries and energies. The spatial conformation ultimately obtained is then a natural adjustment of geometry to minimize the total internal energy. A molecule is considered as a collection of masses centered at the nuclei (atoms) connected by springs (bonds); in response to inter and intramolecular forces, the molecule stretches, bends, and rotates about those bonds (Fig. 8.5). This simple description of a molecular system as a mechanical body is usually associated with a “classical” system. Remarkably, this classical mechanics description — an appropriate characterization even as the amount of quantum-mechanical information used to derive force fields increases — works generally well for describing molecular structures and processes, with the exception of bond-breaking events.

In practical terms, molecular mechanics involves construction of a potential energy, a function of atomic positions, from a large body of molecular data (crystal structure geometries, vibrational and microwave spectroscopy, heats of formation, etc.). Entropic contributions are either neglected or approximated by various techniques. Minimization of this function can then be used to compute favorable regions in the multidimensional configuration space, and molecular dynamics simulations can further explain the system’s thermally accessible states.

#### Early Days

As introduced in Chapter 1, the first molecular mechanics implementations date to the 1940s, but only in the late 1960s/early 1970s did the availability of digital

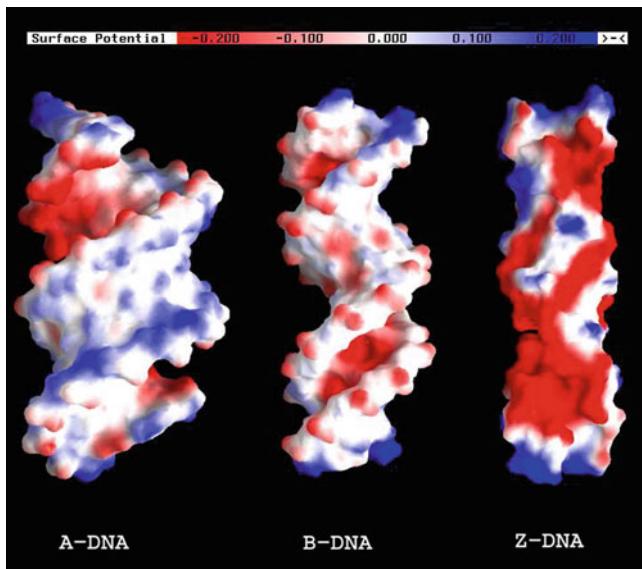


Figure 8.2. The electrostatic potential surface of the electronic response density of A, B and Z-DNA calculated by the linear-scaling electronic structure and solvation methods of York and coworkers [638]. The electronic response density is defined as  $\Delta\rho(\mathbf{r}) = \rho_{\text{sol}}(\mathbf{r}) - \rho_{\text{gas}}(\mathbf{r})$  where  $\rho_{\text{gas}}(\mathbf{r})$  is the relaxed electron density in the gas phase and  $\rho_{\text{sol}}(\mathbf{r})$  is the relaxed electron density in solution (approximated by a continuum dielectric model).

computers make such calculations tractable. One of the force-field pioneers, the late Shneior Lifson, who developed the Consistent Force Field approach with doctoral student Arieh Warshel [766], recalled that “*the empirical [force field] method did not always enjoy a high prestige value among theoretical chemists, particularly those engaged in quantum chemistry*” [765]. But Lifson went on to suggest that in the early days of the field the notions of force-field *consistency* and *transferability* were neither always carefully treated nor well understood. The power of the empirical force-field approach to describe collectively properties of related molecules — a description not possible by quantum mechanics — was also not fully appreciated.

A classic example illustrating the early success of molecular modeling computations is the correct interpretation of peculiar experimental results. In 1967, consistent force-field calculations predicted two stable conformations for a cycloalkane (1,1,5,5-tetramethylcyclodecane). Oddly, neither structure resembled the crystal form [135]. However, the *average* of the two computed molecular mechanics structures matched the experimental data *exactly!* Thus, empirical calculations revealed that the regular crystal contained a distribution of

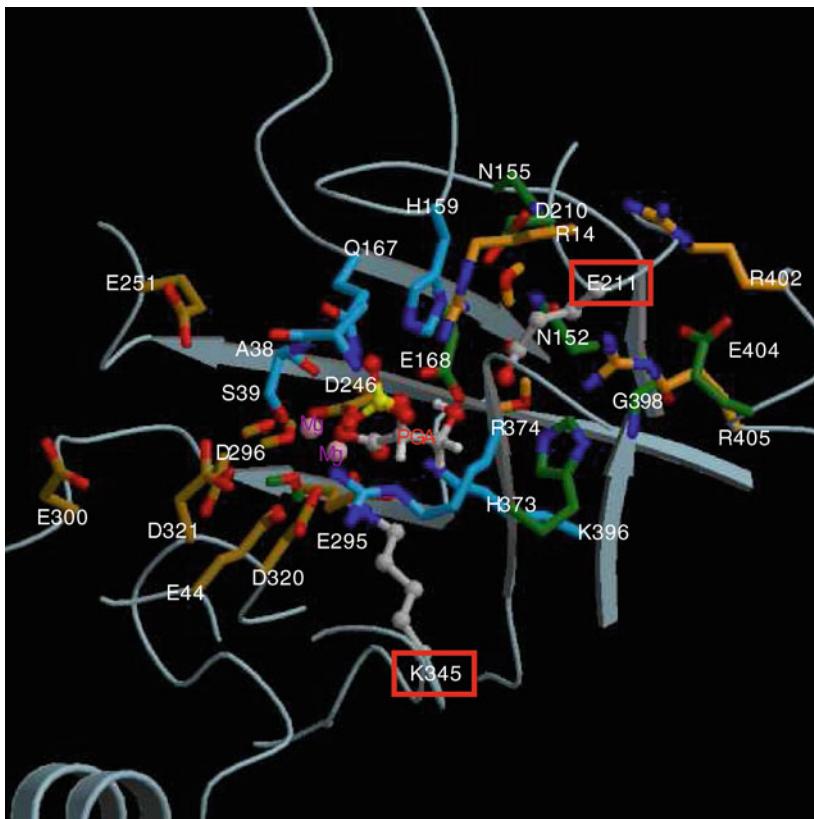


Figure 8.3. Catalytically important protein residues and water molecules at the active site of enolase as identified by *ab initio* QM/MM calculations by Weitao Yang and coworkers [778]. The color coding is according to categories of energetic contribution to stabilization of the transition state in either the first or second step of the reaction (e.g., sky blue, dark golden, green, orange); oxygens are colored red, nitrogens are dark blue, and hydrogen atoms of amino acid residues are omitted. Lys345 (K345, bottom center) is the general base that captures a proton from the substrate (PGA, 2-phospho-D-glycerate) in the first reaction step. Glu211 (E211, top right) is the general acid that assists departure of a hydroxyl group in the second reaction step. Residues in sky blue are hydrogen bond donors interacting with the substrate. Those in dark golden and orange are positioned to counteract the effects of the two magnesium ions (labeled) in the second reaction step but have small energetic effects on the first reaction step. Residues colored green are others found to have energetic effects on the reaction.

two conformers. This was indeed confirmed later by experimentation. Today, this theme emerges often from molecular dynamics simulations, where the spatial and temporal conformational-ensemble average corresponds to the experimental structure [209].

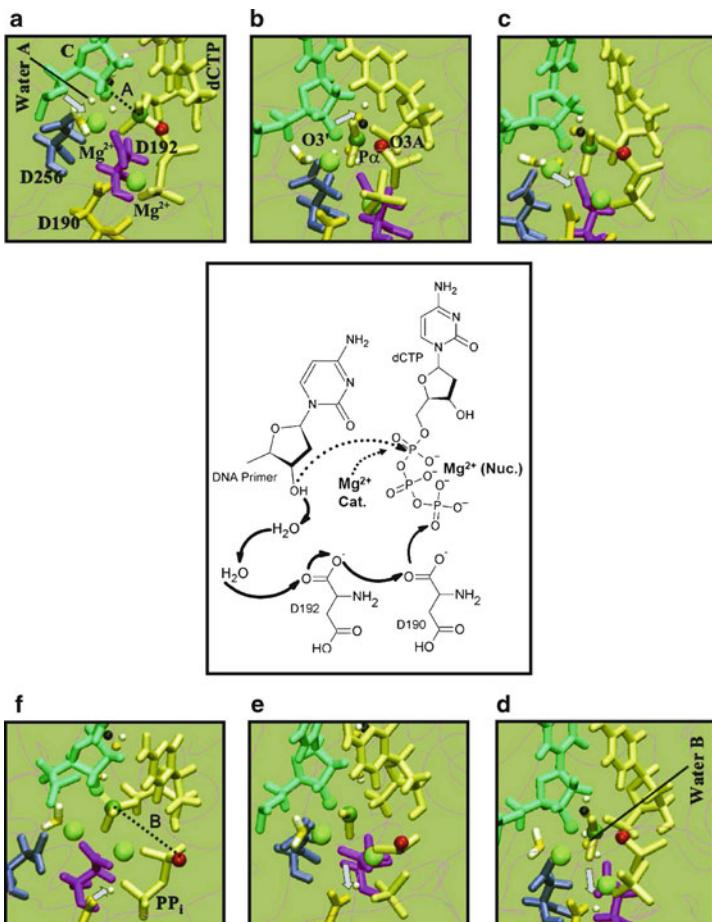


Figure 8.4. Mechanism of Grotthuss hopping chemical reaction along with key intermediates captured during phosphoryl transfer in pol  $\beta$  common to the G:C and G:A systems [1033]. Solid arrows in the scheme indicate the migration path of the proton, and the dotted lines represent the nucleophilic attack and formation of pyrophosphate ( $PP_i$ ) [1033]. The reaction intermediates shown are: (a) reaction state of the closed nucleotide-bound enzyme state; (b) de-protonation of the the O<sub>3'</sub>H to water; (c,d) proton transfers to Asp192; (e) proton transfers to Asp190; (f) proton reaches the pyrophosphate unit to obtain the final product. The colors represent: blue (D256), gold (D190), magenta (D192), yellow (dCTP), cyan (CYT: terminal DNA primer), black (the O<sub>3'</sub>H-proton), green (the O<sub>3'</sub> oxygen, attacking nucleophile), dark green (central phosphorus), red (leaving O<sub>3A</sub> oxygen), and light green (Magnesium). The oxygen and hydrogen atoms of water molecules are in gold and white, respectively. The key distances A = O<sub>3'</sub>-P <sub>$\alpha$</sub>  and B = P-O<sub>3A</sub> are shown; in the reaction, A decreases from 2.90 Å in the reactant state to A = 1.70 Å in the product, and B increases from 1.7 Å in the reactant state to 5.8 Å in the product state. The rate limiting step in the chemical pathway is found to be the initial de-protonation of the O<sub>3'</sub>H. The activation energy was found to be about 17 kcal/mol for G:C and  $\geq$  21 kcal/mol for G:A. The QM/MM procedure involves a novel protocol using energy minimization, dynamics simulations, and quasi-harmonic free-energy calculations.

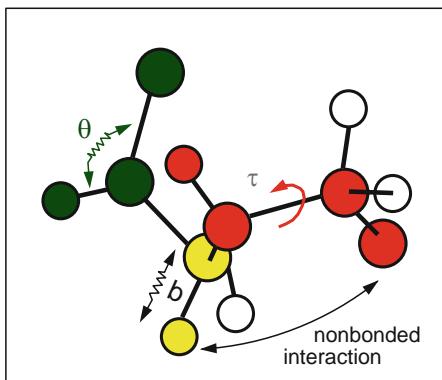


Figure 8.5. A molecule is considered a mechanical system in which particles are connected by springs, and where simple physical forces dictate its structure and dynamics.

## 8.3 Molecular Mechanics: Underlying Principles

In theory, the overall effectiveness of molecular mechanics depends on the validity of its three underlying principles: (1) basic thermodynamic assumption, (2) additivity of the effective energy potentials, and (3) transferability of these potentials.

### 8.3.1 The Thermodynamic Hypothesis

#### Does Sequence Imply Structure?

The thermodynamic hypothesis assumes that many macromolecules are driven naturally to their folded native structure (i.e., with minimal or no intervention of catalysts) largely on the grounds of thermodynamics. This behavior is in contrast to other processes in the central dogma that require the crucial interplay of enzyme machinery to lower activation barriers. Though it is important to consider these complex factors when modeling macromolecules, the basic thermodynamic hypothesis of the empirical approach appears reasonable for many biomolecules under study by molecular mechanics and dynamics.

Indeed, experiments have supported the intrinsic folding/entropy connection for many small globular proteins [52]. Thus, a strong thermodynamic force drives ‘scrambled’ conformations of high free energy to the native state of low free energy. Folding in many cases is reversible (denatured  $\rightleftharpoons$  native state) and attainable from many different configurations.

Theoretical analyses based on statistical studies of spin glasses further suggest that the probability that a randomly synthesized protein will exhibit a thermodynamically dominant fold (i.e, the global minimum of the free energy) increases rapidly as temperature decreases [1167].

As discussed in Chapter 3, the resilience in general of overall protein structures to small mutations — in some cases even at critical regions, as in the Rop turn ([205] and Figure 3.10) — is another indicator of the strong thermodynamic tendency of proteins to adopt and retain native structure/folds. For higher organizational forms of DNA, the intrinsic topology also appears to induce a folded configuration determined in large part by base composition and the ionic medium [269].

The kinetics involved in such folding processes represent another subject of intense interest (see [861, 1257], the discussion in Chapter 2, and Box 8.2).

---

### Box 8.2: The Thermodynamic Hypothesis as Exemplified by The Levinthal Paradox

The concept introduced by Cyrus Levinthal in 1969 [744], which became termed the “Levinthal paradox”, is that a protein cannot find its native state by a random search through all its possible conformations. This is because such an exhaustive enumeration would in theory require eons of years, while real proteins fold on the sub-second and second timeframe. Levinthal thus hypothesized that well defined pathways must exist for protein folding in nature [743].

Recently, protein-folding kinetics have been analyzed and interpreted on the basis of simple simulations of polypeptide models. The “old view” that emerged from Levinthal’s work is now being merged with the “new view” [313, 314, 707], which promotes many competing folding pathways rather than a unique pathway with well-defined intermediates. The two views can be merged when the folding pathways all share a unifying pattern of native contacts and can be described within a common multidimensional energy landscape [1385]. (See also discussion in Chapter 2).

While this protein folding problem is generally formulated purely in terms of “sequence dictates structure”, it is clear that in many cases the folding process *in vivo* is enhanced or facilitated through the external intervention of additional molecules such as chaperones. Such agents may stabilize intermediates, prevent aggregates of misfolded proteins, and/or reduce the activation energies effectively [561, 833]. Members of the chaperone class that includes the heat-shock protein hsp70 bind to newly-synthesized proteins and short linear peptides, possibly preventing aggregates. Other chaperones assist in protein translocation across membranes. Members of the *chaperonin* class include the well-studied GroEL/GroES complex from bacteria [387] that bind to partially-folded polypeptides and assist in the folding. Many questions are now being investigated regarding the folding kinetics in all these systems (see [519, 1326], for example).

---

#### 8.3.2 Additivity

The molecular mechanics principle of additivity assumes that the effective molecular energy can be expressed as a sum of potentials derived from simple physical forces: van der Waals, electrostatic, mechanical-like strains arising from

“ideal” bond length and angle deviations, and internal torsion flexibility (rotation of two chemical groups about the bond joining them). The forces can be separated into local (bonded) and nonbonded (nonlocal) terms.

### Local Terms

The local components for macromolecules can typically be written as:

$$E_{\text{local}}(X) = \sum_{\substack{\text{bonds} \\ i}} E_{\text{bond}}(b_i) + \sum_{\substack{\text{bond angles} \\ i}} E_{\text{bang}}(\theta_i) + \sum_{\substack{\text{dihedral angles} \\ i}} E_{\text{tor}}(\tau_i), \quad (8.3)$$

where summations extend over the sets of bonds  $\{b_i\}$ , bond angles  $\{\theta_i\}$ , and dihedral angles  $\{\tau_i\}$  (see Fig. 8.5). Functional forms are typically harmonic (e.g.,  $S[b - \bar{b}]^2$ ) and trigonometric (e.g., functions of  $\cos(\theta)$ ), as discussed in the next chapter. Note that all internal variables are functions of interatomic distances,  $r_{ij}$ . Cross terms (like  $E(b, b')$  or  $E(b, \theta)$ ) are not commonly used for proteins and nucleic acids because the associated force constants ('off-diagonals' in the force-constant matrix) usually have much smaller values than those associated with the 'diagonal' terms. Such force constants, for example  $S_{bb'}$  for bond/bond interactions or  $K_{b\theta}$  for bond/bond-angle terms, are associated with potentials of the form  $S_{bb'}(b - \bar{b})(b' - \bar{b'})$  and  $K_{b\theta}(b - \bar{b})(\theta - \bar{\theta})$ , respectively, which are common in small-molecule force fields (e.g., [30]).

### Nonlocal Terms

The nonlocal components can be written as:

$$E_{\text{nonlocal}}(X) = \sum_{\substack{\text{nonbonded pairs} \\ (i,j), i < j}} E_r(r_{ij}), \quad (8.4)$$

where the functions are often rational (e.g.,  $\sum_{i=1}^{N_i} r^{-m}$ ). These energies are usually comprised of van der Waals and Coulombic contributions between nonbonded atom pairs. (Nearby atom pairs counted in the local energy may be omitted to avoid double counting).

### Benefits of Separability

The natural separability into bonded (local) and nonbonded (nonlocal) terms can be exploited in the design of minimization algorithms that use function structure to accelerate convergence (see Chapter 11). This separability is also the basis for multiple-timestep protocols for molecular dynamics that update local and nonlocal forces at different frequencies (see Chapters 13 and 14). This difference in force updating is reasonable because the local terms generally change rapidly, while the nonlocal terms change more slowly, with distance and time.

While the number of nonbonded terms grows quadratically  $\mathcal{O}(N^2)$  with the number of atoms, the number of local terms — involving pairs, triplets, and quadruplets of bonded atomic sequences — grows only linearly with size.

This computational complexity of the nonbonded terms is especially a burden in simulation protocols for large systems that require updating for many iterations, as in macromolecular dynamics. Fortunately, some work can be reduced since the nonbonded Coulomb forces change more slowly with distance than the bonded terms, and hence can be updated less often than the local forces. Chapter 10 is devoted to the nonbonded forces.

### Multibody Potentials

In this typical energy formulation, nonbonded interactions are “effective pair potentials” (i.e., additive two-body forces). For electrostatic interactions, for example, these effective pair potentials only reflect in some average sense the charge distribution due to molecular polarizability. It is clear that many-body contributions are important for accurate reproduction of certain molecular properties. For example, accurate account of dispersion forces for polar molecules (i.e., molecules in which the charges are nonuniformly distributed), such as water, require three-body interaction potentials:

$$E_{ijk}(r_{ij}, r_{jk}, r_{ik}).$$

These potentials can be expressed as a composite of trigonometric and rational functions in terms of three bond lengths and three bond angles [131]. Indeed, the importance of water as a solvent for proteins and nucleic acids was realized in the 1970s [1344] and has prompted development of ‘polarized’ water potentials [244, 1343, for example]. In practice, these models increase the number of *interaction sites* per water molecule.

#### 8.3.3 Transferability

The principle of transferability assumes that potentials can be developed to incorporate all experimental data for *representative* structures and then be applied successfully to the prediction of large biological molecules composed of the same chemical subgroups. This is basically a reasonable assumption since bond lengths and bond angles tend to adopt similar values in different molecular species under normal conditions. However, under special straining forces, such as in cycloalkanes, these values may vary significantly from ‘ideal’ or average values. Modeling the structures and properties of complex aromatic or conjugated systems<sup>3</sup> also requires special treatment, such as using sophisticated quantum-mechanical calculations to obtain bond orders of conjugated bonds, which can be related to bond lengths and stretching constants, and in this way reducing the problem to molecular mechanics ([29, 899, 1210, 1246, 1341], for example).

---

<sup>3</sup>Aromatic compounds are benzene-like in structure and properties. Conjugation is a structural feature caused by the overlap of the  $\pi$  orbital with other orbitals in the molecule.

### Functional Variations in Geometry

In small molecules, the environment-dependence of geometric trends can be modeled by a *function* rather than a constant. For example, Allinger and coworkers devised functions for bond lengths in small-molecule force fields (MM3/MM4) to account for the *electronegativity* of attached substituents [1210] and for *hyperconjugation* [26]. The former parameterization [1210] can, for example, accurately model the *shorter* C–C bond in fluoroethane ( $\text{C}_2\text{H}_5\text{F}$ ) relative to ethane ( $\text{C}_2\text{H}_6$ ), since the fluorine is electronegative; similarly, it can also account for a *longer* C–O bond in an alcohol where the electropositive hydrogen is attached to an oxygen (like ethyl alcohol,  $\text{CH}_3\text{CH}_2\text{OH}$ ), relative to an analogous molecule in which a carbon rather than hydrogen is attached (as in dimethyl ether,  $\text{C}_2\text{H}_5\text{OC}_2\text{H}_5$ ). A functional dependence for reference bond lengths used for hyperconjugated systems [26] can account for bond-length trends in molecular species in which bond orbitals overlap and lead to resonance, like longer C–H or C–C bonds in carbonyl compounds (e.g.,  $\text{H}-\text{C}-\text{C}=\text{O} \leftrightarrow \text{H}^+ + \text{C}=\text{C}-\text{O}$ ).

### Proliferation of Atom Types

An alternative approach for incorporating the environment dependence of geometric tendencies is to increase the number of ‘atom types’. Thus, atom types reflect the molecular environment (e.g., aromatic carbon in a nitrogenous base) and hybridization (e.g.,  $\text{sp}^2$  or  $\text{sp}^3$ ) [175, 805, 1359].

There are around 160 atom types, for example, in the CHARMM force field for proteins (version 22) and nucleic acids (version 27) [415, 804–806]: 62 carbons, 31 hydrogens, 28 nitrogens, 18 oxygens, 4 sulfurs, 3 phosphorus atoms, 6 fluorines, one heme iron, and 7 different ions; see Table 8.1 for some examples of these atom types and Figure 8.6 for illustrations for selected residues [805]. The proliferation of atom types in modern force fields thus attempts to improve compatibility with experiment. Alternatively, force fields may be restricted to certain families of molecules such as alkanes, amides, and carboxylic acids [185, 764].

I emphasize that *individual* functions *should not be transferred* from one force field to another, since the entire potential is parameterized as a whole to reproduce consistently experimental data.

Overall, while the transferability assumption is inherent in this empirical science, molecular mechanics has steadily gained recognition through many important contributions. Today’s force fields are excellent for deducing structures and properties of many molecular systems, especially for small molecules using specialized force fields. One advantage of theoretical calculations is that the thermal motions and lattice effects that influence crystallographically-determined structures may not be a problem when accurate force fields are used to predict molecular structures and properties.

Table 8.1. Examples of atom types defined in CHARMM 22 and 27 [415, 804–806].

Atom	Symbol	Atom modifier
Carbon	C	polar (carbonyl, peptide backbone)
	CA	aromatic
	CC	carbonyl (Asn, Asp, Gln, Glu)
	CPT	inter-ring in tryptophan
	CP1, CP2, CP3	special tetrahedral, in proline
	CT1	aliphatic sp <sub>3</sub> in CH
	CT2	aliphatic sp <sub>3</sub> in CH <sub>2</sub>
	CT3	aliphatic sp <sub>3</sub> in CH <sub>3</sub>
	CN1	nucleic acid carbonyl carbon
	CN3	nucleic acid aromatic carbon
Oxygen	O	carbonyl
	OC	carboxylate
	OH1	hydroxyl
	ON6	nucleic acid deoxyribose ring oxygen
	ON6B	nucleic acid ribose ring oxygen
Nitrogen	N	proline
	NH1	peptide
	NH2	amide
	NN1	nucleic acid amide nitrogen
	NN2	nucleic acid protonated ring nitrogen
Hydrogen	H	polar
	HA	nonpolar
	HP	aromatic
	HS	thiol
	HN1	nucleic acid amine proton
	HN2	nucleic acid ring nitrogen proton

## 8.4 Molecular Mechanics: Model and Energy Formulation

In practice, the challenge and success of molecular mechanics rely on both effective formulation of the potential energy function and the application of suitable search algorithms (e.g., multivariate minimization, sampling, dynamic simulations). With the well known difficulties of large-scale minimization (see Chapter 11) and the timestep problem in dynamics (Chapters 13 and 14), the structural outcome — and hence biological implications of any calculation — depends on the combination of modeling and algorithmic techniques employed. Indeed,

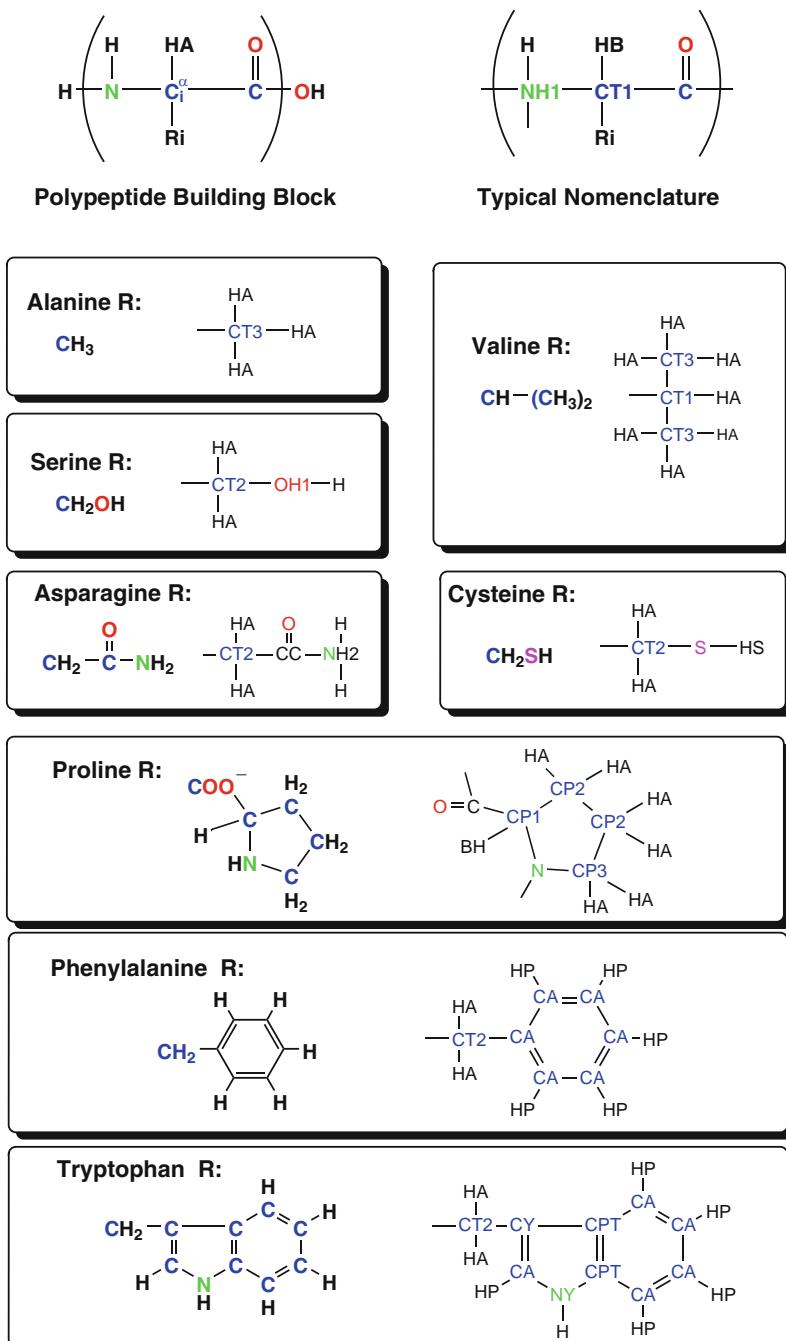


Figure 8.6. Examples of atom types as used in polypeptides in the CHARMM program [805, 806].

some strengths and weaknesses of molecular mechanics have been debated openly in a series of papers [455, 668, 1068, 1069]; see also Homework Assignment 7. Some valid questions are:

- How accurate are quantum-mechanically derived partial charges?
- How do Cartesian and dihedral-angle representations affect results?
- Is it appropriate to introduce arbitrary scale factors in energy coefficients to enhance agreement with experiment?

The cautions and possible pitfalls of force field derivations and parameter choices are thus important to emphasize, even for state-of-the-art force fields.

In formulation of the energy, three basic and important decisions are involved: (1) representative configuration space, (2) functional form, and (3) numerical values for the parameters. We discuss each in turn.

### 8.4.1 Configuration Space

#### A Question of Size

The atomic configuration<sup>4</sup> space for a molecular system consists of  $3N - 6$  degrees of freedom, where  $N$  is the number of atoms. (Six degrees of freedom are removed for rigid-body translation and rotation invariance of the energy). Thus, the number of degrees of freedom for proteins and nucleic acids typically ranges in order from  $10^3$  to  $10^4$ . With explicit representation of water molecules, this number rises by at least an order of magnitude (see, for example, solvated-macromolecular system sizes in Table 1.2 of Chapter 1).

The complete set of  $3N - 6$  degrees of freedom can be described directly by lists of Cartesian coordinates for all the atoms in the system. Alternatively, internal variables such as bond lengths, bond angles, and dihedral angles may be used; this internal representation has been quite successful for the study of proteins [1068, for example].

Other representations have been used for biomolecules to reduce this number of variables. Reductions are possible by fixing bond lengths and angles and working in dihedral angle space, or by restricting energetic pathways to approximate formulations. However, these representations do not usually lead to enhanced efficiency unless the work involved in the nonbonded computations — the real computational ‘bottleneck’, see Chapter 10 — is reduced significantly.

---

<sup>4</sup>*Configuration*, a more general term than *conformation*, is often used by mathematicians to describe the shapes of objects in space. The more chemical term *conformation* often refers to configurations that differ from one another through rotations of groups of atoms about the bond connecting them (i.e., dihedral angles). Note, however, that in organic chemistry a *configuration* refers to stereoisomers, molecules with different connectivity arrangements which cannot be converted into one another through rotations about bonds.

### The Pseudorotation Description

An example of a parameter reduction approach is the pseudorotation path developed for nucleic acids [35, 271, 750]. This energetic path, initially developed for the hydrocarbon cyclopentane [766] and later extended to ribose and deoxyribose sugars, constrains the energy of five-membered sugar rings to a wavelike motion from a mean plane. This plane can be defined in various ways by positions of the five skeletal ring atoms. See Chapter 5 for a discussion of these descriptions in the section on furanose conformations.

While conceptually simple, the reduction in degrees of freedom from 9 (3·5–6) to 2 is clearly approximate. The pseudorotation approximation was noted to produce anomalies in overall nucleic acid structures [937, 944], inconsistencies with ring closure, and mathematical difficulties in expressing the energy derivatives in terms of the independent conformational variables. More generally, while for some systems simplifications in the representative conformation space may be acceptable, it is usually advantageous to avoid constraints altogether [1101]. (The pseudorotation concept remains a useful analysis tool, however).

### Cartesian Space

Cartesian coordinate space is most convenient for direct differentiation of the energy in terms of the independent parameters, as realized in the early days of force field development by Lifson and Warshel [766], and therefore application of efficient second-derivative Newton minimization methods (see Chapter 11). Cartesian space is also most natural for implementation of molecular dynamics, since the generated molecular trajectories

$$\{X(t_0), X(t_0 + \Delta t), \dots, X(t_0 + n \Delta t), \dots\},$$

where  $t_0$  is the initial time reference and  $\Delta t$  is the timestep (see molecular dynamics chapters), generally rely on recursive expressions. That is, the new positions and velocities

$$X(t_0 + n \Delta t) \text{ and } V(t_0 + n \Delta t)$$

are explicit functions of the previous positions and velocities

$$X(t_0 + (n - 1) \Delta t) \text{ and } V(t_0 + (n - 1) \Delta t).$$

With all these considerations, it is usually advantageous to enforce constraints — if necessary — by means of penalty functions (i.e., *soft* constraints). Harmonic penalty functions can be used, for example, to keep bond lengths and angles close to their observed values.

#### 8.4.2 Functional Form

##### Composition

The potential energy  $E$  of a molecular model is typically constructed as the sum of contributions from the following types of terms: bond length and bond angle strain

terms ( $E_{\text{bond}}$  and  $E_{\text{bang}}$ ), a torsional potential ( $E_{\text{tor}}$ ), a Lennard-Jones potential to model repulsion at short interatomic separations and attraction at long distances ( $E_{\text{LJ}}$ ), and a Coulombic potential among the pairs of charged particles in the system ( $E_{\text{coul}}$ ):

$$E = E_{\text{bond}} + E_{\text{bang}} + E_{\text{tor}} + E_{\text{LJ}} + E_{\text{coul}} \quad (8.5)$$

$$\begin{aligned} E_{\text{bond}} &= \sum_{i,j \in S_B} S_{ij} (r_{ij} - \bar{r}_{ij})^2 \\ E_{\text{bang}} &= \sum_{i,j,k \in S_{BA}} K_{ijk} (\cos \theta_{ijk} - \cos \bar{\theta}_{ijk})^2 \\ E_{\text{tor}} &= \sum_{ijkl \in S_{DA}} \sum_n \left( \frac{V n_{ijkl}}{2} [1 \pm \cos(n \tau_{ijkl})] \right) \\ E_{\text{LJ}} &= \sum_{i,j \in S_{NB}} \left( \frac{-A_{ij}}{r_{ij}^6} + \frac{B_{ij}}{r_{ij}^{12}} \right) \\ E_{\text{coul}} &= \sum_{i,j \in S_{NB}} \left( \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}} \right) \end{aligned}$$

These equations represent first approximations to the physical potentials but are nonetheless reasonable for biomolecules. In these general expressions, the symbols  $S_B$ ,  $S_{BA}$ , and  $S_{DA}$  denote the sets of all bonds, bond angles, and dihedral angles. The nonbonded set,  $S_{NB}$ , typically includes all  $(i,j)$ ,  $i < j$ , atom pairs separated by three bonds or more. Bond and angle variables capped by bar symbols denote reference values associated with these quantities.

The 6/12 Lennard Jones potential above (i.e., attraction of form  $-A/r^6$  and repulsion of form  $B/r^{12}$ ) is typical for large-molecule force fields because of its mathematical convenience; a Buckingham potential (with the same functional form for attraction but an exponential repulsion term of form  $B \exp(-B'r)$ ) [766] is in principle closer to the electronic structure of the atom and thus provides a better potential fit over a broader range of interparticle distances [30].

### Molecular Geometry

To be more precise, we define the analytic expressions for the geometric quantities in the potential energy.

**Positions and distances.** For a molecular system of  $N$  atoms (possibly including atom groups) in Cartesian coordinate space, let

$$\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}), \quad i = 1, \dots, N,$$

denote the position vector of atom  $i$ , and

$$\mathbf{r}_{ij} \equiv \mathbf{x}_j - \mathbf{x}_i$$

denote the distance vector from atom  $i$  to  $j$ . Our potential energy function  $E$  then depends on all Cartesian variables of the atoms:

$$E = E(X) \equiv E(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$$

where  $X$  is the collective vector in the Euclidean space  $\mathcal{R}^n$  of dimension  $n = 3N$ .

For any vector  $\mathbf{a}$  we denote its standard Euclidean magnitude by  $\|\mathbf{a}\|$ . For convenience later in writing the potential energy function, we also denote an interatomic vector magnitude  $\|\mathbf{r}_{ij}\|$  as  $r_{ij}$ , in nonbold type.

**Bond angles.** A bond angle  $\theta_{ijk}$  formed by a bonded triplet of atoms  $i-j-k$  is expressed as an inner (or dot) product:

$$\cos \theta_{ijk} = \frac{(\mathbf{x}_k - \mathbf{x}_j) \bullet (\mathbf{x}_i - \mathbf{x}_j)}{r_{ij} r_{kj}}. \quad (8.6)$$

**Dihedral angles.** A dihedral angle  $\tau_{ijkl}$ , defining the rotation of bond  $i-j$  around  $j-k$  with respect to  $k-l$ , is expressed as (see Fig. 3.14 in Chapter 3 and Fig. 8.5):

$$\cos \tau_{ijkl} = \mathbf{n}_{ab} \bullet \mathbf{n}_{bc}. \quad (8.7)$$

The vectors  $\mathbf{n}_{ab}$  and  $\mathbf{n}_{bc}$  denote unit normals to planes spanned by vectors  $\{\mathbf{a}, \mathbf{b}\}$  and  $\{\mathbf{b}, \mathbf{c}\}$ , respectively, where  $\mathbf{a} = \mathbf{r}_{ij}$ ,  $\mathbf{b} = \mathbf{r}_{jk}$ , and  $\mathbf{c} = \mathbf{r}_{kl}$ . Denoting  $\theta_{ab}$  and  $\theta_{bc}$  as angles  $\theta_{ijk}$  and  $\theta_{jkl}$ , respectively, we write:

$$\cos \tau_{ijkl} = \frac{\mathbf{a} \times \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\| \sin \theta_{ab}} \bullet \frac{\mathbf{b} \times \mathbf{c}}{\|\mathbf{b}\| \|\mathbf{c}\| \sin \theta_{bc}}. \quad (8.8)$$

The sign of  $\tau_{ijkl}$  is set by the sign of the triple scalar product  $\mathbf{a} \bullet (\mathbf{b} \times \mathbf{c})$ .

**Lagrange's identity.** To simplify potential energy equations and differentiation [1103], it is convenient to work with inner product expressions and use Lagrange's identity:

$$(\mathbf{a} \times \mathbf{b}) \bullet (\mathbf{c} \times \mathbf{d}) = (\mathbf{a} \bullet \mathbf{c})(\mathbf{b} \bullet \mathbf{d}) - (\mathbf{b} \bullet \mathbf{c})(\mathbf{a} \bullet \mathbf{d}). \quad (8.9)$$

This produces the alternative expression for  $\cos \tau$ :

$$\begin{aligned} \cos \tau_{ijkl} &= \frac{(\mathbf{a} \times \mathbf{b}) \bullet (\mathbf{b} \times \mathbf{c})}{[(\mathbf{a} \times \mathbf{b}) \bullet (\mathbf{a} \times \mathbf{b}) (\mathbf{b} \times \mathbf{c}) \bullet (\mathbf{b} \times \mathbf{c})]^{1/2}} \\ &= \frac{(\mathbf{a} \bullet \mathbf{b})(\mathbf{b} \bullet \mathbf{c}) - (\mathbf{a} \bullet \mathbf{c})(\mathbf{b} \bullet \mathbf{b})}{\{[(\mathbf{a} \bullet \mathbf{a})(\mathbf{b} \bullet \mathbf{b}) - (\mathbf{a} \bullet \mathbf{b})^2] [(\mathbf{b} \bullet \mathbf{b})(\mathbf{c} \bullet \mathbf{c}) - (\mathbf{b} \bullet \mathbf{c})^2]\}^{1/2}} \quad (8.10) \end{aligned}$$

According to this convention,  $\tau = 0^\circ$  defines a *cis* coplanar orientation for atoms  $i-j-k-l$ ,  $\tau = 180^\circ$  defines a *trans* coplanar orientation, and a positive sign corresponds to a clockwise rotation of the far bond with respect to the near bond (when viewed along the  $j-k$  bond).

**Derivatives.** First and second derivatives of these expressions are often needed for structure minimization, molecular dynamics, and various conformational analyses (e.g., normal modes). The derivative expressions are tedious to derive (see Homework Assignment 8) and care must be used to avoid singularities, but various algorithmic procedures have been developed to simplify the task in practice [1103, for example].

### 8.4.3 Some Current Limitations

Parameterization details for each potential energy term, including functional variations, are described in the next chapter. We conclude this chapter by mentioning some general limitations of current molecular force fields:

1. **Many Force Field Choices.** At present, there is no “universal force field”, nor are the many force fields in use close to converging to one another in some sense. Essentially, users around the world develop a preference for one force field over another on the basis of their target application and practical factors such as cost and convenience.

For example, the MM2/3/4 force-field family developed by Norman Allinger and coworkers is a popular choice for small molecular systems [25, 30, 696]. Protein modelers often use the CHARMM package developed by Martin Karplus and coworkers [174]. Nucleic acid modelers might prefer the AMBER program developed by the late Peter Kollman and coworkers [204]. Harold Scheraga’s team developed the dihedral-angle ECEPP family of force fields for protein modeling [58, 897]. Other force fields are also available, such as GROMOS [950], OPLS-AA [618, 629], CFF [614], and CVFF [371]; see also [489, p. 76] and [4] for lists of available force fields and associated molecular modeling packages, and Table 11.1 of Chapter 11 for minimization algorithm information for some of these programs.

A current effort is the development of programs which utilize force fields that cover a wider range of chemical systems. This, however, involves a trade-off between force field accuracy and versatility [1123]. The MMFF force field, for example, developed at Merck in the late 1990s [497–500], was designed to model a wide range of organic molecules and proteins. Still, the recent trend has been to incorporate force field segments for specialized systems in existing modeling packages, for example to model a complex between small organic and drug-like molecules with biomolecular systems. The OPLS-AA force field, in particular, covers a wide variety of organic molecules. Recently, general parameters for organic molecules have been developed for usage in combination with the AMBER and CHARMM force fields through easy extensions [1300, 1331].

2. **Variability in Functional Form and Numerical Parameters.** As detailed in the next chapter, a great deal of variability exists in the functional forms used for each potential energy term as well as in the numerical values for the associated parameters.
3. **Inclusion of Explicit Hydrogen Bonding.** Some macromolecular force fields use explicit hydrogen-bonding potentials, though these potentials were more common in older versions that did not consider hydrogens explicitly. Note that the strength of a hydrogen bond is determined by its geometry, which depends on the distances associated with the Donor-Hydrogen ··· Acceptor sequence and the Donor-Hydrogen ··· Acceptor angle ( $\theta_{dha}$ ) formed about the hydrogen atom. For a linear hydrogen bond (usually strongest),  $\theta_{dha}$  has the ideal value of 180°. In the current biomolecular AMBER and CHARMM force fields, for example, the proper dependence of hydrogen bonding on distance and angular orientations is adequately treated with the electrostatic and Lennard-Jones terms.

In theory, classical electrostatic and (quantum) bonding forces can account for hydrogen-bonding interactions. Thus much work has gone into formulating appropriate hydrogen-bond potentials for small-molecule force fields.

Hydrogen-bond potentials can be derived from crystal packing studies or quantum-mechanical calculations, or introduced to correct weak interactions [898]. A re-optimization of hydrogen-bond potentials for MM3 [768] based on *ab initio* calculations concluded that hydrogen-bond interactions are more complex than previously thought. In that work, the hydrogen bond potential was formulated as a complex function of the Hydrogen ··· Acceptor distance, the Donor-Hydrogen bond length, and the cosine of the  $\theta_{dha}$  angle. While this [768] and earlier works treated the overall nuclear charges, more recent work has shown that a better representation involves placing the center of the electron density where the lone pairs are and not where the nucleus is [796].

4. **Electrostatic Approximations.** In the simple Coulomb potential described above, interactions between pairs of atoms often consider only point charges for each atom. However, the charge distribution about each nucleus is clearly more complex, and in some cases induced dipole effects (i.e., distortion of the electron distribution) are important to consider. This notion of modeling electronic polarizability (roughly a measure of the distortion of an electronic charge cloud) is a continuing focus in force field design (see [223, 243, 501, 630, 970, 1335, 1343] for example).
5. **Limited Use of Quantum-Mechanical Information.** In theory, quantum-mechanical theory from molecular orbital techniques can be applied to small molecules to improve functional form and assign more accurate

parameters (e.g., to fit parameters to *ab initio* relative energies or to the quantum-mechanical energy surface curvature). In practice, this fitting is not easy to perform and depends on the quality of the quantum calculations (e.g., basis set). Yet, quantum-based calculations have been used to assign the electrostatic partial charges, and the newer generation of force fields relies on quantum calculations wherever possible (e.g., [497, 630, 768, 805, 842]), in combination with increasingly accurate experimental measurements. Thus, using both experimental and quantum calculations — despite each technique's limitations — can provide the best overall results.

# 9

## Force Fields

### Chapter 9 Notation

SYMBOL	DEFINITION
<b>Scalars &amp; Functions</b>	
$c$	speed of light
$c_s$	ionic concentration
$e$	electron charge
$\hbar$	Planck's constant over $2\pi$
$k$	harmonic spring constant
$m$	particle mass
$m_e$	electron rest mass
$n$	periodicity of rotational barrier (torsional potential)
$q_i$	Coulomb partial charge of atom $i$
$r$	bond length
$\bar{r}$	reference bond length
$r_{ij}$	interatomic distance (between atoms $i$ and $j$ )
$x$	displacement
$A_{ij}, B_{ij}$	Lennard-Jones coefficients for atom pair $i, j$ (attraction, repulsion)
$D$	Morse well depth parameter
$E$	potential energy
$E_{\text{coul}}$	Coulomb energy
$E_{\text{LJ}}$	Lennard-Jones energy
$E^r$	bond length energy
$E^{rr'}$	stretch/stretch energy
$E^{r\theta}$	stretch/bend energy
$E^{r\theta r'}$	stretch/bend/stretch energy
$E^\theta$	bond angle energy
$E^{\theta\theta'}$	bend/bend energy

Chapter 9 Notation Table (continued)

SYMBOL	DEFINITION
$E^\rho$	Urey-Bradley energy
$E^\tau$	dihedral (torsional) angle energy
$E^{\tau\theta\theta'}$	torsion/bend/bend energy
$E^{\tau\theta}$	torsion/bend energy
$E^\chi$	improper torsion energy
$E^{\chi\chi'}$	improper/improper torsion energy
$F$	force
$F_{\text{coul}}$	Coulomb force
$K_{\text{coul}}$	Coulomb potential constant
$K_h$	harmonic bending constant
$K_t$	trigonometric bending constant
$N$	number of atoms
$N_e$	number of outer shell electrons
$S_h$	harmonic stretching constant
$S_m$	Morse stretching constant (well width parameter)
$S_q$	stretching force constant for special quartic potential
$V_{ij}, r_{ij}^0$	Lennard-Jones coefficients for atom pair $i, j$ (energy minimum/interaction distance)
$V_n$	barrier height of torsional potential associated with periodicity $n$
$\alpha$	atomic polarizability
$\epsilon$	dielectric constant
$\epsilon_0$	permittivity of vacuum
$\theta$	bond angle
$\bar{\theta}$	reference bond angle
$\theta_{\text{tet}}$	tetrahedral bond angle, $109.47^\circ$ , or $\cos^{-1}(-1/3)$
$\kappa$	Coulomb screening parameter
$\lambda$	wave number (wavelength of absorption)
$\mu$	reduced mass
$\nu$	characteristic frequency
$\tau$	torsion (or dihedral) angle
$\tau_0$	reference torsion (or dihedral) angle
$\chi$	Wilson angle (for improper torsion potential)
$\omega$	angular frequency

The purpose of models is not to fit the data but to sharpen the questions.

Samuel Karlin, 1983 (1923–2007).

## 9.1 Formulation of the Model and Energy

In this chapter, we discuss only basic functional expressions of the potential energy function, emphasizing the simple forms typically used for biomolecules. For biomolecular systems, computational speed is premium, and the use of more

complex terms (higher-order expansions, cross terms, etc.), as employed for accurate modeling of smaller systems, is not practical. The next chapter discusses important topics related to this computational complexity of the nonbonded terms: spherical cutoff techniques, fast electrostatic evaluation techniques (Ewald and fast multipoles), and implicit solvation alternatives.

While improvement of potential energy functions — both in terms of functional form and parameters — has been an ongoing enterprise, the current, “*second-generation*” molecular mechanics and dynamics force fields are more sophisticated than those originating from pioneering works in the 1960s and 1970s. Specifically, parameterization depends quite significantly now on quantum mechanical calculations. “*Third generation*” force fields that account more accurately for electronic polarizabilities are already emerging (see [223, 501, 630] for example).

Parameterization of force fields ensures that calculations produce appropriate molecular geometries and interaction energies for a set of model compounds that are appropriate for the force field. For proteins, test compounds are peptides or model peptides [805]. For nucleic acids, deoxyribonucleosides [218, 265] and compounds containing the furanose ring and oligonucleotide crystals [415] are appropriate.

Force field parameters are optimized in a ‘*self consistent*’ fashion [185, 766, 1152] so as to reproduce the increasing body of experimental information on molecular geometries (from crystal and solution studies) to many other properties: measurements of vibrational frequencies, heats of formation, intermolecular energies and geometries, torsional barriers, and more. The use of dynamic simulations to assess the quality of the force field and to refine parameters — so as to better reproduce structural properties and molecular interaction energies — is also an improvement over procedures that utilize energy minimization alone [218, 804, 805].

Before describing each potential-energy term in turn, we review the fundamental molecular motions, called *normal modes*, that form the basis for parameterization of bonded deformations.

## 9.2 Normal Modes

### 9.2.1 Quantifying Characteristic Motions

Molecular vibrational spectra of small molecules form the basis for deriving various force constants, internuclear distances, and bond dissociation energies for bonded nuclei [766]. Such bond motions describe vibrations about equilibrium states. Specifically, all possible vibrations of a molecule can be described as a superposition of the fundamental oscillations (termed *normal modes*) for that molecule. Each molecule of  $N$  atoms has  $3N - 6$  normal modes: 3 degrees of freedom per atom (giving  $3N$ ) minus 3 translational and 3 rotational degrees of freedom for the molecule as a whole.

## Experimental Determination

The vibrational energy levels of molecules can be detected experimentally by spectroscopic techniques, for example through the vibrational absorption of infrared radiation (IR) and by Raman scattering.

IR spectroscopy is a powerful technique that captures information on the transitions between vibrational quantum states, since these transitions lead to absorption and emission of infrared radiation. The IR wavelength range is 1–100  $\mu\text{m}$ ; this spectral range can be compared with the shorter wavelength of the visible spectrum, which has the range 400–750 nm. IR transitions are ‘allowed’ if there is a change in the dipole moment of the molecule during the transition.

The complementary technique of Raman spectroscopy captures transitions between vibrational levels, but the selection rules are different compared to IR: Raman bands appear only if there is a change in the polarizability of a system during the transition.

IR absorption and Raman spectroscopy are complementary techniques since some transitions that have changing dipole moments absorb light, whereas others have changing polarizability and scatter light. Thus, certain light-induced transitions can be weak, or even absent, in one technique, but of high intensity in the other. For small symmetric molecules, the observed transitions are complementary. For larger and asymmetric molecules, however, the selection rules are not rigidly obeyed, and Raman and IR spectra are essentially the same.

## Frequency Units

Rather than expressing these frequencies in inverse seconds (or hertz, Hz, units), these fundamental frequencies are typically reported in *wavenumbers* of inverse centimeters, that is, the number of waves per centimeter. The higher the frequency, the more difficult (i.e., energetically costly) the deformation. For this reason, bond-stretching modes generally have higher frequencies than angle-bending modes, which in turn have higher frequencies than torsion-angle modes.

Note that stretching a bond significantly amounts to breaking it, so the energy is high; angle bending can be considered to have smaller effects on bonding (hence energy barriers are lower); torsion barriers are small for rotations about single bonds since the effects on bonding are smaller still. For double bonds, torsional motion corresponds to bond breaking and the frequencies are higher than for torsional barriers about single bonds.

## Illustration

For larger molecules and complex mixtures, vibrational spectroscopy is a powerful analytical technique for determining which chemical groups are present. To illustrate, Figure 9.1 shows the three fundamental vibrations of a water molecule: an *asymmetric stretch* (around  $3750\text{ cm}^{-1}$ ), a *symmetric stretch* (around  $3650\text{ cm}^{-1}$ ), and an *angle-bending* mode around  $1600\text{ cm}^{-1}$ . The symmetric

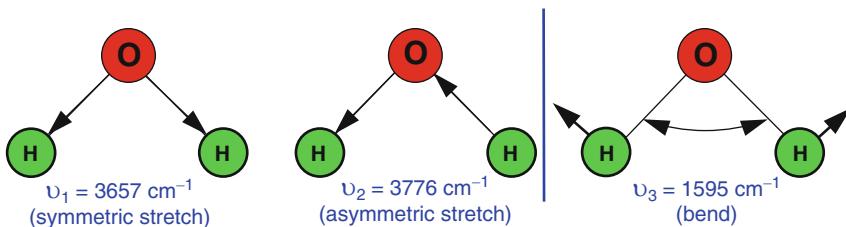


Figure 9.1. Normal modes of a water molecule.

stretch describes the contraction or elongation of both O–H bonds in concert, while the asymmetric mode involves this stretching motion in alternate fashion. The latter has a higher vibrational frequency than the symmetric mode (by about 100 wavenumbers) since the asymmetric vibration is slightly more energetically costly. The more facile angle-bending deformation has the lowest frequency in water among these three modes.

### 9.2.2 Complex Biomolecular Spectra

As the number of atoms in a molecule increases, so does the number of modes, as well as the associated complexity of the vibrational spectrum. Assigning normal modes to observed peaks in the experimental spectra becomes more difficult. Help is available from the characteristic modes of small molecules, which serve as excellent references for interpretation. In addition, the intensities in the vibrational spectrum can be calculated, at least to a rough approximation, by theoretical techniques [767].

While vibrational frequencies for the same bond type (e.g., O–H) vary depending on the molecular context, general values can be assigned to basic two-atom and three-atom sequences separated into distinct bond types (single, double, hydrogen-bonded, etc.).

For example, the symmetric O–H stretch in one water molecule (water vapor) is about 50 wavenumbers higher than that in the H–O–Cl molecule, but 300 wavenumbers higher than the O–H stretching frequency of a hydrogen-bonded water molecule in liquid water and ice (O–H · · · O), where  $\nu = 3400 \text{ cm}^{-1}$ . This reduction is due to the attractive force acting on the hydrogen atom in hydrogen-bonded species, since such attraction reduces the energy and hence the frequency of the O–H stretching motion.

### 9.2.3 Spectra As Force Constant Sources

Such spectroscopic measurements and analyses are used to derive appropriate force constants for biomolecular force fields. Tables 9.1 and 9.2 display examples of approximate stretching (Table 9.1) and bending and torsional (Table 9.2) frequencies.

Table 9.1. Characteristic **stretching** vibrational frequencies.

Vibrational Mode	Frequency [cm <sup>-1</sup> ]
H–O stretch	3600–3700
H–N stretch	3400–3500
H–C stretch	2900–3000
H–Br stretch	2650
C≡C, C≡N stretch	2200
C=C, C=O stretch	1700–1800
C–N stretch	1250
C–C stretch	1000
C–S stretch	700
S–S stretch	500

Table 9.2. Characteristic **bending and torsional** vibrational frequencies.

Vibrational Mode	Frequency [cm <sup>-1</sup> ]
H–O–H, H–N–H bend	1600
H–C–H bend	1500
H–C–H scissor	1400
H–C–H rock	1250
H–C–H wag	1200
H–S–H bend	1200
O–C=O bend	600
C–C=O bend	500
S–S–C bend	300
C=C torsion	1000
C–O torsion	300–600
C–C torsion	300
C–S torsion	200

Vibrational spectra of alkane molecules are a good source of parameters for C–C and C–H vibrational modes in proteins and nucleic acids. The spectrum of a butane molecule ( $\text{CH}_3\text{—CH}_2\text{—CH}_2\text{—CH}_3$ ), for example, reflects both methyl ( $\text{CH}_3$ ) and methylene ( $\text{CH}_2$ ) stretching and bending modes.

The alkane frequencies can be grouped into two strong stretching modes slightly below  $3000 \text{ cm}^{-1}$ , symmetric and asymmetric bending deformations within the range  $1350\text{--}1500 \text{ cm}^{-1}$ , and a C–C stretching mode around  $1000 \text{ cm}^{-1}$ . Indeed, these ranges of modes are evident in simulation-computed

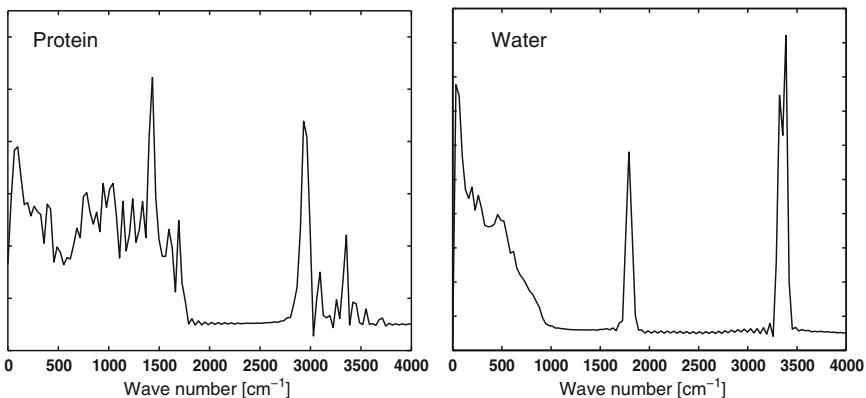


Figure 9.2. Characteristic frequencies calculated over 5 ps molecular dynamics simulations of solvated BPTI for the protein (left) and water (right) atoms. Data are from [1089].

spectra corresponding to a small solvated protein, bovine pancreatic trypsin inhibitor (BPTI) as seen in Figure 9.2 [1089].<sup>1</sup>

#### 9.2.4 In-Plane and Out-of-Plane Bending

Bending modes include two types of in-plane deformations: *scissoring* and *rocking*, and two out-of-plane deformations: *wagging* and *twisting* (see Figure 9.3).

The in-plane *scissoring* deformation of an X–Y–Z sequence makes atoms X and Z move closer together. The *rocking* deformation moves both atoms in one direction while keeping their distance about the same.

The *out-of-plane wagging* bending deformation moves these atoms in the same direction with respect to the reference plane. *Twisting* moves one atom in one direction and the other in the opposite direction.

Figure 9.2 shows the power spectrum of a solvated protein system (BPTI) as computed from Fourier transforms of velocity autocorrelation functions [1089]. The calculated peak *locations* depend sensitively on the force field, assuming the simulation protocol and frequency calculation procedure are sound.<sup>2</sup>

For water, we note in Figure 9.2 characteristic peaks for vibrational modes of stretching around  $3500\text{ cm}^{-1}$  and bending around  $1700\text{ cm}^{-1}$ , as well as the

---

<sup>1</sup>Essentially, vibrational spectra can be computed from molecular dynamics simulations by transforming time-dependent properties, such as velocity autocorrelation functions, into the frequency domain using Fourier transforms. See [1089], for example, for the precise procedure.

<sup>2</sup>The simulation-derived peak *heights* can, at best, reproduce frequency values corresponding to the force field parameters, not the experimental values, though the latter often serve as a reference. For example, the bond stretching force constant used for O–H water bonds may be physically unrealistic, so an unnatural spectral peak may emerge from simulations using unconstrained O–H bonds. (The peak is absent if these bonds are constrained).

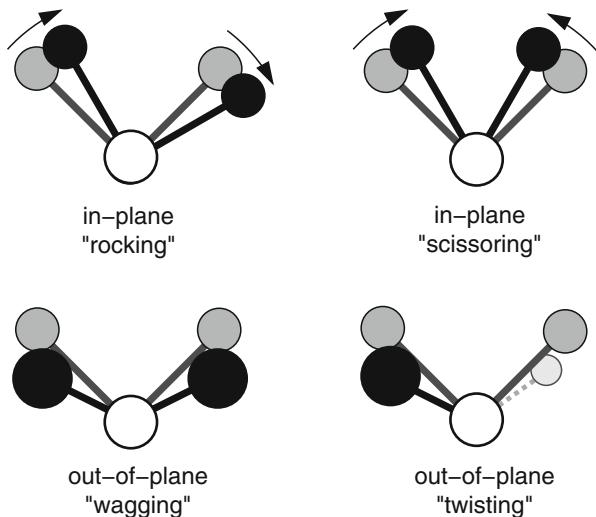


Figure 9.3. Various in-plane and out-of-plane bending vibrational modes. The reference molecular position is shown with grey-shaded atoms, and the position after the move is shown in black (and very-light grey for twisting). For the *wag*, the two nonbonded atoms move out of the paper plane toward us, while for *twist* one atom moves up (toward us) and the other down (away from us, below the paper plane).

slower *tumbling* or rocking (*librational*) motion for the water molecules as a whole in liquid water. For the protein, a wide range of vibrations is captured, from the fastest stretching modes for bonds involving hydrogens (O–H and N–H) around  $3300\text{ cm}^{-1}$  to various angle-bending modes below  $1700\text{ cm}^{-1}$ , to much slower deformations.

### 9.3 Bond Length Potentials

Bond length potentials can be considered as “strain” terms that model small-scale deviations about reference values. The reference values for different chemical bonds can be obtained from solved X-ray crystal structures as well as from quantum mechanical solutions to equilibrium structures of small molecules. For accepted values, see organic chemistry textbooks such as [878] and Handbooks of Chemistry and Physics (e.g., CRC Press Handbook of Chemistry and Physics).

Note that *ab initio* methods calculate *equilibrium* bond lengths whereas experimentally measured values are usually vibrationally averaged bond lengths; the averaging depends on the experiment and hence there are many bond length values in theory. See [795, 797, 798] for a discussion of these different procedures for determining bond lengths and for the interconversion among the bond length values. For macromolecules, these small differences among the reference values are not usually important.

### 9.3.1 Harmonic Term

The harmonic potential modeled after Hooke's law is the simplest molecular-mechanics formulation for bond deformations. According to Hooke's law, the force  $F$  is proportional to the displacement,  $x$ , and the acceleration,  $\ddot{x}$ , as follows:

$$F(x) = -kx = m \frac{d^2x}{dt^2}, \quad k = m\omega^2 > 0. \quad (9.1)$$

Thus, the *angular frequency*  $\omega$  (number of radians per unit time), or  $2\pi$  times the *circular frequency*  $\nu$  ( $= c/\lambda$  where  $c$  is the speed of light and  $\lambda$  is the wavelength), is related to the spring constant  $k$  as:

$$\omega \equiv 2\pi\nu = \sqrt{k/m}. \quad (9.2)$$

The corresponding potential energy  $E$  is:

$$E(x) = \frac{k}{2} x^2. \quad (9.3)$$

More generally, we write this harmonic bond potential as:

$$E_{\text{harmonic}}^r(r) = S_h [r - \bar{r}]^2, \quad (9.4)$$

where  $r$  is the bond length,  $\bar{r}$  is the reference bond-length value, and  $S_h$  is a constant. From the measured mass and frequency for a particular bond vibration, force constants can be derived accordingly ( $k = m\omega^2$ ). For atomic pairs of different species (with masses  $m_1, m_2$ ), the general guideline is

$$k = \mu\omega^2, \quad (9.5)$$

where  $\mu$  is the *reduced mass* defined as:

$$\mu = (m_1 m_2) / (m_1 + m_2). \quad (9.6)$$

The harmonic potential is only adequate for small deviations from reference values, around one vibrational level above the ground state, or on length deformations of the order of 0.1 Å or less. It is not valid for larger deviations from equilibrium since the atoms dissociate and no longer interact; thus the energy levels off, rather than increases, rapidly as the distance increases beyond  $\bar{r}$ . For very small interaction distances, however, the deformation energy is very large.

This physical picture is described by a parabolic potential-well shape for small distance separations and a leveling curve for separations greater than the reference value. Rather than a harmonic function, the precise shape of  $E(r)$  is more adequately represented by higher-order functions that approximate better the experimentally-measured energy trend as a function of distance.

Electronic spectroscopy is necessary to measure this precise distance dependency since frequencies and energies are higher for changes in electronic states. Specifically, the potential energy as a function of internuclear separation for both the ground and excited states is obtained from visible and ultraviolet (UV)

spectroscopic techniques.<sup>3</sup> The commonly used ultraviolet and visible spectrometers measure absorption of light in the range 200–750 nm. Visible and UV absorption bands are broad since several vibrational states are contained in each electronic state, and each vibrational state is further decomposed into many rotational states. Thus, a rigorous assignment of bands to specific transitions is difficult, but the overall shape of the energy well can be deduced.

### 9.3.2 Morse Term

To model bond deformations that exceed very small fluctuations about equilibrium states, the empirical bond potential due to P.M. Morse [879] has proven very successful for reproducing vibrational levels of small molecules [764, 1201]. The Morse function has the form:

$$E_{\text{Morse}}^r(r) = D \{1 - \exp[-S_m(r - \bar{r})]\}^2, \quad (9.7)$$

where the adjustable parameters  $S_m$  and  $D$  characterize the well width and well depth respectively. As seen in Figure 9.4, the Morse potential correctly rises steeply for contraction of the bond length ( $E(r) \rightarrow \infty$  as  $r \rightarrow 0$ ) but levels off to the *dissociation energy*  $D$  at large  $r$ :  $E(r) \rightarrow D$  as  $r \rightarrow \infty$ .

The empirical Morse potential can be written as an infinite series in the powers of the bond displacement ( $r - \bar{r}$ ). Using Taylor series, we expand eq. (9.7) as:

$$\begin{aligned} E_{\text{Morse}}^r(r) &= D \{1 - 2 \exp[-S_m(r - \bar{r})] + \exp[-2S_m(r - \bar{r})]\} \\ &= D \left\{ 1 - 2 \left[ 1 - S_m(r - \bar{r}) + \frac{S_m^2(r - \bar{r})^2}{2} - \frac{S_m^3(r - \bar{r})^3}{3!} \right. \right. \\ &\quad \left. \left. + \frac{S_m^4(r - \bar{r})^4}{4!} + \dots \right] \right. \\ &\quad \left. + \left[ 1 - 2S_m(r - \bar{r}) + \frac{4S_m^2(r - \bar{r})^2}{2} - \frac{8S_m^3(r - \bar{r})^3}{3!} \right. \right. \\ &\quad \left. \left. + \frac{16S_m^4(r - \bar{r})^4}{4!} + \dots \right] \right\} \\ &= DS_m^2(r - \bar{r})^2 - DS_m^3(r - \bar{r})^3 + \frac{7}{12} DS_m^4(r - \bar{r})^4 \\ &\quad + \mathcal{O}(r - \bar{r})^5. \end{aligned} \quad (9.8)$$

Hence we can relate the harmonic stiffness constant  $S_h$  in eq. (9.4) to the Morse constant  $S_m$  of eq. (9.7) as:

$$S_h \approx D(S_m)^2. \quad (9.9)$$

---

<sup>3</sup>UV spectroscopy measures wavelengths just beyond the violet end of the visible spectrum, that is, with  $\lambda < 400$  nm.

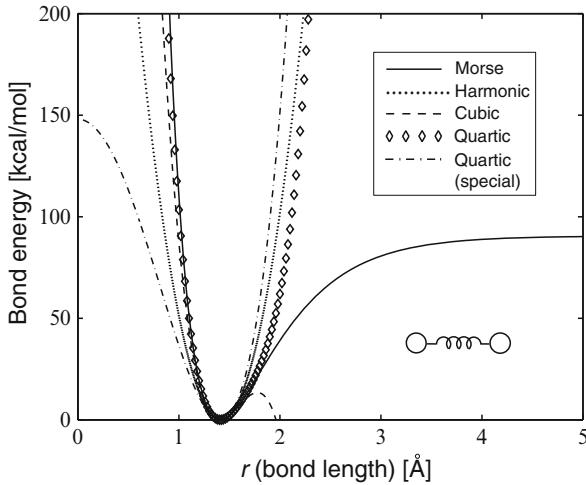


Figure 9.4. Morse, harmonic, cubic, and two quartic bond potentials for H–Br. The Morse, harmonic, and special quartic potentials are given in eqs. (9.7), (9.4), and (9.11), respectively. The cubic and non-degenerate quartic polynomials are given polynomial coefficients to match the Taylor-series expansion of the Morse potential given in eq. (9.8) up to the desired order. Note that the cubic potential causes a problem for significant bond stretches because of the change in curvature.

Put another way, from the relationship between  $S_h$  and the spring constant of a harmonic oscillator, namely  $S_h = k/2 = (\mu\omega^2)/2$ , the value of the Morse well-depth parameter  $S_m$  can be reasonably approximated as  $\sqrt{S_h/D}$  or

$$S_m = \omega \sqrt{\frac{\mu}{2D}} = \pi\nu \sqrt{\frac{2\mu}{D}}. \quad (9.10)$$

Figure 9.4 shows the harmonic and Morse potentials for a hydrogen bromide molecule. The parameters  $D = 90.5$  kcal/mol,  $S_m = 1.814 \text{ \AA}^{-1}$ ,  $\bar{r} = 1.41 \text{ \AA}$ , and  $S_h = 297.8$  (kcal/mol)/ $\text{\AA}^2$  (from eq. (9.9)), are used. We note that the harmonic potential is a good approximation to the energy surface only for small displacements from equilibrium.

### 9.3.3 Cubic and Quartic Terms

To reproduce Morse potentials better than possible with the harmonic potential, cubic and quartic polynomials can be used (through terms added to the quadratic potential) to match the Taylor series expansion of eq. (9.8) up to a desired order, as shown in Figure 9.4. The MM3 force field, for example, uses cubic and quartic bond potentials, which work well for most molecules [30]; a sextic bond potential works even better and is used in MM4. The Merck force field, MMFF [497], uses a quartic bond function. Note that for better optimization of the shape of the bond

length potential function, coefficients of the polynomial can be adjusted; thus they need not coincide with those given by the Morse Taylor expansion of eq. (9.8).

Note that a quartic is preferable to a cubic bond potential because the cubic function has an inflection point at some value  $r > \bar{r}$ ; thus, significant bond stretches lead to negative rather than positive energy ( $E \rightarrow -\infty$  as  $r \rightarrow \infty$ ) (see Figure 9.4). This can cause the molecular energy to have large negative values and the computation (energy minimization, for example) to become nonsensical. Series that end in even powers (like quartic rather than cubic polynomials) can provide better approximations and drive the molecule more rapidly toward the energy minimum.

For large-molecule force fields where computational time is important, a special quartic has been suggested to avoid square root computations [1101, 1103]; it measures the square of the squared bond differences as:

$$E_{\text{quartic}}^r = S_q[r^2 - \bar{r}^2]^2. \quad (9.11)$$

This quartic is special (degenerate) since it does not contain a cubic term. Its shape is therefore similar to the harmonic potential, as shown in Figure 9.4.

At small displacements from equilibrium, we have the relation

$$[r^2 - \bar{r}^2]^2 \equiv [r - \bar{r}]^2 [r + \bar{r}]^2 \approx [r - \bar{r}]^2 (2\bar{r})^2.$$

Hence, comparing the series expansion in eq. (9.11) with the harmonic potential of eq. (9.4), we can relate the quartic-potential force constant to that of the harmonic potential by:

$$S_q \approx S_h / 4\bar{r}^2. \quad (9.12)$$

Figure 9.4 displays this quartic potential with  $S_q$  calculated from  $S_h$  as above. As for the quadratic potential, the energy approximation is good only for very small deviations from equilibrium.

## 9.4 Bond Angle Potentials

The bond angle arrangement around each atom in a molecule is governed by the hybridization of the orbitals around the atom. For example, when an atom has two identical hybrid orbitals ( $sp$ ) around it (e.g., Be in  $\text{BeCl}_2$ ), the bond angle is  $180^\circ$ . When three identical orbitals surround an atom (e.g., B in  $\text{BF}_3$ ,  $sp^2$ ), the arrangement is trigonal and coplanar with bond angles all  $120^\circ$ . When four identical orbitals surround an atom (e.g., C in  $\text{CH}_4$ ,  $sp^3$ ), the arrangement is tetrahedral—all angles are  $109.47^\circ$  ( $\theta_{\text{tet}} = \cos^{-1}[-1/3]$ ).

This simple rule serves as a first approximation for bond angle geometries. However, small deviations from these estimates generally occur, and large deviations sometimes occur. Even small differences of  $1\text{--}2^\circ$  between different bond angles in a molecule can have important global influence on molecular structure, as in riboses.

Indeed, it is important to realize that exact  $sp^3$  orbitals exist only for tetrahedrally-symmetric compounds like methane. Ordinary alkanes already have their orbits deformed: propane, for example, has the C–C–C bond angle of about  $112.5^\circ$  and H–C–H bond angles around  $107.5^\circ$ ; its C–C–H bond angles are approximately tetrahedral. Another common example of bond angle deviations involves ring molecules, like cycloalkanes and riboses; ring-closure constraints can alter the geometry significantly. Finally, electron lone pairs about atoms influence the geometry: in water, the oxygen lone pair forms bond-like orbitals to produce the liquid water bond angle  $\theta(\text{H}–\text{O}–\text{H}) \approx 105^\circ$ . See Figure 9.5 for such illustrations. As for bond potentials, stiffness constants for bond angle bending are determined from measured vibrational frequencies.

#### 9.4.1 Harmonic and Trigonometric Terms

Commonly used bond-angle potentials are harmonic functions that involve the difference between angles and angle cosines:

$$E_{\text{harmonic}}^\theta(\theta) = K_h [\theta - \bar{\theta}]^2, \quad (9.13)$$

$$E_{\text{trig.}}^\theta(\theta) = K_t [\cos \theta - \cos \bar{\theta}]^2. \quad (9.14)$$

As shown above for bond potentials, we can expand the trigonometric function above by a Taylor series in powers of  $\theta - \bar{\theta}$  to relate  $K_t$  to  $K_h$ :

$$E_{\text{trig.}}^\theta(\theta) = K_t \left\{ -\sin \bar{\theta} (\theta - \bar{\theta}) - \frac{\cos \bar{\theta}}{2} (\theta - \bar{\theta})^2 + \dots \right\}^2 \implies K_t \approx K_h \sin^2 \bar{\theta}. \quad (9.15)$$

The advantage of the trigonometric potential is its boundedness and its ease of implementation and differentiation. This is because no inverse trigonometric functions need to be calculated, and singularity problems for linear bond angles can be avoided [1102, 1103]. As a compromise between a quadratic and infinite series in  $\theta - \bar{\theta}$ , the MMFF force field uses a cubic bond-angle function of form [497]

$$E_{\text{cubic}}^\theta = K_1(\theta - \bar{\theta})^2 + K_2(\theta - \bar{\theta})^3 \quad (9.16)$$

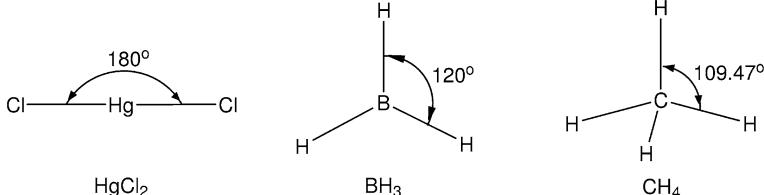
for non-collinear atom orientations. For linear (or near linear) reference angles, the function

$$E_{\text{trig.}}^\theta = K_3(1 + \cos \theta) \quad (9.17)$$

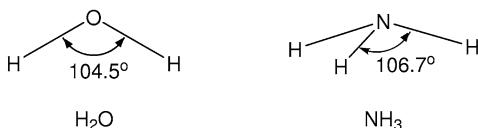
is used instead.

Note that, as for the bond potential, it is hazardous to use a function that ends in an odd power of the deformation since these terms have negative coefficients, which dominate for large deviations. Hence, potential functions that end in even powers are preferable.

## Symmetric



## Asymmetric



## Rings

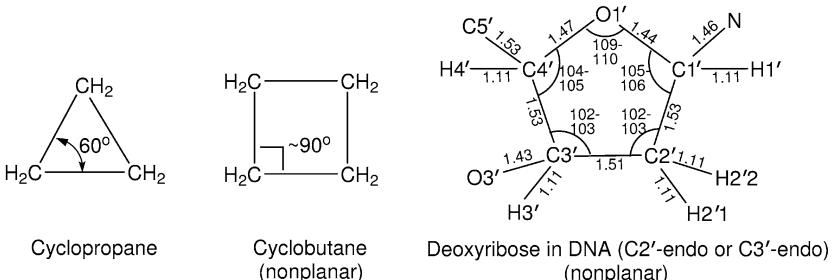


Figure 9.5. Bond angle geometries for simple chain and cyclic molecules. Note that cyclobutane and deoxyribose are nonplanar. The former has bond angles less than  $90^\circ$  by only  $1\text{--}2^\circ$ , but the dihedral angle is substantial, around  $30^\circ$ , which relieves the eclipsing of the hydrogens substantially. The geometry shown for deoxyribose (bond lengths in Å and bond angles in degrees) corresponds to observed C3'-endo and C2'-endo conformations in B-DNA. The five endocyclic deoxyribose dihedral angles  $\nu_0$  through  $\nu_5$  have values (as computed from solvated dodecamers in CHARMM) of approximately  $-24, 38, -38, 24$ , and  $0$  degrees for C2'-endo, and  $0, -24, 38, -38$ , and  $24$  degrees for the C3'-endo sugar pucker.

Figure 9.6 displays harmonic bond angle potentials of the forms given in eqs. (9.13) and (9.14). Note that the harmonic cosine potential is very similar to the harmonic potential for a small range of fluctuations. It should thus be preferred in practice if computational time is an issue.

#### 9.4.2 Cross Bond Stretch / Angle Bend Terms

Cross terms are often used in force fields targeted to small molecular systems (e.g., MM3, MM4) to model more accurately compensatory trends in related bond-length and bond-angle values. These cross terms are considered correction terms to the bond-length and bond-angle potentials.

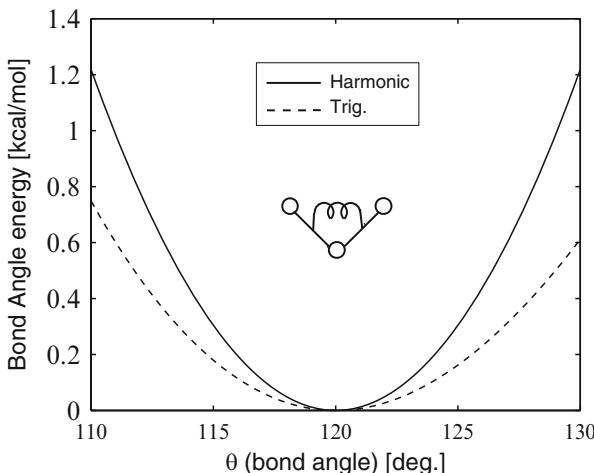


Figure 9.6. Harmonic bond-angle potentials of the form (9.13) and (9.14) for an aromatic C–C–C bond angle (CA–CA–CA atomic sequence in CHARMM) with parameters  $K_h = 40 \text{ kcal}/(\text{Mol}\cdot\text{rad}^2)$  and  $\bar{\theta} = 2.1 \text{ rad}$  ( $120^\circ$ ). The  $K_t$  force constant for eq. (9.14) is calculated via eq. (9.15).

For example, a stretch/bend term for a bonded-atom sequence  $ijk$  allows bond lengths  $i-j$  and  $j-k$  to increase/decrease as  $\theta_{ijk}$  decreases/increases. A bend/bend potential couples the bending vibrations of the two angles centered on the same atom appropriately, so the corresponding frequencies can be split apart to match better the experimental vibrational spectra [30].

These correlations can be modeled via stretch/stretch, bend/bend, and stretch/bend potentials for such  $ijk$  sequences (see Figure 9.7), where the distances  $r$  and  $r'$  are associated with bonds  $ij$  and  $jk$ , and  $\theta$  is the  $ijk$  bond angle:

$$E^{rr'}(r, r') = S[r - \bar{r}][r' - \bar{r}'], \quad (9.18)$$

$$E^{\theta\theta'}(\theta, \theta') = K[\theta - \bar{\theta}][\theta' - \bar{\theta}'], \quad (9.19)$$

$$E^{r\theta}(r, \theta) = SK[r - \bar{r}][\theta - \bar{\theta}]. \quad (9.20)$$

Through addition, a stretch/bend/stretch term of form

$$E^{r\theta r'}(r, \theta, r') = K[S(r - \bar{r}) + S'(r' - \bar{r}')][\theta - \bar{\theta}], \quad (9.21)$$

can be mimicked, as in the Merck force field [497].

A variation of a stretch/bend term, devised to obtain better agreement of calculated with experimental vibrational frequencies for a bonded atom sequence  $ijk$  and associated bond angle  $\theta$  [301], is a potential known as Urey-Bradley.

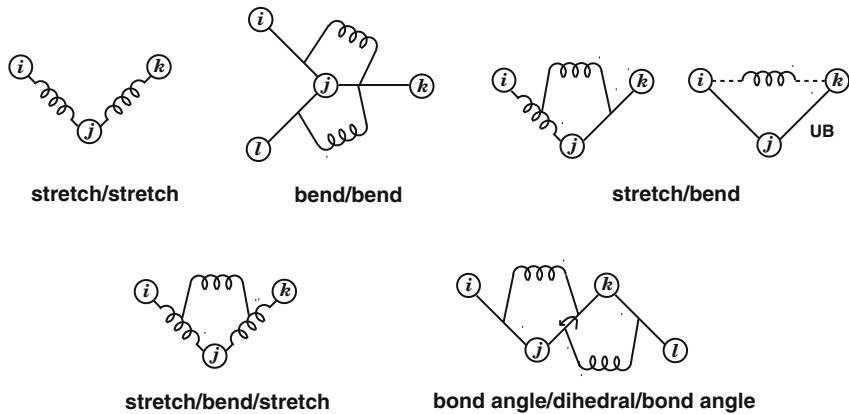


Figure 9.7. Schematic illustrations for various cross terms involving bond stretching, angle bending, and torsional rotations.

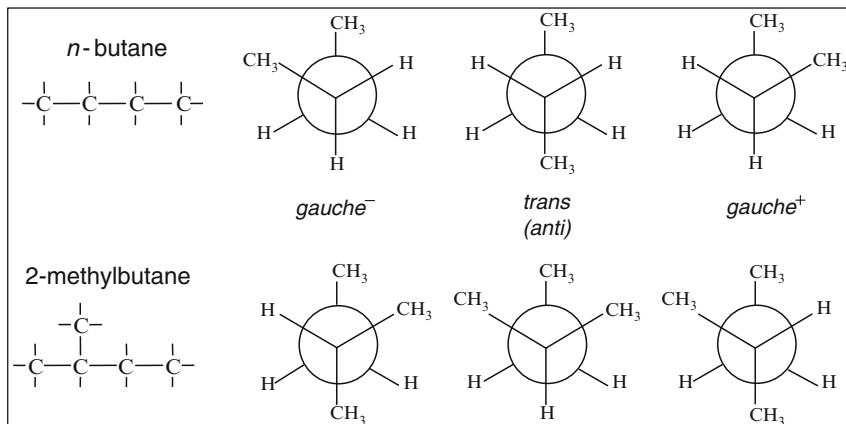


Figure 9.8. Torsional orientations for *n*-butane and 2-methylbutane, as illustrated by Newman projections viewed along the central C–C bond. Note that while *n*-butane has the global minimum at the *trans* (or *anti*) configuration, the *gauche* states of 2-methylbutane are lower in energy than the corresponding *trans* (symmetrical) state since the methyl groups are better separated and hence the molecule is less congested.

It is commonly used in molecular mechanics force fields, like CHARMM. The potential is a simple harmonic function of the *interatomic* (not bond) distance  $\rho$ , between atoms  $i$  and  $k$  in the bonded  $ijk$  sequence (see Figure 9.7):

$$E^\rho(\rho) = S [\rho - \bar{\rho}] . \quad (9.22)$$

## 9.5 Torsional Potentials

### 9.5.1 Origin of Rotational Barriers

Torsional potentials are included in potential energy functions since all the previously described energy contributions (nonbonded, stretching, and bending terms) do not adequately predict the torsional energy of even a simple molecule such as ethane. In ethane, the C–C bond provides an axis for *internal rotation* of the two methyl groups.

Although the origin of the barrier to internal rotation has not been resolved [472, 1010, 1360],<sup>4</sup> explanations involve the relief of steric congestion and optimal achievement of resonance stabilization (another quantum-mechanical effect arising from the interactions of electron orbitals) [1010]. This latter consideration represents a newer hypothesis. For a long time, the primary interactions that give rise to rotational barriers have been thought to be repulsive interactions, caused by overlapping of bond orbitals of the two rotating groups [1002].

In ethane ( $\text{C}_2\text{H}_6$ ), for example, the torsional strain about the C–C bond is highest when the two methyl groups are nearest, as in the *eclipsed* or *cis* state, and lowest when the two groups are optimally separated, as in the *anti* or *trans* state (see Figure 9.8 and the related illustration for *n*-butane in Figure 3.13 of Chapter 3). Without a torsional potential in the energy function, the eclipsed form for ethane is found to be higher in energy than the staggered form, but not sufficiently higher to explain the observed staggered preference [24]. See [27] for a meticulous theoretical treatment of the torsional potential of butane, showing that a large basis set combined with a treatment of electron correlations in *ab initio* calculations is necessary to obtain accurate equilibrium energies relative to the *anti* conformation.

Accounting accurately for rotational flexibility of atomic sequences is important since rotational states affect biological reactivity. Such torsional parameters are obtained by analyzing the energetic profile of model systems as a function of the flexible torsion angle. The energy differences between rotational isomers are used to parameterize appropriate torsional potentials that model internal rotation.

### 9.5.2 Fourier Terms

The general function used for each flexible torsion angle  $\tau$  has the form

$$E^\tau(\tau) = \sum_n \frac{V_n}{2} [1 \pm \cos n\tau], \quad (9.23)$$

where  $n$  is an integer. For each such rotational sequence described by the torsion angle  $\tau$ , the integer  $n$  denotes the *periodicity* of the rotational barrier, and

---

<sup>4</sup>Goodman *et al.* [472] refer to the barrier origin as “the Bermuda Triangle of electronic theory”, reflecting a complex interdependence among three factors: electronic repulsion, relaxation mechanisms, and valence forces.

$V_n$  is the associated *barrier height*. The value of  $n$  used for each such torsional degree of freedom depends on the atom sequence and the force field parameterization (see below). Typical values of  $n$  are 1, 2, 3 (and sometimes 4). Other values (e.g.,  $n = 5, 6$ ) are used in addition by some force fields (e.g., CHARMM [805]). A reference torsion angle  $\tau_0$  may also be incorporated in the formula, i.e.,

$$E^\tau(\tau) = \sum_n \frac{V_n}{2} [1 + \cos(n\tau - \tau_0)]. \quad (9.24)$$

Often,  $\tau_0 = 0$  or  $\pi$  and thus the cosine expression of form (9.23) suffices; this is because eq. (9.24) can be reduced to the form of eq. (9.23) in such special cases, from relations like:

$$1 + \cos(n\tau - \pi) = 1 - \cos(n\tau).$$

Experimental data obtained principally by spectroscopic methods such as NMR, IR (Infrared Radiation), Raman, and microwave, each appropriate for various spectral regions, can be used to estimate barrier heights and periodicities in low molecular weight compounds. According to a theory developed by Pauling [972], potential barriers to internal rotation arise from exchange interactions of electrons in adjacent bonds; these barriers are thus similar for molecules with the same orbital character. This theory has allowed tabulations of barrier heights as class averages [871, 898, 1152, 1153]. Since barriers for rotations about various single bonds in nucleic acids and proteins are not available experimentally, they must be estimated from analogous chemical sequences in low molecular weight compounds.

### 9.5.3 Torsional Parameter Assignment

In current force fields, these parameters are typically assigned by selecting several classes of model compounds and computing energies as a function of the torsion angle using *ab initio* quantum-mechanical calculations combined with geometry optimizations. The final value assigned in the force field results from optimization of the combined intramolecular and nonbonded energy terms to given experimental vibrational frequencies and measured energy differences between conformers of model compounds.

This procedure often results in several Fourier terms in the form (9.23), that is, several  $\{n, V_n\}$  pairs for the same atomic sequence; see examples in Table 9.3. Moreover, parameters for a given quadruplet of atoms may be deduced from more than one quadruplet entry. For example, the quadruplet C1'-C2'-C3'-O3' in nucleic acid sugars may correspond to both a C1'-C2'-C3'-O3' entry and a  $\star$ -C-C- $\star$  entry, the latter designating a general rotation about the endocyclic sugar bond; here,  $\star$  designates any atom. When general rotational sequences are involved (e.g.,  $\star$ -C-N- $\star$ ), CHARMM may list a *pair* of  $\{V_n, \tau_0\}$  values (i.e., different  $\tau_0$  for the same rotational term); only one  $\tau_0$  is used when a specific quadruplet atom sequence is specified.

### Twofold and Threefold Sums

Twofold and threefold potentials are most commonly used (see Figure 9.9 for an illustration). Inclusion of a one-fold torsional term as well is a force-field dependent choice.

A threefold torsional potential exhibits three maxima at 0°, 120°, and 240° and three minima at 60°, 180°, and 300°. Ethane has a simple 3-fold torsional energy profile, as seen in the 3-fold energy curve in Figure 9.9, with all maxima energetically equivalent and corresponding to the eclipsed or *cis* form, and all minima corresponding to the energetically equivalent staggered or *trans* form.

In related hydrocarbon sequences, the local minima may not be of equal energies. For *n*-butane, for example, the *anti* (or *trans*) conformer is roughly 1 kcal/mol lower in energy than the two *gauche* states. In 2-methylbutane, the *gauche* states are lower in energy than the *anti* state due to steric effects. The torsional conformations of these molecules are illustrated in Figure 9.8.

In general, the *gauche* forms may be higher in energy than the *trans* form, as in *n*-butane [1153], or lower in energy, as in 2-methylbutane due to steric effects. Certain X–C–C–Y atomic sequences where the X or Y atoms are electronegative like O or F (e.g., 1,2-difluoroethane, certain O–C–C–O linkages in nucleic acids) [532] also have lower-in-energy *gauche* conformations, but this is due to a fundamentally different effect than present in hydrocarbons — neither steric, nor dipole/dipole — called the *gauche effect* [532].

### Reproduction of *Cis/Trans* and *Trans/Gauche* Energy Differences

A combination of a twofold and a threefold potential can be used to reproduce both the *cis/trans* and *trans/gauche* energy differences. Figure 9.9 illustrates the case for different parameter combinations of the twofold and threefold parameters. One interaction is modeled after the O–C–C–O sequence in nucleic acids (e.g., O3'–C3'–C2'–O2' in ribose) [937], showing a minimum at the *trans* state. The other is a rotation about the phosphodiester bond (P–O) in nucleic acids, showing a shallow minimum at the *trans* state. Note that for small molecules, torsional potentials with  $n = 1, 2, 3, 4$  and 6 are often used to reproduce torsional frequencies accurately.

To see how to combine twofold and threefold torsional potentials, for example, denote the experimental energy barrier by  $\Delta V$ , the total empirical potential energy function as  $E$ , and let  $E^\tau$  represent the following combination of twofold and threefold torsional terms:

$$E^\tau = \frac{V_2}{2} [1 + \cos 2\tau] + \frac{V_3}{2} [1 + \cos 3\tau]. \quad (9.25)$$

The torsional parameters  $V_2$  and  $V_3$  for a given rotation  $\tau$  are then computed from the relations:

$$\begin{aligned} \Delta V_{\text{cis/trans}} &= E_{\tau=0^\circ} - E_{\tau=180^\circ} \\ &= V_3 + [E - E^\tau]_{\tau=0^\circ} - [E - E^\tau]_{\tau=180^\circ}, \end{aligned} \quad (9.26)$$

$$\begin{aligned}\Delta V_{\text{trans/gauche}} &= E_{\tau=180^\circ} - E_{\tau=60^\circ} \\ &= (3V_2)/4 + [E - E^\tau]_{\tau=180^\circ} - [E - E^\tau]_{\tau=60^\circ},\end{aligned}\quad (9.27)$$

since

$$V_3 = E_{\tau=0^\circ}^\tau - E_{\tau=180^\circ}^\tau$$

and

$$\frac{3}{4}V_2 = E_{\tau=180^\circ}^\tau - E_{\tau=60^\circ}^\tau.$$

Thus, Cartesian coordinates of the molecule must be formulated in terms of  $\tau$ . A simplification can be made by assuming fixed bond lengths and bond angles and calculating nonbonded energy differences only. In theory, then, every different parameterization of the nonbonded coefficients requires an estimation of the torsional potentials  $V_2$  and  $V_3$  to produce a consistent set.

*This general requirement in development of consistent force fields explains why energy parameters are not transferable from one force field to another.*

In general, many classes of model compounds must be used to represent the various torsional sequence in proteins and nucleic acids. Since rotational barriers in small compounds have little torsional strain energy compared with large systems, rotational profiles are routinely computed for a series of substituted molecules. Substituted hydrocarbons are used, for example, to determine rotational parameters for torsion angles about single C–C bonds in saturated species.

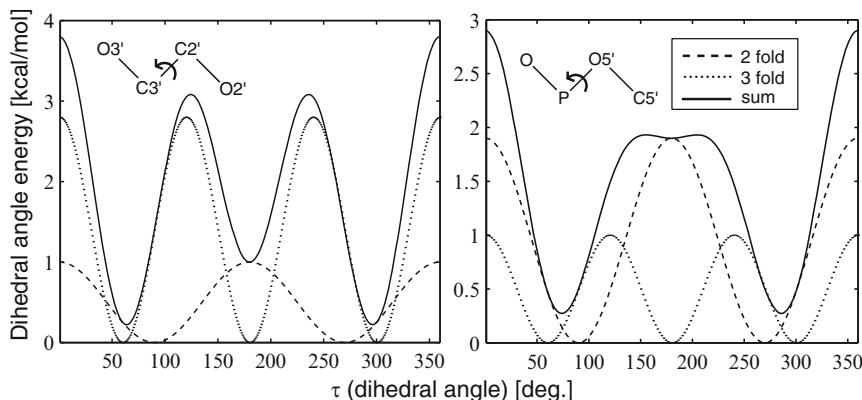


Figure 9.9. Twofold and threefold torsion-angle potentials and their sums for an O–C–C–O rotational sequence in nucleic-acid riboses ( $V_2 = 1.0$  and  $V_3 = 2.8$  kcal/mol, reproducing the known *trans/gauche* energy difference [937]) and a rotation about the phosphodiester (P–O) bond in nucleic acids ( $V_2 = 1.9$  and  $V_3 = 1.0$  kcal/mol, from CHARMM [805]).

## Model Compounds

Examples of model compounds used for assigning torsional parameters to nucleic acids and proteins include the following (see Figure 9.10).

- Hydrocarbons like **ethane**, **propane**, and ***n*-butane**, and the more crowded environments of **2-methylbutane** and **cyclohexane**: for rotations about single C–C bonds in saturated species (i.e., each carbon is approximately tetrahedral), such as C–C–C–C, H–C–C–H, and H–C–C–C;
  - The **ethylbenzene** ring: for rotations about CA–C bonds in ring systems (CA denotes an aromatic carbon);
  - Alcohols like **methanol** and **propanol**: to model H–C–O–H, C–C–O–H, H–C–C–O, and C–C–C–O sequences (e.g., in the amino acids serine and threonine);

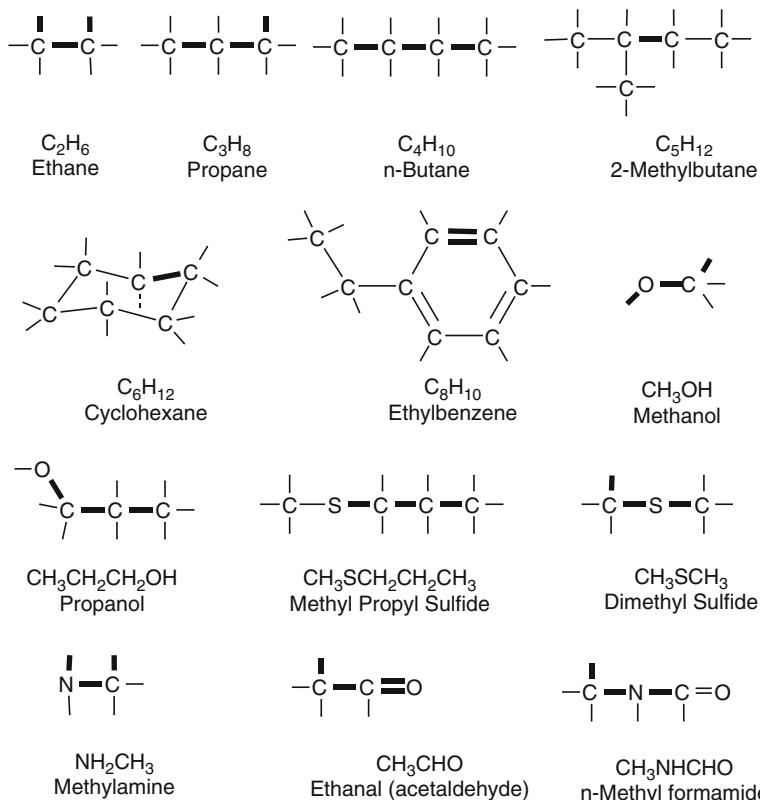


Figure 9.10. Model compounds for determining torsional parameters for various atomic sequences in biomolecules. Illustrated in **bold** is  $\tau$ : the bond about which the  $\tau$  rotation occurs, or the complete  $\tau$  sequence (3 bonds), as needed for an unambiguous definition.

- Sulfide molecules (e.g., **methyl propyl sulfide**): to model rotations involving sulfur atoms (e.g., C–C–C–S, C–C–S–C, and H–C–S–C) as in methionine, or **dimethyl disulfide** (C–S–S–C), to model disulfide bonds in proteins;
- Model amine, aldehydes, and amides (e.g., **methylamine**, **ethanal**, **N-methyl formamide**): for sequences involving nitrogens such as H–C–N–H, H–C–C=O, and H–C–N–C.

See [842] for a comprehensive description of such a parameterization for peptides based on *ab initio* energy profiles.

In the CHARMM and AMBER force fields, the parameters  $n$  and  $V_n$  are highly environment dependent. Furthermore, more than one potential form can be used for the same bond about which rotation occurs; these different torsion-angle terms may be weighted proportionately. Some examples of torsion angle parameters from the AMBER [265] and CHARMM [805, 809] force fields are given in Table 9.3.

Table 9.3. Parameters for selected torsional potentials from the AMBER (first row for each sequence) [218] and CHARMM (second row of that sequence) [415, 805] force fields. Barrier heights are in units of kcal/(mol rad<sup>2</sup>), angles are in radians, and \* represents any atom.

SEQUENCE	$\frac{V_1}{2}$	$\tau_0$	$\frac{V_2}{2}$	$\tau_0$	$\frac{V_3}{2}$	$\tau_0$	DESCRIPTION
*–C–C–*					1.4	0	alkane C–C (e.g., Lys or Leu C <sup>α</sup> –C <sup>β</sup> –C <sup>γ</sup> –C <sup>δ</sup> )
O4'–C1'–N9–C8	2.5 1.1	0 0					purine glycosyl C1'–N9 rotation (A, G)
C–C–S–C					1.0 0.37	0 0	rotation about C–S in Met
O3'–P–O5'–C5'	1.2	$\pi$	1.2 0.1	0 $\pi$	0.5 0.1	0 $\pi$	P–O5' rotation in nucleic-acid backbone

#### 9.5.4 Improper Torsion

Harmonic ‘improper torsion’ terms, also known as ‘out-of-plane bending’ potentials, are often used in addition to the terms described above to improve the overall fit of energies and geometries. They can enforce planarity or maintain chirality about certain groups. The potential has the form

$$E^\chi(\chi) = (V'/2) \chi^2, \quad (9.28)$$

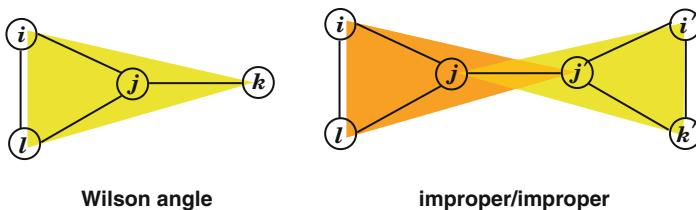


Figure 9.11. Wilson angle definition (left) and geometry for a cross improper/improper torsion term (right).

where  $\chi$  is the improper *Wilson angle*. It is defined for the four atoms *i*, *j*, *k*, *l* for which the central *j* is bonded to *i*, *l*, and *k* (see Figure 9.11) as the angle between bond *j*-*l* and the plane *i*-*j*-*k*.

The following are examples of atom centers used to define sequences of atoms counted in improper torsion potentials (taken from CHARMM):

- Polypeptide backbone nitrogen and carbonyl carbons attached to the  $C^\alpha$  atom (i.e., N and C in the sequence  $-N-C^\alpha-C=O$ ), for maintaining planarity of the groups around the peptide bond;
- Terminal side-chain carbons in the  $CO_2$  unit of the Asp and Glu amino acids;
- Terminal side-chain carbons in the  $O=C-NH_2$  unit of the Asn and Gln amino acids;
- Terminal side-chain carbon in the  $NH_2-C-NH_2$  unit of the Arg amino acid;
- Various carbons and nitrogens in the side-chain rings of the Phe; Tyr, and Trp amino acids (not used in all CHARMM versions);
- Glycosyl nitrogen (N1 in pyrimidines or N9 in purines) in nucleic acid bases;
- Exocyclic nitrogen of the  $NH_2$  unit attached to the aromatic ring of the adenine and cytosine bases in nucleic acids (N6 in adenine and N4 in cytosine).

### 9.5.5 Cross Dihedral/Bond Angle and Improper/Improper Dihedral Terms

Cross terms are used in certain force fields (e.g., CVFF used in the Insight/Discover program) to model various relationships between quadruplet sequences (see Figure 9.11) or to couple torsion angles with bond angles or torsion angles with bond lengths (see Figure 9.7). For example, the association between a torsion angle and related bond angles can take the form

$$E^{\tau\theta\theta'}(\tau, \theta, \theta') = K V^{\tau\theta\theta'} \cos \tau [\theta - \bar{\theta}] [\theta' - \bar{\theta}']. \quad (9.29)$$

This kind of potential couples the two bending motions (e.g., symmetric and antisymmetric wagging of the two methyl groups in ethane) and has an important affect on spectra (vibrational frequencies). A cross term that relates the dihedral angle to one bond angle can, however, have instead a large effect on geometry (with a small effect on spectra):

$$E^{\tau\theta}(\tau, \theta) = K V^{\tau\theta} \cos \tau [\theta - \bar{\theta}]. \quad (9.30)$$

(More generally,  $\cos \tau$  above may be replaced by a Fourier series in the form of eq. (9.23) or eq. (9.24)).

This torsion/bend effect may be important to reproduce fine features like a small C–C–H bond-angle opening in the eclipsed configuration in ethane relative to the staggered configuration.

Torsion/stretch potentials can similarly be used for fine tuning, for example to reproduce a bond-stretching tendency for ethane in the eclipsed torsional configuration (relative to the staggered-configuration bond length) [30].

The association between neighboring Wilson angles, in which the two centers ( $j$  and  $j'$ ) are bonded, can be represented as:

$$E^{\chi\chi'}(\chi, \chi') = V^{\chi\chi'} \chi \chi'. \quad (9.31)$$

## 9.6 The van der Waals Potential

### 9.6.1 Rapidly Decaying Potential

The van der Waals potential for a nonbonded distance  $r_{ij}$  has the common 6/12 Lennard-Jones form in macromolecular force fields:

$$E_{\text{LJ}}(r_{ij}) = \frac{-A_{ij}}{r_{ij}^6} + \frac{B_{ij}}{r_{ij}^{12}}, \quad (9.32)$$

where the attractive and repulsive coefficients  $A$  and  $B$  depend on the type of the two interacting atoms. The attractive  $r^{-6}$  term originates from quantum mechanics, and the repulsive  $r^{-12}$  term has been chosen mainly for computational convenience. Together, these terms mimic the tendency of atoms to repel one another when they are very close and attract one another as they approach an optimal internuclear distance (Figure 9.12).

As mentioned in the last chapter, the 6/12 Lennard Jones potential represents a compromise between accuracy and computational efficiency; more accurate fits over broader ranges of distances are achieved with the Buckingham potential [30, 766]:  $-A_{ij}/r_{ij}^6 + B_{ij} \exp(-B'_{ij} r_{ij})$ .

The steep repulsion has a quantum origin in the interaction of the electron clouds with each other, as affected by the Pauli-exclusion principle, and this combines with internuclear repulsions.

The weak bonding attraction is due to London or dispersion force (in quantum mechanics this corresponds to electron correlation contributions). Namely, the rapid fluctuations of the electron density distribution around a nucleus create a transient dipole moment and induce charge reorientations (dipole-induced dipole interactions), or London forces.

Note that as  $r \rightarrow \infty$ ,  $E_{\text{LJ}} \rightarrow 0$  rapidly, so the van der Waals force is short range. For computational convenience, van der Waals energies and forces can be computed for pairs of atoms only within some “cutoff radius”.

### 9.6.2 Parameter Fitting From Experiment

Parameters for the van der Waals potential can be derived by fitting parameters to lattice energies and crystal structures, or from liquid simulations so as to reproduce observed liquid properties [265]. From crystal data, minimum *contact distances*  $\{r_i\}$  of atom radii (see below) can be obtained. Liquid simulations, such as by Monte Carlo, are typically performed on model liquid systems (e.g., ethane and butane) to adjust empirical minimum contact distances  $\{r_i\}$  (see below) so as to reproduce densities and enthalpies of vaporization of the liquids.

### 9.6.3 Two Parameter Calculation Protocols

To determine the attractive and repulsive coefficients for each  $\{ij\}$  pair, two main procedures can be used for each different type of pairwise interaction (e.g., C–C, C–O).

Energy Minimum/Distance Procedure ( $V_{ij}, r_{ij}^0$ )

In the first approach, coefficients can be obtained by requiring that an energy minimum  $V_{ij}$  will occur at a certain distance,  $r_{ij}^0$ , equal to the sum of van der Waals radii of atoms  $i$  and  $j$ . This requirement between the minimum energy and distances produces the relations:

$$A_{ij} = 2(r_{ij}^0)^6 V_{ij} \quad (9.33)$$

and

$$B_{ij} = (r_{ij}^0)^{12} V_{ij}. \quad (9.34)$$

The latter equation can be written using eq. (9.33) in terms of  $A_{ij}$  and  $r_{ij}^0$  as

$$B_{ij} = \frac{A_{ij}}{2} (r_{ij}^0)^6. \quad (9.35)$$

The van der Waals radii can be calculated from the measured X-ray *contact distances* — which reflect the distance of closest approach in the crystal. These contact distances are appreciably smaller than the sum of the van der Waals radii of the atoms involved. This relationship — between the X-ray contact distances, which crystallographers refer to as “van der Waals radii”, and the molecular mechanics meaning of van der Waals radii — was recognized in the early days of molecular mechanics [28, 1345] but still may not be widely appreciated.

The CHARMM and AMBER force fields employ this energy-minimum / distance procedure outlined above, using the following definitions for  $V_{ij}$  and  $r_{ij}^0$ . For each atom type, the pair  $\{\epsilon_i, r_i\}$  is specified as a result of the parameterization procedure (crystal fitting or liquid simulations). The parameters for a given  $ij$  interaction are then obtained by setting  $V_{ij}$  and  $r_{ij}^0$  as the geometric and arithmetic (see below) means, respectively:

$$V_{ij} = \sqrt{\epsilon_i \epsilon_j}, \quad (9.36)$$

$$r_{ij}^0 = (r_i + r_j), \quad (9.37)$$

and then using eqs. (9.33) and (9.34) to define the attractive and repulsive coefficients, respectively. The presence/absence of the  $\frac{1}{2}$  factor in the eq. (9.37) depends on the original definition of  $\{r_i\}$  (i.e., given radii may have already been divided by 2).

### Slater-Kirkwood Procedure ( $A_{ij}, B_{ij}$ )

The second parameterization procedure is based on the Slater-Kirkwood equation. The attractive coefficient  $A_{ij}$  is determined on the basis of the atomic properties of the interacting atom pair [701, 898, 973]:

$$A_{ij} = \frac{365 \alpha_i \alpha_j}{(\alpha_i/N_{e_i})^{1/2} + (\alpha_j/N_{e_j})^{1/2}}, \quad (9.38)$$

where  $\alpha$  represents the experimentally determined atomic polarizability, and  $N_e$  denotes the number of outer-shell electrons. The numerical factor in the Slater-Kirkwood equation (that is, 365) is derived from universal constants so that the energies are produced in kcal/mol. This factor is computed from the relation  $3e\hbar/2m_e$ , where  $e$  = electron charge,  $\hbar$  = Planck's constant divided by  $2\pi$ , and  $m_e$  = electron rest mass.

Slater and Kirkwood have shown, based on experiments for noble gases, that this derivation for  $A_{ij}$  produces a London attraction potential in good agreement with experiment.

When  $A_{ij}$  is obtained from the Slater-Kirkwood equation (9.38), the coefficient  $B_{ij}$  of the repulsive term can then be obtained from  $A_{ij}$  using equation (9.35). This leads to a combined van der Waals potential that produces an energy minimum at  $r_{ij}^0$ .

The MMFF force field van der Waals parameterization employs the Slater-Kirkwood equation to define  $V_{ij}$  via  $V_{ij} = A_{ij}/[2(r_{ij}^0)^6]$  (from eq. (9.33)) where the minimum-energy separation  $r_{ij}^0$  is determined from special combination rules. In addition, a “buffered 7/14 attraction/repulsion” van der Waals term is used, found to better fit rare gas interactions [498]. The form of this potential is:

$$E_{\text{LJ}}^{\text{MMFF}}(r_{ij}) = \frac{-A_{ij}}{(r_{ij} + \gamma_1 r_{ij}^0)^7} + \frac{B_{ij}}{(r_{ij} + \gamma_1 r_{ij}^0)^7 [r_{ij}^7 + \gamma_2 (r_{ij}^0)^7]}, \quad (9.39)$$

where the  $ij$ -dependent  $A$  and  $B$  values are functions of  $V_{ij}$  and  $r_{ij}^0$ , and  $\gamma_1$  and  $\gamma_2$  are constants.

Such more complex nonbonded terms (also the 6/exponential combination used by the MM2/3/4 force field) represent a tradeoff between the accuracy of fitting and the computational expense involved in potential evaluation and differentiation.

## 9.7 The Coulomb Potential

### 9.7.1 Coulomb's Law: Slowly Decaying Potential

Ionic interactions between fully or partially charged groups can be approximated by Coulomb's law<sup>5</sup> for each atom pair  $\{i, j\}$ :

$$F_{\text{coul}}(r_{ij}) \propto q_i q_j / r_{ij}^2, \quad (9.40)$$

where  $F$  is the force and  $q_i$  is the effective charge on atom  $i$ . The force is positive if particles have the same charge (repulsion) and negative if they have opposite signs (attraction). This leads to the Coulomb potential of form

$$E_{\text{coul}} = K_{\text{coul}} \frac{q_i q_j}{\epsilon r_{ij}}, \quad (9.41)$$

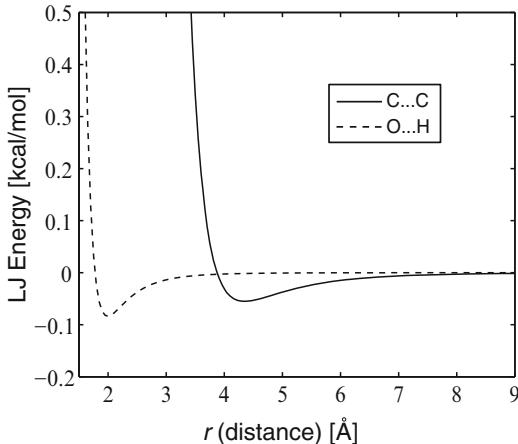


Figure 9.12. Van der Waals nonbonded potentials for C...C and water O...H distances using the CHARMM parameters  $-0.055$ ,  $-0.1521$ , and  $-0.046$  for  $\epsilon_i$  and  $2.1750$ ,  $1.7682$ , and  $0.2245$  for the distances  $r_i$ . The coefficients  $A$  and  $B$  are computed as:  $A_{ij} = 2(r_i + r_j)^6 \sqrt{\epsilon_i \epsilon_j}$ ,  $B_{ij} = (r_i + r_j)^{12} \sqrt{\epsilon_i \epsilon_j}$ , where coefficients produce energies in kcal/mol.

<sup>5</sup>Charles Augustin de Coulomb (1736–1806) was a French physicist who formulated around 1785 the famous inverse square law, now named after him. This result was anticipated 20 years earlier by Joseph Priestley, one of the discoverers of oxygen and author of a comprehensive book on electricity.

where  $\epsilon$  is the dielectric constant and  $K_{\text{coul}}$  is the conversion factor needed to obtain energies in units of kcal/mol with the charge units used (see Box 9.1).

Unlike the van der Waals potential, the Coulomb interactions decay slowly with distance (see Figure 9.13). In fact, electrostatic interactions are important for stabilizing biomolecular conformations in solvent and associating distant residues in the primary sequence closer in the folded structure. However, because of the  $\mathcal{O}(N^2)$  complexity in evaluating all pairwise terms directly, prior calculations have often only counted interactions within a cutoff radius (e.g., 10 Å).

To account for all long-range electrostatic interactions in macromolecules, implementation of fast-electrostatics algorithms, such as the multipole method and Ewald methods, are needed (see next chapter). Such methods can reduce the computational complexity to  $\mathcal{O}(N)$ . The Ewald and Particle Mesh Ewald treatments have been found to be more amenable to biomolecular dynamics simulations (e.g., [751, 1133]), especially for parallel platforms; this trend may change as hardware and software evolve. See next chapter for details.

For details of many concepts in electrostatic energies of macromolecules, see Warshel's textbook [1342]; also notable is the work [1344] which introduces consistent electrostatic calculations in proteins as well a quantum/molecular mechanics approach, and the comprehensive review [1346], which relates microscopic and macroscopic aspects of biomolecules, such as dielectric constants and dipole moments.

### 9.7.2 Dielectric Function

The reduction of the force or energy by a dimensionless factor of  $1/\epsilon$  is appropriate if the charged particles are immersed in any medium other than a vacuum. That is, a weaker interaction occurs in a polarizable medium (e.g., water) than in vacuum, since then charges become “screened”. The distance dependent dielectric function was first introduced in the late 1970s [1344]. Very simple approximations to the dielectric function, such as  $\epsilon(r) = r$  or  $\epsilon(r) = \bar{D}r$  where  $\bar{D}$  is a constant, have been used in first-generation force fields when water molecules were not represented explicitly in the model [1359].

#### Sigmoidal Function

More sophisticated formulations rely on a large body of theoretical and experimental data, which suggest screening functions in the general *sigmoidal* form [528, 548, 854, 1420] that can also account for different ionic concentrations. Such a sigmoidal function uses a screening parameter  $\kappa$  related to the ionic strength of the medium [1215]:

$$\epsilon(r) = \tilde{D} \exp(\kappa r),$$

where  $\kappa$  has dimensions of inverse length. Other forms employ distance-dependent dielectric functions like

$$\epsilon(r) = (D_s + D_0)/(1 + k \exp[-\kappa(D_s + D_0)r]) - D_0,$$

where  $D_s$  is the permittivity of water,  $D_0$  and  $\kappa$  are parameters, and  $k$  is a constant [855]. These screened Coulomb potentials simultaneously model two effects between two charges in a dielectric medium, like water: a substantially-damped electrostatic field when the charges are separated by moderate distances (e.g., several Ångstroms apart), and the diminished dielectric screening, approaching toward vacuum values, when the charges are in close proximity.

Such screened Coulomb potentials have been used for molecular dynamics simulations and have been incorporated into a procedure to calculate pH-dependent properties in proteins ( $pK_a$  shifts, i.e., shifts in hydrogen dissociation constants) with excellent accuracy [855, 1346].

The sigmoidal functions are more suitable than the simpler forms above for implicit solvation models of biomolecules. Still, for detailed structural and thermodynamic studies of biomolecules, the trend has been to model water molecules explicitly (i.e., unit value  $\epsilon$ ). Nonetheless, screened Coulomb potentials are of great usage in other applications, such as macroscopic simulations of long DNA in a given monovalent ionic concentration [1125, for example]. In such applications, the coefficient  $\tilde{D}$  is a function of the salt concentration through a salt-dependent linear charge density for DNA, and the screening parameter  $\kappa$  is the inverse of the Debye length [197].<sup>6</sup> See end of Chapter 10 (section on continuum solvation) and [107], for example, for further details.

### Box 9.1: Coulomb Potential Constant ( $K_{\text{coul}}$ )

Chemists typically use an *electrostatic charge unit* (esu) to define the unit of charge. This unit is defined as the charge repelled by a force of 1 dyne when two equal charges are separated by 1 cm. In these units, the electron charge is  $4.80325 \times 10^{-10}$  esu.

In cgs units, the constant of proportionality in the Coulomb energy is unity; in the SI international system of units, the unit of charge is the *coulomb* (C) = 1 Ampere second (As). In coulomb units, the electron charge is  $1.6022 \times 10^{-19}$  C. The corresponding constant of proportionality in Coulomb's law is:

$$K_{\text{coul}} = 1/(4\pi\epsilon_0), \quad (9.42)$$

where the permittivity of a vacuum  $\epsilon_0$  is

$$\begin{aligned} \epsilon_0 &= 8.8542 \times 10^{-12} \text{ kg}^{-1} \text{ m}^{-3} \text{ s}^4 \text{ A}^2 \\ &= 8.8542 \times 10^{-12} \text{ J}^{-1} \text{ m}^{-1} \text{ C}^2. \end{aligned}$$

<sup>6</sup>The electrostatic energy is typically expressed as the sum over pairwise interactions between hydrodynamic DNA beads separated by segments of length  $l_0$  as:  $[(\nu l_0)^2/\epsilon] \sum_{i < j} [\exp(-\kappa r_{ij})/r_{ij}]$ , where  $\nu$  is an effective linear charge density of the DNA,  $\epsilon$  is the dielectric constant of water, and  $1/\kappa$  is the Debye length at the monovalent salt concentration  $c_s$ , given in Molar units ( $\kappa \approx 0.33\sqrt{c_s}$  inverse Ångstrom units at room temperature for 1:1 electrolyte solutions).

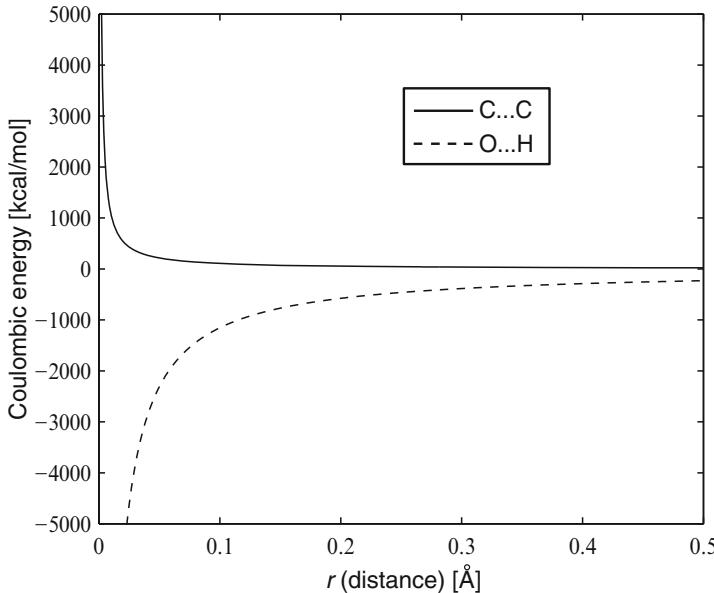


Figure 9.13. Coulomb potentials  $K_{\text{coul}} q_i q_j / r_{ij}$  for C...C and O...H interactions using the CHARMM partial charge parameters (in esu units) of  $-0.1800$  (Arg carbon),  $-0.8340$  (water oxygen), and  $0.4170$  (water hydrogen).

Thus, to obtain energy in kcal/mol using Ångstroms for distance and esu for partial charges, the conversion constant is

$$\begin{aligned}
 K_{\text{coul}} &= \frac{1 \text{ J m}}{4\pi (8.8542 \times 10^{-12}) \text{ C}^2} \\
 &= \frac{(6.0221 \times 10^{23} \text{ mol}^{-1})(10^{10} \text{ Å})(4184^{-1} \text{ kcal})(1.6022 \times 10^{-19})^2}{4\pi (8.8542 \times 10^{-12}) \text{ esu}^2} \\
 &\approx 332 \frac{\text{kcal}}{\text{mol}} \cdot \frac{\text{Å}}{\text{esu}^2}. \tag{9.43}
 \end{aligned}$$

### 9.7.3 Partial Charges

Selection of partial charges to fit atom *centers* has been a difficult issue in molecular mechanics calculations, since different quantum mechanical approaches yield significantly different values [1195]; earlier, determinations of atomic partial charges based on X-ray diffraction data were suggested [980]. However, extensive developments in high-level quantum-derived electrostatic potentials [218, 265, 805] are being applied to determine partial atomic charges for isolated model compounds, with various adjustments to correct for the neglect of many-body polarization effects in liquid water and other factors (e.g., [1330]).

For reference, the MMFF force field uses a simple modification of the Coulomb potential of form [498]

$$E_{\text{coul}}^{\text{MMFF}}(r_{ij}) = K_{\text{coul}} \frac{q_i q_j}{\epsilon (r_{ij} + \delta)^n}, \quad (9.44)$$

where  $\epsilon$  is the dielectric constant (unity by default),  $\delta = 0.05 \text{ \AA}$  (the *buffering constant*), and  $n$  is 1 (default) or 2 (distance-dependent model). This form is required to dampen the attraction between oppositely charged atoms in combination with the buffered van der Waals term used.

## 9.8 Parameterization

### 9.8.1 A Package Deal

The general parameterization process for potential energy functions is a difficult task. Several important decisions must be made regarding choices for the functional form and numerical values for the parameters. Even if one is given a specific energy form and a set of structural and energetic data to reproduce, the combinations of parameters that can be used are endless. Unrealistic choices for one group of parameters can be compensated for by adjustment of another.

In theory, the energy terms should have clear physical significance with parameters calibrated by empirical fitting of crystal data, rotational barriers of analogous small molecules, and vibrational frequencies. However, an approximation is inherent in the extension of data from small to large systems. Moreover, interaction with solvent and counterions, reflected in the experimental data, must be interpreted and incorporated in the energy model. In summary, much freedom and manipulation are possible in constructing empirical energy surfaces. Only if constructed and parameterized correctly will the energy model generate reliable structural predictions, as reviewed in [620, 802].

In the case of nucleic acid sugars, the importance of parameter choices in the energy function has already been realized. For example, different potential energy models have produced results that are *qualitatively* different regarding sugar pseudorotation [522, 750, 909, 937, 1359]. This is particularly possible when choosing appropriate equilibrium values for endocyclic bond angles [937] or torsion-angle parameters about sugar atoms [218]. Since the unusual puckering geometry and ring closure constraints produce significant deviations from tetrahedral bond angle arrangements, it is not clear what equilibrium values should be used in the harmonic bending terms, more appropriate for small fluctuations.

### 9.8.2 Force Field Comparisons

Some force-field dependent conformations for DNA have been discussed with respect to the AMBER and CHARMM force fields [379, 381, 415, 803, 804, 1330].

With some earlier versions, average structures tended to be more B-like with AMBER and intermediate between A and B-DNA with CHARMM in large part due to differences in sugar conformations; newer force fields better balance local geometry with global helical propensities and can reproduce the equilibrium between A and B-DNA in solution (both when Ewald and nonbonded cutoffs are used for electrostatics) [221, 804, 1330]. Differences in backbone and base geometries, as well as dynamic properties, are also believed to be force-field dependent. Such disparities, rectifiable with improved parameters, warrant a particularly cautionary note in interpreting structural transitions (such as between A and B-DNA [219]) in molecular dynamics simulations.

With recent improvements in molecular dynamics algorithms, enhanced sampling methods, computational speed, and rapid performance on parallel architecture, long-time dynamics simulations of proteins and DNA, as well as folding studies, have become possible. Such simulations have revealed force field inaccuracies not observed in shorter-length simulations. For example, AMBER was found to over-stabilize  $\alpha$ -helices [445, 934], while CHARMM tends to prefer  $\pi$ -helices [378]. More generally, evidence for conformational discrepancies was presented over a wide range of protein force fields [882, 1411].

To overcome these problems, peptide backbone parameters were reparameterized for AMBER [568], CHARMM [806, 807], and OPLS-AA [629]. A new AMBER force field was also created for protein simulations [341]. Parameter improvements for nucleic acids in AMBER [988] and GROMOS [1206] and for lipid parameters in CHARMM [556, 652] were also reported.

Still, recent long-time peptide folding studies have raised concerns about conformational biases in existing force fields (e.g., overstabilization of  $\alpha$  helices) that even the most recent force field corrections have not resolved [132, 424, 426]. Indeed, it is impossible for any force field to accurately reproduce all the complex interactions and properties of real systems, and this stems not only from force field approximations but also from hardware and software limitations, time constraints, convergence issues, and the limited resolution or possible errors of some experimental data. Nonetheless, it is interesting that very different computational approaches (all atom on one hand versus coarse grained with implicit solvation) lead to similar results (e.g., folded structures with high  $\alpha$ -helical content for a  $\beta$  protein) that depart from experimental predictions [365, 424, 426, 841]; this can either indicate strong force field biases that persist through various levels of approximation or perhaps physical explanations that modeling can illuminate and eventually help resolve. Recent modeling of folding processes has certainly indicated that conformational space may be more heterogeneous than originally believed [365, 427].

Much work continues to further refine and develop force fields to mirror more accurately complex systems and produce realistic simulation data for biological systems over longer time-scales to resolve such issues.

### 9.8.3 Force Field Performance

Several discussions of force field performance [426, 455, 668, 1039, 1068, 1069, 1172] highlight general issues that remain unresolved and in need of improvement:

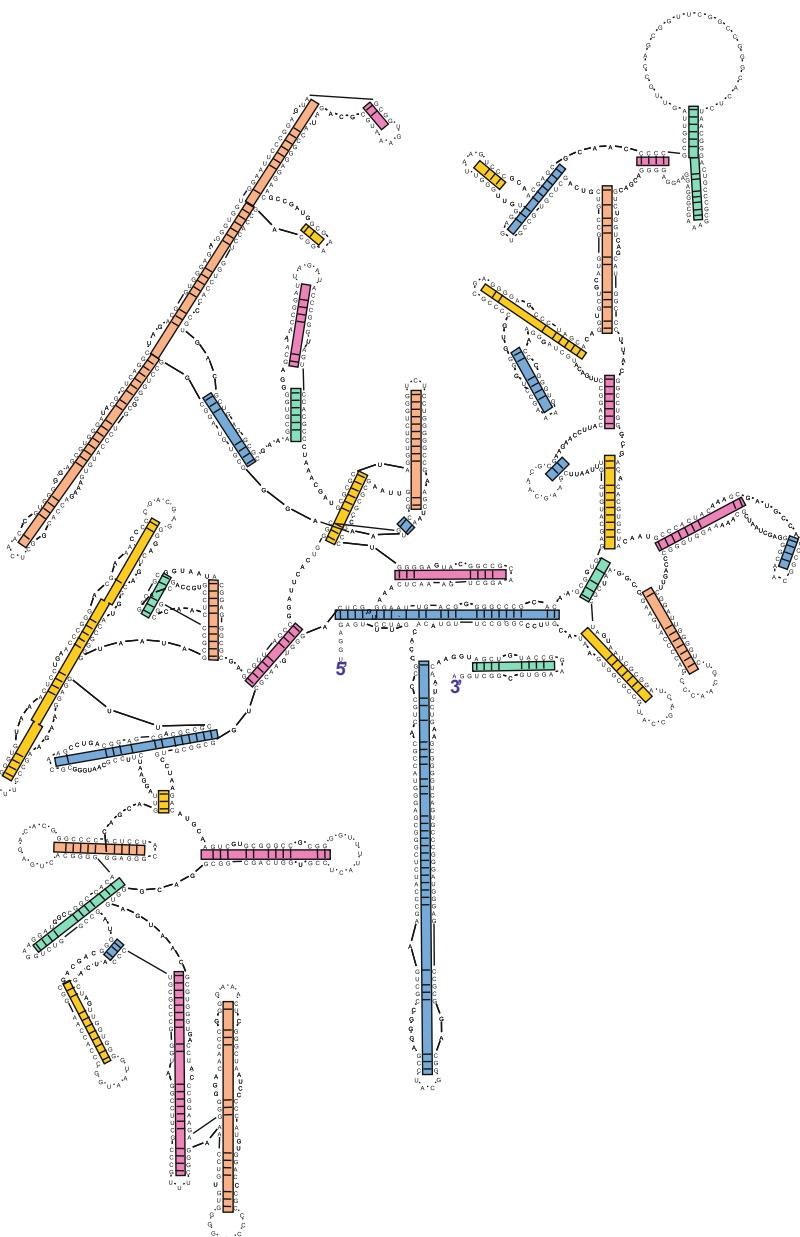
- Determination of partial charges;
- Improvement of electrostatic potentials (e.g., use of distributed multipole analysis);
- Methods for solvent representation;
- Interpretation of results in the absence of solvent;
- The approximation reflected by Cartesian vs. torsion space representation; and
- General interpretation of conflicting results by different models and potentials.

Frequent comparisons among force-field results with respect to experimental and high-accuracy *ab initio* data, such as done in [105, 500], reveal the sizable errors made by most force fields and inherent deficiencies. Improvements in the electrostatic formulation that refines the simple atom-centered charged models are repeatedly urged by computational chemists.

As force fields improve, some of these issues may be resolved, but force fields cannot be said to be “converging” to one another [1123]. Still, modern force fields perform comparably in molecular dynamics simulations [1016].

*Still, I emphasize that force fields need not be perfect to be useful!* Their overall utility is in generating qualitative and quantitative insights into structural, energetic, and dynamic properties of complex systems through systematic studies (for example by varying critical parameters), especially when trends are compared for a related group of biological systems (like single base-pair variants of DNA).

Ultimately, predictions or interpretations can be tested experimentally.



# 10

## Nonbonded Computations

### Chapter 10 Notation

SYMBOL	DEFINITION
<b>Matrices</b>	
<b>D</b>	diffusion tensor (also defined for Coulomb manipulations)
<b>H</b>	Hessian of potential function
<b>M</b>	mass matrix
<b>T</b>	hydrodynamic tensor
<b>Vectors</b>	
<b>m</b>	reciprocal lattice vector, components $\{L/n_x, L/n_y, L/n_z\}$
<b>n</b>	periodic domain vector, components $\{n_x L, n_y L, n_z L\}$
<b>q</b>	arbitrary vector (defined for matrix/vector product)
<b>r<sub>ij</sub></b>	interatomic distance vector
<b>s</b>	scattering vector, components $\{h, k, l\}$ (indices of reflection)
<b>v(x)</b>	gravitational field vector at point <b>x</b>
<b>x<sub>j</sub></b>	position vector of atom <i>j</i> , components $x_{j1}, x_{j2}, x_{j3}$
<b>F<sub>ij</sub></b>	force term
<b>F<sub>s</sub></b>	complex-valued structure factor, $F_s = A_s + iB_s$
<b>F<sub>s<sub>j</sub></sub></b>	scattering vector at atom <i>j</i>
<b>F<sub>NB</sub></b>	nonbonded force component, $-\nabla E_{NB}$
<b>R</b>	Langevin random force
<b>X, x</b>	collective position vector
<b>μ</b>	dipole moment
<b>∇Φ</b>	gradient of potential function
<b>Scalars &amp; Functions</b>	
<i>a, b, c, ē</i>	distance parameters
<i>a<sub>r</sub></i>	hydrodynamic radius
<i>c<sub>s</sub></i>	ionic concentration

Chapter 10 Notation Table (continued)

SYMBOL	DEFINITION
$dv, dv^*$	volume elements
$e$	protomic charge
$f_j$	scattering amplitude for atom $j$
$l$	integer, level of refinement in multipole expansion
$m$	particle mass
$m^0$	zeroth moment of $\Phi$
$m_1^1, m_2^1, m_3^1$	components of dipole moment $\mu$ of $\Phi$
$m_{11}^2, m_{12}^2, \dots, m_{33}^2$	components of quadrupole moment of $\Phi$
$n_x, n_y, n_z$	integers
$p$	integer (power of expansion)
$q_j$	Coulomb partial charge for atom $j$
$r, r_{ij}$	interatomic distance
$\{r, \theta, \phi\}$	spherical coordinates
$u(\mathbf{x})$	gravitational potential function
$A_{ij}, B_{ij}$	Lennard-Jones coefficients for atom pair $i, j$ (attraction, repulsion)
$C_{ij}, D_{ij}$	coefficients for modified Lennard-Jones potential
$D_t$	translational diffusion constant
$E_{\text{coul}}$	Coulomb potential
$E_{\text{LJ}}$	Lennard-Jones potential
$E_{\text{NB}}$	nonbonded potential
$F_s$	scattered amplitude of whole crystal
$G$	gravitational constant
$K_{\text{coul}}$	Coulomb potential constant
$L$	box size dimension
$\{M_n^m\}$	moments of the multipole expansion
$N$	number of variables (atoms)
$N_A$	Avogadro's number
$\{P_n^m\}$	Associated Legendre polynomials of degree $n$
$R_e$	earth's radius
$R_{ij}$	Squared interatomic distance (between atoms $i$ and $j$ ), $(r_{ij})^2$
$S(r)$	shift/switch function
$T$	temperature
$V$	volume of unit cell (associated with volume element $dv$ )
$V^*$	volume of reciprocal space (associated with volume element $dv^*$ )
$\{Y_n^m(\theta, \phi)\}$	spherical harmonics functions
$\alpha, \beta, \gamma$	angles
$\beta$	Gaussian screening parameter
$\gamma$	Langevin damping constant
$\epsilon$	dielectric constant
$\epsilon(\mathbf{x})$	position-dependent dielectric function
$\epsilon_{\text{acc}}$	desired accuracy
$\eta$	solvent viscosity
$\kappa$	Debye screening parameter
$\rho(\mathbf{x})$	electron (or charge) density
$\rho_G$	screening Gaussian
$\phi(\mathbf{s})$	phase angle associated with structure factor $F_s$
$\omega_{ij}$	weight

Chapter 10 Notation Table (continued)

SYMBOL	DEFINITION
$\Phi$	electrostatic potential
$\Phi_{\text{real}}$	real (or direct-space) component of $\Phi$
$\Phi_{\text{recip}}$	reciprocal-space component of $\Phi$
$\Phi_{\text{cor,ex}}$	correction term for excluded nonbonded interactions
$\Phi_{\text{cor,self}}$	correction term for self nonbonded interactions
$\Phi_{\text{cor},\epsilon}$	correction term for finite dielectric

Hofstadter's law: It always takes longer than you expect, even when you take into account Hofstadter's law.

Douglas R. Hofstadter, in *Godel, Escher, Bach*, 1979 (1945–).

## 10.1 A Computational Bottleneck

Reducing the cost of the nonbonded energy and force computations is of primary importance in molecular mechanics and dynamics simulations of biomolecules. This is because the direct evaluation of these nonbonded interactions involving all atom pairs has the complexity of  $\mathcal{O}(N^2)$  where  $N$  is the number of atoms. Recall that the bonded terms are local and thus have a linear computational complexity; see homework assignment 8 for a related exercise.

The rapid, quadratic growth in CPU time when all nonbonded interactions are summed directly versus the linear growth associated with the “cutoff” procedures (consideration of interactions within a limited distance range) is shown in Figure 10.1; see CPU scale at left for the evaluation of an energy and force. The implication of these CPU times on total times for 1 ns trajectories (of one million steps) is also shown (see scale at right), explaining the urgency in reducing the nonbonded-term evaluation cost.

The data in Figure 10.1 and Table 10.1 show, for example, that computing all the nonbonded energy and force interactions directly for a hen egg-white (HEW) lysozyme protein (1960 atoms or 5880 Cartesian variables) in vacuum requires about 0.18 seconds on a single Intel Xeon/3GHz processor of a Dell Linux machine. One million such steps to span one nanosecond by molecular dynamics with a 1 femtosecond timestep would require approximately 1.4 days. A system that is seven times larger (e.g., size range of a small solvated protein) requires roughly a factor of 45 more computational time, or nearly 62 days of CPU to span a single nanosecond!

Fortunately, techniques have been developed to reduce this cost dramatically without destroying the value of simulations of biomolecules in solvent.

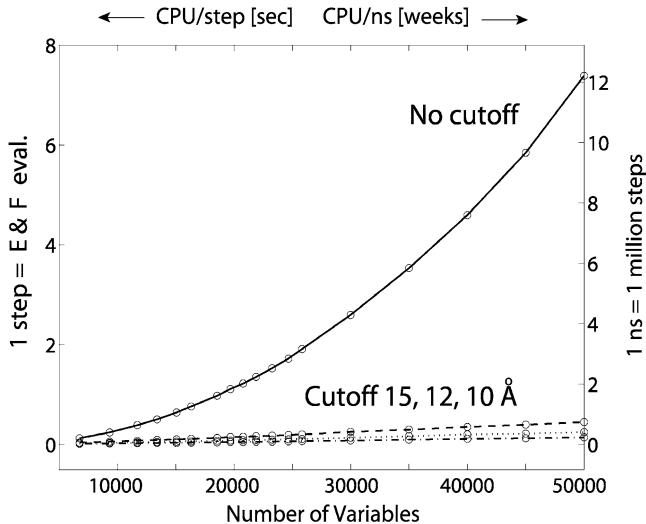


Figure 10.1. CPU time per step (energy plus force evaluation) for water clusters of various sizes modeled in CHARMM when cutoffs are used (at 10, 12, and 15 Å) versus no cutoffs (see left vertical scale) by direct calculation, and corresponding times required for 1 ns trajectories assuming  $10^6$  steps of 1 fs (see right vertical scale). The number of variables is nine times the number of water molecules. Timings were obtained on a single Xeon/3GHz processor of a Dell Linux machine.

This chapter introduces novice modelers to three fundamental techniques for handling the nonbonded interactions of large biomolecular systems: spherical cutoffs, particle-mesh Ewald, and fast multipole schemes. We conclude with a brief mention of alternative continuum solvent models, such as Langevin and Brownian dynamics and Poisson-Boltzmann calculations.

## 10.2 Approaches for Reducing Computational Cost

### 10.2.1 Simple Cutoff Schemes

The spherical cutoff techniques introduced in the next section are easy to implement as well as computationally cheap ( $\mathcal{O}(N)$ ). More sophisticated than straightforward truncation, these methods can yield reasonable approximations to the energy and force functions up to some threshold separation-distance value. They are particularly suitable for van der Waals interactions which decay rapidly with distance and therefore can be considered zero beyond some interatomic separation distance. Spherical cutoff methods have been used by necessity for simulations of large systems

Table 10.1. Computational requirements of nonbonded calculations on CHARMM version 28a2 by cutoffs, direct nonbonded computations ('all nonbonded'), or all nonbonded by particle-mesh Ewald (PME).

Model <sup>a</sup>	Atoms/ Variables	CPU/step <sup>b</sup> [sec.]	CPU/ns <sup>c</sup> [days]
HEW Lysozyme, 12 Å cutoffs HEW Lysozyme, all nonbonded	1960/	0.03	0.29
	5880	0.18	1.38
Solvated DNA, 12 Å cutoffs Solvated DNA, PBC, 12 Å cutoffs Solvated DNA, all nonbonded	12389/	0.26	3.08
	37167	0.63	7.35
		4.04	46.83
Solvated BPTI, 12 Å cutoffs Solvated BPTI, PBC, 12 Å cutoffs Solvated BPTI, PBC/PME Solvated BPTI, all nonbonded	14275/	0.30	3.49
	42825	0.76	8.82
		0.83	9.65
		5.35	61.91

<sup>a</sup> (For lysozyme): nonbonded cutoffs via **group**-based electrostatic switch and **atom**-based Lennard-Jones switch functions, switch buffer 10 to 12 Å, pairlist buffer 12 to 13 Å, and SHAKE used for all bonds involving hydrogens.

(For DNA dodecamer and BPTI): nonbonded cutoffs with periodic boundary conditions (PBC) via **atom**-based electrostatic and Lennard-Jones switch functions, switch buffer 10 to 12 Å, pairlist buffer 12 to 13 Å, and SHAKE used for all bonds involving hydrogens.

<sup>b</sup>Each step entails an energy and gradient evaluation performed in the CHARMM program. The timings were made on a single Intel Xeon/3GHz processor of a Dell Linux machine.

<sup>c</sup>We assume 1 fs timesteps.

### 10.2.2 Ewald and Multipole Schemes

Of course, cutoff methods neglect long-range interactions beyond some distance and therefore are poor approximations for highly charged systems where large-scale conformational arrangements of linearly-distant residues are involved. The alternative techniques mentioned in the following sections are more suitable for these cases.

Approximating all nonbonded interactions can be accomplished by fast electrostatic techniques based on multipole expansions or Ewald lattice techniques. These schemes have revolutionized the biomolecular simulation field in the past decade since their computational complexity is only  $\mathcal{O}(N \log N)$ , a dramatic computational saving with respect to the direct value of  $\mathcal{O}(N^2)$ .

For example, computing the energy and force using spherical cutoffs of range 12 Å for a solvated small protein (BPTI) of size 14275 atoms (42825 variables) modeled in a periodic domain requires about 0.76 seconds on a Xeon/3GHz processor of a Dell Linux machine; one million such steps to span one nanosecond by molecular dynamics with a 1 fs timestep would require about 9 days. If all Coulomb interactions for the same system are computed with the

particle-mesh Ewald technique available in CHARMM, the computing time per step (and nanosecond) would increase slightly (but of course accuracy will increase; see Table 10.1). In comparison, direct evaluation of all nonbonded terms would make the project untenable!

An efficient three-dimensional version of the fast multipole method is rather involved to implement; this might explain the preference to date for Ewald techniques in the biological simulation community. The Ewald approach is applied to periodic domains, and this has been known to produce nonphysical long-range correlations for the system [510, 580, 581, 1203]. These effects may, however, be considered secondary in general in comparison with truncation artifacts.

*Such problems remind practitioners that rarely in the field of biomolecular simulations are pure gains involved due to improving methodologies; there is often a balance between the approximations made and the physical reality of the resulting models.* See [1081] for an overview of Ewald and multipole methods for computing long-range electrostatic effects in biomolecular dynamics simulations.

Before we introduce the Ewald and fast multipole techniques, we present spherical cutoff methods. Continuum solvation models based on the Poisson-Boltzmann equation are also discussed at the end of this chapter.

*The notation used in this chapter (e.g., lattice vectors, scattering factors), though different from some other parts of this text, follows presentations elsewhere on the Ewald summation; these conventions originated in the crystallographic community and have been adopted by the molecular simulation community.*

## 10.3 Spherical Cutoff Techniques

### 10.3.1 Technique Categories

There are three basic categories of cutoff techniques: *truncation*, *switch*, and *shift* formulations. All approaches set the distance-dependent nonbonded function to zero beyond some distance value  $r = b$ ; however the functional values for  $r < b$  are treated differently (see Figure 10.2; the mathematical formulas mentioned in the caption are discussed below):

- The simplest approach, **truncation**, abruptly defines values to be zero at  $b$  and does not alter the values of the energies and forces for distances  $r < b$ .
- **Switching** schemes begin to change values at a nonzero value  $a < b$  but leave values for  $r < a$  unchanged.
- **Shift** functions alter the function more gradually for all  $r < b$ .

These three general categories can be applied to either the *energy* or the *force* function of the nonbonded potential (van der Waals or electrostatic). When the force rather than energy function is altered, the energy value is obtained by integration.

In addition, *atom-based* or *group-based* schemes can be used. In the latter, distance thresholds are applied to distances between group *centers*. Group-based cutoffs can better maintain charges associated with entire residues. They can thus avoid potential instabilities in the energy or force that arise when only a subset of atoms of a particular residue is altered.

Besides choosing the particular approach (e.g., atom-based potential switch, group-based force switch), care is required in specifying the distance parameters  $a$  and  $b$ .

The cutoff techniques described here are also employed when multiple-timestep integration schemes are applied to different force classes (see Chapter 14); in these applications, *force switching* techniques are often used.

### 10.3.2 Guidelines for Cutoff Functions

In developing nonbonded cutoff functions, we are guided by the following considerations.

1. The short-range energies and forces should be altered as minimally as possible (while satisfying other criteria below).
2. The energies should be altered gradually rather than abruptly to avoid the introduction of artificial minima (where the potential energy and gradient values are suddenly zero).
3. The cutoff approach should avoid introducing large forces around the cutoff region (spikes in right panels of Figure 10.2). This is especially important for molecular dynamics simulations.
4. Also for molecular dynamics, it is important that the cutoff approach alters the functions in a way to approximately conserve the energy.

Truncation schemes satisfy criterion 1 above but violate all others and are removed from further consideration. Switching schemes alter the potential less than shift schemes (since function values for  $r < a$  are not altered) but can introduce artificial minima and large sudden forces (Fig. 10.2), violating criterion 3. Energy conservation (criterion 4) can be problematic for certain group-based implementations when polar groups are involved near the cutoff region. Improved (force) shift and force switch functions are often preferred for molecular dynamics applications [1220] with a sufficiently wide buffer region  $[a, b]$  (e.g., 8 to 12 Å or 11 to 15 Å) and a large enough cutoff value  $b$  ( $\geq 12$  Å).

In general, the choice of the best spherical cutoff scheme to use depends on the force field and the system being studied. See [923] and references cited therein for studies on the effect of different cutoff and long-range electrostatic models on the stability and accuracy of biomolecular simulations.

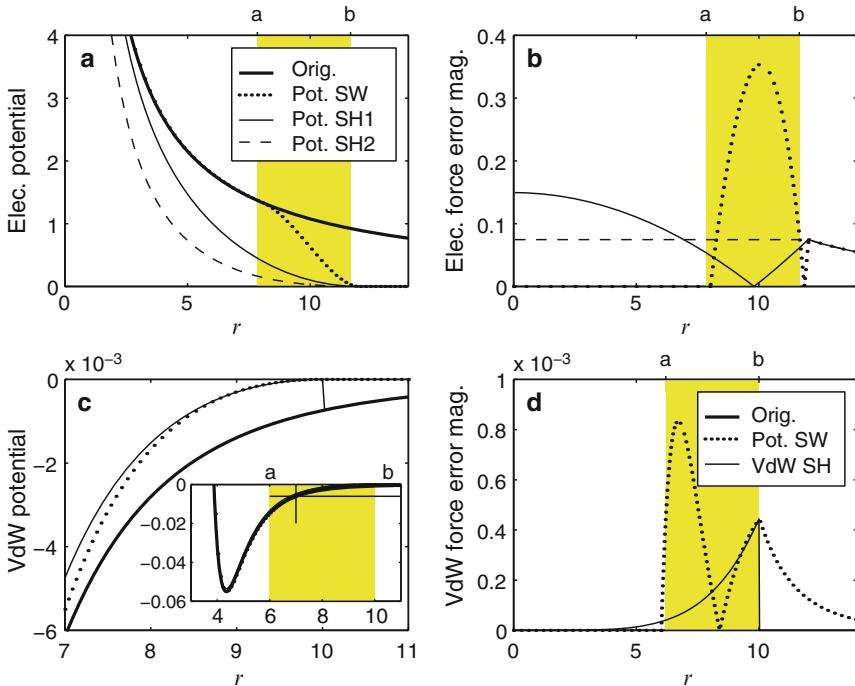


Figure 10.2. Various cutoff schemes — potential switch, eq. (10.4), and two types of potential shift, eqs. (10.10), (10.11) — with buffer regions of 8–12 Å (electrostatic) and 6–10 Å (van der Waals). The altered potentials  $E(r)$  are shown on the left panels (a,c), and the corresponding errors in the associated forces ( $\Delta|F_{\text{mod}} - F_{\text{orig}}|$ ), reflecting the magnitude of the difference between the original and modified force, are shown on the right panels (b,d). The potential shift function for the van der Waals interaction corresponds to eq. (10.14). Parameters from the CHARMM program are used, modeling a C $^\beta$ –C $^\beta$  interaction in peptides. The corresponding charge for this atom type is  $-0.18$  esu, and the van der Waals parameters  $\{\epsilon_i, r_i^0\}$  are  $-0.055$  kcal/mol and  $2.175$  Å, respectively (see eqs. (9.36), (9.37) of Chapter 9 for deriving  $V_{ij}$  and  $r_{ij}^0$ ), producing:  $A_{ij} = 745.295$  [kcal/mol] Å $^6$  and  $B_{ij} = 2.525 \times 10^6$  [kcal/mol] Å $^{12}$ ; the values of  $C_{ij}$  and  $D_{ij}$  from eqs. (10.15) and (10.16) are  $C_{ij} = -7.403 \times 10^{-10}$  [kcal/mol] / Å $^6$  and  $D_{ij} = -0.0015$  kcal/mol.

### 10.3.3 General Cutoff Formulations

Consider the following general modification to the nonbonded energy function  $E_{\text{NB}}$  by a distance-dependent switch or shift function  $S(r)$ :

$$E_{\text{NB}}(X) = \sum_{i,j} \omega_{ij} S(r_{ij}) \left[ \frac{-A_{ij}}{r_{ij}^6} + \frac{B_{ij}}{r_{ij}^{12}} + \frac{q_i q_j}{\epsilon r_{ij}} \right]. \quad (10.1)$$

Here  $X$  is the collective position vector,  $r_{ij}$  represents the distance between atoms  $i$  and  $j$ , and the parameters  $0 \leq \omega_{ij} \leq 1$  are weights. These weights can be

used to exclude bonded terms (i.e., 1–2 interactions) or bond-angle interactions (1–3) (i.e., with  $\omega_{ij} = 0$ ), or to scale other interactions such as those involving a sequence of four bonded atoms (1–4).

### Truncation

Simple truncation can be expressed by the switch function  $S$  defined as

$$S(r) = \begin{cases} 1 & r < b \\ 0 & r \geq b \end{cases}. \quad (10.2)$$

This function introduces a force discontinuity at  $r = b$  and fails to conserve energy. In general,  $S(r)$  is a distance-dependent distance function which assigns a constant value or a function of  $r$  depending on the value of  $r$  with respect to the distance parameters  $a$  and  $b$ . Thus  $S(r)$  may be set separately for the three cases:  $r \leq a$ ,  $a < r \leq b$ , and  $r > b$ .

### Switch/Shift

The switch/shift function can be different for the van der Waals and electrostatic terms; in that case eq. (10.1) should be written as two terms with different functions  $S(r)$ . The CHARMM program, for example, uses the same switch functions for both van der Waals and Coulomb interactions but different shift functions for these terms when selected.

### Atoms/Groups

The above formulation is atom-based. Group-based formulations apply the distance-dependent switch/shift functions based on the separation of *group centers*. For example, if such an intergroup distance is in the buffer region  $[a, b]$ , the modification that  $S$  applies in this range (see below) is used for all atoms in the two groups; similarly, this intergroup distance determines the modifications applied to all atoms in the two groups when the distance falls in the other two regions:  $r < a$  and  $r > b$ .

### Energy/Force Modifications

When a modification is applied to the force  $F_{\text{NB}}$  instead of the potential energy of the nonbonded terms, each  $r_{ij}$ -dependent force term, namely  $F_{ij}(r_{ij})$ , is modified as:

$$\widehat{F}_{ij}(r_{ij}) = \omega_{ij} S(r_{ij}) F_{ij}(r_{ij}) \quad (10.3)$$

where the force rather than the potential is switched by the operator  $S(r)$ . The corresponding energy must then be obtained by integration.

#### 10.3.4 Potential Switch

For potential switch functions,  $S(r)$  is a polynomial of  $r$  that alters the nonbonded energy smoothly and gradually over the buffer region  $[a, b]$  so that  $E(b) = 0$ ,

while leaving values of the energy function  $E(r)$  for  $r \leq a$  unchanged. The polynomial degree must be sufficiently high to ensure that both the energy and its gradient are continuous functions. A satisfactory function is the following cubic polynomial of  $r^2$  (see Figure 10.3, part (a)):

$$S(r) = \begin{cases} 1 & r < a \\ 1 + y(r)^2 [2y(r) - 3] & a \leq r \leq b \\ 0 & r > b \end{cases}, \quad (10.4)$$

where

$$y(r) = (r^2 - a^2)/(b^2 - a^2). \quad (10.5)$$

The  $S(r)$  expression above for  $a \leq r \leq b$  can also be written as (following algebra):

$$S(r) = \frac{(b^2 - r^2)^2 (b^2 + 2r^2 - 3a^2)}{(b^2 - a^2)^3} \quad \text{for } a \leq r \leq b. \quad (10.6)$$

From this form, it is clear that  $S(r)$  decreases monotonically from 1 to 0 as  $r$  increases from  $a$  [ $y(a) = 0$  and  $S(a) = 1$ ] to  $b$  [ $y(b) = 1$  and  $S(b) = 0$ ]. Note that for  $a < r < b$  the derivative from eq. (10.6) is:

$$S'(r) = 12r y(r) [y(r) - 1]/(b^2 - a^2), \quad (10.7)$$

and thus both the right derivative of  $S(r)$  at  $a$  and the left derivative of  $S(r)$  at  $b$  (where  $y(r) = 0$  and 1, respectively) are zero. Since these derivatives are also zero from the left of  $a$  and the right of  $b$  (where  $S(r)$  is a constant function), both the potential and the gradient functions are continuous. This function  $S(r)$  also yields continuous second derivatives when the force, rather than the energy, is switched on or off.

### 10.3.5 Force Switch

Rather than switching the potential, force switching is used in multiple-timestep schemes to gradually separate the short from long-range forces.<sup>1</sup> Often, force classes are defined according to a distance parameter  $b$  closely related to the cutoff distance above. For example, interactions within a region of  $b$  Å can be considered ‘fast’ and those beyond that value ‘slow’ (e.g.,  $b = 6$  Å). In this case, the short-range forces are turned off gradually by a switch function  $S(r)$  that decreases from 1 to zero as  $r$  increases from  $a$  to  $b$  in a manner closely resembling that of  $S$  defined in eq. (10.4). Here the region  $[a, b]$  is the switching buffer region, and  $c = b - a$  is the size of this buffer.

---

<sup>1</sup>These two components are also termed ‘fast’ and ‘slow’ because the short-range terms are rapidly varying in time while the long-range terms change more slowly with time; see Chapter 14.

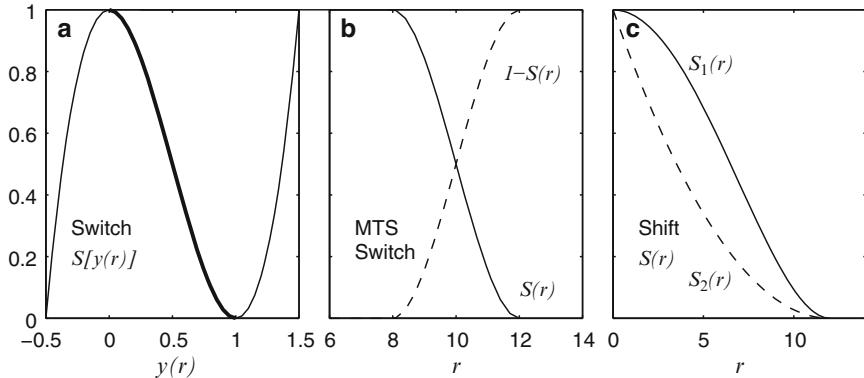


Figure 10.3. Switch and shift functions: (a) the potential switch function of eq. (10.4) as a function of  $y(r)$  given in eq. (10.5); only the heavier part of the curve is relevant since  $y(r)$  increases from 0 to 1 as  $r$  increases from  $a$  to  $b$ ; (b) the switch function  $S(r)$  of eq. (10.9) that is applied to the fast force component in multiple-timestep schemes (solid curve), along with  $1 - S(r)$  (dashed curve) that is applied to the slow force component (see eq. (10.8)); (c) the shift functions  $S_1(r)$  and  $S_2(r)$  of eqs. (10.10) and (10.11).

The switching function is applied as follows to the fast and slow force components of  $F_{\text{NB}}$ :

$$\begin{aligned} F_{\text{NB}, \text{fast}} &= S(r) F_{\text{NB}}(r) \\ F_{\text{NB}, \text{slow}} &= [1 - S(r)] F_{\text{NB}}(r), \end{aligned} \quad (10.8)$$

where the function  $S(r)$  is defined as:

$$S(r) = \begin{cases} 1 & r < a \\ 1 + r^2(2r - 3) & a \leq r \leq b \\ 0 & r > b \end{cases}. \quad (10.9)$$

See Figure 10.3, part (b), for an illustration.

### Buffer Parameters

In addition to the buffer length  $c$  (which can range from 1 to 4 Å, for example), another buffer parameter  $\tilde{c}$  is typically used for bookkeeping purposes. Namely, to keep track of the nonbonded atom pairs  $\{i, j\}$  used to associate each pair of atoms with a force term in which it is calculated (i.e., fast or slow), the pairlist is monitored up to a distance  $b + \tilde{c}$ . Implementations vary from program to program, but the main idea is to use this buffer size  $\tilde{c}$  to monitor changes in interatomic distances that would require a new pairlist generation (see [1025] for example).

### 10.3.6 Shift Functions

Shift-type functions can avoid the sudden changes in force that occur with truncation and switch methods at the cost of underestimating the short-range forces

(see Fig. 10.2). This requires alteration of the nonbonded function over the larger region  $r \leq b$  (rather than  $0 < a \leq r \leq b$ ).

Shift functions include the following two formulations:

$$S_1(r) = [1 - (r/b)^2]^2, \quad \text{for } r \leq b, \quad (10.10)$$

or

$$S_2(r) = [1 - r/b]^2, \quad \text{for } r \leq b. \quad (10.11)$$

Both decrease monotonically from 1 to 0 as  $r$  increases from 0 to  $b$ , but their curvature is different (Fig. 10.3, part (c)). These two functions work well for Coulomb interactions.

In fact,  $S_2(r)$  above augments the true Coulombic force by a constant ( $1/b^2$ ) for  $r \leq b$ . This can be seen by writing the modified energy  $\tilde{E}(r)$  of the scalar Coulomb potential  $E(r) = 1/r$  as

$$\tilde{E}(r) \equiv S_2(r) \cdot E(r) = \frac{1}{r} \cdot \left( \frac{b-r}{b} \right)^2 = \frac{1}{b^2} \cdot \frac{(b-r)^2}{r}. \quad (10.12)$$

The associated derivative is:

$$\tilde{E}'(r) = \frac{r^2 - b^2}{b^2 r^2} = \frac{1}{b^2} - \frac{1}{r^2} = E'(r) + \frac{1}{b^2}, \quad (10.13)$$

showing a derivative augmentation by the constant  $1/b^2$ .

The potential shift approach by  $S_2(r)$  is often termed *force shift* for this reason. Note, however, that this name is misleading since  $S_2$  is applied to the *potential* and not to the force.

CHARMM, for example, employs  $S_1(r)$  above for group-based potential shift and  $S_2(r)$  for atom-based shifts. For the van der Waals interactions, the potential shift function is additive rather than multiplicative, applied as an auxiliary term so as to dampen the force monotonically to zero:

$$E_{\text{LJ}}(X) = \begin{cases} \sum_{i,j} \omega_{ij} \left[ \left( \frac{-A_{ij}}{r_{ij}^6} + \frac{B_{ij}}{r_{ij}^{12}} + C_{ij}r_{ij}^6 + D_{ij} \right) \right] & r \leq b \\ 0 & r > b \end{cases}. \quad (10.14)$$

Here,  $C$  and  $D$  are chosen so that both the van der Waals potential and force functions are zero at  $r = b$ . This leads to the expressions in terms of  $A_{ij}$  and  $B_{ij}$  as:

$$C_{ij} = -A_{ij} b^{-12} + 2 B_{ij} b^{-18} \quad (10.15)$$

$$D_{ij} = 2 A_{ij} b^{-6} - 3 B_{ij} b^{-12}. \quad (10.16)$$

See Figure 10.2 (lower panels) for an illustration of this shift function approximation.

## 10.4 The Ewald Method

In this section, we only sketch the efficient particle-mesh Ewald (PME) method, popular for biomolecular dynamics. See [284, 369, 1413] for technical details, and [71, 283, 1081], for example, for works which emphasize the impact of PME on nucleic acid simulations. An outline of the Ewald summation technique can also be found in [428].

Throughout this section, we use notation consistent with other works in this subject (coming from crystallography), though different from notation used elsewhere in this text. The symbol  $X$  for the collective Cartesian vector used elsewhere is identical to the collective position vector  $\mathbf{x}$  used here. The position vector of atom  $i$  is  $\mathbf{x}_i$ , and the interparticle distance vector from atom  $j$  to  $i$  is  $\mathbf{r}_{ij} \equiv \mathbf{x}_i - \mathbf{x}_j$ ; we express the magnitude of this distance as the Euclidean norm of the vector, or  $|\mathbf{r}_{ij}|$ .

### 10.4.1 Periodic Boundary Conditions

The Ewald method is a technique for calculating the electrostatic energy of a system on a lattice with periodic boundary conditions. By periodic conditions, we mean that the modeled system (biomolecules and solvent molecules) is placed in the *unit cell* and considered to have infinitely many images in space. This replication forms an infinite lattice in 3D space. Figure 10.4 shows a standard square lattice in 2D of dimension  $L$ ; this lattice generalizes to a cubic geometry in 3D.

#### Space-Filling Polyhedra

Non-cubic periodic domains can also be modeled in 3D, such as the truncated octahedron, hexagonal prism, and face-centered cube, as shown in Figure 10.5. These lattice geometries reflect variations in the relationships among the three side lengths ( $a, b, c$ ) and the three angles between these sides ( $\alpha, \beta, \gamma$ ) of the domain. For some commonly used space-filling lattices we have the following relations:

1. **Cubic:**  $a = b = c$  and  $\alpha = \beta = \gamma = 90^\circ$ .
2. **Hexagonal:** (e.g., hexagonal prism):  $a = b$  and  $\alpha = \beta = 90^\circ$  and  $\gamma = 120^\circ$ .
3. **Rhombic Dodecahedron** (e.g., face-centered cube):  $a = b = c$  and  $\alpha = \gamma = 60^\circ$  and  $\beta = 90^\circ$ .

More generally, other space-filling polyhedra are shown in Figure 10.6.

#### Minimum-Image Convention

Only coordinates of the unit cell need to be recorded and propagated. As an atom leaves the unit cell by crossing the boundary, an image enters to replace it, and hence the total number of particles is conserved. In biomolecular simulations, the unit cell is usually set to be large enough to limit such occurrences and

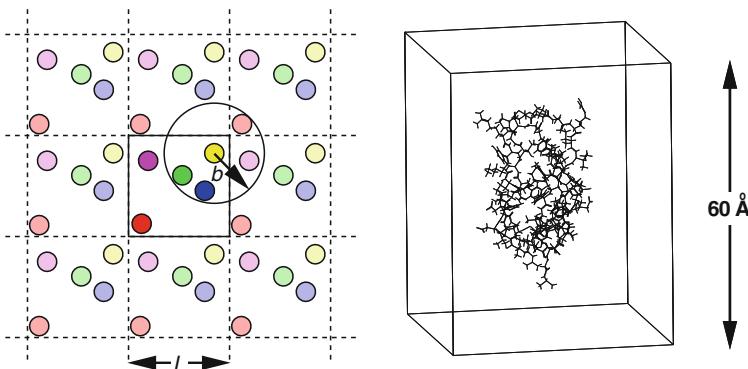


Figure 10.4. Periodic domain in 2D, showing the unit cell (center) and its images (surrounding replicas), and in 3D (rectangular), as used for a solvated protein (BPTI) system [1089].

to avoid artifacts that can be caused by artificial periodic boundary conditions [510, 580, 581, 1203]. Such artifacts may be more pronounced in a solvent of low dielectric permittivity (not water), for a solute having a large overall charge, and when the solute cavity is relatively large compared to the size of the unit cell (due to artificially small solvation of extended solute). Specifically, it has been found that artificial periodicity perturbs the potentials of mean force associated with conformational equilibria of solvated biomolecules and can tend to overstabilize compact folded forms. A reasonable cell size involves a solvated biomolecule with at least a 10 Å layer of water molecules [291].

In practical terms, a *minimum-image convention* is typically used so that each atom  $i$  interacts only with the *closest* periodic image of the other  $N - 1$  particles. In addition, a spherical cutoff (as described in the last section) is applied to restrict this number of calculated interactions further. For a consistent combination of spherical cutoffs and the minimum-image convention, the cutoff distance ( $b$  of last section) is at most  $L/2$ , where  $L$  is the dimension of the side of the box. Note from Table 10.1 that applying periodic boundary conditions only about doubles the computational work with respect to using 12 Å cutoffs (for a finite system). Further, implementing the particle-mesh Ewald fast summation scheme (to approximate all nonbonded interactions in the periodic model) requires only slightly more work than the periodic summation with this cutoff.

### Choice of Geometry

The geometry of the periodic domain used also affects the total number of pairwise interactions considered. Note from Figure 10.5, for example, that a truncated octahedron may be more compact than a cubic lattice for a solvated protein, while a hexagonal prism is appropriate for solvated DNA systems. Programs to optimize such models, that is, to orient and solvate the solute macromolecule in a given domain subject to a minimal water layer thickness (e.g., 10 Å) so as to

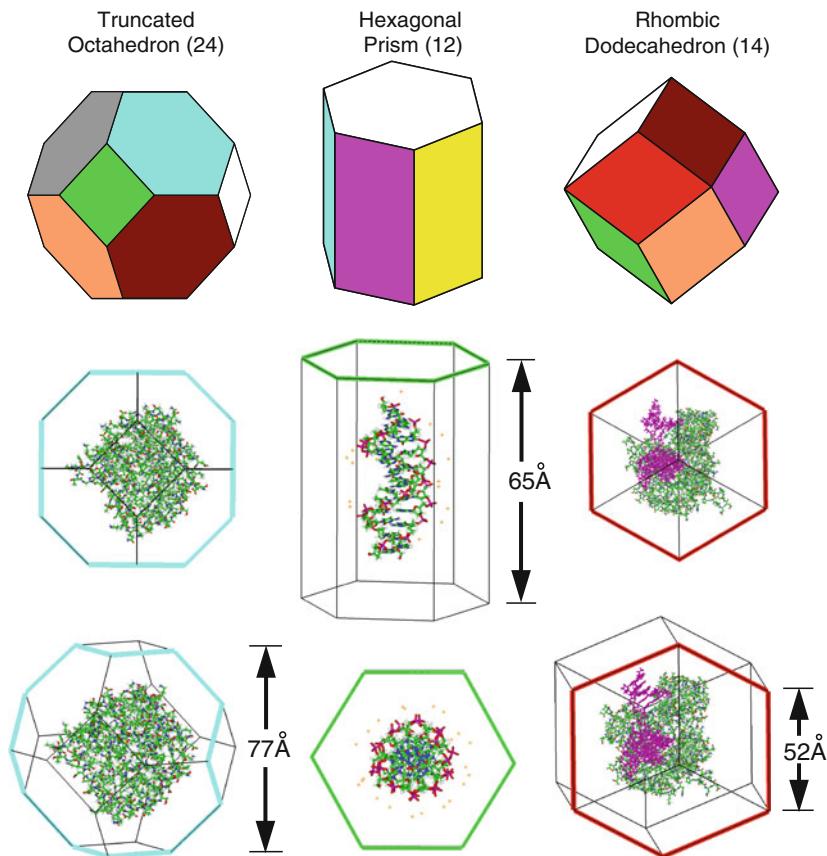


Figure 10.5. Examples of nonrectangular periodic domains in 3D used for biomolecular simulations: *truncated octahedron*, containing a solvated protein (villin) [338]; *hexagonal prism*, containing a solvated DNA dodecamer [1230]; and *rhombic dodecahedron* (face-centered cube), containing a polymerase/DNA complex [1134]; water molecules (not shown) fill the domain. A side and top view is shown for each geometry, as well as a space-filling rendering (at top) indicating the number of vertices in each case.

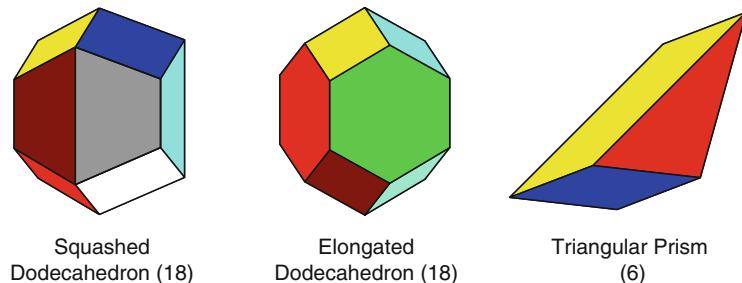


Figure 10.6. Three additional space-filling polyhedra are shown, with the number of vertices for each domain indicated.

yield the lowest number of water molecules, were pioneered by Mihaly Mezei [857] (see references and details for program **Simulaid** on the author's website) and recently generalized, with enhanced efficiency, for additional space-filling polyhedra in the program **PBCAID** [1027].

### 10.4.2 Ewald Sum and Crystallography

The Ewald construct has its roots in the crystallographic community. There are three specific commonalities between crystallography and the Ewald summation for fast electrostatics:

1. **Reciprocal Term.** The notion of the *reciprocal space* is used in addition to the direct space. In crystallography, an Ewald sphere describes the diffraction conditions in terms of the reciprocal rather than direct lattice. The reciprocal lattice is an orthogonal system related to the orthogonal system associated with the atoms in the unit cell (termed the real-space lattice). This lattice is used to express the scattering vector in crystals (see Box 10.1). In the Ewald summation, two terms arise for the computation of the electrostatic energy, for atom pairs in the *direct lattice* and for atoms pairs corresponding to interactions with *images* of the unit-cell atoms. The former is evaluated by direct calculation, and the latter (smooth and long-range) by Fourier transforms.
2. **Fast Fourier transforms.** FFTs are key ingredients in both applications. In crystallography, FFTs are used to express the amplitude of diffraction for the whole crystal from the electron density function associated with one unit cell of the crystal (see Box 10.1). In the Ewald sum, FFTs are used to evaluate the electrostatic energy corresponding to the reciprocal lattice.
3. **Electron Density Distributions.** The notion of *electron density distribution* is key in both applications. Crystallographers compute the total electron density to obtain a spatial arrangement of the atoms in space in terms of FFTs of the *structure factors*, which describe the scattering pattern of each atom with respect to an electron at the cell origin (see Box 10.1). Modelers use an analogous concept to describe the electrostatic energy in terms of charge densities rather than point charges to make the infinite sums convergent. Terms analogous to the crystallographic structure factors also arise in the Ewald sum FFTs.

#### Box 10.1: Some Tricks from X-Ray Crystallography

A crystal is a 3D periodic arrangement of atoms (see [1047], for example, for an introduction to crystallography, and Chapter 1 for an introduction to the phase problem).

The structural subunit of crystals is the *unit cell*. The X-ray diffraction pattern by the electrons of a molecular crystalline system is interpreted in terms of both the *direct lattice*, the orthogonal system associated with the unit-cell atoms, and the *reciprocal lattice*, an auxiliary orthogonal system used to express the scattering vector in crystals. Specifically, the scattering of each atom  $j$  with respect to an electron at the origin of the cell it occupies is:

$$F_{\mathbf{s}_j} = f_j \exp(2\pi i \mathbf{x}_j \cdot \mathbf{s}), \quad (10.17)$$

where  $i = \sqrt{-1}$ ,  $f_j$  is an experimentally-measured amplitude,  $\mathbf{x}_j = \{x_{j1}, x_{j2}, x_{j3}\}$  denotes the Cartesian position vector of atom  $j$  in the direct lattice, and  $\mathbf{s}$  is the scattering vector corresponding to a triplet of indices (*indices of reflection*), such as  $\mathbf{s} = \{h, k, l\}$ , in the reciprocal lattice. Thus for  $N$  atoms in the unit cell, we write the complex-valued *structure factor*  $F_{\mathbf{s}}$  as:

$$F_{\mathbf{s}} = \sum_{j=1}^N f_j \exp(2\pi i \mathbf{x}_j \cdot \mathbf{s}) = A_{\mathbf{s}} + iB_{\mathbf{s}}, \quad (10.18)$$

$$A_{\mathbf{s}} = \sum_{j=1}^N f_j \cos(2\pi i \mathbf{x}_j \cdot \mathbf{s}), \quad B_{\mathbf{s}} = \sum_{j=1}^N f_j \sin(2\pi i \mathbf{x}_j \cdot \mathbf{s}). \quad (10.19)$$

From these quantities, the phase angles associated with  $F_{\mathbf{s}}$  are computed as  $\phi_{\mathbf{s}} = \tan^{-1}(B_{\mathbf{s}}/A_{\mathbf{s}})$ . The intensity of the diffracted X-ray beam is proportional to the square of the structure factor's amplitude and depends only on the interatomic vectors, not on the actual, origin-dependent atomic coordinates:

$$|F_{\mathbf{s}}|^2 = A_{\mathbf{s}}^2 + B_{\mathbf{s}}^2 = \sum_{i=1}^N \sum_{j=1}^N f_i f_j \cos[2\pi (\mathbf{x}_i - \mathbf{x}_j) \cdot \mathbf{s}]. \quad (10.20)$$

The relationship between crystallography and the electrostatic potential arises by relating the diffracted amplitude to a continuous distribution of electron density,  $\rho(\mathbf{x})$ , expressed as electrons per unit volume. The scattered amplitude for the whole crystal,  $F(\mathbf{s})$ , from a small volume element  $dV$ , results from an effective point charge of  $\rho(\mathbf{x}) dV$  electrons:

$$F(\mathbf{s}) = \int_V \rho(\mathbf{x}) \exp(2\pi i \mathbf{s} \cdot \mathbf{x}) dV, \quad (10.21)$$

where the integration is taken over the volume  $V$  of the unit cell. This is the Fourier transform of the electron density in one unit cell of the crystal, sampled at points  $\mathbf{s}$  on the reciprocal lattice. The inverse transformation yields the electron density via integration over the entire volume of the reciprocal space in which  $\mathbf{s}$  is defined:

$$\rho(\mathbf{x}) = \int_{V^*} F(\mathbf{s}) \exp(-2\pi i \mathbf{s} \cdot \mathbf{x}) dV^*. \quad (10.22)$$

Crystallographers compute a discrete analog of this electron density to obtain the atomic arrangement in the crystal in which each point of the reciprocal lattice (where  $F(\mathbf{s})$  is defined) has an associated weight of  $F_s/V$  where  $F_s$  is the structure factor defined in eq. (10.18):

$$\rho(\mathbf{x}) = \frac{1}{V} \sum_{h,k,l=-\infty}^{\infty} F_s \exp(-2\pi i \mathbf{s} \cdot \mathbf{x}). \quad (10.23)$$

This computation is performed using Fast Fourier transforms. The amplitudes of these structure factors are known from the observed intensities of the X-ray reflections, but the phase angles cannot be measured. This is the heart of the problem of deducing structure from X-ray diffraction patterns.

---

### 10.4.3 Mathematical Morphing of a Conditionally Convergent Sum

#### Coulomb Energy in Periodic Domains

The following Ewald sum describes the total Coulomb energy corresponding to a system in an infinite periodic domain; the conversion factor  $K_{\text{coul}}$  (see eq. (9.41) of Chapter 9) is omitted for simplicity, and the dielectric constant  $\epsilon = 1$ :

$$E_{\text{coul}} = \frac{1}{2} \sum_{i,j=1}^N \sum'_{\substack{\text{images } |\mathbf{n}|}} \frac{q_i q_j}{|\mathbf{r}_{ij} + \mathbf{n}|} = \frac{1}{2} \sum_{j=1}^N q_j \Phi(\mathbf{x}_j), \quad (10.24)$$

where

$$\Phi(\mathbf{x}_j) = \sum_{i=1}^N \sum'_{|\mathbf{n}|} \frac{q_i}{|\mathbf{r}_{ij} + \mathbf{n}|}. \quad (10.25)$$

The absolute value signs above denote vector magnitudes, and the summation extends over all images of the unit cell. The vector  $\mathbf{n}$  denotes the vector  $\mathbf{n} = (n_x L, n_y L, n_z L)$ , where  $n_x, n_y$ , and  $n_z$  are integers and  $L$  is the box size, and where the triplet of indices  $(n_x, n_y, n_z)$  are related to the magnitude of  $\mathbf{n}$ , namely  $|\mathbf{n}|$ , by  $|\mathbf{n}| = |n_x| + |n_y| + |n_z|$ .

Thus, for example, the ‘image’ corresponding to  $|\mathbf{n}| = 0$  has only one triplet:  $\mathbf{n} = (0, 0, 0)$ ;  $|\mathbf{n}| = 1$  has all triplets that can be written as:  $\mathbf{n} = (\pm L, 0, 0)$ ,  $(0, \pm L, 0)$ , and  $(0, 0, \pm L)$ ; and  $|\mathbf{n}| = 2$  is associated with the vectors that can be expressed as  $\mathbf{n} = (\pm 2L, 0, 0)$ ,  $(0, \pm 2L, 0)$ ,  $(0, 0, \pm 2L)$ ,  $(\pm L, \pm L, 0)$ ,  $(\pm L, 0, \pm L)$ , and  $(0, \pm L, \pm L)$ . Similarly, the reciprocal lattice vectors  $|\mathbf{m}|$  defined below correspond to vectors  $(L/n_x, L/n_y, L/n_z)$ .

The sum over the images extends from  $|\mathbf{n}| = 0$  to  $\infty$ . The prime symbol in the sum ( $\sum'_{\mathbf{n}}$ ) indicates that for  $\mathbf{n} = 0$  we omit the  $i = j$  interaction (so the denominator is well defined).

#### Conditional Convergence

The sum in eq. (10.24) is only *conditionally convergent* since the terms decay as  $1/|\mathbf{n}|$ , like the harmonic series.<sup>2</sup> The value of the sum also depends on the

---

<sup>2</sup>A series  $S = \sum_{n=1}^{\infty} a_n$  is conditionally convergent if the infinite sum  $\sum_n a_n$  converges but  $\sum_n |a_n|$  diverges. It is also known that the sum for a conditionally convergent series depends on the

nature of the surrounding medium (dielectric constant  $\epsilon$ ) [293]. For a unit cell surrounded by vacuum, a dipolar layer exists on the surface and its contribution must be canceled.

### Ewald's Trick

Ewald noted a trick that can be used to convert this sum into an expression involving a sum of two absolutely and rapidly convergent series in direct and reciprocal space. This conversion is accomplished by representing each point charge as a Gaussian charge density, producing an exponentially decaying function. This Gaussian transformation must be counteracted by an analogous subtraction to leave the net result of an effective point charge (delta function charge, see Fig. 10.7). This canceling distribution is summed in the reciprocal space (and transformed back to real space) because it is a smoothly-varying periodic function which can be represented by a rapidly convergent Fourier series.

Essentially, the basic idea can be written for the radial function  $\Phi(r)$  as the splitting

$$\Phi(r) \equiv \frac{1}{r} = \Phi_{\text{real}}(r) + \Phi_{\text{recip}}(r), \quad (10.26)$$

where

$$\Phi_{\text{real}}(r) = \frac{1}{r} - \frac{\text{erf}(\beta r)}{r} = \frac{\text{erfc}(\beta r)}{r}, \quad (10.27)$$

and

$$\Phi_{\text{recip}}(r) = \frac{\text{erf}(\beta r)}{r}. \quad (10.28)$$

The first term ( $\Phi_{\text{real}}$ ) is short-range with singularity at the origin, and the second ( $\Phi_{\text{recip}}$ ) is a smooth and long-range term (which can be Fourier transformed). The error function erf is defined as

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) dt, \quad (10.29)$$

$$= \frac{2}{\sqrt{\pi}} \left( x - \frac{x^3}{3} + \frac{1}{2!} \frac{x^5}{5} - \frac{1}{3!} \frac{x^7}{7} + \dots \right), \quad (10.30)$$

and the complementary error function erfc is

$$\text{erfc}(x) = 1 - \text{erf}(x). \quad (10.31)$$

order of summation. For  $a_n = (-1)^{n+1}/n$ , the ‘alternating harmonic series’  $1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \dots$  converges, but the harmonic series  $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \dots$  diverges, though  $a_n \rightarrow 0$  as  $n \rightarrow \infty$ . To see that the sum for finite  $n$  terms can be as large as we please for the harmonic series, note that terms can be grouped so that each subsum is larger than  $\frac{1}{2}$ :  $\sum_{n=1}^{\infty} \frac{1}{n} = 1 + (\frac{1}{2} + \frac{1}{3}) + (\frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7}) + (\text{next 8 terms}) + \dots > 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \dots$ .

In practice,  $\text{erf}(x)$  can also be evaluated from its relationships to other known functions, like incomplete gamma functions [1015] or the normal probability function  $f(t) = \exp(-t^2/2)/\sqrt{2\pi}$ , since

$$\text{erf}(x/\sqrt{2}) = 2 \int_0^x f(t) dt,$$

and this integral (area under the standard normal curve from 0 to  $x$ ) is often tabulated (see CRC Standard Mathematical Tables).

The Screening Gaussian  $\rho_{G_j}$

More generally (for the periodic potential), we perform this splitting using a screening spherical Gaussian  $\rho_G$ , centered at each point charge, parameterized by  $\beta$ . The parameter  $\beta$  controls the width of the distribution (width =  $\sqrt{2}/\beta$ ) and also the rate of convergence of the sum (see below). For each point charge, the surrounding distribution is given by:

$$\rho_{G_j}(\mathbf{x}) = -q_j \left( \frac{\beta}{\sqrt{\pi}} \right)^3 \exp[-\beta^2 |\mathbf{x}|^2]. \quad (10.32)$$

Thus, instead of the cumulative point charge density described by a sum of delta functions:

$$\rho(\mathbf{x}) = \sum_{j=1}^N q_j \delta(\mathbf{x} - \mathbf{x}_j), \quad (10.33)$$

we use a sum of localized densities based on the Gaussian of eq. (10.32) to define the total Gaussian screening charge density  $\rho_G(\mathbf{x})$ :

$$\rho_G(\mathbf{x}) \equiv \sum_{j=1}^N \rho_{G_j}(\mathbf{x}) = - \sum_{j=1}^N q_j \left( \frac{\beta}{\sqrt{\pi}} \right)^3 \exp[-\beta^2 |\mathbf{x} - \mathbf{x}_j|^2]. \quad (10.34)$$

The real-space sum can be written in terms of the complementary error function using Poisson's equation<sup>3</sup> for the electrostatic potential, due to the Gaussian charge cloud followed by integration [428]. (See Box 10.2 for a definition of the divergence and Laplace operators). Namely, the solution of Poisson's equation in the form  $\nabla^2 \Phi_{G_j}(r) = -4\pi \rho_{G_j}(r)$  leads to  $\Phi_{G_j}(r) = \frac{q_j}{r} \text{erf}(\beta r)$  after two steps of integration. Thus, the sum in eq. (10.24) is converted into the pair of terms:

$$E_{\text{coul}} = \frac{1}{2} \sum_{j=1}^N \sum'_{|\mathbf{n}|} \frac{q_j \text{erfc}(\beta |\mathbf{x}_{ij} + \mathbf{n}|)}{|\mathbf{r}_{ij} + \mathbf{n}|} + \frac{1}{2} \sum_{j=1}^N \sum'_{|\mathbf{n}|} \frac{q_j \text{erf}(\beta |\mathbf{x}_{ij} + \mathbf{n}|)}{|\mathbf{r}_{ij} + \mathbf{n}|}. \quad (10.35)$$

---

<sup>3</sup>Simeon Denis Poisson (1781–1840) was a brilliant mathematician whose name frequently appears in text books. The Poisson equation (1812) in potential theory is a result of his discovery that Laplace's equation for the gravitational force holds only at points where no mass is located (see also Box 10.2).

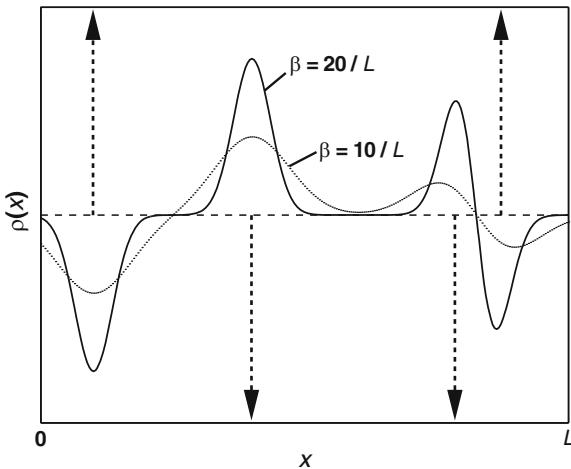


Figure 10.7. Schematic illustration of Ewald's trick for converting a conditionally convergent sum into two terms that are rapidly and absolutely convergent: each point charge (represented by delta functions, dashed spikes) is masked by a Gaussian function; correspondingly, a canceling Gaussian function must be added (not shown). The charge density when point charges are screened by Gaussians is shown for four point charges for two values of  $\beta$ , the parameter controlling the width of the Gaussian distribution.

The first term ( $E_{\text{real}}$ ) is short-range, and the second ( $E_{\text{recip}}$ ) is the smooth long-range component of the Coulomb potential.

### Box 10.2: The Divergence and Laplacian Operators

The divergence symbol, as used in the Poisson and Poisson-Boltzmann equations, is a differential operator on a continuously differentiable vector  $\mathbf{w} \in \mathcal{R}^n$  whose components  $\{w_i\}$  are functions of the coordinate variable  $\mathbf{x}$  with components  $\{x_i\} \in \mathcal{R}^n$ . The divergence of  $\mathbf{w}$  is written as  $\text{div } \mathbf{w}$  or  $\nabla \cdot \mathbf{w}$  and defined as:

$$\text{div } \mathbf{w} \equiv \nabla \cdot \mathbf{w} \equiv \frac{\partial w_1(\mathbf{x})}{\partial x_1} + \frac{\partial w_2(\mathbf{x})}{\partial x_2} + \cdots + \frac{\partial w_n(\mathbf{x})}{\partial x_n}.$$

For example, for the gravitational field vector  $\mathbf{v}(\mathbf{x})$  at a point  $\mathbf{x}$  in space whose distance from the origin is  $r$ , we have  $\mathbf{v}(\mathbf{x}) = -(GR_e^2/r^3)\mathbf{x}$  where  $R_e$  is the earth's radius and  $G$  is the gravitational constant. For  $r > R_e$ , it can be shown that the divergence of  $\mathbf{v}$  is zero:

$$\begin{aligned} \nabla \cdot \mathbf{v}(\mathbf{x}) &= -\frac{GR_e^2}{r^3} \nabla \cdot \mathbf{x} + \nabla \left( \frac{-GR_e^2}{r^3} \right) \cdot \mathbf{x} \\ &= -GR_e^2 \left[ \frac{1}{r^3} \nabla \cdot \mathbf{x} + \frac{d}{dr} \left( \frac{1}{r^3} \right) \nabla r \cdot \mathbf{x} \right] = 0, \end{aligned}$$

since  $\nabla \cdot \mathbf{x} = 3$  and  $\nabla r \cdot \mathbf{x} = (\mathbf{x} \cdot \mathbf{x})/r = r$ . The corresponding potential function  $u(\mathbf{x}) = GR_e^3/r$ , where  $\nabla u = \mathbf{v}(\mathbf{x})$ , thus satisfies  $\nabla \cdot \nabla u = \nabla \cdot \mathbf{v} = 0$  or Laplace's

equation, also written as  $\nabla^2 u = 0$  or  $\Delta u = 0$ . The *Laplacian* is a differential operator from scalar to scalar fields whereas the *divergence* is a differential operator from vector to scalar fields.

---

#### 10.4.4 Finite-Dielectric Correction

The decomposition above, while convergent, is strictly correct only for an infinite dielectric ( $\epsilon = \infty$ ) medium or for unit cells with a zero dipole moment. The correction term ( $E_{\text{cor},\epsilon}$ ) for a nonuniform field associated with a macroscopic crystal in a dielectric continuum with external dielectric constant  $\epsilon$  was only derived 60 years after Ewald's original derivation [293], and yields:

$$E_{\text{cor},\epsilon} = E_{\text{coul}}(\epsilon = 1) - E_{\text{coul}}(\epsilon = \infty) = \frac{2\pi}{3L^3} \left| \sum_{j=1}^N q_j \mathbf{x}_j \right|^2. \quad (10.36)$$

Though reliable, Ewald's algorithm as corrected in [293] was still  $\mathcal{O}(N^2)$  in computational complexity. This is because the long-range reciprocal-space Fourier sum requires  $\mathcal{O}(N^2)$  to be a sufficiently accurate approximation for large  $\beta$  (see below).

#### 10.4.5 Ewald Sum Complexity

The key breakthroughs for reducing the computational complexity of the Ewald sum came in two steps.

##### Optimization of $\beta$

First, by optimizing the parameter  $\beta$  that controls the width of the screening Gaussians (see Figure 10.7), the relative convergence rates of the real and reciprocal-space series are adjusted to optimize the work involved [402]. Namely, as  $\beta$  increases, the real-space sum converges more rapidly and the reciprocal sum more slowly. Both sums are truncated in practice (finite number of terms). Thus, for example, a sufficiently large  $\beta$  yields an accurate direct-space sum with an appropriate truncation, resulting in  $\mathcal{O}(N)$  work for the direct-space sum rather than  $\mathcal{O}(N^2)$ ; however, the reciprocal-space sum still requires  $\mathcal{O}(N^2)$  work. The optimal work balance between the two components can yield an overall  $\mathcal{O}(N^{3/2})$  method by adjusting  $\beta$  [991]. This is much better than  $\mathcal{O}(N^2)$  but still considerably expensive for large biomolecular systems.

##### Mesh Interpolation

The second breakthrough in the Ewald sum came by noting that the trigonometric-function values in the Fourier series used to represent the reciprocal-space term

can be evaluated through a smooth interpolation of the potential over a regular grid. The resulting particle-mesh Ewald (PME) method reduces the overall computation to  $\mathcal{O}(N \log N)$ . The smoothing can be done by Lagrange interpolation [284] or by B-spline interpolation [369].

### Variations

Credit for this interpolation idea is due to Hockney and Eastwood who developed several methods in the early 1970s for simulations of hot gas plasmas in fusion machines. Included is a ‘particle-particle particle-mesh’ ( $P^3M$ ) scheme, as detailed in their text [554]. The  $P^3M$  method for evaluating long-range forces in large systems splits the interparticle force summation into a short-range, rapidly-varying part and a smooth, slowly-varying remainder. A direct ‘particle-particle’ sum is used to compute the former, while a ‘particle-mesh’ interpolation procedure is used to approximate the latter on a uniform grid. A ‘Q-minimizing’ method is also used to optimize the scheme’s parameters (given a desired accuracy), such as mesh size, cutoff radius, and charge assignment scheme.

The  $P^3M$  method and its cousins are thus closely related to the fast PME schemes used in biomolecular dynamics. Application of the  $P^3M$  scheme as described in [554] to molecular dynamics has also demonstrated good performance [792, 793], with guidelines for the scheme’s parameters obtained by optimizing an auxiliary function.

#### 10.4.6 Resulting Ewald Summation

Following a series of mathematical manipulations (see [1267], for example), the resulting composition of the Ewald summation for  $E_{\text{coul}}$  of eq. (10.24) at  $\epsilon = 1$  has five terms:

$$E_{\text{coul}} = E_{\text{real}} + E_{\text{recip}} + E_{\text{cor},\text{self}} + E_{\text{cor},\text{ex}} + E_{\text{cor},\epsilon}. \quad (10.37)$$

The first term,  $E_{\text{real}}$ , corresponds to a *real-space* (or direct) sum of the electrostatic energy due to the point charges screened by oppositely-charged Gaussians. The second sum,  $E_{\text{recip}}$ , is the *associated canceling term* (periodic sum of Gaussians) summed in reciprocal space using smooth interpolation of Fourier-series values. The last three terms are correction terms.

The first correction (which is position independent) subtracts the *self-interaction term* (each point charge and its corresponding Gaussian charge cloud) which is included in the first two terms. The second correction subtracts the Coulomb contribution from the *nonbonded pairs excluded* from the Coulomb energy (denoted as pairs  $i, j \in \text{Ex}$  below) since they are separately accounted for in bonded, bond-angle, and possibly dihedral-angle terms. The third correction accounts for the *non-infinite dielectric medium* (eq. (10.36)).

Following algebraic manipulations, the resulting Ewald sum can be expressed as follows:

$$E_{\text{real}} = \frac{1}{2} \sum_{i,j=1}^N q_i q_j \sum'_{|\mathbf{n}|} \frac{\operatorname{erfc}(\beta |\mathbf{r}_{ij} + \mathbf{n}|)}{|\mathbf{r}_{ij} + \mathbf{n}|}; \quad (10.38)$$

$$E_{\text{recip}} = \frac{1}{2\pi L^3} \sum_{|\mathbf{m}| \neq 0} \frac{\exp(-\pi^2 |\mathbf{m}|^2 / \beta^2)}{|\mathbf{m}|^2} S(\mathbf{m}) S(-\mathbf{m}), \quad (10.39)$$

$$\text{where } S(\mathbf{m}) = \sum_{j=1}^N q_j \exp[2\pi i \mathbf{m} \cdot \mathbf{x}_j]; \quad (10.40)$$

$$E_{\text{cor, self}} = \frac{-\beta}{\sqrt{\pi}} \sum_{j=1}^N q_j^2; \quad (10.41)$$

$$E_{\text{cor, ex}} = -\frac{1}{2} \sum_{i,j \in \text{Ex}}^N q_i q_j \frac{\operatorname{erf}(\beta |\mathbf{r}_{ij}|)}{|\mathbf{r}_{ij}|}; \quad (10.42)$$

$$E_{\text{cor, } \epsilon} = \frac{2\pi}{(1+2\epsilon)L^3} \left| \sum_{j=1}^N q_j \mathbf{x}_j \right|^2. \quad (10.43)$$

The  $\operatorname{erfc}$  function decays to zero with increasing values of the independent variable. If  $\beta$  is sufficiently large, the only term that contributes to the real-space sum is for  $|\mathbf{n}| = 0$ , and it can be computed in practice with a spherical cutoff of order  $b = 10 \text{ \AA}$ .

The reciprocal space or Fourier term corresponds to a summation over the reciprocal vectors  $\mathbf{m}$ , where the Fourier terms  $S(\mathbf{m})$  are *charge-weighted structure factors* (see Box 10.1). Eqs. (10.39) and (10.40) can also be written as:

$$E_{\text{recip}} = \frac{1}{2\pi L^3} \sum_{i,j=1}^N q_i q_j \sum_{|\mathbf{m}| \neq 0} \frac{\exp(-\pi^2 |\mathbf{m}|^2 / \beta^2)}{|\mathbf{m}|^2} \exp[2\pi i \mathbf{m} \cdot (\mathbf{x}_j - \mathbf{x}_i)]. \quad (10.44)$$

When the charges are interpolated over a uniform set of grid points, these structure factors are easily computed by FFTs.

#### 10.4.7 Practical Implementation: Parameters, Accuracy, and Optimization

##### Gaussian Width

In practice, the Gaussian width parameter  $\beta$  is determined so that the real-space term achieves a desired accuracy tolerance. In typical solvated proteins or DNA simulations,  $\beta$  has an order of magnitude of roughly  $10/L$  ( $L$  ranges roughly

from 60 to 100 Å; values exceeding 70 Å have been implicated with smaller artifacts from the enforced periodicity [580]). The effective cutoff used for the real-space interactions is around 10 to 12 Å. When multiple-timestep schemes are implemented for molecular dynamics, the parameter  $\beta$  (or the cutoff for the direct-space term) may be further optimized to distribute the work for the real and reciprocal terms appropriately, so as to yield the greatest overall speedup; see [97], for example.

### Grid Size and Accuracy

The reciprocal-space uses multidimensional piecewise interpolation (e.g., B-splines) for evaluating the Fourier terms, with grid size and number of terms chosen to achieve the desired accuracy. For instance, moderate accuracy (e.g.,  $10^{-4}$  relative force error) might be achieved with a coarse interpolation grid (1 to 2 Å). Very high accuracy (e.g.,  $10^{-10}$  relative force error) might be obtained with a finer grid ( $\sim 0.5$  Å). The reciprocal-space energy and force terms are expressed as *convolutions* and can thus be evaluated quickly using FFTs.<sup>4</sup> Work has also shown that the finite number of wave vectors used in the discrete approximation gives rise to truncation errors due to the exclusion of intramolecular interactions [1019]; this, and the related existence of fast terms in the reciprocal-force component [97, 98, 1025, 1236], create a problem for development of efficient multiple-timestep integrators for molecular dynamics simulations using PME approximations [89].

### Computer Architecture Considerations

In the implementation of the PME method on multiprocessors, the cutoff radius used for the direct sum may be increased to balance the work between the real-space and reciprocal-space components. This accelerates the computations because 3D FFTs are challenging to parallelize well. In contrast, the direct sum parallelizes easily by spatial decomposition. Using a larger cutoff in the real-space sum reduces the number of lattice vectors (Fourier terms) needed for the same accuracy in the reciprocal sum and therefore improves the overall performance.

Experience to date shows that PME implementations for biomolecules — especially when tightly adapted to the computer architecture — are very fast; the best implementations require about the same work as needed to evaluate the same periodic version of the potential but with cutoffs in the range of 10–12 Å [1133]. Figure 10.8 shows performance times for an efficiently distributed PME code by John Board and co-workers at Duke University for a huge water system.

---

<sup>4</sup>For two functions of time  $f(t)$  and  $g(t)$  and corresponding Fourier transforms  $F(f)$  and  $F(g)$ , we define the *convolution* of the two original functions  $f$  and  $g$ ,  $f * g$ , as:  $f * g = \int_{-\infty}^{\infty} f(\tau) g(t - \tau) d\tau$ . It can be shown that  $F(f * g) = F(f) F(g)$ . That is, the Fourier transform of the convolution of two functions ( $f * g$ ) is just the product of the individual Fourier transforms of those functions.

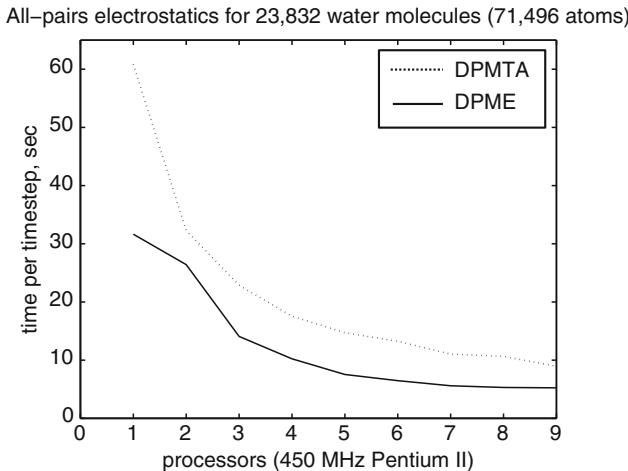


Figure 10.8. CPU time for evaluating the electrostatic energy for 23,832 water molecules (71,496 atoms) by codes developed at Duke by John Board and co-workers for the distributed PME (DPME) and distributed parallel multipole tree algorithm (DPMTA), run at modest accuracy of  $10^{-3}$  relative force accuracy, on a tightly coupled network of 450 MHz Pentium II processors.

See also a review on modeling electrostatic effects in proteins [1347], which discusses stability and accuracy issues with traditional PME, or non-spherical Ewald methods as described above. These problems can be alleviated by using spherical boundary conditions in combination with a local field approach. The above review and others argue that only spherical Ewald methods rigorously give correct results for charged systems.

## 10.5 The Multipole Method

In this section, we present a brief introduction to the efficient fast multipole technique. For details, see [479, 481] and other references cited below.

### 10.5.1 Basic Hierarchical Strategy

Fast multipole techniques are powerful alternatives to the Ewald schemes introduced above for evaluating the pairwise interactions in large molecular systems. They are widely used for many important problems in applied mathematics, engineering, physics, chemistry, and biology [478, 480]. Examples in astrophysics include evaluation of gravitational potentials, and examples in chemistry include electrostatic potentials in molecular dynamics and quantum mechanics (Hartree-Fock calculations).

## Series Expansion

Multipole schemes rely on a power-series expansion that describes the interaction between groups of particles (charged bodies in molecular dynamics) [1226]. The term *multipole* refers to the *moments*  $m^k = \sum_{i=1}^N q_i x_i^k$  that appear in the expansion, such as dipole and quadrupole. To see this, consider the electrostatic potential  $\Phi$  at  $\mathbf{x}$  for unit dielectric written in terms of the charge distribution about each atom  $j$  as (see also eqs. (10.24), (10.25) for the periodic version):

$$\Phi(\mathbf{x}) = \Phi(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N) = \frac{1}{2} \sum_{\{i,j\}, i \neq j} q_i q_j / |\mathbf{r}_{ij}| = \frac{1}{2} \sum_j q_j \Phi(\mathbf{x}_j), \quad (10.45)$$

where

$$\Phi(\mathbf{x}_j) = \sum_{i \neq j} q_i / |\mathbf{r}_{ij}|. \quad (10.46)$$

An expansion of  $\Phi(\mathbf{x}_j)$  yields (see Box 10.3):

$$\Phi(\mathbf{x}_j) = \Phi(0) + \nabla \Phi^T \mathbf{x} + \frac{1}{2} \mathbf{x}^T \mathbf{H} \mathbf{x} + \mathcal{O}(|\mathbf{x}|^3) \quad (10.47)$$

where the derivatives are evaluated at  $\mathbf{x} = 0$ . This expansion produces expressions for  $\Phi(0)$  based on the *zeroth moment* ( $\Phi(0) = m^0 / |\mathbf{x}_j|$ ), for  $\nabla \Phi(\mathbf{x})$  in terms of the *dipole moment* (vector  $\mu = [m_1^1, m_2^1, m_3^1]$ ), and for the second-derivative term based on the *quadrupole moment* (elements  $m_{11}^2, m_{12}^2, m_{13}^2, m_{21}^2, \dots, m_{33}^2$ ). See Box 10.3 for details.

Note that multipole expansions can be expressed using Cartesian coordinates, as in Boxes 10.3 and 10.4, or spherical coordinates, as in Subsection 10.5.3. Such power series can save computational work since the target function can be written as a linear combination of moments (see Box 10.4 for a simple example).

## Domain Decomposition

Accelerated multipole algorithms use a hierarchy of approximations to represent these interactions on the basis of a spatial particle partitioning in a tree-like structure (typically oct-tree): the original domain (level 0) is subdivided into 8 level-1 domains, leading to 64 regions in level 2, and so on (see also illustration in [316]). Thus, each successive refinement of the computational domain produces ‘offspring’ corresponding to the ‘parents’ at the prior level. This recursive partitioning, like FFTs, works together with a systematic formulation and manipulation of power series expansions of the target potential to reduce the evaluation computational complexity from  $\mathcal{O}(N^2)$  to the more modest  $\mathcal{O}(N \log N)$  and  $\mathcal{O}(N)$ , depending on the implementation (see Figure 10.9).

Multipole expansions represent one suitable choice of power series, since they converge well when groups of particles (charges) are well-separated, allowing a small number of coefficients in the expansions while maintaining good accuracy;

other expansions are possible [45, 316, 317, 460, 1329]. These power series expansions are used to compute interactions between well-separated pairs (contained in well-separated clusters). Interactions between nearby particles are computed directly.

### Summation Protocol

The various algorithms differ by how the tree structure is generated and by the protocol used to determine which power-series approximation to invoke at every stage (on the basis of distance separations monitored by interaction lists). Figure 10.9, for example, illustrates the definition of interaction lists for multipole expansions by boxes beyond first neighbors with respect to a given particle (shown in red); starting from second neighbors is an alternative definition.

Generally, good efficiency — linear complexity for moderate-sized systems — for 3D implementations requires meticulous programming and algorithmic structure [481]. This programming sophistication is especially important for applications where the particles are distributed *heterogeneously* through the computational domain (usually not the case for molecular dynamics applications). For heterogeneous cases, *adaptive* schemes are needed to distribute the computations in a balanced and efficient manner since regular subdivisions may generate some empty cells and empty cells need not be divided.

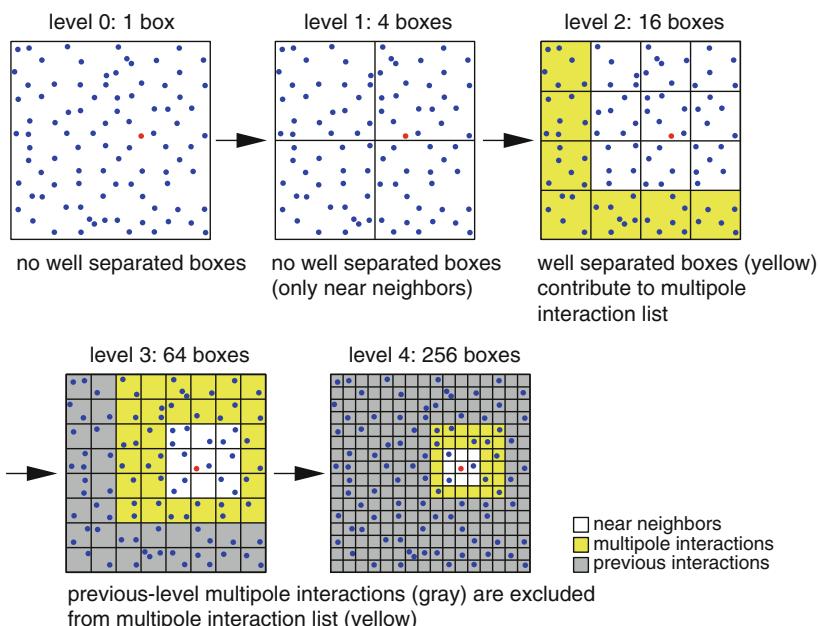


Figure 10.9. Hierarchical domain partitioning approach for fast multipole schemes. The partitioning is illustrated with respect to a central particle (shown in red).

The seasoned programming required for fast multiple algorithms partly explains why Ewald codes have been more popular in the molecular dynamics community over the last few years. It is possible that as larger systems (of order  $10^5$  atoms) become more common in macromolecular simulations, the difference between the  $\mathcal{O}(N \log N)$  (best Ewald implementation) and  $\mathcal{O}(N)$  (best multipole implementation) may become significant, and the effort required for multipole codes may well be worth the programming investment.

Besides Coulomb interactions, multipole (and Ewald) methods apply more generally to any function that can be approximated by a converging power-series expansion, such as functions of  $1/r^n$  or exponentials as in screened Coulomb (Debye-Hückel) expressions ( $\exp(-\kappa r)/r$ ) [154, 385] (see below).

### Box 10.3: Multipole Expansion

Consider a charge distribution about each atom  $j$  (in a system of  $N$  atoms) as defined in eq. (10.46), where  $\mathbf{r}_{ij} = \mathbf{x}_i - \mathbf{x}_j = [x_{i1} - x_{j1}, x_{i2} - x_{j2}, x_{i3} - x_{j3}]$ ,  $r_{ij} = |\mathbf{r}_{ij}|$ , and  $R_{ij} = (r_{ij})^2$ . An expansion at position  $\mathbf{x}$  about the origin can be written as

$$\begin{aligned}\Phi(\mathbf{x}_j) &= \sum_{i=1}^N \frac{q_i}{[(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + (x_{i3} - x_{j3})^2]^{1/2}} \\ &= \Phi(0) + \sum_{i=1}^N \sum_{k=1}^3 \frac{\partial \Phi}{\partial x_{ik}} x_{ik} + \frac{1}{2} \sum_{i=1}^N \sum_{l=1}^3 \sum_{k=1}^3 \frac{\partial^2 \Phi}{\partial x_{ik} \partial x_{il}} x_{ik} x_{il} + \text{higher order terms} \\ &\equiv \Phi(0) + \nabla \Phi^T \mathbf{x} + \frac{1}{2} \mathbf{x}^T \mathbf{H} \mathbf{x} + \mathcal{O}(|\mathbf{x}|^3)\end{aligned}$$

where derivatives are evaluated at  $x_{i1} = x_{i2} = x_{i3} = 0$ . The gradient vector  $\nabla \Phi$  and Hessian matrix  $\mathbf{H}$  above contain partial derivatives corresponding to the  $3N$  components of the vector  $\mathbf{x}$ . Namely,  $\nabla \Phi$  has the  $3N$  components  $\{\partial \Phi / \partial x_{11}, \partial \Phi / \partial x_{12}, \dots, \partial \Phi / \partial x_{N3}\}$ , and the Hessian is:

$$\mathbf{H} = \begin{pmatrix} \partial^2 \Phi / \partial x_{11}^2 & \partial^2 \Phi / \partial x_{11} \partial x_{12} & \dots & \partial^2 \Phi / \partial x_{11} \partial x_{N3} \\ \partial^2 \Phi / \partial x_{12} \partial x_{11} & \partial^2 \Phi / \partial x_{12}^2 & \dots & \partial^2 \Phi / \partial x_{12} \partial x_{N3} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \partial^2 \Phi / \partial x_{N3} \partial x_{11} & \partial^2 \Phi / \partial x_{N3} \partial x_{12} & \dots & \partial^2 \Phi / \partial x_{N3}^2 \end{pmatrix}.$$

The *moments*  $\{m^k\}$  for integers  $k$  corresponding to the partial charges  $\{q_i\}$  are defined by  $m^k = \sum_{i=1}^{3N} q_i x_i^k$  where the summation extends over all components of  $\mathbf{x}$ . The first term in the above expansion is written in terms of the zeroth moment  $m^0$ :

$$\Phi(0) = \sum_{i=1}^N \frac{q_i}{[x_{j1}^2 + x_{j2}^2 + x_{j3}^2]^{1/2}} = \sum_{i=1}^N \frac{q_i}{|\mathbf{x}_j|} = \frac{m^0}{|\mathbf{x}_j|}.$$

The gradient, or the *dipole* potential, has 3 components:  $m_1^1, m_2^1, m_3^1$ . They represent the *dipole moment*  $\mu$ . The expression for the gradient in terms of the dipole moments can be derived following the relation for the derivatives:

$$\frac{\partial R_{ij}^{-1/2}}{\partial x_{ik}} = -\frac{1}{2} R_{ij}^{-3/2} \frac{\partial R_{ij}}{\partial x_{ik}} = -r_{ij}^{-3} (x_{ik} - x_{jk})|_{x_{ik}=0} = x_{jk} / |\mathbf{x}_j|^3.$$

We then have:

$$\begin{aligned}\Phi'(\mathbf{x}_j) &= \nabla \Phi^T \mathbf{x} = \sum_{i=1}^N q_i \left( \frac{\partial R_{ij}^{-1/2}}{\partial x_{i1}} x_{i1} + \frac{\partial R_{ij}^{-1/2}}{\partial x_{i2}} x_{i2} + \frac{\partial R_{ij}^{-1/2}}{\partial x_{i3}} x_{i3} \right) \\ &= \sum_{i=1}^N \left( q_i x_{i1} \frac{x_{j1}}{|\mathbf{x}_j|^3} + q_i x_{i2} \frac{x_{j2}}{|\mathbf{x}_j|^3} + q_i x_{i3} \frac{x_{j3}}{|\mathbf{x}_j|^3} \right) \\ &= \frac{1}{|\mathbf{x}_j|^3} (m_1^1 x_{j1} + m_2^1 x_{j2} + m_3^1 x_{j3}) = \frac{(\mu^T \cdot \mathbf{x}_j)}{|\mathbf{x}_j|^3}.\end{aligned}$$

The second-derivative term, the *quadrupole potential*, has six terms (because of symmetry), again with all partial derivatives evaluated at  $x_{i1} = x_{i2} = x_{i3} = 0$ :

$$\begin{aligned}\Phi''(\mathbf{x}_j) &= \frac{1}{2} \mathbf{x}^T H \mathbf{x} = \frac{1}{2} \sum_{i=1}^N q_i \left( \frac{\partial^2 R_{ij}^{-1/2}}{\partial x_{i1}^2} x_{i1}^2 + \frac{\partial^2 R_{ij}^{-1/2}}{\partial x_{i2}^2} x_{i2}^2 + \frac{\partial^2 R_{ij}^{-1/2}}{\partial x_{i3}^2} x_{i3}^2 \right. \\ &\quad \left. + 2 \frac{\partial^2 R_{ij}^{-1/2}}{\partial x_{i1} \partial x_{i2}} x_{i1} x_{i2} + 2 \frac{\partial^2 R_{ij}^{-1/2}}{\partial x_{i1} \partial x_{i3}} x_{i1} x_{i3} + 2 \frac{\partial^2 R_{ij}^{-1/2}}{\partial x_{i2} \partial x_{i3}} x_{i2} x_{i3} \right) \\ &= \frac{1}{2} \left( m_{11}^2 \frac{\partial^2 R_{ij}^{-1/2}}{\partial x_{i1}^2} + m_{22}^2 \frac{\partial^2 R_{ij}^{-1/2}}{\partial x_{i2}^2} + \dots + 2 m_{23}^2 \frac{\partial^2 R_{ij}^{-1/2}}{\partial x_{i2} \partial x_{i3}} \right).\end{aligned}$$


---



---

#### Box 10.4: Example of Work Reduction by Multipole Expansion

Let our potential be defined as:

$$\Phi(x) = \sum_{j=1}^N \Phi(x_j) = \sum_{j=1}^N \sum_{i=1}^N q_i (x_j - y_i)^3.$$

That is,  $\Phi(x)$  is the sum of the following components:

$$\begin{array}{lll}\{\Phi(x_1)\} & q_1 (x_1 - y_1)^3 + q_2 (x_1 - y_2)^3 + \dots + q_N (x_1 - y_N)^3 \\ \{\Phi(x_2)\} & q_1 (x_2 - y_1)^3 + q_2 (x_2 - y_2)^3 + \dots + q_N (x_2 - y_N)^3 \\ & \vdots & \vdots \\ \{\Phi(x_N)\} & q_1 (x_N - y_1)^3 + q_2 (x_N - y_2)^3 + \dots + q_N (x_N - y_N)^3.\end{array}$$

Evaluation by *straightforward* summation requires  $\mathcal{O}(N^2)$  work, but our function of  $x$  and  $y$  simplifies as:

$$(x - y)^3 = x^3 - 3x^2y + 3xy^2 - y^3.$$

This is a finite power series, with degree  $p = 3$  and moments  $m^k = \sum_{i=1}^N q_i y_i^k$ . Therefore we can rewrite each  $\Phi(x_j)$  as a linear combination of the moments  $\{m^k\}$  for  $k = 0, 1, 2, 3$ :

$$\Phi(x_j) = x_j^3 m^0 - 3x_j^2 m^1 + 3x_j m^2 - m^3.$$

The work can thus be substantially reduced: once the moments are computed in  $\mathcal{O}(Np)$  work, evaluating each  $\Phi(x_j)$  requires only  $\mathcal{O}(p)$  work, where  $p$  is the power of the expansion, 3 here. *This reduction  $\mathcal{O}(N^2)$  to  $\mathcal{O}(3N)$  is significant!*

---

### 10.5.2 Historical Perspective

#### Hierarchical Refinements

Coming from the astrophysics community, Appel first introduced the multipole approach for solving the  $N$ -body problem by a hierarchical, power-series approach [53]. Barnes and Hut accelerated the association scheme between particles at successive refinement levels to produce a method that is asymptotically  $\mathcal{O}(N \log N)$  [90]. To see this let  $\mathbf{D}_{N \times N}$  be the matrix defined by

$$\mathbf{D}_{ij} = \begin{cases} q_i / |\mathbf{r}_{ij}| & i \neq j \\ 0 & i = j \end{cases} \quad (10.48)$$

and  $\mathbf{q}$  be the vector of  $N$  partial charges. Hence, the potential for each atom  $j$  due to the charges induced by all other atoms is

$$q_j \Phi(\mathbf{x}_j) = q_j \sum_{i \neq j} q_i / |\mathbf{r}_{ij}| = \{\mathbf{D}\mathbf{q}\}_j,$$

the  $j$ th component of the matrix/vector product  $\mathbf{D}\mathbf{q}$ . Summing up the values of the components of the resultant product ( $N$  potentials evaluated at  $N$  points) yields the desired potential. This is clearly  $\mathcal{O}(N^2)$  computation by a straightforward matrix/vector multiplication. However, if the potential  $\Phi$  is sufficiently smooth when distances are sufficiently large, the matrix  $\mathbf{D}$  can be approximated by low-rank submatrices; in this case, an  $\mathcal{O}(Np)$  scheme results where  $p$  is the rank of the approximating matrices. See also the simple example in Box 10.4 and Figure 10.9.

#### Hierarchical Protocol

This  $\mathcal{O}(N \log N)$  method works roughly as follows.

A hierarchy of boxes is introduced in the computational domain so that each refinement level  $l$  is a subdivision (e.g., into 8) of the domain in level  $l - 1$ . Boxes at the same refinement level that share a boundary are considered *near neighbors*, and those at the same refinement level that do not share a boundary are considered *well-separated*.

As the algorithm sweeps through refinement levels, a multipole expansion is associated with each box at that level to describe the far-field potential contribution from the particles in that box. (The clusters that contribute to this far field

are determined by an *interaction list*). These multipole expansions for the clusters are then used to compute interactions between distant clusters of particles, with nearby interactions computed directly.

This clustering is at the heart of such hierarchical methods: the computation for the interactions between particles is recast as work between clusters containing groups of particles.

### $\mathcal{O}(N \log N)$ Work

The recursive computation described above is completed in about  $\log_8 N$  steps (levels), leading to total work of  $\mathcal{O}(N \log N)$ . However, the *constant* associated with  $(N \log N)$  depends on the number of operations required to form the expansions. This cost of the local expansions depends on the number of coefficients  $p$  used (like the number of Fourier terms in FFTs). This number can be specified based on the desired accuracy,  $\epsilon_{\text{acc}}$ , for the resulting potential.

It can be shown that  $p$  can be related to  $\epsilon_{\text{acc}}$  as

$$p = \log_{\sqrt{3}}(1/\epsilon_{\text{acc}})$$

at sufficiently large separations. The corresponding total work is approximately  $200 N p^2 \log_8 N$  [481].

### Fast Multipole Machinery ( $\mathcal{O}(N)$ )

Unfortunately, these estimates do not produce substantial speedups in practice with respect to direct calculations (considering also the additional bookkeeping work required for multipole schemes), since  $p \approx 20$  for 7 digits of accuracy. Significant speedups are achieved only for very large system sizes.

Independently of the Barnes and Hut scheme, the fast multipole method was also developed in the mid 1980s by Rokhlin [1065]. Greengard and Rokhlin then made a seminal contribution [479] by developing further mathematical machinery that exploits the smoothness of the far-field potential and works with the tree codes. Relying on translation operators that act on the distant (multipole) as well as local expansions (harmonics), they lumped and converted expansions associated with several clusters into local expansions. The cost of translating the multipole-to-local expansions is reduced via fast schemes for application of rotation matrices. This combined clever mathematical and numerical machinery yields an asymptotically  $\mathcal{O}(N)$  method. The linear constant depends sensitively on the practical algorithmic implementation.

#### 10.5.3 Expansion in Spherical Coordinates

The multipole expansion is most conveniently written in spherical coordinates. A point  $\mathbf{x}$  in 3D space can be represented by the triplet  $\{r, \theta, \phi\}$  instead of the Cartesian triplet  $\{x, y, z\}$ , where:

$$r = \sqrt{x^2 + y^2 + z^2}, \quad \theta = \cos^{-1}(z/r), \quad \phi = \tan^{-1}(y/z).$$

Consider  $N$  charges located at points represented in spherical coordinates as  $\{\rho_i, \alpha_i, \beta_i\}$  lying inside a sphere of radius  $a$ , i.e., with  $|\rho_i| < a$  for all  $i$ . Then for any 3D point  $\mathbf{x} \equiv \{r, \theta, \phi\}$  outside that sphere, i.e., with  $r > a$ , the potential at  $\mathbf{x}$  can be written in terms of *spherical harmonics* functions,  $\{Y_n^m\}$ , solutions of the Laplace equation in spherical coordinates, as follows:

$$\Phi(\mathbf{x}) = \sum_{n=0}^{\infty} \sum_{m=-n}^{+n} \frac{M_n^m}{r^{n+1}} Y_n^m(\theta, \phi). \quad (10.49)$$

The functions  $Y_n^m(\theta, \phi)$  are called the *spherical harmonics of degree  $-(n+1)$* , or *multipoles*, and the coefficients  $M_n^m$  are known as *moments of the expansion*:

$$M_n^m = \sum_{i=1}^k q_i \rho_i^n Y_n^{-m}(\alpha_i, \beta_i). \quad (10.50)$$

The spherical harmonics functions can be expressed in terms of partial derivatives of  $1/r$  from the following relations [477]:

$$Y_n^m(\theta, \phi) = \begin{cases} r^{n+1} A_n^m \frac{\partial^n}{\partial z^n} \left( \frac{1}{r} \right) & m = 0 \\ r^{n+1} A_n^m \left( \frac{\partial}{\partial x} + i \frac{\partial}{\partial y} \right)^m \left( \frac{\partial}{\partial z} \right)^{(n-m)} \left( \frac{1}{r} \right) & m = 1, 2, \dots \\ r^{n+1} A_n^{-m} \left( \frac{\partial}{\partial x} - i \frac{\partial}{\partial y} \right)^{-m} \left( \frac{\partial}{\partial z} \right)^{(n+m)} \left( \frac{1}{r} \right) & m = -1, -2, \dots \end{cases} \quad (10.51)$$

where

$$A_n^m = \frac{(-1)^n}{\sqrt{(n-m)!(n+m)!}}. \quad (10.52)$$

The spherical harmonic functions  $\{Y_n^m(\theta, \phi)\}$  are also related to the *associated Legendre polynomials of degree  $n$* ,  $\{P_n^m\}$ , defined as

$$P_n^m(x) = (-1)^m (1-x^2)^{m/2} \frac{d^m}{dx^m} P_n(x)$$

by

$$Y_n^m(\theta, \phi) \equiv \sqrt{\frac{(n-|m|)!}{(n+|m|)!}} P_n^{|m|}(\cos \theta) \exp(im\phi).$$

The expansion power  $p$  is chosen to achieve the desired accuracy so that the remainder

$$\left| \sum_{n=p+1}^{\infty} \sum_{m=-n}^{+n} \frac{M_n^m}{r^{n+1}} Y_n^m(\theta, \phi) \right| \leq \frac{\sum_{i=1}^N |q_i|}{r-a} \left( \frac{a}{r} \right)^{p+1} = \epsilon_{\text{acc}}. \quad (10.53)$$

Thus to achieve greater accuracy, we can increase the expansion power  $p$  or decrease the radius  $a$ . For example, if  $r=2a$ , the remainder above is  $(2^{-(p+1)}/a) \sum_i |q_i|$ . Setting  $p = \log_2(\epsilon_{\text{acc}}^{-1})$  yields an accuracy of  $\epsilon_{\text{acc}}$  relative to  $\sum_i |q_i|/a$ .

The spherical harmonics functions are believed to provide a more efficient basis than Cartesian expansions, since they involve  $p^2$  coefficients rather than  $p^3$ .

#### 10.5.4 Biomolecular Implementations

Thorough comparisons of the balance among speed, accuracy, and scalability of Ewald and fast multipole schemes are ongoing areas of research. The fast multipole approach was first implemented for biomolecular dynamics in the early 1990s [142, 1176], and later parallelized to achieve good speedup on a large number of processors [140, 141, 1133], both on workstation clusters and on supercomputers like Cray T3D/T3E. A periodic version of the fast multipole method relies on machinery similar to that used for the periodic Ewald summation [1136], namely obtaining a convergent sum by masking the original sum by Gaussians.

Implementations in 3D of the periodic fast multipole method were adapted for, and applied to, molecular dynamics by Board and co-workers in the late 1990s [693, 1133] and by Figueirido *et al.* [399], as well as compared to the PME alternative.

In 3D applications, implementors have found it particularly difficult to achieve good performance with high accuracy. For example, in [399] it has been reported that the multipole approach is three times *slower* than PME for about 6000 atoms and slightly slower than PME for about 22,000 atoms; it is suggested, however, that fast multipoles will be competitive for larger sizes.

In serial test implementations, Board and collaborators found that the periodic version of the fast multipole method is still significantly slower than PME by roughly a factor of two for a large water system of 71,496 atoms [1133]; this performance involves modest accuracy (4–5 digits of accuracy in the potential and 3–4 in the corresponding force) and required 8 multipole terms. This performance result may be explained by the very fast, ‘custom’ version of PME of Board and co-workers which uses a table lookup for the erfc function and yields equivalent run times for calculations involving spherical cutoffs of 10–12 Å. Work for distributed PME versus distributed parallel multipole tree code, run at the modest accuracy, still shows PME to be faster on a small number of processors (Figure 10.8).

The newer version of the fast multipole method extends the 2D scheme of translation operators acting on harmonic functions to produce a new diagonal form for these operators [481]. This approach produces a more modest break-even system size in 3D when compared to direct calculations, such as 2000 particles for single-precision accuracy or 5000 for 10-digit accuracy. Application of this fast multipole method to biomolecular dynamics should be forthcoming.

### 10.5.5 Other Variants

Besides these commonly-used fast multipole and Ewald methods, scientists have also developed variants as well as new methods. Summations based on multi-grid techniques developed for partial differential equations, as implemented for the fast evaluation of integral transforms [164], were extended to molecular electrostatic potentials [1088, 1198]. A variable-order extension of Appel's algorithm in Cartesian coordinates was also reported, with adaptive tree structure and error control [342, 343].

## 10.6 Continuum Solvation

### 10.6.1 Need for Simplification!

The fast electrostatic techniques described above have been designed and are necessary for ‘deluxe’ models of macromolecular systems — very detailed solvated biomolecular representations involving thousands of explicit water molecules surrounding the biomolecule or biomolecules. This large size mimics the cellular environment and is needed to reproduce the bulk properties of water as well as the effects of local solvation on important associations such as those between two proteins, proteins and ligands, and proteins and nucleic acids.

While such elaborate models can yield invaluable information on the rich complexity of biomolecular environments, they are clearly computationally-intensive and, unfortunately, free neither from approximations nor artifacts. As described previously, artificial periodicity can lead to noncanceling errors associated with the Coulomb and solvation contributions to the electrostatic free energy [580] (see also below). Trajectories are also highly sensitive to various force-field approximations, propagation protocols, and modeling choices (e.g., positions of ions and water molecules); see discussion in the chapters on molecular dynamics.

Implicit solvent approaches reduce the number of degrees of freedom drastically and account for them in an average sense, in the form of solvation free energy estimates. Such approaches can be effective, especially when combined with coarse grained models of molecular systems. Practical methods that involve clever engineering constructs of various approximations have been developed by many researchers, including McCammon, Case, Karplus, Roux, Honig, Truhlar, and many others, as recently reviewed [223, 270, 377, 587]. However, Chen and Brooks caution that current surface-area based nonpolar models have significant limitations and thus could benefit by the incorporation of several nonpolar solvation aspects [223].

Thorough introductions to implicit solvation models can be found in [758, 1073], the latter with emphasis on quantum-mechanical based approaches with a literature survey of relevant works. See [223] for a more recent review including many technical details, Cramer's textbook [270] for a thorough treatment, and [376, 1190, 1263] for reviews of Poisson-Boltzmann and generalized Born models.

### 10.6.2 Potential of Mean Force

Implicit solvent models based on *continuum electrostatics* treatments form an alternative to this deluxe approach to modeling [813, 1073, 1190, 1263]. These methods essentially approximate a *potential of mean force* [648] for the solvated biomolecular environment and can provide a broad, qualitative and quantitative view of overall biomolecular structure, recognition, and functional properties.

#### Balancing Biophysics with Numerics

The potential of mean force is an effective free energy potential that depends on state variables like temperature and pressure [648]. It augments the potential energy used to describe the intramolecular interactions by indirectly incorporating the effects of rearrangement in the medium on the biomolecule (solute). This effect is important to consider since water molecules and ions in the solute environment perturb, in a cooperative manner, the pairwise forces acting on pairs of solute particles through their own locations. The basic idea is to construct a function that captures in an average sense solute configurations without explicit consideration of the solvent degrees of freedom. Additionally, hydrodynamics can be applied to provide continuum approximations to the effects on the solute of the motions of the solvent molecules. In this way, the model complexity can be compromised with physical reproducibility/reliability.

#### Electrostatic and Non-Electrostatic Components

When such potentials of mean force are constructed, they are generally separated into two components: *electrostatic* and *non-electrostatic* (or nonpolar) interactions [580, 1073]. This decomposition is useful for dissecting the different microscopic factors that influence molecular conformations and solvation. The electrostatic component can also be related to various continuum electrostatic approximations (see below).

#### Variations

There are many ways to construct the effective potential of mean force for this broader structural and functional description, both empirically and theoretically.

Empirical constructs lump all effects into an ‘information based’ or ‘statistical’ potential, derived, for example, from observed protein structures in solution [912].

Other approaches involve approximations of integral equations [1014], free energy simulations, solvent-accessible surface models, various combinations of implicit/explicit solvent representations [150, 785, 1314], generalized Born models [96, 1273], solutions based on classical continuum electrostatics (i.e., Poisson-Boltzmann equation) [565] and quantum mechanics [758], and phenomenological dynamic models such as Langevin and Brownian dynamics (BD) [813]. Generalized Born models, in particular, construct a dielectric cavity surrounding the molecular charges to approximate solvation and salt-screening effects. In all cases, the dielectric constant must be chosen with care [1146].

A good review on implicit solvent models can be found in [223]. See also the special volume in *Biophysical Chemistry*, volume 78 (5 April 1999), guest edited by Benoît Roux and Thomas Simonson, with updates in [1190, 1263]. The study in [1439] shows that implicit solvent models yield reasonable agreement in terms of protein/ligand binding free energies when compared to their much more computational-costly explicit-solvent models. Applications to membrane systems [1263] reveal the progress of modeling such heterogeneous materials by implicit solvation models but also some computational problems (e.g., artificial periodicity imposed by Ewald sums and related methods) that require further work to characterize and resolve.

Below, we sketch two approaches, based on Langevin dynamics and continuum electrostatics.

### 10.6.3 Stochastic Dynamics

The theory of hydrodynamics has a long history in molecular modeling, originating from the study of liquids, polymers, and simple molecular reactions [184, 966, 1072, 1407, 1456]. These theories have been applied to macromolecular dynamics via Langevin and Brownian dynamics simulations that generate *stochastic trajectories*, so called because the governing dynamic equations include stochastic forces that mimic solvent effects, in addition to the systematic force (negative gradient of the potential energy).

See classic statistical mechanics texts such as [853] for extensive background, and Chapter 14 of this text for a brief discussion of generating stochastic trajectories by Langevin and Brownian dynamics algorithms; a thesis by Hongmei Jian [607] nicely summarizes the theory and numerical applications of Langevin and Brownian dynamics to long DNA molecules.

In the stochastic treatment, the influence of solvent particles on the solute is incorporated through additional frictional and random terms in a manner consistent with physical laws regarding equilibrium and nonequilibrium processes (e.g., equilibrium conformation distributions, fluctuation/dissipation theorem) [853]. Applications of Langevin and Brownian dynamics simulations have been particularly successful for macroscopic models of biomolecules, such as long DNA of thousands of base pairs [107, 233, 608, 1106]. In such applications, polymer theory has been used to guide parameterization (for hydrodynamic radii, timesteps, etc.) through governing macroscopic polymer properties [410] such as persistence length, radius of gyration, and diffusion constants. Protein applications include long-timescale enzyme catalysis events such as the loop opening/closing motion in triosephosphate isomerase (TIM) [298, 299, 1321, 1322].

#### The Langevin Equation

In the simplest form of the Langevin equation, the friction kernel is taken to be space and time independent for each particle, and the influence of the environment on the systematic, internal force is represented in an average sense. Thus, explicit

hydrodynamic interactions are ignored, and the internal force is augmented by a frictional term, proportional to the velocity, and a random force  $R$ , which crudely mimics molecular collisions and viscosity in the realistic cellular environment:

$$\mathbf{M} \frac{d^2\mathbf{x}}{dt^2} = -\nabla E(\mathbf{x}) - \mathbf{M}\gamma \frac{d\mathbf{x}}{dt} + R(t). \quad (10.54)$$

Here  $E$  is the potential energy governing the solute,  $\gamma$  is the damping constant (or collision frequency), and  $R(t)$  is the ‘white noise’ vector with mean,  $\langle R(t) \rangle$ , of zero.

From classic theories of Brownian motion, it can be shown that although molecular collisions are random, the ensemble of these collisions produces a systematic effect. In other words, random motions exist at thermal equilibrium as a fluctuation. It follows that the frictional force and the random force are related by the *fluctuation/dissipation theorem* [684]. This relation can be expressed by the  $\gamma$ -dependence of the covariance of  $R$ :

$$\langle R(t)R(t')^T \rangle = 2\gamma k_B T \mathbf{M} \delta(t - t') \quad (10.55)$$

( $k_B$  is Boltzmann’s constant and  $T$  is the temperature). Here the covariance matrix is diagonal since hydrodynamic interactions between particles have been discounted. The damping constant  $\gamma$  controls both the magnitude of the frictional force and the variance of the random forces. It thus ensures that the system converges to a Boltzmann distribution characterized by the temperature  $T$ . The larger the value of  $\gamma$ , the greater the influence of the surrounding fluctuating force (solvent).

In the limit of small  $\gamma$ , the motion is termed *inertial*, and in the limit of large  $\gamma$  it is *diffusive* or *Brownian*; see [1036] for vivid illustrations of those regimes for models of supercoiled DNA. Different integration algorithms are generally applied depending on the relevant regime (e.g., [180, 1200, 1293]). See also Chapter 14 for a simple discretization [180] of the above Langevin equation and other discretizations in [1200, 1293].

The stochastically modeled system reaches the same equilibrium as the original system obeying  $\mathbf{M} (d^2\mathbf{x}/dt^2) = -\nabla E(\mathbf{x})$ , but the *rate* at which equilibrium is reached depends on the viscous coupling to the environment. Since the number of degrees of freedom in the Langevin model is the number corresponding to the solute particles, the model is computationally much cheaper than the corresponding all-atom representation which includes explicit solvent.

### Langevin Parameters from Hydrodynamic and Other Considerations

Guidelines for choosing numerical values for  $\gamma$  come from hydrodynamic theory. Stokes’ law describes how the frictional resistance of a spherical particle in solution varies linearly with its radius: the effective force magnitude is  $6\pi\eta a_r$  times the particle’s velocity, where  $a_r$  is the hydrodynamic radius of each spherical particle and  $\eta$  is the solvent viscosity. Stokes’ law is frequently applied to particles

of molecular size. Hence,  $\gamma$  in the Langevin equation can be set to

$$\gamma = 6\pi\eta a_r/m,$$

where  $m$  is the particle's mass.

From a modeling point of view, it is also possible to choose  $\gamma$  so that the resulting translational diffusion constants  $D_t$  match the experimental values in the *diffusive limit* (large  $\gamma$ ). This follows the relationship

$$D_t = k_B T / \sum_i m_i \gamma,$$

where the summation extends over all masses  $m_i$  in the system. Note that this expression for  $D_t$  reduces to

$$D_t = k_B T / 6\pi\eta a_r$$

when Stokes' law for setting  $\gamma$  is used. This is the Stokes-Einstein law of diffusion for a Brownian particle; see homework assignment 12 for an example.

A *computationally efficient* value for  $\gamma$  can also be selected from practical considerations, since an optimal coupling of the system to the thermal reservoir can accelerate configurational sampling during finite-length simulations [783, 1036, 1037]. This choice of  $\gamma$  implies relinquishing the idea of approximating the 'true' dynamics in favor of efficient sampling.

### The Brownian Limit

In the diffusive limit of the Langevin equation, the motion is more random or *Brownian* in character. Such motion characterizes in a global sense a dense system in which the solute collides often with the surrounding fluid particles, and is thus continuously and significantly reoriented by the solvent molecules. Specifically, the Brownian regime assumes that the velocity relaxation time is much more rapid than position relaxation time.

Theories for Brownian motion, dating from Einstein's work (circa 1905), have been based on the generalized Langevin equation and on the Fokker-Planck equation, a partial differential equation that describes the evolution of a system in a probabilistic sense [853].

Practical BD algorithms [106, 367, 813] are similar to Monte Carlo procedures (see Chapter 12) in that the current position is perturbed by a random displacement vector. However, this random displacement is more complicated to formulate, as it depends on the forces and the diffusion tensor. See the brief section in Chapter 14 on Brownian dynamics algorithms and illustrations at the end of Chapter 6 for long DNA and polynucleosomes.

When very large biomolecular polymers are modeled, such as long DNA of thousands of base pairs, a more detailed account of hydrodynamic interactions is required to accurately represent the changes in solute forces induced by the flow of the surrounding fluid particles. This is done by formulating a position-dependent hydrodynamic tensor  $T$  related to the diffusion tensor  $D$  instead of the  $\gamma$ -dependent friction term in the simple Langevin equation.

The Brownian methods that incorporate hydrodynamic effects via a tensor that is configuration dependent yield better descriptions of large polymers in solution and can reach longer time frames such as milliseconds [106, 107, 608]. Still, the computational complexity grows rapidly, as  $\mathcal{O}(N^3)$ , with the number of modeled beads ( $N$  particles) when hydrodynamic effects are included [607].

Various approximations can be used to make such BD simulations with hydrodynamics applicable to long-time macroscopic models of biological polymers like DNA [107]. An idea of Marshall Fixman [404] to use Chebyshev polynomial approximations for matrix/vector products was developed and applied in [1119], demonstrating that an  $\mathcal{O}(N^2)$  complexity is possible for large systems (see Chapter 14).

#### 10.6.4 Continuum Electrostatics

For introductions into the study of polyelectrolyte solutions and related theories, such as the Poisson-Boltzmann equation and Debye-Hückel theory, readers may wish to consult the book [537] for a good treatment of statistical thermodynamics and kinetics, the text [547] for statistical thermodynamics, and the volume [1048] for an early overview of how Poisson-Boltzmann and Debye-Hückel theories have been applied to polyelectrolyte solutions. Classic references on electrolyte solutions are [517, 1056]. Applications of the Poisson-Boltzmann equation for analysis of biomolecular electrostatics are reviewed in [412] for example.

##### Gauss' Law for the Electrostatic Potential

Continuum electrostatic approximations are based on numerical solutions to the Poisson-Boltzmann (PB) equation (see [537, 547], for example, for introductions). This second-order differential equation combines theories from statistical mechanics — the Boltzmann distribution for a charge density  $\rho$  — with an equation from electrostatics — Gauss' law (or the Poisson equation),<sup>5</sup> which relates the second derivative of the electrostatic potential  $\Phi$  to the charge density.

Gauss' law describes the electrostatic potential  $\Phi$  at position  $\mathbf{x}$  in terms of the fixed charged density of the solute,  $\rho_{\text{solute}}(\mathbf{x})$  and the position-dependent dielectric function  $\epsilon(\mathbf{x})$  as:

$$\nabla \cdot [\epsilon(\mathbf{x}) \nabla \Phi(\mathbf{x})] = -4\pi \rho_{\text{solute}}(\mathbf{x}). \quad (10.56)$$

(See Box 10.2 for the definition of the divergence operator  $\nabla$  used above). This equation reduces to Coulomb's law when the dielectric constant is uniform throughout space and the charges are modeled as point charges.

For a polar solvent like water, the effective dielectric constant increases as the point  $\mathbf{x}$  moves farther away from the solute. This spatially dependent dielectric function — low  $\epsilon$  near the solute and high  $\epsilon$  for the bulk solvent — allows a better

---

<sup>5</sup>Physicists often refer to this equation as Gauss' law while mathematicians tend to favor the term Poisson's equation.

description than Coulomb's law: it takes into account the difference in electric polarizability between the macromolecule and the solvent. (See [1420], for example, for estimates of local dielectric constants in the environment of B-DNA in solution). Poisson's equation cannot be solved analytically for arbitrary geometries and must be solved numerically by finite-difference or other methods (see below).

When mobile ions are also present in the solution, the charge density is delocalized from the solute/solvent boundary. The charge atmosphere is thus position-dependent, since the ions redistribute in the solution in response to the electric potential. A better approximation is obtained by considering a charge density around the solute resulting from the *distribution of charges* in the medium.

Assume that we have an electrolyte solution occupying a volume  $V$  and containing  $N_i$  ions of corresponding charge  $q_i$  for  $n_i$  ion species  $i$ . Let  $c_i = N_i/V$  denote the bulk concentration of the ionic species  $i$ . The total *charge density* resulting from the sum of all charge densities of the ions can be described by a Boltzmann distribution as:

$$\rho(\mathbf{x}) = \sum_{i=1}^{n_i} q_i c_i \exp[-\tilde{E}_i(\mathbf{x})/k_B T], \quad (10.57)$$

where  $\tilde{E}_i(\mathbf{x})$  is the ‘effective potential of mean force’ for electrolytes of type  $i$  at position  $\mathbf{x}$  for a *given* solute configuration, or the energy change required to bring the ion from infinity to the position  $\mathbf{x}$ .

In practice, it is assumed that  $\tilde{E}_i(\mathbf{x})$  is approximated by the product charge times the potential, that is, the distribution of ions is determined by the electrostatic field:

$$\tilde{E}_i(\mathbf{x}) \approx q_i \Phi(\mathbf{x}).$$

Thus, the Boltzmann distribution leads to an exponential relation between the charge density  $\rho$  and the electrostatic potential.

### The Poisson-Boltzmann Equation

Combining Gauss' law (eq. (10.56)) with the Boltzmann charge density described by eq. (10.57) yields the *Poisson-Boltzmann (PB) equation*:

$$\nabla \cdot [\epsilon(\mathbf{x}) \nabla \Phi(\mathbf{x})] = -4\pi \rho_{\text{solute}}(\mathbf{x}) - 4\pi \sum_{i=1}^{n_i} q_i c_i \exp[-q_i \Phi(\mathbf{x})/k_B T]. \quad (10.58)$$

The PB equation is the basis for modern electrolyte theory. The solution of this nonlinear equation for  $\Phi(\mathbf{x})$  yields thermodynamic properties in the electrolyte solution. Since the electrostatic potential does not vary linearly in space with the source charges, properties of linear systems do not generally apply. As for Poisson's equation, analytic solutions are not available in general and various numerical methods are used in practice.

### Linear Approximations to the PB Equation; Debye-Hückel Theory

For the case of dilute or moderate solutions (e.g., molar concentrations  $c_s < 10^{-3}$  M) and low fixed charge, it is possible to approximate the PB equation using results from *Debye-Hückel theory* (see [131, 275, 537, 547], for example, for introductions). Specifically, a linearized approximation to the PB equation can be used to represent the ionic atmosphere of a solute immersed in aqueous solution and counterions.

A linearized version of the PB equation (eq. (10.58)) can be obtained by a Taylor-series expansion of the Boltzmann factor with a truncation beyond the first-order terms:

$$\exp[-q_i \Phi(\mathbf{x})/k_B T] \approx 1 - [q_i \Phi(\mathbf{x})/k_B T]. \quad (10.59)$$

The linearized version is justified when

$$q_i \Phi(\mathbf{x}) \ll k_B T,$$

that is, when energies are much smaller than the thermal energy. This often holds for monovalent electrolytes and weak source charges (dilute solution), as mentioned above.

In the special case of spherical symmetry about the origin for the distribution of charges, the potential  $\Phi$  depends on the distance  $r$  of the point from the origin. Linearization of the PB equation yields the following *linear* second-order differential equation for  $\Phi(r)$  [537]:<sup>6</sup>

$$\nabla^2 \Phi(r) = \kappa^2 \Phi(r), \quad (10.60)$$

where

$$\kappa^2 = \frac{8\pi N_A e^2 \rho_A}{1000 \epsilon k_B T} c_s. \quad (10.61)$$

Here  $N_A$  is Avogadro's number,  $\rho_A$  is the solvent density,  $e$  is the protonic charge ( $4.803 \times 10^{-10}$  esu), and  $\epsilon$  is the solvent dielectric constant. The ionic concentration  $c_s$  is measured in molar units as a sum over all molar concentrations per liter of solution,  $c_i$ , associated with charges (or valences)  $q_i$ :  $c_s = \frac{1}{2} \sum_{i=1}^{n_i} c_i q_i^2$ .

The linearized PB equation as expressed in eq. (10.60) with (10.61) can be solved to determine the effective potential  $\Phi$  at a distance  $r$  from a central ion. Recall that the Coulomb potential  $q/(\epsilon r)$  is produced at a point separated by distance  $r$  from an isolated central ion, which is represented as a uniform charged sphere of radius  $a_r$  in a medium of dielectric constant  $\epsilon$ . The Debye-Hückel solution for the modified electrostatic potential representing the influence of the ionic atmosphere is (e.g., [537]):

$$\Phi(r) = B(\kappa) \frac{q \exp(-\kappa r)}{\epsilon r}, \quad (10.62)$$

---

<sup>6</sup>We can also write  $\frac{1}{r^2} \frac{d}{dr} \left( r^2 \frac{d}{dr} \right) \Phi(r) = \kappa^2 \Phi(r)$ .

where  $B(\kappa)$  is the salt-dependent coefficient

$$B(\kappa) = \exp(\kappa a_r)/(1 + \kappa a_r). \quad (10.63)$$

Thus, Debye-Hückel theory produces an effective electrostatic potential in which the Coulomb interactions are *screened* by ions. The theory predicts the range of electrostatic influence of a central ion to be the Debye screening length  $\kappa^{-1}$ . In other words,  $\kappa^{-1}$  is the characteristic distance of exponential screening.

From eq. (10.61), we see that the *screening parameter*  $\kappa$  is proportional to the square root of the ionic concentration  $c_s$ . For 1:1 electrolytes (monovalent:monovalent salts like NaCl),

$$\kappa \approx 0.33\sqrt{c_s} \text{ \AA}^{-1} \quad (10.64)$$

at room temperature ( $25^\circ$ ), with  $\epsilon = 78.5$ .

For dilute solutions, or  $\kappa a_r \ll 1$ , we have  $B(\kappa) \approx 1$ ; it follows that the Coulomb screening — reduction of the unscreened potential by the factor  $\exp(-\kappa r)$  — reflects the reduced effective charge  $\rho$  on the central ion due to counterion accumulation. For physiological ionic strengths, such as 0.15M, the Debye length is approximately 8 Å; this distance represents considerable damping of Coulomb interactions.

Figure 10.10 compares the Coulomb potential ( $1/\epsilon r$ ) to the screened Coulomb potential ( $B(\kappa) \exp[-\kappa r]/\epsilon r$ ) for two values of  $\kappa$  corresponding to  $c_s = 0.15\text{M}$

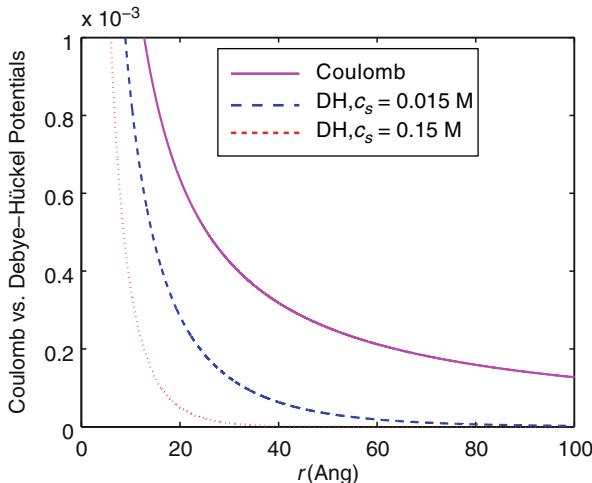


Figure 10.10. Screened Coulomb potentials at two values of the Debye length. The Coulomb potential ( $1/\epsilon r$ ) is compared to the screened Coulomb potential of form  $B(\kappa) \exp[-\kappa r]/\epsilon r$  as a function of distance  $r$  for two values of  $\kappa$ . These two values are set according to eq. (10.64) for  $c_s = 0.15\text{M}$  and  $0.015\text{M}$  monovalent salt concentrations, with coefficients  $B(\kappa)$  computed according to eq. (10.63) (the coefficients are very close to 1). The two  $\kappa$  values (0.128 and 0.040 Å, respectively) correspond to Debye lengths of about 8 and 25 Å.

and 0.015M monovalent salt, where the Debye lengths are about 8 and 25 Å, respectively.

The DH approximation has been applied to long DNA for exploring conformational stability and mobility as a function of monovalent ionic concentrations in the natural cellular environment (e.g., [1125, 1309]). Models are based on the pioneering work of Stigter, who modeled DNA as charged cylinders [1223] and reproduced the experimentally-observed dependence of the effective diameter of DNA on salt concentration by a tail approximation to the PB equation. Extensions of such DH approximations to macroscopic models of protein/DNA systems have been described [109] and applied to model chromatin [108].

### General Solutions to the Poisson-Boltzmann Equation

Many practical procedures have been developed to solve the PB equation numerically, including solutions of its various approximations (e.g., [16, 80, 563, 610, 813, 903, 1058, 1178]); see also references cited in [81], and [303, 1023] for applications to calculate charge distributions. The linearized approximations are useful in many cases, as mentioned above. For high charge density (such as for polyelectrolyte DNA) and high salt concentrations, the nonlinear PB version is preferred.

Numerical solutions are typically obtained through finite-difference or finite-element (or boundary element) methods. Both involve a discretization of the (irregular) biomolecular domain in 3D, so that the potential, charge density, and dielectric constant are defined at grid points ( $\epsilon$  is usually defined over broader domain regions). The dielectric constant is assigned appropriate values depending on its proximity to the solute, though a two-dielectric model is typically used — to distinguish the inside region, near the solute, from the outside region, far from the solute. The finite-difference solution can be obtained iteratively by various linear algebra solvers (e.g., linear conjugate gradient method, Gauss-Seidel, or Successive Over Relaxation methods) [16, 903, 1058].

The resulting quality of the numerical solutions depends on the various assumptions made and settings used. The convergence of the solvers also depends on such parameters as the grid size, initial charge and  $\epsilon$  assignments, and algorithmic parameters.

An analytical gradient minimization method based on a finite-element discretization for solving the PB equation has also been presented [429]. While the DelPhi program uses a regular cubic grid, this gradient-based approach uses an adaptive grid so as to include more grid points at the solvent-accessible surface, where the dielectric value is changing rapidly. Overall, preliminary comparisons show that the computational efficiency of both approaches is comparable [429]; the gradient-based method is also used for geometry optimization applications, serving as an improvement to gas-phase molecular mechanics minimizations since solvation effects are included.

Popular packages used in the biomolecular community are those developed at research groups at Columbia University (DelPhi and GRASP) and the University

of Houston (UHBD, now at UCSD, which includes segments for Brownian dynamics). The latter has been used to study huge macromolecular systems [81] (see also Figure 10.11).

The APBS (Adaptive PB Solver) program is also popular [81] and can be used on its own or in conjunction with other molecular modeling packages like CHARMM or AMBER, with results visualized with popular graphics programs like VMD or PYMOL. In addition, web tools like PBEQ based on the PB solver in CHARMM [610] or Mark-US (a functional assessment tool for proteins) from the Honig lab make applications easier to users by combining PB solvers with interactive visualization modules.

Classical electrostatics solutions can provide useful information on charge distributions around macromolecules (e.g., [303, 1023]), specific localization of charged and polar groups in biomolecular systems (e.g., [790, 1253]), the shape of molecular surfaces, and the relation between salt effects and conformational changes (e.g., [1125]). Effects of the ionic atmosphere on intermolecular binding associations, such as between proteins and ligands, can also be analyzed [565, 1057].

Effective numerical packages such as DelPhi and UHBD, combined with advanced graphical rendering programs, offer valuable modeling tools for exploring the effects of electrostatics and solvation on molecular structure and function.

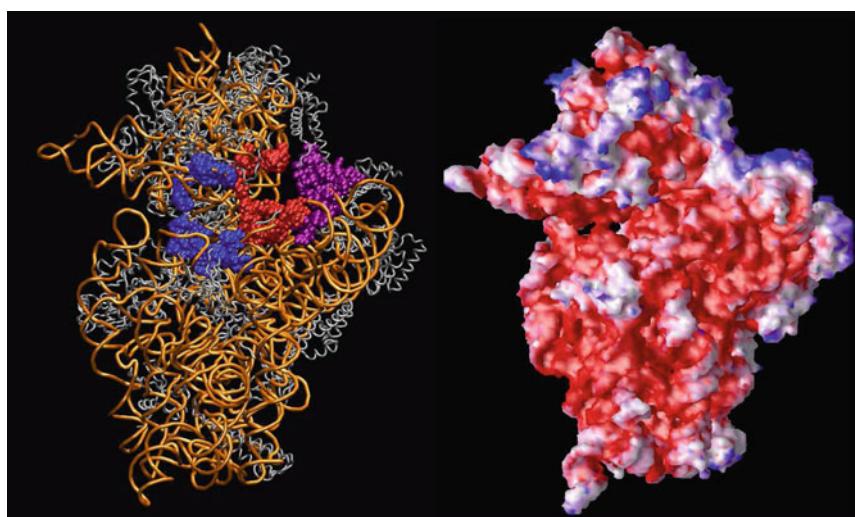


Figure 10.11. The electrostatic environment of the 30S ribosomal unit as computed by the linear PB equation at 150 mM ionic strength with a solute dielectric of 2 and solvent dielectric of 78.5 [81]. The protein (silver) and nucleic acid (gold) atoms are distinguished on the left, with selected components of A, P, and E sites shown in blue, red, and purple respectively. At right, the electrostatic potential mapped on the 30S surface is rendered in blue (positive) and red (negative) to illustrate regions with values greater than  $2.6 \text{ } k_B T/e$  and less than  $-2.6 \text{ } k_B T/e$ , respectively.

The electrostatic analysis of a ribosomal unit is shown in Figure 10.11. (See also analysis of the nucleosome core particle in Chapter 6).

### Algorithmic Challenges

Recent algorithmic advances — by adaptive finite element methods [80, 563] tailored to multiprocessor systems [81] — allow solutions of the linear PBE for very large systems of order one million atoms. Figure 10.11 illustrates the PB solution for the 30S ribosome subunit.

A future goal in the field is repeated solution of the PBE, for example in the course of a dynamic simulation.

# 11

## Multivariate Minimization in Computational Chemistry

### Chapter 11 Notation

SYMBOL	DEFINITION
<b>Matrices</b>	
$\mathbf{A}$	symmetric matrix, components $\{A_{ij}\}$
$\hat{\mathbf{B}}_k$	approximation to Hessian inverse at $\mathbf{x}_k$ (QN methods)
$\mathbf{D}_k$	scaling matrix at step $k$ of minimization method (trust region approach)
$\mathbf{H}$	Hessian matrix, components $H_{ij}(\mathbf{x}) \equiv \partial^2 f(\mathbf{x}) / \partial x_i \partial x_j$
$\mathbf{I}$	identity matrix
$\mathbf{M}$	preconditioning matrix, related to $\mathbf{H}$ (TN methods)
$\mathbf{M}_k$	preconditioning matrix at step $k$ of TN method
$\mathbf{U}_k$	QN low-rank update matrix at step $k$
<b>Vectors</b>	
$\mathbf{b}, \mathbf{y}$	constant vectors
$\mathbf{e}_j$	unit vectors
$\mathbf{g}$	gradient vector of $f$ , components $g_i(\mathbf{x}) \equiv \partial f(\mathbf{x}) / \partial x_i$
$\mathbf{g}_k$	gradient vector at $\mathbf{x}_k$ (short hand for $\mathbf{g}(\mathbf{x}_k)$ )
$\mathbf{p}_k$	search vector at step $k$ of minimization method
$\mathbf{p}_k^j$	inner-loop CG iterate $j$ for outer-loop search vector $\mathbf{p}_k$ (TN method)
$\mathbf{r}$	residual vector defined in TN methods ( $\mathbf{Mz} = \mathbf{r}$ )
$\mathbf{s}_k$	displacement vector at step $k$ , $\mathbf{x}_{k+1} - \mathbf{x}_k$ (QN method)
$\mathbf{x}$	vector of $n$ components $\{x_i\}$
$\mathbf{x}_0$	starting point vector for minimization
$\mathbf{x}_k$	minimization iterate at step $k$ of method
$\mathbf{x}^*$	local minimum point of objective function
$\mathbf{y}_k$	gradient difference vector at QN step $k$ , $\mathbf{g}_{k+1} - \mathbf{g}_k$
$\mathbf{z}$	solution vector defined in TN methods ( $\mathbf{Mz} = \mathbf{r}$ )

Chapter 11 Table (continued)

SYMBOL	DEFINITION
<b>Scalars &amp; Functions</b>	
$a, b$	numbers
$c_i(\mathbf{x})$	constraint function $i$
$c_r$	small positive number (in TN methods)
$f_0$	constant (function value)
$f(\mathbf{x})$	objective function, dependent on vector $\mathbf{x}$
$h$	small number (finite difference interval)
$n$	problem dimension
$p$	convergence order
$q(\mathbf{x})$	quadratic function
$q_k(\mathbf{s})$	quadratic model of objective function
$r_k$	residual norm at step $k$ of TN methods
$\alpha$	line search parameter for sufficient decrease condition
$\beta$	line search parameter for sufficient decrease of curvature (also convergence ratio)
$\beta_k$	scheme-dependent scale parameter of search vector $\mathbf{p}_k$ (CG and QN methods)
$\epsilon_f, \epsilon_g$	small positive numbers
$\epsilon_m$	small positive number, machine precision
$\eta_k$	forcing sequence in TN methods
$\lambda$	line search steplength
$\lambda_t$	trial line search steplength
$\xi$	variable in the neighborhood of $x$ for a univariate function $f(x)$
$\phi(\lambda)$	polynomial of steplength $\lambda$
$\Delta_k$	size bound in QN methods at step $k$

'Pon my word Watson, you are coming along wonderfully. We have really done very well indeed. It is true that you have missed everything of importance, but you have hit upon the method.

Arthur Conan Doyle (1859–1930), in *A Case of Identity* (1891).

Economic forecasting makes weather forecasting look like physics.

Ben Bernanke, 26 July 2009.

[Bernanke, the US Federal Reserve Chairman, spoke about the current economic crisis in a town-hall forum in Kansas City, MO, hosted by the PBS NewsHour's Jim Lehrer.]

## 11.1 Ubiquitous Optimization: From Enzymes to Weather to Economics

Optimization is a fundamental component of molecular modeling. The determination of a low-energy conformation for a given force field can be the final objective of the computation. It can also serve as a starting point for subsequent calculations, such as molecular dynamics simulations or normal-mode analyses.

Both local and global optimization problems lie at the heart of numerous scientific and engineering problems — from the biological and chemical disciplines to architectural and industrial design to economics. Optimization is part of our everyday life — responsible for our weather forecasts, flight planning, telephone routing, microprocessor design, and the functioning of enzymes in our bodies.

### 11.1.1 Algorithmic Sophistication Demands Basic Understanding

The mathematical techniques developed to address these optimization problems are just as robust and varied as the target problems themselves. The algorithmic complexity of such techniques has led to many available computer programs that require minimal input from the user (e.g., the starting point and a routine for function evaluation).

However, the prudent user of these canned software modules — even within standard molecular mechanics and dynamics packages — should understand the fundamental structure of the optimization algorithms and associated performance issues to make their application both efficient and correct, in terms of the physical interpretations.

This chapter introduces key optimization concepts for this purpose. We also highlight the fundamentals of local optimizers for *large-scale nonlinear unconstrained problems*, an important optimization subfield relevant to biological macromolecules. We describe the most promising approaches among them, and discuss practical issues, such as parameter variations and termination criteria. Of course, the latter are best learned by experimentation in the context of real problems. To illustrate behavior for complex problems, some comparisons among three competitive minimizers are also included, for molecular models minimized in the molecular mechanics and dynamics program CHARMM.

### 11.1.2 Chapter Overview

Specifically, Section 11.2 introduces optimization fundamentals such as problem formulation and terminology. Section 11.3 describes the basic algorithmic framework of iterative minimization protocols (based on line search and trust region methods); it also discusses convergence criteria and line search procedures and introduces the key concept of descent directions.

In Section 11.4, we present the Newton method, including a historical perspective, and one-dimensional implementations for nonlinear equations as well

as optimization. This presentation familiarizes readers with the Newton method framework — the basis for formulating many other optimization methods — and with performance and convergence issues relevant to minimization of multivariate functions.

In Section 11.5, we mention effective methods for large-scale nonlinear optimization, namely quasi-Newton (QN), nonlinear conjugate gradient (CG), and truncated Newton (TN) schemes. Section 11.6 outlines available software and presents comparative performance in CHARMM for two molecular models (a small model system and a protein). Finally, in Sections 11.7 and 11.8, we summarize recommendations to optimization practitioners and offer a future perspective to field developments.

For details, as well as for other categories of the rich and exciting field of optimization, I refer readers to classic texts [297, 407, 459, 789, 918], some reviews [417, 915, 1104, 1108, 1131], and a perspective [1390]. In recent years, optimization methods have become increasingly important in material science and nanotechnology, for finding the favored structure of materials, including biological systems. Such problems are typically solved by global optimization problems, as recently reviewed [196].

## 11.2 Optimization Fundamentals

The methods for solving an optimization task depend on the problem classification. Since the value of the independent variable that maximizes a function  $f$  also minimizes the function  $-f$ , it suffices to deal with minimization.

The optimization problem is classified according to the *type* of independent variables involved (real, integer, mixed), the *number* of variables (one, few, many), the *functional characteristics* (linear, least squares, nonlinear, non-differentiable, separable, etc.), and the problem *statement* (unconstrained, subject to equality constraints, subject to simple bounds, linearly constrained, nonlinearly constrained, etc.). For each category, suitable algorithms exist that exploit the problem's structure and formulation.

### 11.2.1 Problem Formulation

For a vector  $\mathbf{x}$  of  $n$  components  $\{x_i\}$ , we write the minimization problem as:

$$\min_{\mathbf{x}} \{f(\mathbf{x})\}, \quad \mathbf{x} \in \mathcal{D}, \quad (11.1)$$

where  $f$  is the objective function and  $\mathcal{D}$  is a given region (which can be the entire Euclidean space  $\Re^n$ ). The problem can be subject to  $m$  constraints, which can be written more generally as a combination of equality and inequality constraints:

$$\begin{aligned} c_i(\mathbf{x}) &= 0 && \text{for } i = 1, \dots, m', \\ c_i(\mathbf{x}) &\leq 0 && \text{for } i = m' + 1, \dots, m. \end{aligned} \quad (11.2)$$

This general formulation can be obtained for problems with bound constraints in the form

$$c_i(\mathbf{x}) = x_i,$$

where  $x_i$  is the  $i$ th component of the vector  $\mathbf{x}$ , or for problems with two-sided constraints such as

$$l_i \leq c_i(\mathbf{x}) \leq u_i.$$

In this chapter, we only cover unconstrained optimization formulations. For a comprehensive review of interior methods for continuous nonlinear optimization problems subject to constraints, see [417].

### 11.2.2 Independent Variables

In most computational chemistry problems,  $\mathbf{x}$  is a real vector in Euclidean space, i.e.,  $\mathbf{x} \in \mathbb{R}^n$ , and  $f$  defines a transformation to a real number, i.e.,  $f(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}$ . When the components of  $\mathbf{x}$  are integers, the optimization problem is classified as *integer programming*. When  $\mathbf{x}$  is a mixture of real and integer variables, the problem is of *mixed-integer programming* type. Common examples of integer programming are network optimization and the ‘traveling salesman problem’,<sup>1</sup> also classified as combinatorial optimization. See [1390], for example, and references cited therein.

### 11.2.3 Function Characteristics

The nature of the function  $f$  is the next step in problem classification. Many application areas such as finance and management-planning tackle *linear* or *quadratic* objective functions.

#### Linear and Quadratic Functions

Linear objectives can be written in vector form as

$$f(\mathbf{x}) = \mathbf{b}^T \mathbf{x} + f_0, \quad (11.3)$$

where  $\mathbf{b}$  is a column vector of dimension  $n$ , and  $f_0$  is a scalar. Quadratic objective functions can be expressed as

$$f(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x} + f_0, \quad (11.4)$$

---

<sup>1</sup>The notorious ‘traveling salesman’ problem seeks to find the optimal travel route that covers a given number of cities, each one only once, and returning to the home town. Visually, imagine drawing such a route on a map, where each city  $k$  for  $k = 0, \dots, n$  is designated by coordinates  $\{x_k, y_k\}$ . The connected route started at  $\{x_0, y_0\}$  covers each city and returns to the original point. Though simple to envision, there are clearly many such routes, and the number of combinations that connect all these cities grows steeply with  $n$ . This problem in fact belongs to a class of very difficult problems (known as *NP-complete*) for which no polynomial-complexity algorithm is known (i.e., the computational time for an *exact* solution of this problem increases exponentially with  $n$ ).

where  $\mathbf{A}$  is a constant *symmetric* matrix of dimension  $n \times n$ . (By definition, the  $n^2$  entries of a symmetric matrix  $\mathbf{A}$  satisfy  $A_{i,j} = A_{j,i}$ ). The superscripts  $T$  above refer to a vector transpose; thus  $\mathbf{x}^T \mathbf{y}$  is an inner product.

*Linear programming* problems refer to linear objective functions subject to linear constraints (i.e., a system of linear equations and inequalities), and *quadratic programming* problems have quadratic objective functions and linear constraints.

### Least-Squares Functions

Nonlinear functions can be classified further. *Least-squares* functions have the form

$$f(\mathbf{x}) = \frac{1}{2} \sum_{i=1}^m f_i(\mathbf{x})^2. \quad (11.5)$$

### Separable Functions

*Separable* functions can be expressed as a sum of subfunctions, namely

$$f(\mathbf{x}) = \sum_{i=1}^m f_i(\mathbf{x}), \quad (11.6)$$

where each subfunction  $f_i$  depends only on a subset of the independent variables. That is, for each subfunction  $f_i$  there are many unit vectors  $\mathbf{e}_j$  (with 1 in component  $j$  and 0 elsewhere) for which  $f_i(\mathbf{x} + \mathbf{e}_j) = f_i(\mathbf{x})$ . All molecular mechanics potential functions arising from the local, bonded interactions can be written this way.

### Nonsmooth Functions

Because most optimization algorithms exploit derivative information to locate optima, nonsmooth functions pose special difficulties, and very different algorithmic approaches must be used. See [153] and [407, Chapter 14] for a general introduction to nonsmooth optimization, and the two-volume set [549, 550] for the special case of nonsmooth convex problems. Optimization of nonsmooth functions requires new mathematical machinery (e.g., *subdifferentials*) that extends ordinary differentiation and leads to counterparts of most results in differential calculus (Taylor expansions, mean value theorem, etc.).

### Potential Energy Functions

Geometry optimization problems for molecular potential functions in the context of standard all-atom force fields in computational chemistry are typically of the multivariate, continuous, and nonlinear type [1108]. They can be formulated as constrained (as in adiabatic relaxation, an example of which was shown in Chapter 5) or unconstrained. Discontinuities in the derivatives may be a problem in certain formulations involving truncation, such as of the nonbonded terms (see Section 11.6).

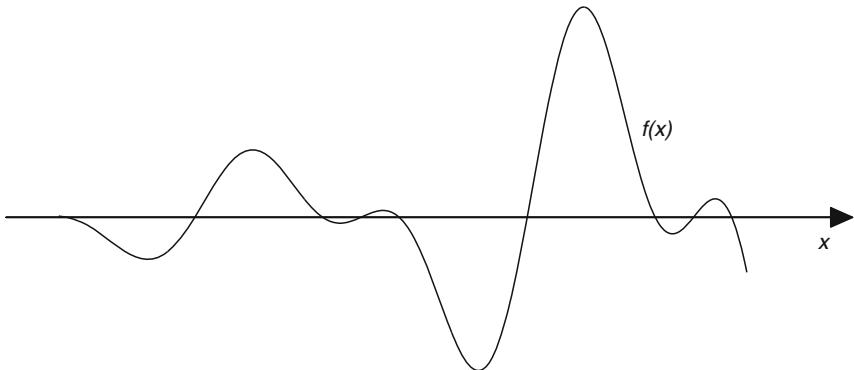


Figure 11.1. A one-dimensional function with several minima. This function was constructed from the actual univariate function at one line search step of the truncated Newton algorithm (see later in chapter) applied to minimization of a small protein's potential energy function.

The large number of independent variables for biomolecules, in particular, warrants their classification as *large-scale* and rules out the use of many algorithms that are effective for a small number of variables. However, as we will discuss, effective techniques are available today that achieve rapid convergence even for large systems. In practice, for macromolecular applications these optimization algorithms must be modest in storage requirements and economical in computations, which are dominated by the function and derivative evaluations.

#### 11.2.4 Local and Global Minima

##### Definitions

The *local* unconstrained optimization problem in the Euclidean space  $\Re^n$  can be stated as in eq. (11.1) for  $\mathbf{x} \in \mathcal{D} \subset \Re^n$  where  $\mathcal{D}$  denotes a neighborhood of the starting point,  $\mathbf{x}_0$ . The *global* optimization problem is much more difficult because it requires finding the global minimum among all the local minima, and the number of minima can be exponentially large.

A (strong) *local minimum*  $\mathbf{x}^*$  of  $f(\mathbf{x})$  satisfies

$$f(\mathbf{x}^*) < f(\mathbf{y}) \quad \text{for all } \mathbf{y} \in \mathcal{D}, \mathbf{y} \neq \mathbf{x}^*. \quad (11.7)$$

The point  $\mathbf{x}^*$  is a *weak local minimum* if  $f(\mathbf{x}^*) \leq f(\mathbf{y})$ .

A *global minimum*  $\mathbf{x}^*$  satisfies the stringent requirement that

$$f(\mathbf{x}^*) < f(\mathbf{y}) \quad \text{for all } \mathbf{y} \neq \mathbf{x}^*. \quad (11.8)$$

See Figure 11.1 for an illustration of a one-dimensional function with several minima. The function corresponds to the actual univariate function minimized in the line search substep of the TN method (see later in chapter for details).

## Convergence

Finding a *local minimum* is a challenging task for a large biological system governed by a nonlinear potential energy function. This is because the optimization scheme must find a minimum from any point along the potential surface, even one associated with a very high-energy, and should not get trapped at local maxima or saddle points. Finite-precision arithmetic and various errors that accumulate over many operations also degrade practical performance in comparison to theoretical expectations (which can be described as *convergence order*; see Box 11.1). Nonetheless, the local optimization problem is solved in a mathematical sense: convergence to a local minimum can be achieved on modern computers. In the mathematical literature, this is referred to as *global convergence* to a local minimum. Still, though many algorithms are available in widely-used molecular mechanics and dynamics packages, performance and solution quality vary considerably and depend greatly on the user-specified algorithmic convergence parameters and the starting point.

The global optimization problem, by contrast, remains unsolved in general. This is because the exponentially-growing number of minima with system size cannot be exhaustively surveyed. Certainly, effective strategies have been developed in specific application contexts (e.g., for polypeptides) and work well for moderately-sized systems. See [196,411], for example, for reviews, the website at [www.mat.univie.ac.at/~neum/glopt.html](http://www.mat.univie.ac.at/~neum/glopt.html) for general information, and homework 13 for the deterministic global optimization approach based on the *diffusion equation* [997].

Global minimization algorithms differ from the local schemes in that they do not necessarily require the energy to decrease systematically, making possible escape from local potential wells and entry into others. Global optimization methods can be stochastic or deterministic, or a combination thereof; they often rely on local optimization components.

### Box 11.1: Convergence Definitions

A sequence  $\{x_k\}$  converging to  $x^*$  has *order p* if  $p$  is the largest number such that a finite limit  $\beta$  (the “convergence ratio”, not to be confused with the line search parameter  $\beta$ ) exists, where:

$$0 \leq \lim_{k \rightarrow \infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|^p} = \beta < \infty. \quad (11.9)$$

When  $p = 2$ , we have *quadratic convergence*. When  $p = 1$ , we refer to the convergence as *superlinear* if  $\beta = 0$  and as *linear* if the nonzero  $\beta$  is less than 1.

For example, the reader can verify that the sequences  $\{2^{-2^k}\}$ ,  $\{k^{-k}\}$ , and  $\{2^{-k}\}$  converge, respectively, quadratically, superlinearly, and linearly. Quadratic convergence is faster than superlinear, which in turn is faster than linear.

### 11.2.5 Derivatives of Multivariate Functions

#### Gradient

When  $f$  is a smooth function with continuous first and second derivatives, we define its *gradient vector* of first derivatives by  $\mathbf{g}(\mathbf{x})$ , where each component of  $\mathbf{g}$  is

$$g_i(\mathbf{x}) = \partial f(\mathbf{x}) / \partial x_i. \quad (11.10)$$

#### Hessian and Curvature

The  $n \times n$  symmetric matrix of second derivatives,  $\mathbf{H}(\mathbf{x})$ , is called the *Hessian*. Its components are defined as:

$$H_{i,j}(\mathbf{x}) = \partial^2 f(\mathbf{x}) / \partial x_i \partial x_j. \quad (11.11)$$

At a *stationary* point, the gradient is zero. At a minimum point  $\mathbf{x}^*$ , in addition to stationarity, the curvature is positive. For higher dimensions, convexity is expressed as *positive-definiteness* of the Hessian. A multivariate function is positive-definite at a point  $\mathbf{x}^*$  if

$$\mathbf{y}^T \mathbf{H}(\mathbf{x}^*) \mathbf{y} > 0 \quad \text{for all nonzero } \mathbf{y}. \quad (11.12)$$

In particular, positive definiteness guarantees that all the eigenvalues are positive at  $\mathbf{x}^*$ . A *positive semi-definite* matrix has nonnegative eigenvalues; a *negative semi-definite* matrix has nonpositive eigenvalues; and a *negative-definite* matrix has only negative eigenvalues. Otherwise, the matrix is *indefinite*. The utilization of curvature information is important for formulating effective multivariate optimization algorithms.

Figure 11.2 illustrates this notion of curvature for quadratic functions of two variables:

$$q(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x}.$$

Namely, it displays the *contours* of these functions — curves on which the function is constant — in four cases. These cases are defined by different properties of the matrix  $\mathbf{A}$ : (a) indefinite, (b) positive definite, (c) negative definite, and (d) singular (i.e., not invertible). Figure 11.3 displays corresponding three-dimensional views of the functions, with circles and a line indicating stationary points. We use similar contour plots later (Figure 11.10) to illustrate paths of different minimization algorithms.

### 11.2.6 The Hessian of Potential Energy Functions

#### Sparsity

A matrix is termed *sparse* if it has a large percentage of zero entries; otherwise it is *dense*. (There is no specific threshold percentage of zero elements below which

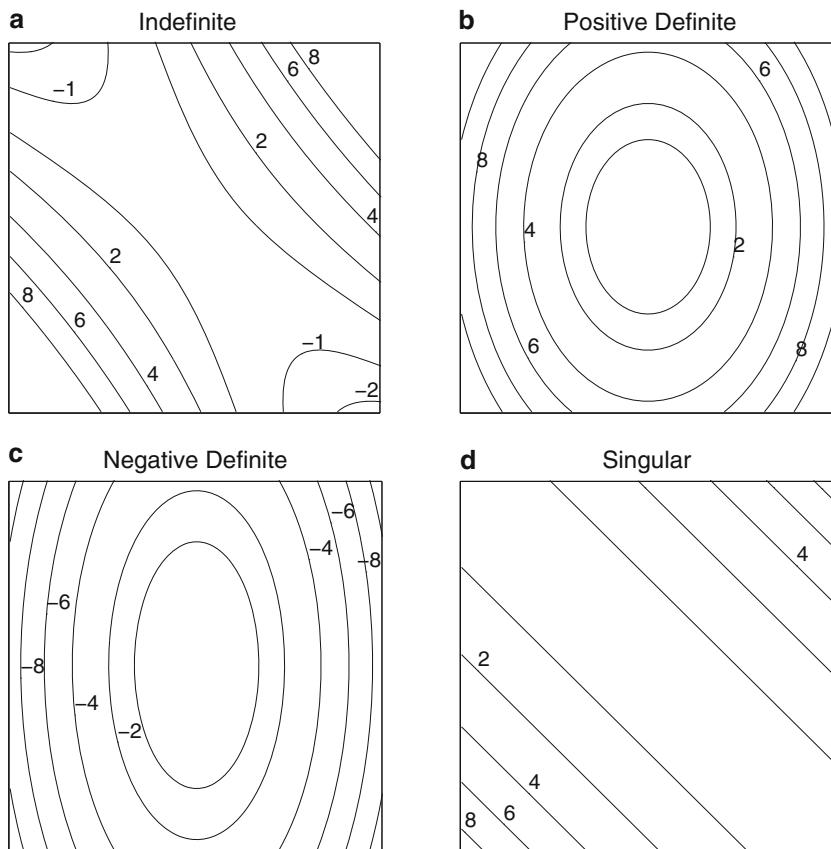


Figure 11.2. Two-dimensional contour curves for the quadratic function  $q(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x}$  of two variables, where  $\mathbf{A}$  is: (a) indefinite, with entries by row 1,2,2,2; (b) positive definite, entries 4,0,0,2; (c) negative definite, entries  $-1,0,0,-4$ ; and (d) singular, entries 1,1,1,1. See also Figure 11.3.

a matrix is considered ‘sparse’). A sparse matrix can be *structured*, as in a banded matrix of bandwidth  $p$  where there are zeros for  $|i - j| > p$ . Alternatively, a sparse matrix can be *unstructured*, as shown in Figures 11.4 and 11.5.

In these figures, the matrix indices are the independent variables (three times the number of atoms) of the potential energy function for molecular systems. A point in the matrix position  $\{i, j\}$  indicates a nonzero Hessian element for the second-derivative term of the potential energy objective function. Examples are shown for various molecular systems. The left-column matrices correspond to the Hessian pattern resulting when 8 Å cutoffs are used for the nonbonded terms. The right-column patterns correspond to only the local, bonded second-derivative

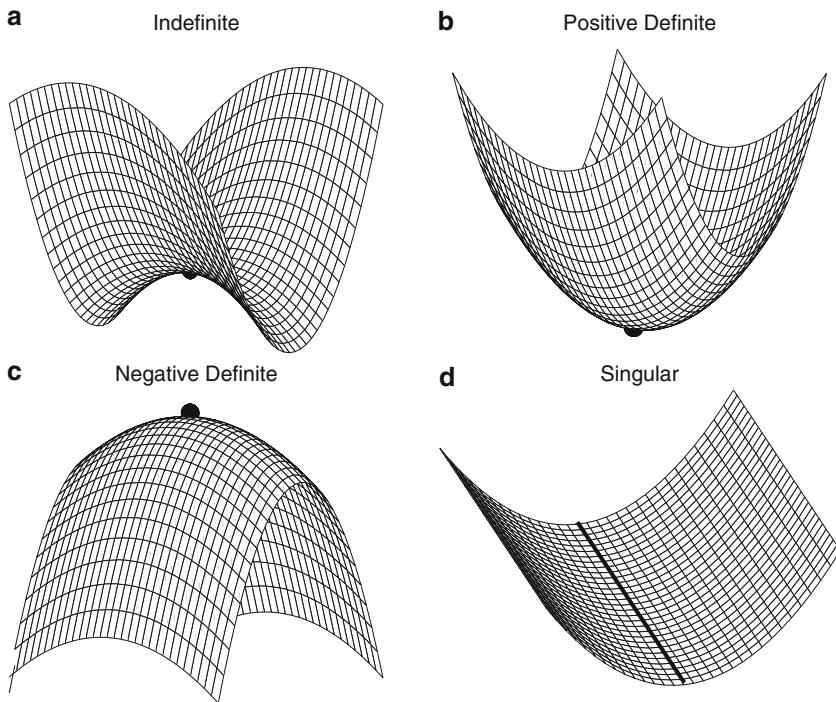


Figure 11.3. Three-dimensional curves for the quadratic functions as described for Figure 11.2. Critical points are shown by thick circles (a–c) and a line (d).

terms (bond length, bond angle, and dihedral angle). The insets zoom on two sub-matrices and illustrate how the sparsity pattern repeats in triplets (for the  $x$ ,  $y$ , and  $z$  components), and how nearly banded the local Hessian structure is due to the finite range of the bonded interactions.

We also see that although the matrices corresponding to 8 Å cutoffs are sparse for the larger systems, the atom ordering used determines the resulting pattern. For example, the X pattern for the DNA system results from the consecutive ordering of atoms down one strand and up the complementary strand; the water atoms are numbered following the DNA atoms.

### Memory Intensity

Because the formulation of a dense Hessian ( $n^2$  entries) is both memory and computation intensive, many Newton techniques for minimization approximate curvature information implicitly and often progressively, i.e., as the algorithm proceeds. Limited-memory versions reduce computational and storage requirements so that they can be applied to very large problems and/or to problems where second derivatives are not available.

## Exploitation of Derivatives

In most molecular mechanics packages, the second derivatives are programmed, though sparsity (when relevant) is not often exploited in the storage techniques for large molecular systems. The optimizer should utilize some of this second-derivative information to make the algorithm more efficient. Truncated Newton methods, for example, are designed with this philosophy.

## 11.3 Basic Algorithmic Components

### 11.3.1 Greedy Descent

The basic structure of an iterative local optimization algorithm is one of “greedy descent”. Namely, a sequence  $\{\mathbf{x}_k\}$  is generated from a starting point  $\mathbf{x}_0$  in such a way that each iterate attempts to further reduce the value of the objective function.<sup>2</sup>

#### Two Frameworks

Two algorithmic frameworks are available for such algorithms: line-search or trust-region methods. Both are found throughout the literature and in software packages and are essential components of effective descent schemes that guarantee convergence to a local minimum from any starting point. No clear evidence has emerged to render one class superior over another.

In describing iterative minimization techniques, it is convenient to use short hand notation for quantities used at each step  $k$  of the minimization algorithm. Namely, associated with each iterate  $\mathbf{x}_k$ , we denote the gradient and Hessian at  $\mathbf{x}_k$ , namely  $\mathbf{g}(\mathbf{x}_k)$  and  $\mathbf{H}(\mathbf{x}_k)$ , as  $\mathbf{g}_k$  and  $\mathbf{H}_k$ . The initial guess for the iterative minimization process ( $\mathbf{x}_0$ ) can be derived from experimental data, where available, or from results of conformational search techniques.

#### Algorithmic Parameters

The final stopping criteria must be chosen with care to ensure a sufficiently accurate solution and, at the same time, avoid wasting computational effort when further progress is not realized. For example, the norm of the gradient alone (i.e.,  $\|\mathbf{g}_k\|$ ) may not be a satisfactory stopping criterion in unconstrained optimization, as it often exhibits oscillations in the course of the optimization [917]; see also Figure 11.11.

The line search framework requires careful implementation of convergence criteria of its own at each step, for a one-dimensional optimization procedure. This segment is a tricky part of minimization methods and requires well tested software

---

<sup>2</sup>This does not imply that the reduction in the gradient norm is *monotonic*; see Figure 11.11 for example.

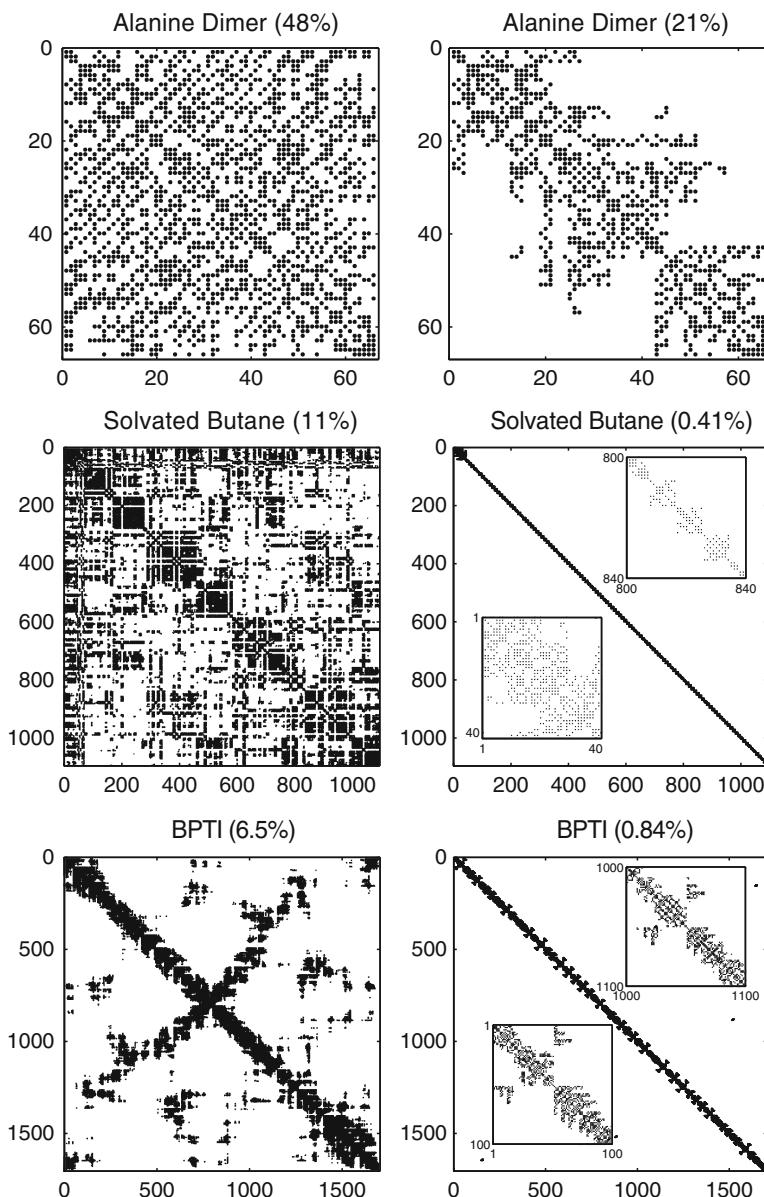


Figure 11.4. Hessian patterns from the potential energy functions of various molecular systems corresponding to 8-Å cutoffs (matrices at left column) or to local terms (right column; bond-length, bond-angle, and dihedral-angle components). The percentage sparsity is shown for each case, and insets show enlargements of some Hessian submatrices. The matrix axes label Cartesian coordinates, i.e., the  $x, y, z$  coordinates of each atom in turn; the atom ordering comes from the molecular mechanics package (CHARMM used here).

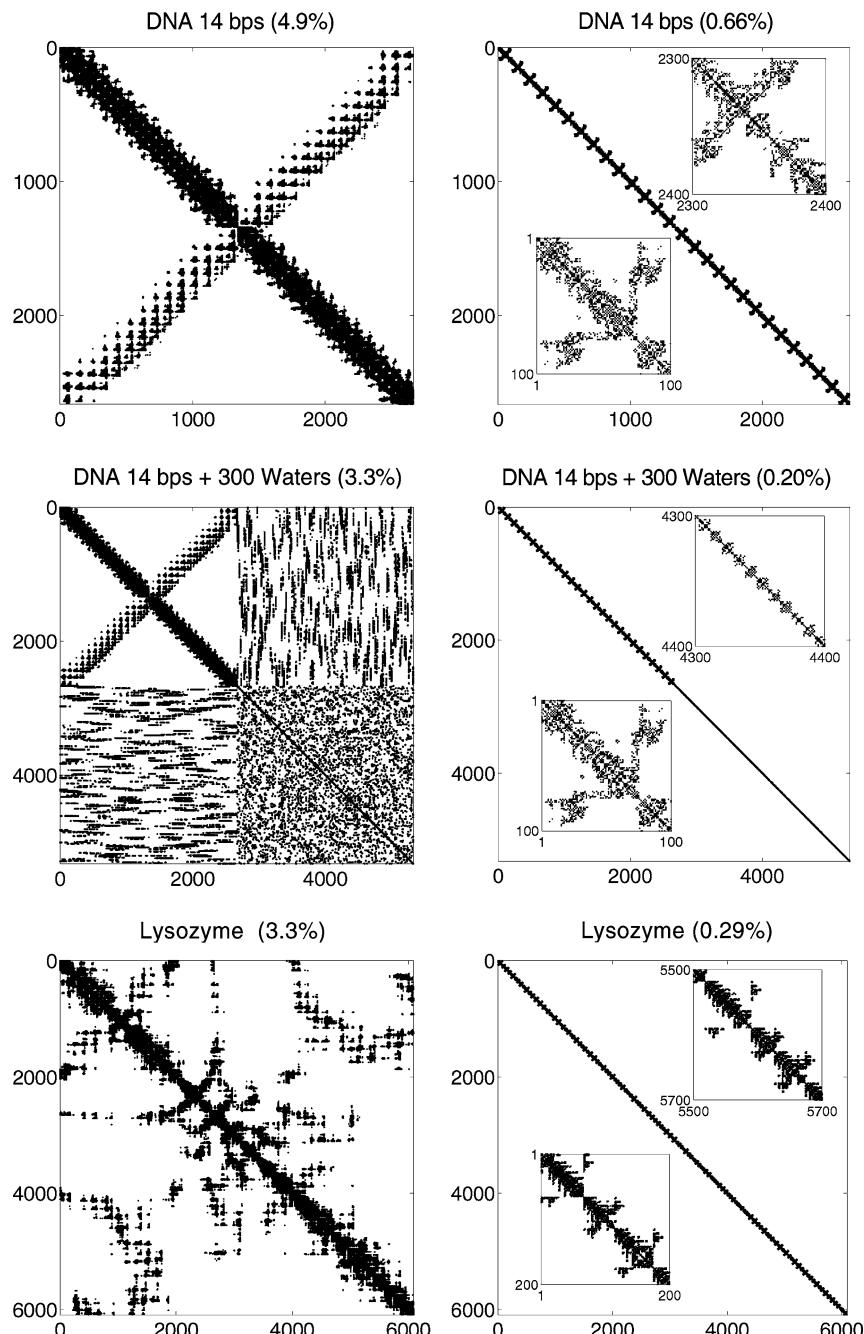


Figure 11.5. Sparse Hessian patterns, continued (see caption to Figure 11.4).

with safeguards against many undesirable situations that can occur in practice, like very small steplengths and failure to bracket the univariate minimum (see [297, 918, 1400], for example).

We now describe in turn the line search and trust-region frameworks for minimization (Subsections 11.3.2 and 11.3.3); this is followed by a discussion of convergence criteria for the minimization process (Subsection 11.3.4).

### 11.3.2 Line-Search-Based Descent Algorithm

#### Algorithm [A1]: Basic Descent Using Line Search

From a given point  $\mathbf{x}_0$ , perform for  $k = 0, 1, 2, \dots$  until convergence:

1. Test  $\mathbf{x}_k$  for convergence (see subsection 11.3.4).
2. Calculate a *descent* direction  $\mathbf{p}_k$  (method dependent).
3. Determine a steplength  $\lambda_k$  by a one-dimensional line search so that the new position vector,  $\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{p}_k$ , and corresponding gradient  $\mathbf{g}_{k+1}$ , satisfy:

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) + \alpha \lambda \mathbf{g}_k^T \mathbf{p}_k \quad [\text{"sufficient decrease"}] \quad (11.13)$$

and

$$|\mathbf{g}_{k+1}^T \mathbf{p}_k| \leq \beta |\mathbf{g}_k^T \mathbf{p}_k| \quad [\text{"sufficient directional derivative reduction"}] \quad (11.14)$$

where  $0 < \alpha < \beta < 1$

(e.g.,  $\alpha = 10^{-4}$ ,  $\beta = 0.9$  in Newton methods).

4. Set  $\mathbf{x}_{k+1}$  to  $\mathbf{x}_k + \lambda_k \mathbf{p}_k$  and  $k$  to  $k + 1$  and go to step 1.
- 

#### Step 2: Descent Direction

A *descent direction*  $\mathbf{p}_k$  is one along which the function must decrease locally. Formally, we define such a vector as one for which the directional derivative is negative:

$$\mathbf{g}_k^T \mathbf{p}_k < 0. \quad (11.15)$$

To see why this property implies that  $f$  can be reduced, approximate the nonlinear objective function  $f$  at  $\mathbf{x}$  by a linear model along the descent direction  $\mathbf{p}$ , assuming that higher-order terms are smaller than the gradient term. Then we see that the difference in function values is negative:

$$\begin{aligned} f(\mathbf{x} + \lambda \mathbf{p}) - f(\mathbf{x}) &= \lambda \mathbf{g}(\mathbf{x})^T \mathbf{p} + \frac{\lambda^2}{2} \mathbf{p}^T \mathbf{H}(\mathbf{x}) \mathbf{p} \\ &\approx \lambda \mathbf{g}(\mathbf{x})^T \mathbf{p} < 0, \end{aligned} \quad (11.16)$$

for sufficiently small positive  $\lambda$ .

### Steepest Descent

The descent condition is used to define the algorithmic sequence that generates  $\mathbf{p}_k$ . The simplest way to specify a descent direction is to set

$$\mathbf{p}_k = -\mathbf{g}_k \quad (11.17)$$

at each step. This “steepest descent” direction defines the steepest descent (SD) method. SD methods generally lead to improvements quickly but then exhibit slow progress toward a solution. Though it has become customary to recommend the use of SD for initial minimization iterations when the starting function and gradient-norm values are very large, this approach is not necessary when a more robust minimization method is available.

### Step 3: The One-Dimensional Optimization Subproblem (Line Search)

The line search procedure, a univariate minimization problem, is typically performed via approximate minimization of a quadratic or cubic polynomial interpolant of the one-dimensional function of  $\lambda$  (given  $\mathbf{x}_k$  and  $\mathbf{p}_k$ ):

$$\phi(\lambda) \equiv f(\mathbf{x}_k + \lambda \mathbf{p}_k).$$

See Figure 11.6 for an illustration of the first step of this univariate minimization process, where the minimum ( $\lambda^*$  in figure) is sought in the initial interval  $[0, \lambda_t]$  where  $f(\mathbf{x}_k) = \phi(0)$  and  $f(\mathbf{x}_k + \lambda_t \mathbf{p}_k) = \phi(\lambda_t)$ . For details, consult standard texts (e.g., [918]). This iteration process is generally continued until the  $\lambda$  value that minimizes the polynomial interpolant of  $\phi(\lambda)$  satisfies the line search criteria

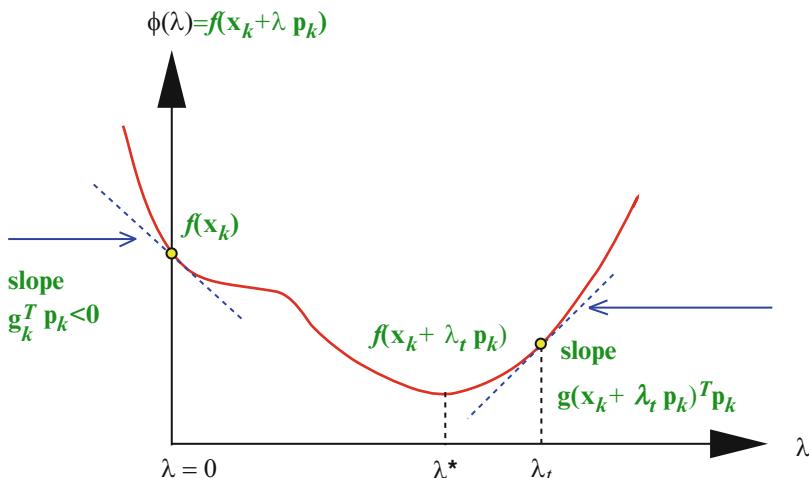


Figure 11.6. One-dimensional line search minimization at step  $k$  of the multivariate method, using polynomial approximation to estimate the optimal steplength  $\lambda^*$  in the region  $[0, \lambda_t]$ .

(eqs. (11.13) and (11.14)). The resulting steplength  $\lambda_k$  defines the next iterate for minimization of the multivariate function by  $\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{p}_k$ .

The line search criteria in Step 3 of Algorithm [A1] have been formulated to ensure sufficient decrease of  $f$  relative to the size of step ( $\lambda$ ) taken. The first condition (eq. (11.13)) prescribes an upper limit on acceptable new function values; recall that the second term on the right is negative by the descent property. The second criterion, eq. (11.14), imposes a lower bound on  $\lambda$ . The control parameters  $\alpha$  and  $\beta$  determine the balance between the computational work performed in the line search and the reduction in function achieved. (See [1104] for illustrations and further discussion). The work in the line search (number of polynomial interpolations) should be balanced with the overall progress realized in the minimization algorithm.

### 11.3.3 Trust-Region-Based Descent Algorithm

**Algorithm [A2]: Basic Descent By A Trust Region Subsearch**

From a given point  $\mathbf{x}_0$ , perform for  $k = 0, 1, 2, \dots$  until convergence:

1. Test  $\mathbf{x}_k$  for convergence (see subsection 11.3.4).
2. Calculate a step  $\mathbf{s}_k$  by solving the subproblem

$$\min_{\mathbf{s}} \{q_k(\mathbf{s})\}, \quad (11.18)$$

where  $q_k$  is the quadratic model of the objective function:

$$q_k(\mathbf{s}) = f(\mathbf{x}_k) + \mathbf{g}_k^T \mathbf{s} + \frac{1}{2} \mathbf{s}^T \mathbf{H}_k \mathbf{s}, \quad (11.19)$$

subject to a size bound,  $\Delta_k$  (a positive value), on  $\mathbf{s}$ . This bound involves a scaling matrix,  $\mathbf{D}_k$ , and requires

$$\|\mathbf{D}_k \mathbf{s}\| < \Delta_k, \quad (11.20)$$

where  $\|\cdot\|$  denotes the standard Euclidean norm.

3. Set  $\mathbf{x}_{k+1}$  to  $\mathbf{x}_k + \mathbf{s}_k$  and  $k$  to  $k + 1$  and go to step 1.
- 

#### Basic Idea

The idea in trust-region methods — the origin of the quadratic optimization subproblem in step 2 above — is to determine the vector  $\mathbf{s}_k$  on the basis of the size of region within which the quadratic functional approximation can be “trusted” (i.e., is reasonable). The quality of the quadratic approximation can be assessed from the following ratio:

$$\rho_k = \frac{f(\mathbf{x}_k) - f(\mathbf{x}_k + \mathbf{s}_k)}{f(\mathbf{x}_k) - q_k(\mathbf{s}_k)}. \quad (11.21)$$

A value near unity implied that the bound  $\Delta_k$  imposed on  $\mathbf{s}$  can be increased; in contrast, a negative or a small positive value for  $\rho_k$  implies that the quadratic model is poor, requiring a decrease in  $\Delta_k$ .

Many Newton methods (see next section for details) based on trust-region approaches determine a candidate  $\mathbf{s}_k$  by solving the linear system

$$\mathbf{H}_k \mathbf{s} = -\mathbf{g}_k \quad (11.22)$$

that results from minimizing  $q_k(\mathbf{s})$ . (A related system may also be formulated). The scaling of this vector  $\mathbf{s}$  is determined according to the quality of the quadratic model at the region of approximation. A good source of a trust-region Newton method is the program LANCELOT [261].

### 11.3.4 Convergence Criteria

The criteria used to define convergence of the minimization algorithm (in step 1 of Algorithms [A1] or [A2]) must be chosen with care. The desire to obtain as accurate a result as possible should be balanced with the amount of computation involved. In other words, it is wasteful to continue a loop when the answer can no longer be improved. A well-structured algorithm should halt the iteration process when progress is poor.

The gradient norm value, together with measures of progress in the function values and the independent variables (e.g., eqs. (11.25) and (11.26)) are used to assess performance. Upper limits for the total number of allowable function and/or gradient evaluations are important safeguards against wasteful computing cycles.

Specifically, reasonable tests for the size of the gradient norm are:

$$\|\mathbf{g}_k\| \leq \epsilon_g (1 + |f(\mathbf{x}_k)|) \quad (11.23)$$

or

$$\|\mathbf{g}_k\| \leq \epsilon_g \max(1, \|\mathbf{x}_k\|), \quad (11.24)$$

where the gradient norm may be set to the Euclidean norm divided by  $\sqrt{n}$  (this introduces a dependency on the number of variables). The parameter  $\epsilon_g$  is a small positive number such as  $10^{-6}$  that might depend on the *machine precision*,  $\epsilon_m$ ;  $\epsilon_m$  is roughly the largest number for which  $1 + \epsilon_m = 1$  in computer representation. The student is encouraged to code a routine for determining  $\epsilon_m$ .<sup>3</sup>

For example, our truncated-Newton package TNPACK [1121, 1122, 1130] checks the following four conditions at each iteration:

$$f(\mathbf{x}_{k-1}) - f(\mathbf{x}_k) < \epsilon_f (1 + |f(\mathbf{x}_k)|), \quad (11.25)$$

$$\|\mathbf{x}_{k-1} - \mathbf{x}_k\| < (\epsilon_f)^{1/2} (1 + \|\mathbf{x}_k\|), \quad (11.26)$$

---

<sup>3</sup>Typically,  $\epsilon_m$  is  $10^{-15}$  and  $10^{-7}$ , respectively, for double and single-precision IEEE arithmetic [954].

$$\| \mathbf{g}_k \| < (\epsilon_f)^{1/3} (1 + |f(\mathbf{x}_k)|), \quad (11.27)$$

$$\| \mathbf{g}_k \| < \epsilon_g (1 + |f(\mathbf{x}_k)|). \quad (11.28)$$

Here, all norms are the Euclidean norm divided by  $\sqrt{n}$ , and  $\epsilon_f$  and  $\epsilon_g$  are small numbers (like  $10^{-10}$  and  $10^{-8}$ , respectively). If the first three conditions above are satisfied, or the fourth condition alone is satisfied, convergence is considered to have been satisfied and the minimization process is halted; otherwise, the loop continues. Note that the first and second conditions test for convergence of the sequences of function values and iterates of the independent variables, respectively, while the third and fourth conditions measure the size of the gradient norm in terms of two different parameters.

---

### Box 11.2: Historical Perspective of ‘Newton’s’ Method

The method’s credit to Sir Isaac Newton is a partial one. Although many references also credit Joseph Raphson, the contributions of mathematicians Thomas Simpson and Jean-Baptiste-Joseph Fourier are also noteworthy. Furthermore, Newton’s description of an algebraic procedure for solving for the zeros of a polynomial in 1664 had its roots in the work of the 16th-century French algebraist François Viète. Viète’s work itself had precursors in the 11th-century works of Arabic algebraists.

In 1687, three years after Newton described a root finder for a polynomial, Newton described in *Principia Mathematica* an application of his procedure to a nonpolynomial equation. That equation originated from the problem of solving Kepler’s equation: determining the position of a planet moving in an elliptical orbit around the sun, given the time elapsed since it was nearest the sun. Newton’s procedure was nonetheless purely *algebraic* and not even iterative.

In 1690, Raphson turned Newton’s method into an *iterative* one, applying it to the solution of polynomial equations of degree up to ten. His formulation still did not use calculus; instead he derived explicit polynomial expressions for  $f(x)$  and  $f'(x)$ .

Simpson in 1740 was first to formulate the Newton-Raphson method on the basis of calculus. He applied this iterative scheme to solve general systems of nonlinear equations. In addition to this important extension of the method to nonlinear systems, Simpson extended the iterative solver to multivariate minimization, noting that the nonlinear solver can be applied to optimization by setting the gradient to zero.

Finally, Fourier in 1831 published the modern version of the method as we know it today in his celebrated book *Analyse des Équations Déterminées*. The method for solving  $f(x) = 0$  was simply written as:  $x_{k+1} = x_k - f(x_k)/f'(x_k)$ . Unfortunately, Fourier omitted credits to either Raphson or Simpson, possibly explaining the method’s name.

Thus, strictly speaking, it is appropriate to title the method as the Newton-Raphson-Simpson-Fourier method.

---

## 11.4 The Newton-Raphson-Simpson-Fourier Method

Newton's method is a classic iterative scheme for solving a nonlinear system  $f(\mathbf{x}) = 0$  or for minimizing the multivariate function  $f(\mathbf{x})$ . These root-finding and minimization problems are closely related.

Though the method is credited to Newton or Newton and Raphson, key contributions were made also by Fourier and Simpson; see Box 11.2 for a historical perspective. For brevity, we refer to the “Newton-Raphson-Simpson-Fourier method” as Newton's method.

### A Fundamental Optimization Tool

Many effective methods for nonlinear, multivariate minimization can be related to Newton's method. Hence, a good understanding of the Newton solver, including performance and convergence behavior, is invaluable for applying optimization techniques in general.

We first discuss the univariate case of Newton's method for obtaining the zeros of a function  $f(x)$ . In one dimension, instructive diagrams easily illustrate the method's strengths and weaknesses. We then discuss the general multivariate formulations. The section that follows continues by describing the effective variants known as quasi-Newton, nonlinear conjugate gradient, and truncated-Newton methods.

#### 11.4.1 The One-Dimensional Version of Newton's Method

##### Iterative Recipe

The modern version of Newton's method (see Box 11.2) for solving  $f(x) = 0$  is:

$$x_{k+1} = x_k - f(x_k)/f'(x_k). \quad (11.29)$$

This iterative scheme can be derived easily by using a Taylor expansion to approximate a twice-differentiable function  $f$  locally by a quadratic function about  $x_k$ :

$$f(x_{k+1}) = f(x_k) + (x_{k+1} - x_k) f'(x_k) + \frac{1}{2}(x_{k+1} - x_k)^2 f''(\xi) \quad (11.30)$$

where  $x_k \leq \xi \leq x_{k+1}$ . Omitting the second-derivative term, the solution of  $f(x_{k+1}) = 0$  yields the iteration process of eq. (11.29). The related *discrete-Newton* and *quasi-Newton* methods [918] correspond to approximating  $f'(x)$  by *finite-differences*, as

$$f'(x_k) \approx [f(x_k + h) - f(x_k)] / h, \quad (11.31)$$

or by the method of *secants*

$$f'(x_k) \approx [f(x_k) - f(x_{k-1})] / (x_k - x_{k-1}), \quad (11.32)$$

where  $h$  is a suitably-chosen small number.

### Geometric Interpretation

Newton's method in one dimension has a simple geometric interpretation: at each step, approximate  $f(x)$  by its tangent at point  $\{x_k, f(x_k)\}$  and take  $x_{k+1}$  as the abscissa of the intersection of this line with the  $x$ -axis (see Figure 11.7).

### Performance

The method works well in the ideal case (Figure 11.7a), when  $x_0$  is near the solution ( $x^*$ ) and  $|f'(\xi)| \geq M > 0$  nearby.

However, difficulties arise when  $x_0$  is far from the solution, or when  $f'(x)$  is close to zero (Figure 11.7b). Further difficulties emerge when  $f'(x)$  is zero at the solution.

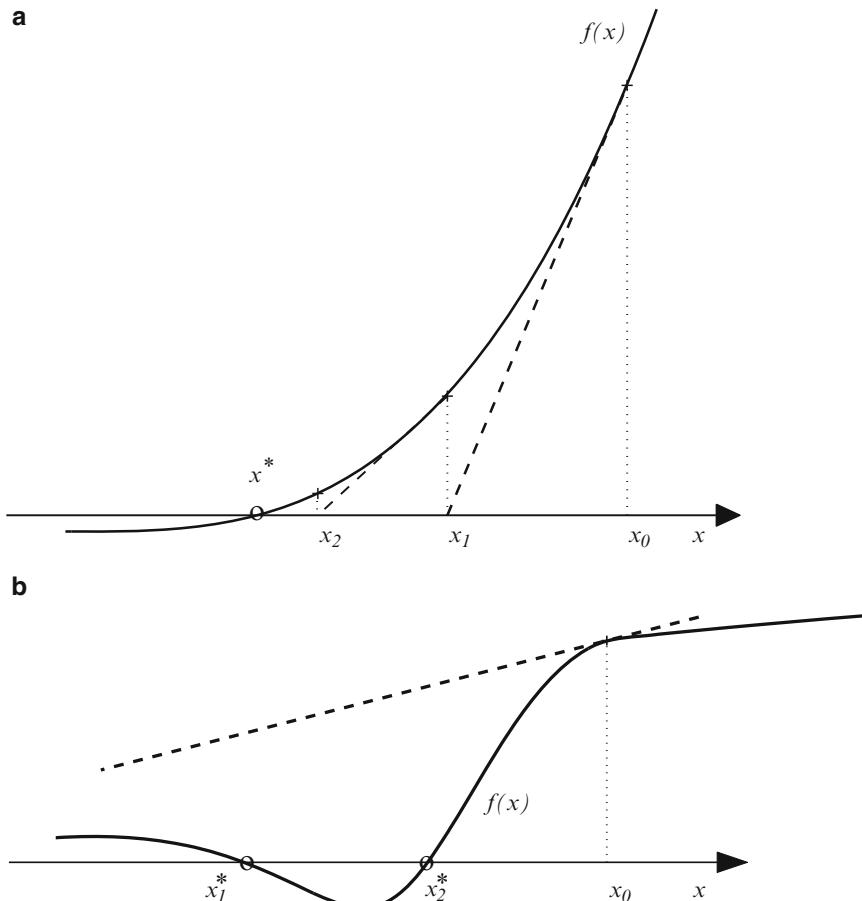


Figure 11.7. Newton's method in one dimension: (a) geometric interpretation and behavior in the ideal case (rapid convergence), and (b) divergent behavior near point where  $f'(x) = 0$ .

Note that the Newton iteration process is undefined when  $f'(x) = 0$  and can exhibit poor numerical behavior when  $|f'(x)|$  is very small, as shown in Figure 11.7b. In general, both performance and attainable accuracy of the solver worsen if any of the above complications arise.

A simple example for solving for the square root of a number by Newton's method in Box 11.3 (with associated data in Figure 11.8) illustrates the rapid convergence for the ideal case when the root of  $f(x)$  is simple and reasonably separated from the other root. In the non-ideal case (e.g.,  $x_0$  far from the solution or  $f'(x)$  close to zero), convergence is slow at the beginning but then improves rapidly until the region of *quadratic convergence* is approached. (See Box 11.1 for a definition of quadratic convergence).

The quadratic convergence of Newton's method for a simple root and for  $x_0$  sufficiently close to the solution  $x^*$  can easily be shown on the basis of the Taylor expansion. Tensor methods based on fourth-order approximations to the objective function can achieve more rapid convergence [1137], but they are not generally applicable. The attainable accuracy for Newton's method depends on the function characteristics, and on whether the root is simple or not.

### Box 11.3: Newton's Method: Simple Examples

We can apply Newton's method to solve for the square root of a number  $a$  by defining

$$f(x) = x^2 - a = 0;$$

the resulting iterative scheme for computing  $x = \sqrt{a}$  is:

$$x_{k+1} = \frac{1}{2} \left[ x_k + \frac{a}{x_k} \right], \quad (11.33)$$

defined for  $x_k \neq 0$ . A computer result from a double-precision program is shown in Figure 11.8 for  $a = 0.01$  with four starting points: 5, -100, 1000, and  $10^{-6}$ .

The rapid, *quadratic* convergence (see Box 11.1) can be noted in all cases at the last 3–4 steps. In these steps, the number of correct digits for the solution is approximately doubled from one step to the next! Note the larger number of iterations for convergence when  $x_0$  is near zero ( $x_0 = 10^{-6}$  shown). Since the derivative of the objective function is zero at  $x = 0$ , the tangent takes the iterates very far, as illustrated in Fig. 11.7b, but then works back systematically toward the solution.

Figure 11.9 further illustrates the notions of of accuracy, convergence, and problem conditioning for solving

$$f(x) = x^3 - bx = 0, \quad b > 0,$$

by Newton's method. The three roots are  $-\sqrt[3]{b}$ , 0, and  $+\sqrt[3]{b}$ . The corresponding iterative scheme for solving this cubic polynomial of  $x$  becomes:

$$x_{k+1} = 2x_k^3/(3x_k^2 - b). \quad (11.34)$$

Near the point where the Newton iteration is undefined ( $x_k = \sqrt[3]{b/3}$ ), the iterative process converges very slowly. When  $b$  is small, as in our example ( $b = 0.0001$ ), the three roots

Newton iterate, $x$	Error: $ x - x^* /x^*$ (for nonzero $x^*$ )	Newton iterate, $x$	Error: $ x - x^* /x^*$ (for nonzero $x^*$ )
$x_0 = 5$		$x_0 = 15$	$155.2521328066817$
0 5.000000000000000 49.00000000000000	7.812926635966156	7 7.812926635966156	77.12926635966156
1 2.501000000000000 24.01000000000000	3.907103283034968	8 3.907103283034968	38.07103283034968
2 1.252499200319872 11.52499200319872	9.1954831361974763	9 1.954831361974763	18.54831361974762
3 0.6302416186767726 5.30230542746187304	10 0.979734463768975	8.799734463768974	8.799734463768974
4 0.3230542746187304 2.230542746187304	11 0.4950889022510404	3.950889022510404	3.950889022510404
5 0.1770044127792565 0.7700441277925646	12 0.2576436474057566	1.576436474057566	1.576436474057566
6 0.1167500897135064 0.1675008971350636	13 0.1482284733538422	0.482284733538422	0.482284733538422
7 0.1012015644103529 1.2015644103528789E-02	14 0.10784594750729798	7.8459475072977791E-02	7.8459475072977791E-02
8 0.1000071330766507 7.1330766507383161E-05	15 0.1002850419724900	2.854019724899588E-03	2.854019724899588E-03
9 0.1000000002543858 2.5438576245484512E-09	16 0.10000004061123768	4.0611237676901890E-06	4.0611237676901890E-06
10 0.1000000000000000 0.0000000000000000	17 0.10000000000008246	8.2461815154033502E-12	8.2461815154033502E-12
$x_0 = -100$	18 0.1000000000000000	0.0000000000000000	0.0000000000000000
$x_0 = 10^{-6}$			
0 -100.0000000000000 999.0000000000000	0 9.999999999999995E-07	0.9999900000000000	0.9999900000000000
1 -50.0000500000000 499.0000000000000	1 5000.00000500000	49999.00000499999	49999.00000499999
2 -25.0001249999000 249.0012499990000	2 25000.00001250000	24999.00001250000	24999.00001250000
3 -12.50026249895001 124.0026249895001	3 12500.00002625000	12499.00002625000	12499.00002625000
4 -6.250531241075214 61.50531241075214	4 625.0000053125000	6249.0000053124999	6249.0000053124999
5 -3.126065552544528 30.26065552544528	5 312.5000106562499	3124.000106562499	3124.000106562499
6 -1.564632230895322 14.64632230895322	6 156.2500213281244	1561.500213281244	1561.500213281244
7 -0.7855117545897353 6.855117545897353	7 78.12504266405783	780.2504266405783	780.2504266405783
8 -0.3991211544161648 2.991211544161648	8 39.06258533199397	389.6258533199396	389.6258533199396
9 -0.2120881016068179 1.120881016068179	9 19.53142066571737	194.3142066571736	194.3142066571736
10 -0.1296191592706879 0.2961915927068787	10 9.765966330621753	96.65966330621752	96.65966330621752
11 -0.1033841239244204 3.3841239244203625E-02	11 4.883495147415947	47.83495147415947	47.83495147415947
12 -0.1000553871053945 5.538710539441011E-04	12 2.442771430566446	23.42771430566446	23.42771430566446
13 -0.100000153301663 1.5330166275306922E-07	13 1.223432570726772	11.23432570726771	11.23432570726771
14 -0.1000000000000012 1.1657341758564144E-14	14 0.6158031472956539	5.158031472956539	5.158031472956539
15 -0.1000000000000000 0.0000000000000000	15 0.3160210514743984	2.160210514743984	2.160210514743984
$x_0 = 100$	16 0.1738322565259314	0.7383225652593136	0.7383225652593136
0 1000.000000000000 9999.000000000000	17 0.1156794895626801	0.1567948956268007	0.1567948956268007
1 500.000000000000 4999.000000000000	18 0.1010626187661944	1.0626187661944286E-02	1.0626187661944286E-02
2 250.0000124999999 2499.0000124999999	19 0.1000055864307498	5.5864307498265653E-05	5.5864307498265653E-05
3 125.0000262499899 1249.0000262499899	20 0.1000000001560323	1.5603232594862959E-09	1.5603232594862959E-09
4 62.50005312499107 624.0005312499106	21 0.1000000000000000	0.0000000000000000	0.0000000000000000
5 31.25010656242754 311.5010656242753			

Figure 11.8. Computer output from the application of Newton's method to solve the simple quadratic  $x^2 = a$ ,  $a = 0.01$ , from four different starting points.

are relatively close. The Newton iterates in Figure 11.9 started from  $-50$ ,  $\sqrt{b/3} + 10^{-10}$ ,  $-1$ ,  $10^{-10}$ ,  $0.009$ , and  $0.011$  show that the solution obtained depends on the starting point.

### 11.4.2 Newton's Method for Minimization

To derive the iteration process of Newton's method for minimization of the one-dimensional  $f(x)$ , we use a quadratic, rather than linear, approximation:

$$f(x_{k+1}) \approx f(x_k) + (x_{k+1} - x_k) f'(x_k) + \frac{1}{2}(x_{k+1} - x_k)^2 f''(x_k). \quad (11.35)$$

Since  $f(x_k)$  is constant, minimization of the second and third terms on the right-hand-side in eq. (11.35) yields the iteration process:

$$x_{k+1} = x_k - f'(x_k)/f''(x_k). \quad (11.36)$$

Thus, we have replaced  $f$  and  $f'$  of eq. (11.29) by  $f'$  and  $f''$ , respectively. This Newton scheme for minimizing  $f(x)$  is defined as long as the second derivative at  $x_k$  is nonzero.

Newton iterate, $x$	Error: $ x - x^* /x^*$ (for nonzero $x^*$ )	Newton iterate, $x$	Error: $ x - x^* /x^*$ (for nonzero $x^*$ )
$x_0 = -50$		$x_0 = .0058 + 10^{-10}$	
0 - 50.00000000000000	4999.000000000000	0 5.7735027918962576E-03	0.4226497208103743
1 - 33.33333377777778	3332.333377777778	1 111111.1159104916	1111110.59104916
2 - 22.222318518520	2221.222318518520	2 74074.07727366130	7407406.727366131
3 - 14.81481645679016	1480.481645679016	3 49382.71818244116	4938270.818244115
4 - 9.876545804526833	986.6545804526833	4 32921.81212162789	3292180.212126278
5 - 6.584366119684733	657.4366119684732	5 21947.874775260	2194786.4775260
6 - 4.389580788123710	437.9580788123710	6 14631.91649850275	1463190.649850275
7 - 2.926392254584279	291.6392254584279	7 9754.610999003351	975460.0999003351
8 - 1.950935763478845	194.0935763478845	8 6503.073999337846	650306.3999337845
9 - 1.300635232964303	129.0635232964302	9 4335.382666228647	433537.2666228647
10 - 0.8671072413145484	85.71072413145484	10 2890.255110824224	289024.5110824224
11 - 0.587097123434421	56.8097123434420	11 1926.836740557172	192682.6740557172
12 - 0.3854365263554411	37.54365263554411	12 1284.557827049647	128454.7827049647
13 - 0.2507153518629270	24.70153518629269	13 856.3718847170644	85636.18847170644
14 - 0.1714300741869435	16.14300741869435	14 570.9145898373255	57090.45898373255
15 - 0.1144164918146208	10.44164918146208	15 380.6097265971409	38059.97265971409
16 - 7.6472379206287938E-02	6.647237920628794	16 253.7398177898131	25372.98177898131
17 - 5.1273843468759718E-02	4.127384346875972	17 169.1598786141209	16914.98786141209
18 - 3.462153071892676E-02	2.462153071892676	18 112.732525407821	11276.32525407821
19 - 2.3741241957549244E-02	1.3741241957549244	19 75.18216855757360	7517.216855757360
20 - 1.6822346583303216E-02	0.6822346583303216	20 50.2144600062744	5011.144600062744
21 - 1.2712265750664345E-02	0.2712265750664345	21 33.4142977711917	3340.42977711918
22 - 1.06772168791739191E-02	6.77216879173919102	22 22.27619918313077	2226.619918313077
23 - 1.0059418708361139E-02	5.9418708361139161E-03	23 14.8508045299750	1484.080045299750
24 - 1.0000522346453959E-02	5.2234645395859980E-05	24 9.900535131697179	989.0535131697179
25 - 1.000000040921884E-02	4.092188393706243E-09	25 6.600358999013134	659.0358999013134
26 - 1.0000000000000000	0.0000000000000000	26 4.400246299498232	439.0246299498232
$x_0 = -1$		27 2.933500183234001	292.3500183234001
0 - 1.00000000000000	99.00000000000000	28 1.955674364178697	194.5674364178697
1 - 0.666688896296543	65.6688896296543	29 1.303794272497512	129.3794272497512
2 - 0.4444925944752626	43.44429544752625	30 0.8692132262697133	85.92132262697133
3 - 0.2963783993367317	28.63783993367316	31 0.5795010512121747	56.95010512121746
4 - 0.1976606072449507	18.76606072449507	32 0.3863723851127763	37.63723851127763
5 - 0.1318862603202570	12.18862603202570	33 0.2576391179580148	24.76391179580147
6 - 8.809292421460537E-02	7.80929242142054	34 0.1718457086006079	16.18457086006079
7 - 5.898200850453032E-02	4.89820085045303	35 0.1146932668316612	10.46932668316612
8 - 3.9701746654566113E-02	2.9701746654566113	36 7.6656423605242732E-02	6.665642360524274
9 - 2.7039652798163824E-02	1.703965279816382	37 5.1395830038388754E-02	4.139583003838875
10 - 1.8887531503436573E-02	0.8887531503436572	38 3.4701786556898713E-02	2.470178655689871
11 - 1.38895100791484E-02	0.3889510079148153	39 2.379313189699858E-02	1.379313189699886
12 - 1.1193786717951709E-02	0.1193786717951709	40 1.6854498439733485E-02	0.6854498439733484
13 - 0.16729368239794E-02	0.1672936823979400E-02	41 1.2730083741629223E-02	0.2730083741629223
14 - 1.0000404037696271E-02	4.0403576962704663E-04	42 1.0684415288974207E-02	6.8441528897420639E-02
15 - 1.0000000244636725E-02	2.4463672546048976E-07	43 1.0060600942778943E-02	6.060094277894651E-03
16 - 1.000000000000899E-02	8.9858676055598607E-14	44 1.000000045191397791E-02	5.4319139779039627E-05
17 - 1.000000000000002E-02	1.7347234759768071E-16	45 1.00000000044252924E-02	4.4252924241705571E-09
$x_0 = 10^{-10}$		46 1.00000000000000E-02	0.0000000000000000
$x_0 = .009$		$x_0 = .011$	
0 1.00000000000000E-10	0.1000000000000000	0 1.0999999999999999E-02	9.999999999999998E-02
1 - 2.00000000000000E-26	0.2000000000000000	1 1.0121673003802283E-02	1.216730038022837E-02
2 - 1.60000000000000E-73	0.1600000000000000	2 1.000215936055317E-02	2.159360553163475E-04
3 - 8.19200000000000E-215	0.0819200000000000	3 1.00000000000000E-02	3.10000000000000E-08
4 - 0.0000000000000000	0.0000000000000000	4 1.00000000000000E-02	4.10000000000000E-15
0 8.9999999999999993E-03	0.1000000000000000	5 9.9999999999999985E-03	5.9999999999999985E-03
1 1.0195804195804197E-02	0.19580419580419658E-02	6 1.7347234759768071E-16	1.7347234759768071E-16
2 1.0005499738015171E-02	0.54997380151706327E-04		
3 1.0000004531252979E-02	0.45312529787372435E-07		
4 1.000000000003079E-02	0.30791341698588326E-13		
5 9.99999999999999985E-03	0.17347234759768071E-16		

Figure 11.9. Computer output from the application of Newton's method to solve  $x^3 - bx = 0$ ,  $b = 10^{-4}$ , from various starting points.

### 11.4.3 The Multivariate Version of Newton's Method

We generalize Newton's method for minimization in eq. (11.36) to multivariate functions by expanding  $f(\mathbf{x})$  locally along a search vector  $\mathbf{p}$  (in analogy to eq. (11.35)):

$$f(\mathbf{x}_k + \mathbf{p}_k) \approx f(\mathbf{x}_k) + \mathbf{g}(\mathbf{x}_k)^T \mathbf{p}_k + \frac{1}{2} \mathbf{p}_k^T \mathbf{H}(\mathbf{x}_k) \mathbf{p}_k. \quad (11.37)$$

Minimizing the right-hand side leads to solving the linear system of equations, known as the *Newton equations*, for  $\mathbf{p}_k$ , as long as  $\mathbf{H}_k$  is positive definite:

$$\mathbf{H}_k \mathbf{p}_k = -\mathbf{g}_k . \quad (11.38)$$

Performing this approximation at each step  $k$  to obtain  $\mathbf{p}_k$  leads to the iteration process

$$\mathbf{x}_{x+1} = \mathbf{x}_k - \mathbf{H}_k^{-1} \mathbf{g}_k . \quad (11.39)$$

Thus, the search vector

$$\mathbf{p}_k = -\mathbf{H}_k^{-1} \mathbf{g}_k \quad (11.40)$$

is used at each step of the *classic* Newton method for minimization. This requires repeated solutions of a linear system involving the Hessian. Not only is this an expensive, order  $n^3$  process for general dense matrices; for multivariate functions with many minima and maxima, the Hessian may be *ill-conditioned* (i.e., have large maximal-to-minimal eigenvalue ratio  $\lambda_{\max}/\lambda_{\min}$ ) or *singular* (zero eigenvalues) for certain  $\mathbf{x}_k$ .

Thus, in addition to the line-search or trust-region modifications that essentially *dampen* the Newton step (by scaling  $\mathbf{p}_k$  by a positive scalar less than unity), effective strategies must be devised to ensure that  $\mathbf{p}_k$  is well defined at each step. Such effective strategies are described in the next section. These include quasi-Newton (QN), nonlinear conjugate gradient (CG), and truncated Newton (TN) methods.

## 11.5 Effective Large-Scale Minimization Algorithms

The popular methods that fit the descent framework outlined in subsections 11.3.2 and 11.3.3 require gradient information. In addition, the truncated-Newton (TN) method may require more input to be effective, such as second-derivative information from components of the objective function that can be computed cheaply.

The steepest descent method (recall  $\mathbf{p}_k = -\mathbf{g}_k$ ) can be viewed as a simple version of the  $\mathbf{p}_k$  definition in eq. (11.40) in which the Hessian replaced by the identity matrix.

Nonlinear CG methods improve upon (the generally poor) convergence of SD methods by using better search directions than SD that are still cheap to compute.

QN methods, which are closely related to nonlinear CG methods, can also be presented as robust alternatives to the classic Newton method ( $\mathbf{p}_k$  by eq. (11.40)) which update curvature information as the algorithm proceeds.

TN methods are another clever and robust alternative to the classic Newton framework that introduce curvature information only when locally warranted, so as to balance computation with realized convergence. Hybrid schemes have also been devised, e.g., limited-memory QN with TN [194]. See also [1451] for applications of truncated Newton minimization to protein structure and protein-ligand docking problems.

The methods described in turn in this section — QN, nonlinear CG, and TN methods — render SD obsolete as a general method.

### 11.5.1 Quasi-Newton (QN)

#### Basic Idea

QN methods avoid using the actual Hessian and instead build-up curvature information as the algorithm proceeds [458, 918]. Actually, it is often the Hessian inverse ( $\widehat{\mathbf{B}}$ ) that is updated in practice so that a term  $\widehat{\mathbf{B}}_k \mathbf{g}_k$  replaces  $\mathbf{H}_k^{-1} \mathbf{g}_k$  in eq. (11.39). Here  $\widehat{\mathbf{B}}_k$  is short hand for  $\widehat{\mathbf{B}}(\mathbf{x}_k)$ .

The Hessian approximation  $\mathbf{B}_k$  is derived to satisfy the *quasi-Newton condition* (see below). QN variants define different formulas that satisfy this condition.

Because memory is considered premium for large-scale applications, the matrix  $\mathbf{B}_k$  or  $\widehat{\mathbf{B}}$  is formulated through several vector operations, avoiding explicit storage of an  $n \times n$  matrix. In practice,  $\mathbf{B}_k$  is updated by adding a *low rank* update matrix  $\mathbf{U}_k$ .

#### Recent Advances

Two important developments have emerged in modern optimization research in connection with QN methodology. The first is the development of *limited-memory* versions, in which the inverse Hessian approximation at step  $k$  only incorporates curvature information generated at the last few  $m$  steps (e.g.,  $m = 5$ ) [456, 774, 914, 918]. The second is the emergence of insightful analyses that explain the relationship between QN and nonlinear CG methods.

#### QN Condition

The QN condition specifies the property that the new approximation  $\mathbf{B}_{k+1}$  must satisfy:

$$\mathbf{B}_{k+1} \mathbf{s}_k = \mathbf{y}_k. \quad (11.41)$$

Here

$$\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k \quad (11.42)$$

and

$$\mathbf{y}_k = \mathbf{g}_{k+1} - \mathbf{g}_k. \quad (11.43)$$

(Note that the ‘step vector’  $\mathbf{s}_k$  can be equated with the displacement from  $\mathbf{x}_k$ , namely  $\lambda_k \mathbf{p}_k$ , used in the basic Algorithm [A1]). If  $f(\mathbf{x})$  were a quadratic function, its Hessian  $\mathbf{H}$  would be a constant and would satisfy (from the Taylor expansion of the gradient) the following relation:

$$\mathbf{g}_{k+1} - \mathbf{g}_k = \mathbf{H} (\mathbf{x}_{k+1} - \mathbf{x}_k). \quad (11.44)$$

This equation makes clear the origin of the QN condition of eq. (11.41).

### Updating Formula

The updating QN formula can be written symbolically as:

$$\mathbf{B}_{k+1} = \mathbf{B}_k + \mathbf{U}_k(\mathbf{s}_k, \mathbf{y}_k, \mathbf{B}_k) \quad (11.45)$$

where  $\mathbf{U}_k$  is a matrix of low rank (typically 1 or 2). Note that a rank 1 matrix can be written as the outer product of two vectors:  $\mathbf{u}\mathbf{v}^T$ . In addition to rank, imposed symmetry and positive-definiteness are used in the formulation of  $\mathbf{U}_k$ .

### BFGS Method

One of the most successful QN formulas in practice is associated with the BFGS method (for its developers Broyden, Fletcher, Goldfarb, and Shanno). The BFGS update matrix has rank 2 and inherent positive definiteness (i.e., if  $\mathbf{B}_k$  is positive definite then  $\mathbf{B}_{k+1}$  is positive definite) as long as  $\mathbf{y}_k^T \mathbf{s}_k > 0$ . This condition is satisfied automatically for convex functions but may not hold in general without the sufficient reduction of curvature criteria (eq. (11.14)) in the line search. In practice, the line search must check for the descent property; updates that do not satisfy this condition may be skipped.

The BFGS update formula is given by

$$\mathbf{B}_{k+1} = \mathbf{B}_k - \frac{\mathbf{B}_k \mathbf{s}_k \mathbf{s}_k^T \mathbf{B}_k^T}{\mathbf{s}_k^T \mathbf{B}_k \mathbf{s}_k} + \frac{\mathbf{y}_k \mathbf{y}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}. \quad (11.46)$$

The corresponding formula used in practice to update the inverse of  $\mathbf{B}$ , namely  $\hat{\mathbf{B}}$ , is:

$$\hat{\mathbf{B}}_{k+1} = \left( \mathbf{I} - \frac{\mathbf{s}_k \mathbf{y}_k^T}{\mathbf{y}_k^T \mathbf{s}_k} \right) \hat{\mathbf{B}}_k \left( \mathbf{I} - \frac{\mathbf{y}_k \mathbf{s}_k^T}{\mathbf{y}_k^T \mathbf{s}_k} \right) + \frac{\mathbf{s}_k \mathbf{s}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}. \quad (11.47)$$

From this  $\hat{\mathbf{B}}$ , the BFGS search vector is defined as

$$\mathbf{p}_k = -\hat{\mathbf{B}}_k \mathbf{g}_k; \quad (11.48)$$

(compare to the Newton search vector defined in eq. (11.40)).

### Practical Implementation

Because we only require the product of  $\hat{\mathbf{B}}$  with the gradient (and not  $\mathbf{B}$  or  $\hat{\mathbf{B}}$  *per se*), effective matrix/vector products have been developed to minimize storage requirements by using low-rank QN updates. This requires  $\mathcal{O}(n)$  memory to store the successive pairs of update vectors ( $\mathbf{s}_k$  and  $\mathbf{y}_k$ ) and the respective inner products  $\mathbf{y}_k^T \mathbf{s}_k$ .

Limited-memory QN methods reduce storage requirements further by only retaining the  $\{\mathbf{s}, \mathbf{y}\}$  pairs from the previous few iterates (3–7). The identity matrix,  $\mathbf{I}$ , or a multiple of it, is typically used for the initial Hessian approximation  $\mathbf{B}_0$ . Updating this scaling at each iteration enhances overall efficiency [456, 774].

The limited-memory BFGS code of Nocedal and co-workers [916] is one of the most effective methods in this class. The combination of modest memory, requiring only gradient information, and good performance in practice makes it an excellent choice for large-scale multivariate minimization [891]. The method has been extended to constrained optimization [192, 193, 1450], used to propose preconditioners for CG methods [876], and combined with TN methods in a QN/TN cyclic fashion [194]. The text of Nocedal and Wright [918] presents a comprehensive description of the limited-memory BFGS method.

### 11.5.2 Conjugate Gradient (CG)

Nonlinear CG methods form another popular type of optimization scheme for large-scale problems where memory and computational performance are important considerations. These methods were first developed in the 1960s by combining the linear CG method (an iterative technique for solving linear systems  $\mathbf{Ax} = \mathbf{b}$  where  $\mathbf{A}$  is an  $n \times n$  matrix [467]) with line-search techniques. The basic idea is that if  $f$  were a convex quadratic function, the resulting nonlinear CG method would reduce to solving the Newton equations (eq. (11.38)) for the search vector  $\mathbf{p}$  when  $\mathbf{H}$  is a constant positive-definite matrix.

#### CG Search Vector

In each step of the nonlinear CG method, a search vector  $\mathbf{p}_k$  is defined by a recursive formula. A line search is then used as outlined in Algorithm [A1]. The iteration process that defines the search vectors  $\{\mathbf{p}_k\}$  is given by:

$$\mathbf{p}_{k+1} = -\mathbf{g}_{k+1} + \beta_{k+1} \mathbf{p}_k, \quad (11.49)$$

where  $\mathbf{p}_0 = -\mathbf{g}_0$ . The scheme-dependent parameter  $\beta_k$  that defines the search vectors is chosen so that if  $f$  were a convex quadratic and the line search exact (i.e.,  $\mathbf{x}_k + \lambda_k \mathbf{p}_k$  minimizes  $f$  exactly along  $\mathbf{p}_k$ ), then the *linear* CG process would result. The reduction to the linear CG method in this special case is important because linear CG is known to terminate in at most  $n$  steps of exact arithmetic. This finite-termination property relies on the fundamental notion that two sets of vectors ( $\{\mathbf{g}\}$  and  $\{\mathbf{p}\}$ ) generated in the CG method satisfy

$$\mathbf{g}_k^T \mathbf{p}_j = 0 \quad \text{for all } j < k.$$

This orthogonality condition implies that the search vectors span the entire  $n$ -dimensional space after  $n$  steps, so that  $\mathbf{g}_{n+1} = 0$  in finite arithmetic.

#### CG Variants

Different formulas for  $\beta_k$  (not to be confused with the line search parameter introduced earlier) have been developed for the nonlinear CG case, though they all reduce to the same expressions for convex quadratic functions. These variants exhibit different behavior in practice.

Three of the best known algorithms are due to Fletcher-Reeves (FR), Polak-Ribi  re (PR), and Hestenes-Stiefel (HS). They are defined by the parameter  $\beta$  (for eq. (11.49)) as:

$$\beta_{k+1}^{\text{FR}} = \mathbf{g}_{k+1}^T \mathbf{g}_{k+1} / \mathbf{g}_k^T \mathbf{g}_k, \quad (11.50)$$

$$\beta_{k+1}^{\text{PR}} = \mathbf{g}_{k+1}^T \mathbf{y}_k / \mathbf{g}_k^T \mathbf{g}_k, \quad (11.51)$$

$$\beta_{k+1}^{\text{HS}} = \mathbf{g}_{k+1}^T \mathbf{y}_k / \mathbf{P}_k^T \mathbf{y}_k. \quad (11.52)$$

(Recall  $\mathbf{y}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$ ).

The PR version is often found in software packages. Still, to be effective PR restarts the iteration process, i.e., sets  $\beta_k$  to zero occasionally, for example when  $\beta_k$  becomes negative.

Some important modifications of this version are due to Powell [1012], available in the IMSL library, and to Shanno & Phua [1168], available in the NAG library. These modifications have slightly more memory requirements but fewer function evaluations. An interesting CG-PR-FR hybrid algorithm might also be an effective alternative [457].

A careful line search is important for nonlinear CG methods.

### CG/QN Connection

Key connections between CG and QN-Newton algorithms for minimization began to emerge in the late 1970s. Essentially, it was found that the CG conjugacy property can be closely related to the QN condition, and thus an appropriate formula for  $\beta_k$  could be obtained from both viewpoints.

The many developments in the 1980s have shown that the limited-memory QN class of algorithms balances the extremely modest storage requirements of nonlinear CG with good convergence properties in practice. The fact that the unit steplength in QN methods is often acceptable leads to greater efficiency in terms of function evaluations and hence less computational time overall.

Still, the linear and nonlinear CG methods play important theoretical roles in the numerical analysis literature, as well as practical roles in many numerical techniques; see the research monograph of [1] for a modern perspective. The linear CG method, in particular, proves ideal for solving the linear subproblem in the truncated Newton method for minimization (discussed next), especially with convergence-accelerating techniques known as *preconditioning*.

### Recent CG Advances

Interesting recent work on nonlinear CG methods has produced several variants of the conjugate gradient formulas for computing the search directions for unconstrained optimization. Such directions possess favorable properties, such as guaranteed directions of descent, some without line searches, and global convergence under mild assumptions. See [490, 1358, 1424, 1438] for example.

Preliminary numerical tests show that the new variants perform well compared to other approaches, but a consensus has thus far not been reached, and research continues on this active front.

### 11.5.3 Truncated-Newton (TN)

#### Approximate Solution of the Newton Equations

In the early 1980s a very simple but important idea emerged in connection with the Newton equations: why solve this linear system for the search vector  $\mathbf{p}_k$  exactly [295]? In the context of large-scale nonlinear optimization, an accurate solution of eq. (11.38) is not warranted! Far away from a minimum, any descent direction that can be computed cheaply may still produce progress toward a minimum. Only near the solution, where the quadratic model is good, should the system be solved more accurately.

In practice, truncated-Newton (TN) methods allow a *nonzero residual*,  $r_k$ , for the Newton equations. For example, we can require

$$r_k \equiv \|\mathbf{H}_k \mathbf{p}_k + \mathbf{g}_k\| \leq \eta_k \|\mathbf{g}_k\|, \quad (11.53)$$

where  $\eta_k$  is the *forcing sequence*.

This condition on the size of the residual  $r_k$  at step  $k$  of the minimization scheme becomes stricter as the gradient norm becomes smaller. Thus, near the solution we solve for  $\mathbf{p}_k$  more accurately, whereas far away we permit a cruder approximation.

Theoretical work further showed that asymptotic quadratic convergence of the method can be realized for a well chosen  $\eta_k$  as  $\|\mathbf{g}_k\| \rightarrow 0$  [295]. For example, an effective setting is:

$$\eta_k = \min \{c_r/k, \|\mathbf{g}_k\|\}, \quad 0 < c_r \leq 1. \quad (11.54)$$

This choice forces the residuals to be progressively smaller as the number of iterations ( $k$ ) increases and as the gradient becomes smaller. Another termination criterion based on the quality of the quadratic approximation has also been suggested [891].

#### Truncated Outer Iteration; Effective Residual

To implement an upper bound on the residual norm in practice, an iterative, rather than direct, procedure that can be “truncated” is required for approximating  $\mathbf{p}_k$  from eq. (11.38) at each outer step  $k$ .

The linear CG method is an excellent candidate since it is simple and very modest in memory. The linear CG algorithm mirrors in structure the general descent method of Algorithm [A1]. That is, it generates search vectors  $\{\mathbf{p}_k^1, \mathbf{p}_k^2, \dots\}$  at each step recursively (as the nonlinear conjugate gradient method of the previous subsection) until the residual (eq. 11.54), or another suitable truncation criterion, is satisfied for the  $j$ th iterate  $\mathbf{p}_k^j$ . However, in place of the line search, an explicit formula for the steplength is used. This expression is derived analytically

by minimizing the quadratic model at the current point along  $\mathbf{p}_k^j$  and then using the conjugacy condition to simplify the formula.

### Preconditioning

To accelerate convergence of this inner iteration process, *preconditioning* is essential in practice. This technique involves modification of eq. (11.38) through application of a closely-related matrix to  $\mathbf{H}_k$ ,  $\mathbf{M}_k$  (effectively, multiplication of both sides by the inverse of  $\mathbf{M}_k$ ).

The preconditioner  $\mathbf{M}$  is typically chosen as a sparse symmetric matrix that is rapid to assemble and factor. Theoretically, convergence improves if  $\mathbf{M}_k^{-1}\mathbf{H}_k$ , the coefficient matrix of the new linear system, has clustered eigenvalues or approximates the identity matrix.

The TN code in CHARMM [302, 1397] uses a preconditioner from the local chemical interactions (bond length, bond angle, and dihedral-angle terms). This sparse matrix is rapid to compute and was found to be effective in practice, whether the matrix is indefinite or not, with an appropriate (unusual) modified Cholesky factorization [1397]. Other possibilities of preconditioners in general contexts have also been developed, such as a matrix derived from the BFGS update (defined in the QN subsection) [876].

### Overall Work

Although more complex to implement than QN or nonlinear CG methods, TN algorithms can be very efficient overall in terms of total function and gradient evaluations, convergence behavior, and solution accuracy, as long as the many components of the algorithm are carefully formulated (truncation, solution process for the inner loop, preconditioning, etc.).

In terms of the computational work per outer Newton step ( $k$ ), TN methods based on preconditioned CG require a Hessian/vector product ( $\mathbf{H}\mathbf{p}$ ) at each inner loop iteration, and one solution of a linear system  $\mathbf{M}\mathbf{z} = \mathbf{r}$  where  $\mathbf{M}$  is the preconditioner. Because  $\mathbf{M}$  may be sparse, this linear solution often takes a very small percentage of the total CPU time (e.g., <3% [1397]). The benefits of faster convergence generally far outweigh these additional costs associated with the preconditioner.

### Hessian/Vector Products

The Hessian/vector products in each linear CG step ( $\mathbf{H}_k \mathbf{p}_k^j$ ) are more significant in terms of computer time. For a Hessian formulated with a nonbonded cutoff radius (e.g., 8 Å), many zeros result for the Hessian (see Figures 11.4 and 11.5); when this sparsity is exploited in the multiplication routine, performance is fast compared to a dense matrix/vector product. However, when the Hessian is dense and large in size, the following forward-difference formula of *two gradients* often works faster (we omit subscripts  $k$  from  $\mathbf{H}$ ,  $\mathbf{p}$ , and  $\mathbf{x}$  for clarity):

$$\mathbf{H}\mathbf{p} \approx [\mathbf{g}(\mathbf{x} + h\mathbf{p}) - \mathbf{g}(\mathbf{x})] / h, \quad (11.55)$$

where  $h$  is a suitably-chosen small number. The central difference approximation,

$$\mathbf{H}\mathbf{p} \approx [\mathbf{g}(\mathbf{x} + h\mathbf{p}) - \mathbf{g}(\mathbf{x} - h\mathbf{p})] / 2h, \quad (11.56)$$

may alternatively be used for greater accuracy at the cost of one more gradient evaluation with respect to the one-sided difference formula.

In either case, finding an appropriate value for the finite-difference stepsize  $h$  is nontrivial, and the accuracy of the product near the solution (where the gradient components are small) can be problematic.

### Performance

Thus, TN methods require more care in implementation details and user interface, but their performance is typically at least as good overall as limited-memory QN Newton methods. If simplicity is at a premium, the latter is a better choice. If partial second-derivative information is available, the objective function has many quadratic-like regions, and the user is interested in repeated minimization applications, TN algorithms may be worth the effort.

In general, though Newton methods may not always perform best in terms of function calls and CPU time, they are the most reliable of methods for multivariate minimization and have the greatest potential for achieving very small final-gradient norms. This can be especially important if normal-mode analysis is performed following minimization.

#### 11.5.4 Simple Example

To illustrate performance of the methods described in this section, we have constructed a nonlinear minimization problem with an objective function dependent on two variables, whose contour lines are shown in Figure 11.10. Though the original problem has more variables, this construct represents a ‘slice’ of the real problem. Illustrations on more realistic, multivariate functions (potential functions of molecular systems) are presented in the next section.

The two-variable problem is derived from our charge optimization procedure [106] that determines electrostatic charge parameters for particles distributed on a virtual surface enclosing a macromolecular system; the electrostatic energy is modeled by a Debye-Hückel potential as an approximation (in the far zone) to a continuum, Poisson-Boltzmann solution to the electrostatic field surrounding the system. The objective function thus reflects the error in electric field (or potential) between the discrete and continuum approximations to the electrostatic potential of the complex macromolecular system.

Specifically, our constructed two-dimensional example seeks to optimize two charge values on the surface of the nucleosome, with the remaining 275 charges fixed.

The contour plots of our function, most readily seen from the darker illustration in Figure 11.10 (bottom right), show the unique minimum lying inside a shallow valley in the function surface.

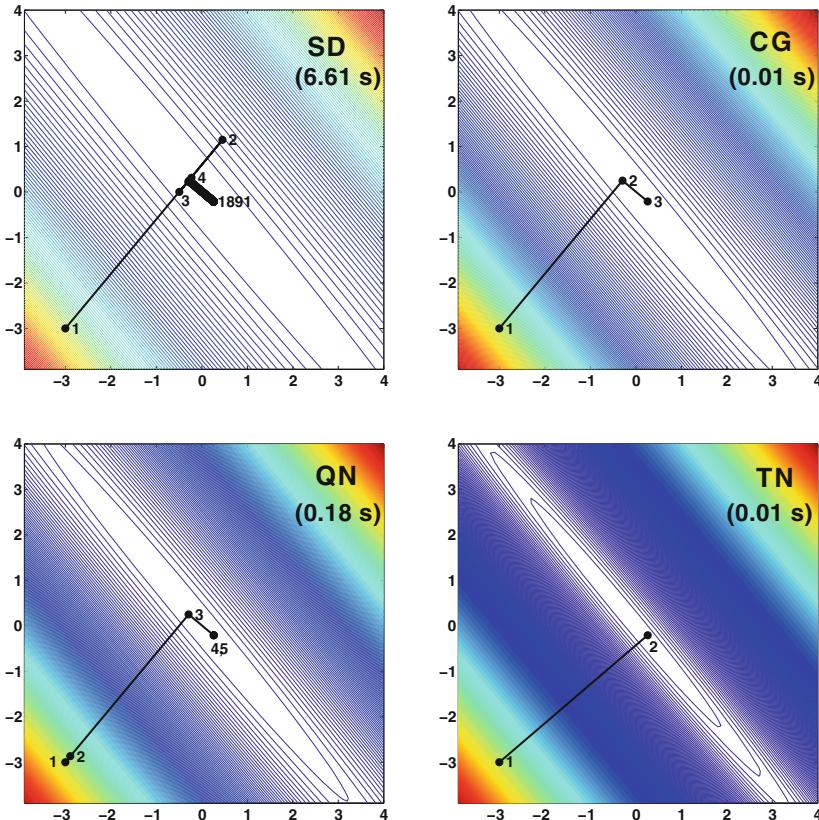


Figure 11.10. Minimization paths (with corresponding CPU times) for a function of two variables shown on top of function contour plots, for steepest descent (SD), nonlinear conjugate gradient (CG), BFGS quasi-Newton (QN), and truncated-Newton (TN) algorithms. See text for functional construction details. All contours are the same, but different levels of resolution are used in each plot to discern both the region near the minimum (darkest contour plot) and the higher-energy regions (lighter plots).

The steepest descent (SD) path (top left) first overshoots the minimum and then slowly approaches it, reaching the high desired gradient accuracy of order  $10^{-12}$  after nearly 2000 iterations. It also requires two orders of magnitude more CPU time than the other methods.

The CG path (top right) is direct and efficient for this problem, likely because of the quadratic nature of the function. Equally direct and efficient are the Newton minimizers: QN BFGS in Matlab and the TN package TNPACK (bottom illustrations). TNPACK, in particular, achieves a low gradient norm in one step.

## 11.6 Available Software

Table 11.1 summarizes the available minimizers in several chemistry and mathematics packages. See [877] and the NEOS (Network-Enabled Optimization System) Guide at [neos.mcs.anl.gov/](http://neos.mcs.anl.gov/) for a compilation of up-to-date mathematical software for optimization and related information.

Table 11.1. Available optimization algorithms.

Package	Minimizers
AMBER	SD, nonlinear CG from the IMSL library (due to Powell), and Newton.
CHARMM	SD, nonlinear CG (FR, and modified PR version, the latter from the IMSL library). <sup>a</sup> Adopted-Basis Newton (ABNR), Newton, truncated-Newton (TNPACK).
DISCOVER /BIOSYM	SD, nonlinear CG (PR, FR versions), <sup>a</sup> quasi-Newton, truncated Newton.
DUPLEX	Powell's coordinate descent method (no derivatives).
ECEPP/2	Calls SUMSL, a quasi-Newton method based on a trust-region approach (by Gay).
GROMACS	SD, nonlinear CG (FR version), <sup>a</sup> both with and without SHAKE constraints, limited-memory BFGS.
IMSL Lib.	Many routines for constrained and unconstrained minimization (nonsmooth, no derivatives, quadratic and linear programming, least-squares, nonlinear, etc.), including a nonlinear CG method of Powell (modified PR version with restarts). <sup>a</sup>
LANCELOT	Various Newton methods for constrained and unconstrained nonlinear optimization, specializing in large-scale problems and including a trust-region Newton method and an algorithm for nonlinear least squares that exploits partial separability.
MATLAB	SD, DFP <sup>b</sup> and BFGS quasi-Newton, simplex algorithm, and others for linear and quadratic programming, least squares, etc.
MMFF94 /94S	Calls OPTIMOL which uses a BFGS quasi-Newton method, with variable-metric updating scheme, but for Cartesian optimization (there is also a torsion-only optimizer) initiates the initial inverse Hessian approximated from the inverse of a $3 \times 3$ block-diagonal Hessian.
MM3	$3 \times 3$ block-diagonal Newton and full Newton.
NAG Lib.	Quasi-Newton, modified Newton and nonlinear CG (CONMIN by Shanno & Phua, modified PR version); also quadratic programming, least squares minimization, and many service routines.
NAMD	Nonlinear CG.
Tinker	Low-storage BFGS method.
X-PLOR	Nonlinear CG (from IMSL library).

<sup>a</sup>FR and PR refer to the Fletcher-Reeves and Polak-Ribi  re nonlinear CG versions.

<sup>b</sup>DFP is a rank-1 QN method, credited to Davidon, Fletcher, and Powell.

### 11.6.1 Popular Newton and CG

Nonlinear CG and various Newton methods are quite popular, but algorithmic details and parameters vary greatly from package to package. In particular, nonlinear CG implementations are quite different. Several comprehensive mathematical libraries, such as IMSL, NAG, and MATLAB are sources of quality numerical software.

### 11.6.2 CHARMM’s ABNR

Of special note is the “adopted-basis Newton-Raphson” method implemented in CHARMM, ABNR. It is a memory-saving adaptation of Newton’s method that avoids analytic second derivatives. The idea is to use SD steps for a given number of iterations,  $m$  (e.g., 5), after which a set of  $m+1$  coordinate and gradient vectors are available. A Hessian is constructed numerically in this  $m \times m$  subspace, and all corresponding eigenvalues and eigenvectors are computed. If all eigenvalues are negative, SD steps are used; if some are negative and some are positive, the search direction is modified by a Newton direction constructed from the eigenvectors corresponding to the positive eigenvalues only. In all cases, the  $n$ -dimensional search vector  $\mathbf{p}_k$  is determined via projection onto the full space. The ABNR algorithm is similar in strategy to limited-memory QN methods in that it uses only recent curvature information and exploits this information to make steady progress toward a solution.

### 11.6.3 CHARMM’s TN

The TN method in CHARMM [302] is detailed elsewhere [1121, 1122, 1130, 1397, 1398]. It uses a preconditioner constructed from the local chemical interactions (see Figures 11.4 and 11.5, right panels) and determines  $\mathbf{p}_k$  from a truncated preconditioned CG loop. When negative curvature is detected, the preconditioned CG loop is halted with a guaranteed direction of descent. Interestingly, numerical analysis and experiments have shown that the method can produce quadratic convergence near a solution regardless of whether the preconditioner is indefinite or not [1397]. As implemented, the method is applicable only to moderate system sizes (due to Hessian memory limitations).

### 11.6.4 Comparative Performance on Molecular Systems

In Table 11.2 we illustrate the minimization performance of three methods in CHARMM — nonlinear CG, ABNR, and TNPACK — for several molecular systems; see [1397, 1398] for details.

Note that the same minimum is obtained for the small systems (butane and *n*-methyl-alanyl-acetamide) but that different minima typically result for the larger systems.

Considerable differences in CPU times can also be noted. The CG method performs much slower and can fail to produce very small gradient norms. Both Newton methods perform well for these problems, though ABNR is relatively expensive for the small system. TNPACK displays much faster convergence overall and yields smaller final gradient norms.

Note also that CG requires about two function evaluations per iteration (in the line search), while ABNR employs only one on average. TNPACK uses more than one function evaluation per outer iteration, since the (unscaled) magnitude of the produced search vector often leads to small steplengths at some

iterations of the line search. The quadratic convergence of TNPACK is evident from Figure 11.11, where the gradient norm per iteration is shown.

## 11.7 Practical Recommendations

In general, geometry optimization in the context of molecular potential energy functions has many possible caveats. Hence, a novice user especially should take the following precautions to generate as much confidence as possible in a minimization result.

1. *Use many starting points.* There is always the possibility that the method will fail to converge from a certain starting point, or converge to a nearby stationary point that is not a minimum.

A case in point is minimization of biphenyl from a planar geometry [773]; many minimizers will produce the flat ring geometry, but this actually corresponds to a maximum! Different starting points will produce the correct nonplanar structure. See the homework assignment on minimization (number 10).

2. *Compare results from different algorithms.* Many packages offer more than one minimizer, and thus experimenting with more than one algorithm is an excellent way to check a computational result. Often, one method fails to achieve the desired resolution or converges very slowly. Another reference calculation under the same potential energy surface should help assess the results.

Minimization of the DNA in vacuum system in Figure 11.11 and Table 11.2 by three algorithms also reveals very different final energies. This is because the DNA strands have separated! Proper solvation and ions remedy this physical/chemical problem. Interestingly, adding only water keeps the strands nearby but untwists the strands; only added ions and water maintain the proper DNA chemistry.

3. *Compare results from different force fields whenever possible.* Putting aside the quality of the minimizer, the local minimum produced by any package is only as good as the force field itself. Since force fields for macromolecules today are far from converging to one another — in fact there are very large differences both in parameters and in functional forms — a better understanding of the energetic properties of various conformations can be obtained by comparing the relative energies of the different configurations as obtained by different force fields. Differences are expected, but the results should help identify the lowest-energy configuration. If significant differences are observed, the researcher could further investigate both the associated force fields (e.g., a larger partial charge, an additional torsional term) and the minimization algorithms for explanations.

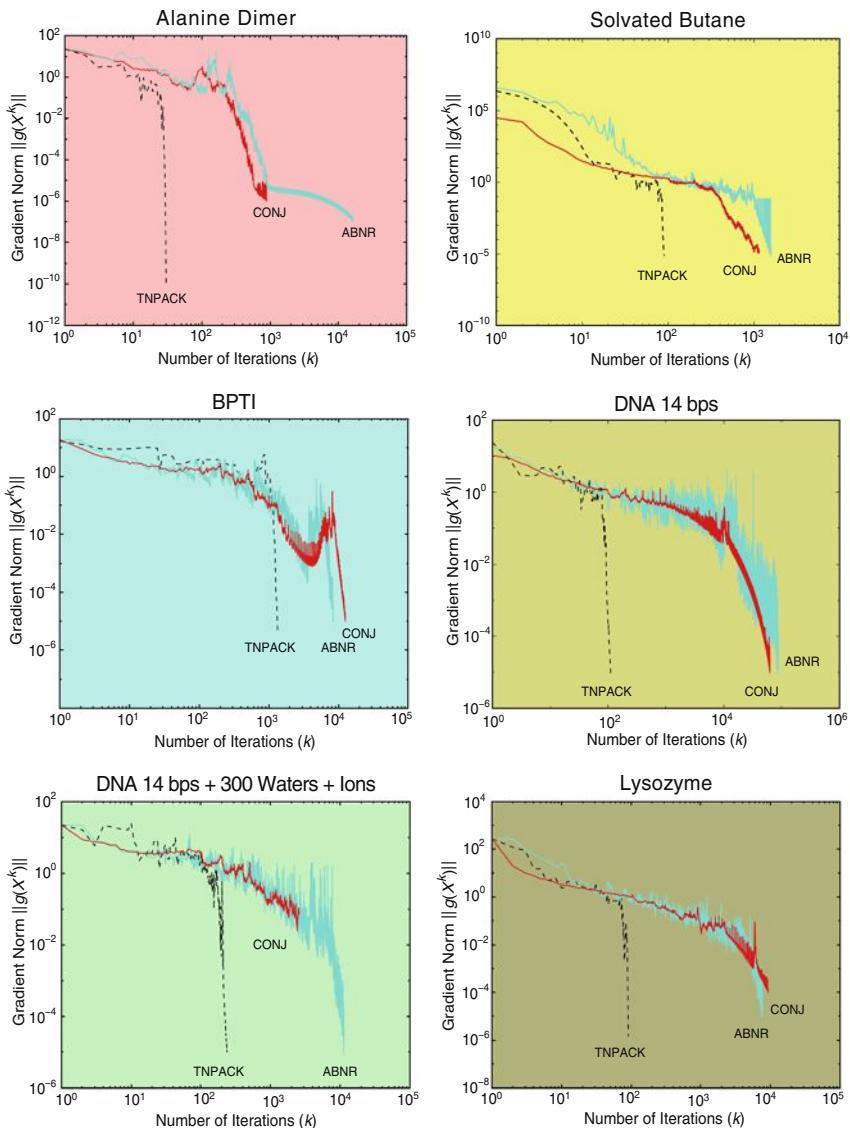


Figure 11.11. Minimization progress (gradient norm) of three CHARMM algorithms (CONJ: nonlinear conjugate gradient, ABNR: adopted-basis Newton-Raphson, and TNPACK: truncated Newton) for various molecular systems. See Figures 11.4 and 11.5 for corresponding sparse matrix patterns; the local Hessians are used as preconditioners for TNPACK. See also Table 11.2 for final function and gradient-norm values and required CPU time.

Table 11.2. Performance of three CHARMM minimizers on various molecular systems. See Figures 11.4 and 11.5 for patterns of the preconditioner used in TNPACK and Figure 11.11 for minimization progress. *Note:* For the DNA system in vacuum, though minimization produces a local minimum for each method, the structures are physically incorrect: without proper solvation, the DNA strands intertwine and separate. This also explains the very different values of final energies. For the DNA system with water and ions, the CG method terminates prematurely with an error message; the final gradient is relatively large.

Method <sup>a</sup>	Final $f$	Final $\ g\ $	Itns. <sup>b</sup>	$f \& g$ Evals	CPU <sup>c</sup>
<b>N-Methyl-Alanyl-Acetamide (<math>n = 66</math>)</b>					
CG	-15.245	$9.83 \times 10^{-7}$	882	2507	2.34 s
ABNR	-15.245	$9.96 \times 10^{-8}$	16466	16467	7.47 s
TNPACK	-15.245	$7.67 \times 10^{-11}$	29 (210)	44	1.32 s
<b>Solvated Butane (<math>n = 1125</math>)</b>					
CG	-2374.00	$1.27 \times 10^{-5}$	1152	3175	49.48 m
ABNR	-2398.22	$7.0 \times 10^{-6}$	1574	1575	48.52 m
TNPACK	-2381.04	$7.7 \times 10^{-6}$	90 (1717)	263	59.44 m
<b>BPTI (<math>n = 1704</math>)</b>					
CG	-2792.93	$9.9 \times 10^{-6}$	12469	32661	97.8 m
ABNR	-2792.96	$8.9 \times 10^{-6}$	8329	8330	25.17 m
TNPACK	-2773.70	$4.2 \times 10^{-6}$	65 (1335)	240	5.21 m
<b>DNA 14 Bps (<math>n = 2664</math>)</b>					
CG	-538.41	$9.72 \times 10^{-6}$	62669	62670	20.42 h
ABNR	-1633.90	$9.23 \times 10^{-6}$	86496	86497	6.69 h
TNPACK	-560.68	$7.62 \times 10^{-6}$	111 (3724)	268	0.54 h
<b>DNA 14 Bps + 300 Waters + Ions (<math>n = 5364</math>)</b>					
CG	-11921.00	$7.52 \times 10^{-2}$	2580	6616	1.62 h
ABNR	-11774.64	$8.19 \times 10^{-6}$	11306	11307	2.78 h
TNPACK	-11928.50	$9.84 \times 10^{-6}$	236 (6555)	687	1.75 h
<b>Lysozyme (<math>n = 6090</math>)</b>					
CG	-4628.362	$9.89 \times 10^{-5}$	9231	24064	19.63 h
ABNR	-4631.584	$9.97 \times 10^{-6}$	7637	7638	6.11 h
TNPACK	-4631.380	$1.45 \times 10^{-6}$	78 (1848)	218	1.49 h

<sup>a</sup>CG: nonlinear conjugate gradient, ABNR: adopted basis Newton Raphson, TNPACK: truncated Newton based on the TNPACK package.

<sup>b</sup>For TNPACK, the total number of inner (preconditioned CG) iterations is indicated in parentheses, following the number of outer iterations.

<sup>c</sup>s: seconds, m: minutes, h: hours.

- Check eigenvalues at the solution when possible. If the significance of the computed minima is unclear, the corresponding eigenvalues may help diagnose a problem. Near a true minimum, the eigenvalues should all be positive (except for the six zero components corresponding to translation and

rotation invariance). In finite-precision arithmetic, “zero” will correspond to numbers that are small in absolute value (e.g.,  $10^{-6}$ ). Values larger than this tolerance might indicate deviations from a true minimum, perhaps even a maximum or saddle point. In this case, the corresponding structure should be perturbed substantially and another trial of minimization attempted.

5. *Be aware of artificial minima caused by nonbonded cutoffs or improper physical models!* When cutoffs are used for the nonbonded interactions, especially in naive implementations involving sudden truncation or potential-switching methods, the energy and/or gradient can exhibit numerical artifacts: deep energy minima and correspondingly-large gradient value near the cutoff region. Good minimization algorithms can find these minima, which are correct as far as the numerical formulation is involved, but unfortunately not relevant physically.

One way to recognize these artifacts is to note their large energy difference with respect to other minima computed for the same structure (as obtained from different starting points or minima). These artificial minima should disappear when all the nonbonded interactions are considered, or improved spherical-cutoff treatments (such as force shifting and switching methods) are implemented instead.

Besides artifacts caused by nonbonded cutoffs, improper physical models — such as that for DNA lacking solvent and ions, as discussed above — also produce artificial minima. For this example of DNA in vacuum, parallel strands rather than intertwined polynucleotide strands are produced as a result of minimization.

## 11.8 Future Outlook

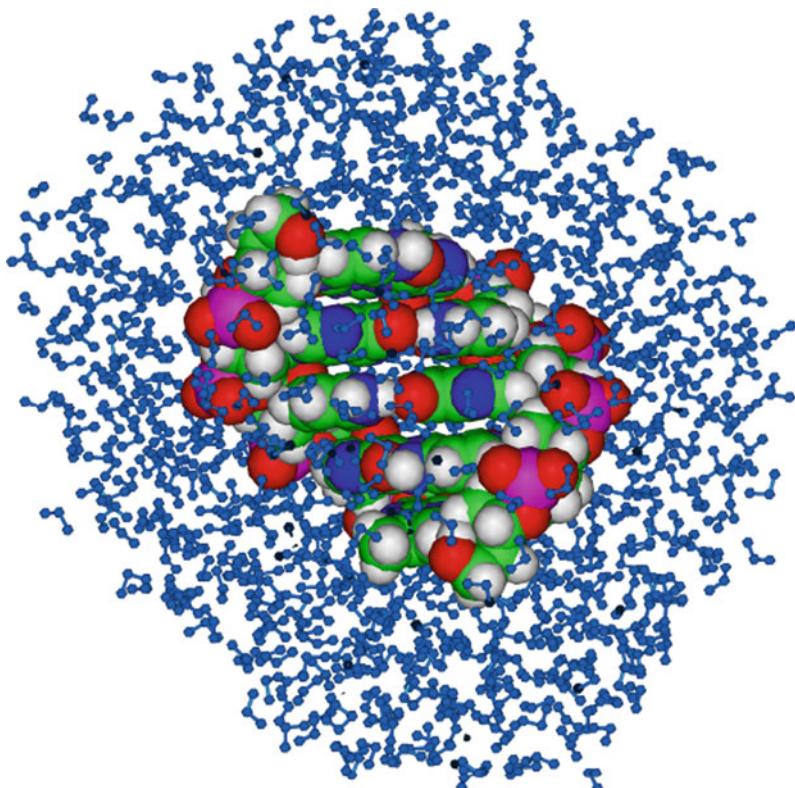
Only a small subset of topics was covered here in the challenging and ever-evolving field of nonlinear large-scale optimization. Interested readers are referred to the comprehensive treatments in the texts cited at the beginning of this chapter. The increase in computer memory and speed, and the growing availability of parallel computing platforms will undoubtedly influence the development of optimization algorithms in the next decade. Parallel architectures can be exploited in many ways: for performing minimization simulations concurrently from different starting points; for evaluating function and derivatives in tandem; for greater efficiency in the line search or finite-difference approximations; or for performing matrix decompositions in parallel for structured, separable systems. In addition, advances in conformational sampling methods (e.g., as reviewed in [1116, 1117]) also address minimization problems in practice.

The increase in computing speed is also making *automatic differentiation* a powerful resource for nonlinear optimization. In this technique, automatic routines are available to construct program codes for function derivatives. The

construction is based on the chain-rule application to the elementary constituents of a function [482]. It is foreseeable that such codes will introduce greater versatility in Newton methods [916]. The cost of differentiation is not reduced, but the convenience and accuracy may increase.

Function separability is a more general notion than sparsity, since problems associated with sparse Hessians are separable but the reverse is not true. It is also another area where algorithmic growth can be expected [916]. (Recall that separable functions are composites of subfunctions, each of which depends only on a small subset of the independent variables; see eq. (11.6)). Therefore, efficient schemes can be devised in this case to compute the search vector, function curvature, etc., much more cheaply by exploiting the invariant subspaces of the objective function.

Such advances in local optimization will certainly lead to further progress in solving the global optimization problem as well; see [196, 411, 958, 1000] for examples. Scientists from all disciplines will anxiously await all these developments.



# 12

## Monte Carlo Techniques

### Chapter 12 Notation

SYMBOL	DEFINITION
<b>Matrices</b>	
<b>M</b>	mass matrix (components $m_i$ )
<b>Vectors</b>	
$p$ (or $P$ )	collective momentum vector
$q$ (or $X$ )	collective position vector
$R$	collective random force vector
<b>Scalars &amp; Functions</b>	
$a$	multiplier (of random number generator, a prime)
$c$	increment (of random number generator)
$i, j, k, m, q, r$	integers
$t$	time
$u, v$	real numbers
$v_i$	velocity component $i$
$\{x_i\}, \{\tilde{x}_i\}$	sequences of numbers
$B_i$	data batch (for MC mean)
$E_k$	kinetic energy
$E_p$	potential energy
$F$	probability distribution function
$H$ (or $E$ )	total energy
$M$	modulus (of random number generator, usually of order of computer word size); also used for number of batches in a sample
$N$	number of particles
$N$	sample size
$R_g$	radius of gyration

Chapter 12 Notation Table (continued)

SYMBOL	DEFINITION
$S, T$	polynomial functions
$T$	temperature
$Wr$	DNA writhing number
$\beta$	Boltzmann factor $1/(k_B T)$
$\gamma$	Langevin damping constant
$\mu$	mean (also chemical potential in MC Carlo sampling section)
$\rho$	probability density function
$\sigma^2$	variance ( $\sigma$ is standard deviation)
$\tau$	period (of random number generator)
$\langle A(x) \rangle_{\mathcal{D}}$	mean of property $A(x)$ over domain $\mathcal{D}$
$\lfloor y \rfloor$	largest integer smaller than or equal to $y$

It is a pollster's maxim that the truth lies not in any one poll but at the center of gravity of several polls.

Michael R. Kagay, *New York Times (Week in Review)*, 19 October 1998.

It is indeed true that the stock market can forecast the business cycle.  
The stock market has called nine of the last five recessions.

Paul A. Samuelson, first American Nobel laureate in economics (1915–2009).

## 12.1 MC Popularity

From Washington D.C. to Wall Street to Los Alamos, statistical techniques termed collectively as Monte Carlo (MC) are powerful problem solvers. Indeed, disciplines as disparate as politics, economics, biology, and high-energy physics rely on MC tools for handling daily tasks.

Many problems that can be formulated as stochastic phenomena and studied by random sampling can be solved through MC simulations. Essentially, a game of chance is played, but with theoretical and practical rules from probability theory, stochastic processes, and statistical physics (Markov chains, Brownian motion, ergodic hypothesis) that lend the ‘sport’ practical utility.

### 12.1.1 A Winning Combination

MC methods are used for numerical integration, global optimization, queuing theory, structural mechanics, and solution of large systems of linear, partial differential or integral equations. MC methods are employed widely in statistical physics and chemistry, where the behavior of complex systems of thousands or more atoms in space and time is studied. Their appeal can be explained by a winning combination of simplicity, efficiency, and theoretical grounding.

### 12.1.2 From Needles to Bombs

Early records of random sampling to solve quantitative problems can be found in the 18th and 19th centuries with needle throwing experiments to calculate geometrical probabilities (George Louis Leclerc, a.k.a. Comte de Buffon, 1777)<sup>1</sup> or to determine  $\pi$  (Simon de Laplace<sup>1</sup>). In 1901, Lord Kelvin also described an important application to the evaluation of time integrals in the kinetic theory of gases. Yet a novel class of MC methods (using Markov chains) provides the modern roots of MC theory, and is largely credited to the Los Alamos pioneers (Von Neumann, Fermi, Ulam, Metropolis, Teller, and others).

These brilliant scholars studied properties of the newly discovered neutron particles in the middle of the 20th century by formulating mathematical problems in terms of probability and solving analogues by stochastic sampling. Their work led to a surge of publications in the late 1940s and early 1950s on solving problems in statistical mechanics, radiation transport, and other fields by carefully-designed sampling experiments.

Most notable among these works was the famous algorithm of Metropolis *et al.* in 1953 [856]. With the rapid growth of computer speed and the development of many techniques to improve sampling, reduce errors, and enhance efficiency, MC methods have become a powerful utility in many areas of science and engineering.

### 12.1.3 Chapter Overview

In this chapter, only the most elementary aspects of MC simulations are described, including the generation of uniform and normal random variables, basic probability theory background (see also Box 12.1), and the Metropolis algorithm (due also to A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller).

Such methods can be used in molecular simulations to generate efficiently conformational ensembles that obey *Boltzmann statistics*, that is, the probability of a configuration  $X$  with energy  $E(X)$  is proportional to  $\exp(-E(X)/k_B T)$  ( $k_B$  is Boltzmann's constant and  $T$  is the temperature). From such ensembles, various geometric and energetic means can be estimated. Low-energy regions can also be identified by decreasing the temperature in the sampling protocol (this is termed “simulated annealing”). Method extensions that are of particular interest to the biomolecular community, such as biased MC, hybrid MC, parallel tempering (also called replica-exchange method, REM), and other variants are also mentioned. The REM method has been particularly effective in the MD incarnation termed Replica Exchange MD (REMD); see MD chapters.

---

<sup>1</sup>Buffon used a Monte Carlo integration procedure to solve the following problem: a needle of length  $L$  is thrown at a horizontal plane ruled with parallel straight lines separated by  $d > L$ ; what is the probability that the needle will intersect one of these lines? Buffon derived the probability as an integral and attempted an experimental verification by throwing the needle many times and observing the fraction of needle/line intersections. It was Laplace who in the early 1800s generalized Buffon's probability problem and recognized it as a method for calculating  $\pi$ .

The codes illustrated in this chapter are provided in Fortran, still the language of choice in some of the popular molecular mechanics and dynamics packages like CHARMM and AMBER. Analogous routines written in the C language can be obtained from the course website.

Since MC methods rely strongly on random number generators, the first section of this chapter is devoted to the subject. Students can skip Sections 12.2 and 12.3 if they wish to read material related to other aspects of MC simulations. To see immediately why random number generators are important and how they are used in MC simulations, students may wish to read at the onset Subsection 12.5.4 and see the example there (subroutine `monte`) for calculating  $\pi$  by MC sampling.

Good general introductions to MC simulations can be found in the texts by Kalos and Whitlock [625], Bratley, Fox and Schrage [165], Frenkel and Smit [428], and Chandler [213]. There are many good general reviews such as [38] and numerous web resources for Monte Carlo tutorials (check Wikipedia), for example obtained through the Molecular Monte Carlo Home Page of [www.cooper.edu/engineering/chemechem/monte.html](http://www.cooper.edu/engineering/chemechem/monte.html). See reviews of MC methods for molecular systems in Section 12.6.

### 12.1.4 Importance of Error Bars

A point which cannot be overstressed in any introduction to MC methods is the fundamental importance of error bars in any MC estimate. Unlike in politics, perhaps, the reliability of any conclusion (e.g., estimate) in science depends on the associated accuracy. Scientists would no doubt have discarded the results of an election whose “margin of error . . . is far greater than the margin of victory, no matter who wins”, an assessment by mathematician John Allen Paulos of the rocky 2000 U.S. Presidential race, between Texas governor George W. Bush (who became President) and former President Clinton’s Vice President Albert Gore.

## 12.2 Random Number Generators

### 12.2.1 What is Random?

The computer sampling performed in MC simulations of stochastic processes relies on generation of “random” numbers. Actually, those numbers are typically *pseudorandom* since a deterministic recursion rule is used to generate a sequence of numbers given an *initial seed*  $x_0$ :  $x_{i+1} = f(x_i, x_{i-1}, x_{i-2}, \dots)$ , where  $f$  is a function.<sup>2</sup> This reproducibility of the sequence is an essential requirement for debugging computer programs. Even sequences obtained via chaos theory (see [469], for example, and references cited therein) are deterministic.

---

<sup>2</sup>We use the term *random* for brevity in most of this chapter, though the terms *pseudorandom* or *quasi-random* are technically correct.

It is essential to use ‘good’ random number generators in MC applications to avoid artifacts in the results. (The statement by Dilbert’s cartoon character, a horned accounting troll, that “you can never be sure” [of randomness] is not an option for scientists! See cartoon posting on the [dilbert.com](http://dilbert.com) archives for 10/25/01). The quality of a generator is determined not only by subjecting the generating algorithm to a large number of established tests (both empirical and theoretical). It is also important to test the *combination* of generator and application. The two examples described in the Artifacts subsection below (following the introduction of generator algorithms) illustrate how the performance of generators is application specific.

Much work has gone into developing random number generators on both serial and parallel computer platforms, as well as associated criteria for testing them (see [722] for example of software for testing generators). Concurrently, work has focused on the careful implementation of the mathematical expressions to ensure good numerical performance (e.g., avoid overflow or systematic loss in accuracy) and efficiency, on both general and special-purpose hardware.

Novices are well advised to use a routine from a reputable library of programs rather than programming a simple procedure reported in the literature, since many such procedures have not been actually tested comprehensively. Still, caveats are warranted even for some library routines; see below.

The reader is referred to classic texts by Kalos and Whitlock [625], Knuth [663], and Law and Kelton [706] for general introductions into random number generators. Some of these books also review basic probability theory. Good reviews by L’Ecuyer can be found in [718, 719, 721, 722] (see also [www.iro.umontreal.ca/~lecuyer](http://www.iro.umontreal.ca/~lecuyer) for updated works) and [837].

### 12.2.2 Properties of Generators

Let

$$\{x_1, x_2, \dots, \dots\}$$

be a sequence of numbers. In theory, we aim for sequences of numbers that exhibit independence, uniformity, and a long period  $\tau$ . In addition, it is important that such generators be as portable and efficient as possible.

#### Uniformity and Subtle Correlations

Most MC algorithms manipulate hypothetical *independent uniformly distributed* random variables (*variates*). That is, the independent variables are assumed to have a *probability density function*  $\rho$  (see Box 12.1) that satisfies  $\rho_u(x) = 1$  for  $x$  in the interval  $[0, 1]$  and  $\rho_u(x) = 0$  elsewhere.<sup>3</sup> From such uniform variates, we can obtain other probability distributions than the uniform distribution, such as the

---

<sup>3</sup>We say that  $x$  lies in  $[a, b]$  if  $a \leq x \leq b$  and that  $x$  lies in  $[a, b)$  if  $a \leq x < b$ ; similarly,  $x$  in  $(a, b)$  means  $a < x < b$ .

normal, exponential, Gamma, or Poisson distributions; see subsection below on normal variates and [706] for generating continuous and discrete random variates from many distributions.

Roughly speaking, independence of two random variables means that knowledge of one random variate reveals no information about the distribution of the other variate. In the strict sense of probability theory,<sup>4</sup> it is impossible to obtain true independence for random numbers. Generating *uncorrelated* random variates is a weaker goal than independence. (This is because independent random variables are uncorrelated but uncorrelated variables are not independent in general). Though correlations exist even in the best known random number generators, quality random number generators can defer correlations to high-order and high-complexity relations.

### Long Period

The period  $\tau$  associated with a sequence of random numbers is the number of sequential random values before the series repeats itself, that is,

$$x_{i+\tau} = x_i \quad \text{for all integers } i \geq 0.$$

We require the sequence to have as *long a period as possible* to allow long simulations of independent measures.

The period length is an important consideration for modern large-scale simulations.<sup>5</sup> For a 32-bit computer, if the generator's state uses only 32 bits, the maximum period is usually  $2^{30} \sim 10^9$  (assuming 2 bits are lost). This number is not a large number by today's standards. More than one million iterations may be performed in dynamics simulations and far more in MC sampling simulations. Moreover, each iteration may require large random *vectors* (e.g., in Langevin dynamics). Thus, the random number generators that might have been adequate only a decade ago on 32-bit machines quickly exhaust their values for the complex applications at present. Unfortunately, many such generators, which experts deem *unacceptable* [715], are often the default methods for many operating systems and software packages.

State-of-the-art generators use more bits for their state than the computer type and employ combinations of methods to defer correlations to high-order and high-complexity relations. This makes possible formulation of sequences with very long periods. For example, the codes given in [716] produce sequences with period lengths of up to order  $2^{200}$  on 32-bit computers and  $2^{400}$  on 64-bit machines!

---

<sup>4</sup>The random variables  $x_1$  and  $x_2$  are independent if the joint probability density function  $\rho(x_1, x_2)$  is equal to the product of the individual probability density functions:  $\rho(x_1, x_2) = \rho_1(x_1)\rho_2(x_2)$ .

<sup>5</sup>Though for complex systems, the state descriptors (e.g., coordinates) are unlikely to be repeated in phase with the cycle of a (short) random number generator, subtle problems may occur in some applications, making the goal of long period generally desirable.

## Portability

Portability and efficiency are also important criteria of generators.

*Portable generators* are those that produce the same sequence across standard compilers and machines, within machine accuracy. Portability permits code comparison and repeatability on different platforms. This requirement is nontrivial because even if the mathematical recipe is identical certain floating-point calculations may involve hardware-wired instructions and branched directives for the sub-operations.

## Efficiency

The issue of *speed* of random number generators can be important for some problems that involve a large number of computationally-intensive iterations. (See Table 12.1 for CPU data on different generators). Even if the relative computational cost of random number generators in large-scale applications is small, it is important to use quality compiler optimization utilities to reduce most of the overhead associated with *calling* the random number generator function itself. For this reason, it is also important to use a subroutine that returns a *vector* of random variates if an array of such numbers is desired, rather than calling the function multiple times for each vector component.

### Box 12.1: The Probability Density and Distribution Functions

Let  $X$  be a random variable that takes on values  $x$ . We say that  $X$  is a *discrete* random variable if it takes on a countable number of values and *continuous* if it takes on an uncountably-infinite number of values.

The distribution function  $F(x)$  (also termed the *cumulative distribution function*) of a random variable  $X$  defined as the probability that  $X$  takes on a values no larger than  $x$  (a real number), that is

$$F(x) = P(X \leq x), \quad -\infty < x < \infty. \quad (12.1)$$

A continuous random variable  $X$  has the closely related *probability density function*  $\rho(x)$ . (For discrete random variables, analogous definitions are formulated using a probability function  $p(x)$ ). This relation is given by:

$$F(x) = P(X \leq x) = \int_{-\infty}^x \rho(y) dy, \quad -\infty < x < \infty. \quad (12.2)$$

Thus,  $\rho(x)$  is closely related to the derivative of  $F(x)$  (under some additional assumptions of regularity, we have  $\rho(x) = F'(x)$ ).

For example, a uniform random variable on  $[0, 1]$  has the probability density function

$$\rho(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}, \quad (12.3)$$

and the corresponding density function  $F$  is defined by:

$$F(x) = \int_0^x \rho(y) dy = \int_0^1 1 dy = x. \quad (12.4)$$

The reader can verify that the mean  $\mu$  (or *expected value*) of this continuous uniform random variable (by definition,  $\mu \equiv E(X) = \int_{-\infty}^{\infty} x \rho(x) dx$ ) is  $\mu = \int_0^1 x \rho(x) dx = \frac{1}{2}$  and that the variance  $\sigma^2$  (by definition,  $\sigma^2 \equiv E(X - \mu)^2 = E(X^2) - \mu^2$ ) is  $\sigma^2 = \int_0^1 x^2 \rho(x) dx - (\frac{1}{2})^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$ .

A Gaussian (or *normal*) random variable with mean  $\mu$  and variance  $\sigma^2$  has the density function

$$\rho(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right), \quad -\infty < x < \infty. \quad (12.5)$$

The associated distribution function is often denoted as  $\mathcal{N}(\mu, \sigma^2)$ . The probability density function for a *standard normal* random variable (with  $\mathcal{N}(0, 1)$ ) is

$$\rho(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2), \quad -\infty < x < \infty. \quad (12.6)$$


---

### 12.2.3 Linear Congruential Generators (LCG)

The simplest type of random number method is a *linear congruential generator* (LCG), first used in 1948 by D. H. Lehmer.

#### Basic Recipe

LCGs compute successive iterates by multiplying the previous iterate by a constant,  $a$ , adding this product to another constant,  $c$ , and then taking the modulus of this result with respect to another large number,  $M$ .

Specifically, the LCG recipe relies on three integers.  $M$  (the *modulus*) is a large positive number;  $a$  (the *multiplier*) is a positive integer less than  $M$  and shares no divisors with  $M$ ; and  $c$  (the *increment*) is less than  $M$ .

We then generate a sequence of variates from an initial integer seed  $\tilde{x}_0$  less than  $M$ , namely  $\{\tilde{x}_1, \tilde{x}_2, \dots\}$ , according to the recursion relation:

$$\tilde{x}_{i+1} = (a\tilde{x}_i + c) \bmod M, \quad i = 0, 1, \dots. \quad (12.7)$$

The uniform variates for this LCG are then obtained by division as:

$$x_i = \tilde{x}_i/M.$$

If  $c = 0$ , these real numbers  $\{x_i\}$  are in the open unit interval  $(0, 1)$ , and if  $c \neq 0$  they are contained in the interval  $[0, 1]$ . When  $c = 0$ , this LCG is called *multiplicative linear congruential generator* (MLCG).

The recurrence relation of eq. (12.7) has a period no greater than  $M$ . If the integers are properly chosen, the period will have the maximal length  $M$ . Judicious choice of  $a$ ,  $c$ , and  $M$  must be made, as well as thorough tests for randomness of the resulting sequences. See [663, pp. 170–171] for specific recommendations.

### Simple Example

As a simple illustration, consider the MLCG sequence with  $M = 11$  and  $a = 8$  ( $c = 0$ ). From  $\tilde{x}_0 = 1$ , we generate the sequence

$$\begin{aligned}\tilde{x}_{i+1} &= (8\tilde{x}_i) \bmod 11 \implies \\ &\{ 1, 8, 9, 6, 4, 10, 3, 2, 5, 7, 1, 8, 9, \dots \}.\end{aligned}\quad (12.8)$$

We see that this sequence has the maximal period length of  $M - 1 = 10$  and that each integer in the interval  $[1, 10]$  is generated exactly once per cycle. The reader can verify that, for this choice of  $M$  (with  $c = 0$ ), the values  $a = 2, 6, 7, 8$  also have these favorable properties; the other values generate sequences with only two or five elements and hence violate the uniformity criteria strongly. However, as will also be discussed below, even the full-length sequences exhibit unacceptable correlations.

Of course, we are interested in much longer sequence lengths in real applications. Often,  $M$  is taken to be the *word size* of the machine and  $a$  is a prime number. However,  $M$  and  $a$  must be chosen with care, and the resulting algorithm carefully programmed, to avoid an integer *overflow* for the product  $a\tilde{x}_i$ ; this is explained further below.

### IBM's SURAND and Unix's rand and drand48

One old and still widely used MLCG method (possibly because its modulus  $M$  is the largest prime that fits in the 32-bit signed integer word used by many computers [46]) is **SURAND**, though it is considered poor by experts [715] (see discussion under Lattice Structure below and Figure 12.2). Developed by IBM for its system/360 series, **SURAND** has the values:

#### SURAND MLCG:

$$a = 7^5 = 16807; \quad M = 2^{31} - 1 = 2147483647; \quad c = 0.$$

A ‘naive’ FORTRAN implementation of this generator might be the simple implementation above, that is, include the two statements:

```
seed = mod (a * seed, m)
ranu = seed / m.
```

However, this would not produce the right sequence because of overflow.

To avoid the overflow in the product  $a\tilde{x}_i$  (or  $a * seed$  in the code), it is necessary to ensure that all intermediate integers are bounded by  $M - 1$ . The basic idea, based on [960], is outlined in Box 12.2.

**Box 12.2: Avoiding Overflow in Linear Congruential Generator Implementation**

To avoid overflow in the computation of the product  $a x$  in the implementation of eq. (12.7) (we suppress subscripts for clarity), let us *assume* for the moment that we could factor  $M$  as

$$M = a q, \quad q = \text{integer}. \quad (12.9)$$

Then, we could write the MLCG recursion relation as

$$f(x) = a x \bmod M = a x \bmod (a q) = a (x \bmod q).$$

Of course  $M$  is a prime, and no such factorization  $M = aq$  exists. However, instead of eq. (12.9), we can *approximately factor*  $M$  as:

$$M = a q + r, \quad 1 \leq r \leq a - 1, \quad (12.10)$$

where

$$q = M \text{ div } a \equiv \lfloor M/a \rfloor, \quad r = M \bmod a. \quad (12.11)$$

Here  $\lfloor y \rfloor$  denotes the largest integer smaller than, or equal to,  $y$ ; in other words,  $\lfloor M/a \rfloor$  is the integer division of  $M$  by  $a$ . If  $r < q$ , this approximate factorization is useful since then the magnitude of the intermediate product is not greater than  $M - 1$ .

For the SURAND MLCG ( $a = 16807$  and  $M = 2^{31} - 1$ ), we obtain  $q = 127773 > r = 2836$ .

---

This better implementation leads to the following correct implementation of SURAND (see [165, 960] for further details):

```

C*****
      double precision function ranu ()
c Good implementation of SURAND. See Park & Miller,
c Comm. ACM31:1192, 1988. Subroutine ranset should be called
c (once) before the first function call.
      integer a, m, q, r, seed
      double precision rm
      parameter (a=16807, m=2147483647, q=127773, r=2836, rm=1d0/m)
      common /random/ seed
      save /random/
      data seed /1/
      seed = a * mod(seed, q) - r * (seed/q)
      if (seed .le. 0) seed = seed + m
      ranu = seed * rm
      return
      end
C*****

```

However, this LCG is not recommended since there are far better procedures today. There are also faster and simpler ways to implement this recursion [716].

Other known MLCG combinations are the default random number generators available at the time of this writing on our SGI's Unix System Library,

`rand` and `drand48`, using 32-bit and 48-bit integer arithmetic, respectively. Their parameters are as follows.

`rand` MLCG:

$$a = 1103515245; \quad M = 2^{31} = 2147483648; \quad c = 12345.$$

`drand48` MLCG:

$$a = 25214903917; \quad M = 2^{48}; \quad c = 11.$$

Note the somewhat confusing online documentation for `drand48`, which reports  $a$  and  $c$  in base 8 rather than 10 ( $273673163155_8$  and  $13_8$ , respectively).

We discuss some of the defects of `rand` and `drand48` below (see also Figure 12.2).

### Lattice Structure in Linear Congruential Generators

Many statistical tests have been formulated to assess the suitability of random number generators. Linear congruential methods, for example, are known to exhibit correlations in certain hyperspaces; this basic defect is termed *coarse lattice structures*. Essentially, this means that when subsets of such sequences are represented in Euclidean space (two dimensions or higher), a lattice structure emerges; in other words, points lie on a number of hyperplanes rather than cover the space in a random-like manner. This pattern indicates that the sequence is not truly as random and uniform as sought. One way to visualize lattice structure is to plot  $k$ -lag pairs of numbers of the sequence, namely  $\{x_i, x_{i+k}\}$  in the unit-square plane for fixed  $k$ . Often, we plot pairs of consecutive numbers in the sequence on the unit square and triplets of consecutive numbers in the sequence on the unit cube.

This defect of LCG methods has been credited to G. Marsaglia in 1968; see also [986]. *Spectral tests* have since been developed to measure such  $k$ -dimensional uniformities. Such tests essentially determine the maximum distance between adjacent hyperplanes; the larger this value, the worse the generator.

To illustrate, consider  $k = 1$ -lag pairs for our simple MLCG above with  $M = 11$  and  $a = 8$  (see expression in (12.8)). If we plot in two dimensions all consecutive pairs of points, that is:

$$\{1, 8\}, \{8, 9\}, \{9, 6\}, \{6, 4\}, \dots, \{7, 1\}, \quad (12.12)$$

we see alarmingly that these points lie on four parallel lines with either positive and negative slopes (see Fig. 12.1). The spectral test would determine the maximum distance between these parallel lines.

Figure 12.1 shows the lattice structure generated from the four  $a$  values that yield full periods for this generator ( $\tilde{x}_{i+1} = (a \tilde{x}_i) \bmod 11$ ). Clearly, defects emerge.

You might think that this toy problem is especially misleading. Unfortunately, even long LCG sequences are known to display such uniform patterns or structures that indicate imperfect uniform sampling.

Figure 12.2 shows the structure obtained for SURAND resulting from generating 5 billion random numbers and plotting pairs (in two dimensions) and triplets

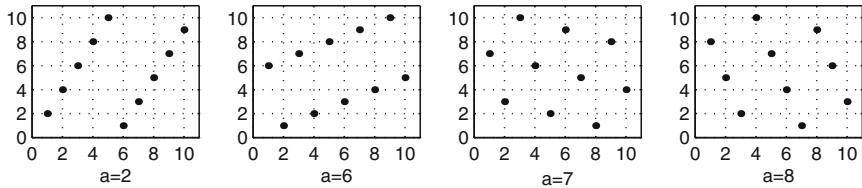


Figure 12.1. Lattice structure in two dimensional space for the multiplicative linear congruential generator (MLCG) generator  $y_{i+1} = (a y_i) \bmod 11$  for various values of  $a$ .

(in three dimensions) of consecutive ( $\text{lag } k = 1$ ) numbers that appear on a sub-region of the unit square. Clearly, defects are evident: a regular pattern emerges, indicating limited coverage. These defects would not have been apparent from a similar plot using far fewer numbers in the sequence — say 50,000 — as done in [46, Figure 3.3].

Figure 12.2 also shows patterns obtained from the Unix default generators `rand` and `drand48` discussed above. For both `rand` and `drand48`, we also generated 5 billion consecutive numbers in the sequence. The corresponding ( $\text{lag } k = 1$ ) plots on subregions of the unit square reveal a lattice pattern for `rand` but not `drand48`.

Most texts and review articles on the subject illustrate such patterns (e.g., [46, Figure 3]). For vivid color illustrations of the artifacts introduced by poor random number generators on the lattice structure of a polycrystalline Lennard-Jones spline, see [559], for example.

See also the related Monte Carlo exercise which involves generating 2D and 3D plots to search for structure of a particular (faulty!) random number generator termed RANDU. The LCG RANDU defined in that exercise can already exhibit a high degree of correlation when a relatively small number of sequence points (e.g., 2500) is generated!

#### 12.2.4 Other Generators

To overcome some of these deficiencies of linear congruential generators, other methods have been designed. Two alternative popular classes are *lagged Fibonacci* and *shift-register* generators.

##### Fibonacci Series

A *Fibonacci series* is one in which each element is the sum of the two preceding values, e.g.,  $\{1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, \dots\}$ . The series is named after the master of unit fractions, made famous for his rabbit breeding question;<sup>6</sup> the answer is 377, the 13th element of the Fibonacci series above.

---

<sup>6</sup>How many pairs of rabbits can be produced in a year from one rabbit pair? Assume that every month each pair produces a new offspring couple, which from the second month also becomes productive.

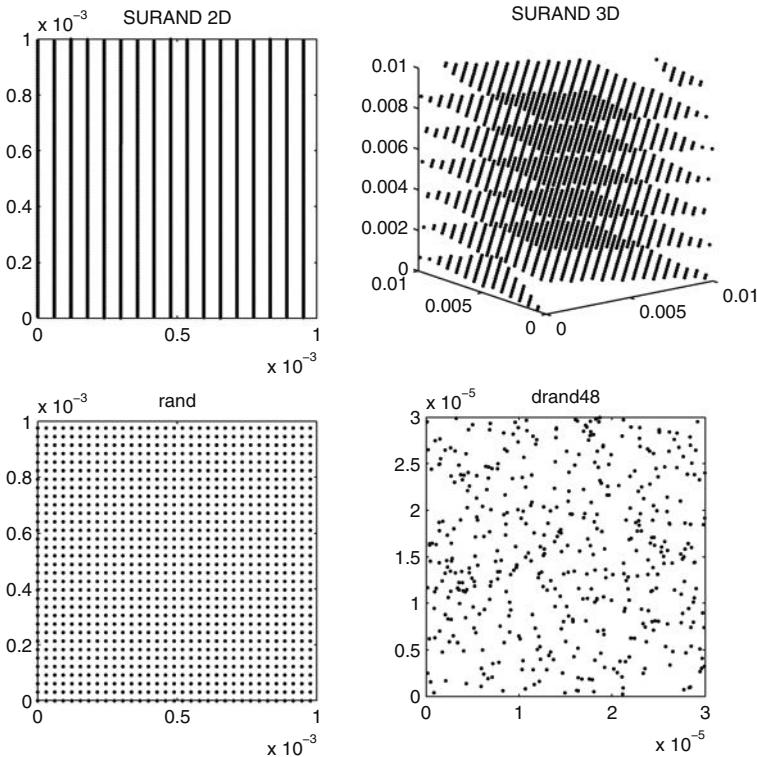


Figure 12.2. Structure corresponding to three linear congruential generators, with basic formula  $y_{i+1} = (ay_i + c) \bmod M$ , and displayed on subregions of the unit square or cube: **SURAND** ( $a = 16807$ ,  $M = 2147483647$ ,  $c = 0$ ); **rand** ( $a = 1103515245$ ,  $M = 2^{31}$ ,  $c = 12345$ ); and **drand48** ( $a = 25214903917$ ,  $M = 2^{48}$ ,  $c = 11$ ). In all cases, 5 billion numbers in the sequence were generated; fewer numbers correspond to the points displayed on the subregions in view (pairs in 2D and triplets in 3D). The code for **SURAND** was based on the one given in this text, and the codes for **rand** and **drand48** were these available internally on our SGI's Unix System Library. The programs used for plotting the data are available on the text's website. See also Tables 12.1 and Figure 12.3 for Monte Carlo averages using **rand** and **drand48**.

A Fibonacci random number generator computes each variate by performing some operation on the previous two iterates. Schematically, we write:

$$\tilde{x}_{i+1} = (\tilde{x}_{i-j} \odot \tilde{x}_{i-k}) \bmod M, \quad j < k, \quad (12.13)$$

where  $\odot$  is an arithmetic or logical operation and  $j$  and  $k$  are integer lags.

Multiplicative or additive lagged Fibonacci generators are common, and their periods can be quite long. Popular, easy-to-implement additive generators in this class have the form

$$\tilde{x}_{i+1} = (\tilde{x}_{i-j} + \tilde{x}_{i-k}) \bmod 2^m, \quad j < k, \quad (12.14)$$

where  $k$  is sufficiently large (e.g., 1279). The maximal period of such a generator is  $(2^k - 1) 2^{m-1}$  [663], and the procedure has the advantage of working directly in floating point, without the usual integer-to-floating point conversion.

An example of an additive lagged Fibonacci generator used by the Thinking Machines Library has  $j = 5, k = 17, m = 32$  (period around  $2^{48}$ ). A multiplicative lagged Fibonacci generator considered better has the form

$$\tilde{x}_{i+1} = (\tilde{x}_{i-j} \times \tilde{x}_{i-k}) \bmod 2^m, \quad j < k, \quad (12.15)$$

though its period,  $(2^k - 1) 2^{m-3}$ , is smaller by a factor of 4. This class of generators is the recommended choice by Knuth [663] and Marsaglia [832], though it was noted [663] that little theory exists to demonstrate their desirable randomness properties. L'Ecuyer later cautioned against the use of these lagged Fibonacci generators, as they display highly unfavorable properties when subjected to certain spectral tests [714].

### Shift-Register Generators

Shift-register (or *Tausworthe*) random number generators have a similar form to lagged Fibonacci series generators but employ  $M = 2$  in eq. (12.13). This means that only binary bits of a variate are generated and then collected into words by relying on a shift register. The operation  $\odot$  is the ‘exclusive or’.

An example of a shift-register generator is a  $k$ -step method which generates a sequence of  $k$  random numbers by splitting this sequence into consecutive blocks and then taking each block as the digit expansion in base  $p$  (typically 2). We can thus write this family of generators as:

$$\tilde{x}_{i+k} = \left[ \sum_{j=0}^{k-1} a_j \tilde{x}_{i+j} \right] \bmod 2, \quad (12.16)$$

where the  $\{\tilde{x}_i\}$  and  $\{a_j\}$  are either 0 or 1. The output values  $\{x_i\}$  are constructed from these bits.

Shift register methods with carefully chosen parameters are very fast, have strong uniformity properties, and possess period lengths that are not bounded by the word size of the machine. Though they may not exhibit lattice structure in the same space as LCGs, quality parameters must be selected based on analysis of lattice structure in a different space (of formal series); see [713, 717] for example. The maximal length of series (12.16) is  $2^{k-1}$ .

### Combination Generators

Many other methods exist, including linear matrix generators, nonlinear recurrence generators such as inversive congruential and quadratic, and various combinations of these generators. Combining output of good basic generators to create new random sequences can improve the quality of the sequence and increase the length. Schematically, we form  $\{\tilde{z}_i\}$  from  $\{\tilde{x}_i\}$  and  $\{\tilde{y}_i\}$  by defining

$$\tilde{z}_i = \tilde{x}_i \odot \tilde{y}_i,$$

where  $\odot$  is typically a logical (e.g., exclusive-or operator) or addition modulo  $M$ . If the associated periods of each sequence,  $\tau_x$  and  $\tau_y$ , are relatively prime, the cycle length of  $\{\tilde{z}_i\}$  can be as large as the product  $\tau_x \tau_y$ . The properties of the combination sequence will be no worse than those of either sequence and typically better.

L'Ecuyer proposes good combined generators based on the addition of linear congruential sequences of higher order [716]. Three such highly efficient methods are offered programmed in the C language. The two sequences for 32-bit machines have lengths of  $2^{191} (\sim 10^{57})$  and  $2^{319} (\sim 10^{96})$ , and the 64-bit version has the impressive length of  $2^{377} (\sim 10^{113})$ . All combine two sequences defined by

$$\tilde{x}_{1,i} = (a_{1,1}\tilde{x}_{1,i-1} + a_{1,2}\tilde{x}_{1,i-2} + \cdots + a_{1,k}\tilde{x}_{1,i-k}) \bmod M_1, \quad (12.17)$$

$$\tilde{x}_{2,i} = (a_{2,1}\tilde{x}_{2,i-1} + a_{2,2}\tilde{x}_{2,i-2} + \cdots + a_{2,k}\tilde{x}_{2,i-k}) \bmod M_2, \quad (12.18)$$

for  $i = 0, 1, \dots$ , where  $M_1$  and  $M_2$  are distinct primes and the two sequences have period lengths  $M_1^k - 1$  and  $M_2^k - 1$ , respectively.

A combined multiplicative linear congruential generator can be formed by adding  $\delta$  multiples of the variates from the two series, or by forming

$$X_i = (\delta_1\tilde{x}_{1,i}/M_1 + \delta_2\tilde{x}_{2,i}/M_2), \quad (12.19)$$

where  $\delta_1$  and  $\delta_2$  are integers, each relatively prime to its associated sequence modulus ( $M_1$  and  $M_2$ ). With properly chosen parameters, the period length of  $\{X_i\}$  will be  $(M_1^k - 1)(M_2^k - 1)/2$ . These formulas can be generalized to more than two sequences.

L'Ecuyer's sequence of length  $2^{191}$  [716] combines two sequences by using  $k = 3$  (the number of prior sequence iterates),  $M_1 = 2^{32} - 209$ ,  $M_2 = 2^{32} - 22853$ , and  $a_{1,1} = a_{2,2} = 0$ ,  $a_{1,2} = 1403580$ ,  $a_{1,3} = -810728$ ,  $a_{2,1} = 527612$ ,  $a_{2,3} = -1370589$ . The second 32-bit-machine generator uses more terms with  $k = 5$ , and the long, 64-bit generator has  $k = 3$  but two larger moduli, of order  $2^{63}$ . See the programs for these generators (in the C language) in [716].

FORTRAN and C codes for another generator of length  $\sim 2^{121}$  with good statistical properties [720] are available on the course website. This generator combines four MLCGs defined by

$$\tilde{x}_{j,i} = (a_j\tilde{x}_{j,i-1}) \bmod M_j, \quad j = 1, 2, 3, 4, \quad (12.20)$$

via

$$w_i = \sum_{j=1}^4 (\delta_j\tilde{x}_{j,i}/M_j) \bmod 1, \quad \delta_j = (-1)^{j+1}. \quad (12.21)$$

The respective multipliers and moduli are set to  $a_1 = 45991$ ,  $a_2 = 207707$ ,  $a_3 = 138556$ ,  $a_4 = 49689$ , and  $M_1 = 2147483647$ ,  $M_2 = 2147483543$ ,  $M_3 = 2147483423$ ,  $M_4 = 2147483323$ , respectively [720].

### 12.2.5 Artifacts

Two interesting examples that illustrate the importance of quality random number generators and their appropriate testing with the application at hand are summarized in Boxes 12.3 and 12.4.

The first, from a real situation in 1989, reflects a coincidental relationship between the problem (matrix) size and the period of the generator (the matrix dimension divides the period  $\tau$ ), as well as short  $\tau$  and limited accuracy (single-precision computer arithmetic). These problems could have been avoided by averting this matrix-dimension/generator-period relationship, increasing the generator period, and using double-precision arithmetic.

The second instructive example of “hidden errors” stemming from apparently good random number generators was reported in 1992 [392]. Essentially, the researchers showed that incorrect results can be produced under certain circumstances by random number generators that have a long period and have passed certain tests for randomness. Thus, a careful testing of the combination of random number generator and application is generally warranted. Though thought to be generators of high-quality, the generators used in [392] are known to have unfavorable lattice structure [714]. Of course, these examples also argue for using generators with as-long-a-period as possible.

In addition to possible systematic errors with high-quality random number generators for some algorithms due to subtle correlations, researchers showed that even good generators can yield inconsistent results regardless of the algorithm [1046]. Though researchers believed that the two used generators had passed all known statistical tests, the generators produced (with the same algorithm) critical temperature estimates for the continuous clock model that differed by 2%, much higher than intrinsic errors. This model has a second-order phase transition, and the exact answer is not known. Thus, it cannot be determined which result is more reliable.

*The best advice, therefore, appears to be not only to use reliable generators with as long periods as possible, but also to experiment with several generators as well as algorithms to the extent possible.*

**Box 12.3: Accidental Relationship Between Problem Size and Generator Length  
(1989 Linpack Benchmarks)**

In 1989, David Hough, a numerical analyst working at Sun Microsystems, noticed very peculiar behavior in benchmark testing of the package Linpack. (The original posting can be found on the archived NA Digest, Volume 89, Issue 1, 1989). The single-precision factorization of  $512 \times 512$  randomly-generated matrices produced a perplexing underflow (roughly around  $10^{-40}$ ) in the diagonal pivot values. (*Sherlocks: note that  $512 = 2^9$* ). However, matrices composed of random data from a uniform distribution are known to be remarkably well conditioned! So how can this seemingly well conditioned matrix be nearly singular?

The answer came upon examination of the random number generator and how it was used to set the matrix elements. Specifically, the matrix elements  $a_{ij}$  were set to be in the range  $[-2, 2]$  according to the following subprogram, which relies on a simple MLCG with  $a = 3125$  and  $M = 65536$ . (*Holmes fans: note that  $M = 2^{16}$* ).

```
C*****
      subroutine matgen(amat,Lda,n,b,norma)
      real amat(Lda,1), b(1), norma, halfm, quartm
      integer a, m
      parameter (a = 3125, m = 65536)
      parameter (halfm = 32768.0, quartm = 16384.0)
      iseed = 1325
      norma = 0.0
      do 30 j = 1, n
         do 20 i = 1, n
            iseed = mod (a * iseed, m)
            amat(i,j) = (iseed - halfm) / quartm
            norma = max (amat(i,j), norma)
20     continue
30     continue
      return
      end
C*****
```

A quick examination first showed that the period of this MLCG is only  $\tau = M/4 = 16384$ ; the full period of  $M$  could have easily been generated by changing one line above, to incorporate the nonzero  $c = -1$  increment value. Still, this would have only delayed the underflow problem to a  $1024 \times 1024$  matrix.

The main problem here lies in the fact that the matrix size chosen,  $512 = 2^9$ , divides the modulus, also a power of 2 ( $M = 2^{16}$ ). Hence, the period  $\tau = 2^{14}$  factors  $\tau = 512 \times 32$ . *This means that the first 32 columns of the “random” matrix are repeated 16 times!* The matrix is subsequently singular, and after each 32 steps of Gaussian elimination the zeros in the lower triangular part of the matrix are reduced by a factor of order  $(10^{-7})$ . Clearly, after six rounds of such transformations (each treating 32 columns), the element size in the lower triangle would drop to order  $\mathcal{O}(10^{-42})$ , explaining the underflow. This problem could have been removed by using matrix sizes that do not divide the period, resorting to double precision arithmetic (underflow threshold of  $10^{-300}$ ), and by increasing the period of the generator substantially.

Note also that besides underflow, the generator could have experienced other problems for matrices smaller than  $512 \times 512$ , since  $512 \times 512 = 2^{18}$ , and  $2^{18}$  is greater than  $2^{14}$ , the generator’s period.

---

### 12.2.6 Recommendations

In sum, high-quality, long-sequence random number generators are easy to find in the literature, but they are not necessarily available on default system implementations.

*For the best results, an MC simulator is well advised to consult the resident mathematical expert for the most suitable generator and computing platforms with respect to the application at hand.*

Certainly, the user should compare application results as obtained for several random number generators. Indeed, L'Ecuyer likens generators to cars [716]: no single model nor size of an automobile can possibly be a universal choice. Good expert advice can be obtained by examining Pierre L'Ecuyer's website (presently at [www.iro.umontreal.ca/~lecuyer](http://www.iro.umontreal.ca/~lecuyer)). Certainly, given today's computationally-intensive biomolecular simulations, it is advisable to use sequences with long periods. Combined multiplicative linear congruential generators are good choices. See [716] for good recommendations. See also [607, pp. 76–78] for an efficient FORTRAN implementation of a good combination MLCG used for DNA work, though with a small period by today's standards ( $10^{18}$ ).

Generators particularly suitable for parallel machines are also available [559, 985], characterized by different streams of variates produced by good seeding algorithms and variations in the parameters of the underlying recursion formulas. See the SPRNG scalable library package (Scalable Parallel Random Number Generation) targeted for large-scale parallel Monte Carlo applications: [sprng.cs.fsu.edu](http://sprng.cs.fsu.edu), for software. The package can be used in C, C++, and Fortran and has been ported to most major computer platforms.

---

#### **Box 12.4: Accidental Relationship Between Simulation Protocol and Generator Structure (Ising Model)**

Ferrenberg *et al.* [392] used different generators in the context of simulating an Ising model, a model characterized by an abrupt, temperature-dependent transition from an ordered to a disordered state. The states, characterized by the spin directionality of the particles, were generated by an algorithm termed Wolff that determines the flips of a cluster on the basis of a random number generator. Surprisingly, the researchers found that the correct answer was approximated far better by the 32-bit multiplicative linear congruential generator SURAND, well recognized to have lattice-structure defects. So why does the apparently-superior shift register generator produce *systematically incorrect results* — energies that are too low and specific heats that are too high?

An explanation to these observations came upon inspection of the Wolff algorithm. Namely, subtle correlations in the random number sequence affect the Wolff algorithm in a specific manner! If the high order bits are zero, they will remain zero according to the spin generation algorithm. This in turn leads to a bias in the cluster size generated and hence the type of equilibrium structures generated.

The main message from this work was a note of caution on the effect of subtle correlations within random number generators on the system generation algorithms used for simulating the physical system. This suggests that not only should a generator be tested on its own; it should be tested together with the algorithm used in the MC simulation to reduce the possibility of artifacts.

---

## 12.3 Gaussian Random Variates

### 12.3.1 Manipulation of Uniform Random Variables

Our uniform random variates computed in the last section,  $U$ , can be used to generate variates  $X$  that correspond to a more general, given probability distribution by a simple transformation. To generate a continuous variate  $X$  with distribution function  $F(x)$  (see Box 12.1) which is continuous and strictly increasing on  $(0, 1)$  (i.e.,  $0 < F(x) < 1$ ), we set  $x$  to  $F^{-1}(u)$  where  $F^{-1}$  is the inverse of the function  $F$ . The challenge in practice is to establish good algorithms for evaluating  $F^{-1}(u)$  to the desired accuracy.

Below we only describe generating variates from a Gaussian (normal) distribution, commonly needed in molecular simulations. For information on generating variates from many other distributions, see [706] and the web page of Luc Devroye, for example.

### 12.3.2 Normal Variates in Molecular Simulations

A vector of normally-distributed random variates satisfying a given mean ( $\mu$ ) and variance ( $\sigma^2$ ) (see Box 12.1, eq. (12.5)) is often required in molecular simulations. One example is the initial velocity vector (of components  $\{v_i\}$ ) in a molecular dynamics simulation corresponding to the target temperature of an  $n$ -atom system,

$$\sum_{i=1}^{3n} m_i v_i^2 = 3 n k_B T .$$

Another example is the Gaussian random force vector  $R$  in Langevin dynamics with zero mean and variance chosen so as to satisfy:

$$\langle R(t)R(t') \rangle = 2\gamma k_B T M \delta(t - t') , \quad (12.22)$$

where  $M$  is the mass matrix and  $\gamma$  is the damping constant.

In the molecular and Langevin dynamics cases above, we first set each component of the vector **vec** from a standard normal distribution (zero mean and unit variance) so that the sum is also a normal distribution with the additive means and variances (see Central Limit Theorem below); to obtain the desired variance  $\sigma^2$  rather than unit variance, we then modify each component according to the vector update relation

$$\text{vec} \leftarrow \sigma \text{ vec} + \mu .$$

Since, by this modification, it is easy to generate normal variates from the normal distribution with mean  $\mu$  and variance  $\sigma^2$  ( $\mathcal{N}(\mu, \sigma^2)$ ) from variates sampled from a standard normal distribution ( $\mathcal{N}(0, 1)$ ), it suffices to generate standard normal variates.

There are several techniques to set a variate  $X_u$  from a Gaussian or normal distribution on the basis of a uniformly distributed variate  $U$  (for which procedures were discussed above). Two are described below.

### 12.3.3 Odeh/Evans Method

One efficient approach was described by Odeh and Evans [928]. For a given  $u$  value in the range  $0 < u < 1$ , the corresponding normal variable  $x_u$  is computed to satisfy:

$$u = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_u} \exp(-t^2/2) dt. \quad (12.23)$$

This is accomplished by approximating  $x_u$  as the sum of two terms:

$$x_u = y + S(y)/T(y), \quad y = \sqrt{\{\ln(1/u^2)\}},$$

where  $S$  and  $T$  are polynomials of degree 4 chosen to yield the minimal degree rational approximation to  $x_u - y$  with maximum error less than  $10^{-7}$ .

In practice, a vector of random variates is first formed ( $U \equiv \{u_1, u_2, \dots, u_n\}$  in the notation above; see subroutine `ranuv` below, based on function `ranu`), from which a standard normal distribution is formulated ( $X_U \equiv \{x_{u_1}, x_{u_2}, \dots, x_{u_n}\}$ ); each component is then adjusted to yield the desired mean and standard deviation (see subroutine `rannv1` below).

```

***** subroutine rannv1 (n, vec, mean, var)
c A vector of n pseudorandom numbers is generated, each from a
c standard normal distribution (mean zero, var. one), based on Odeh
c & Evans, App. Stat. 23:96 (1974). For a nonzero mean MU c and/or
c non unity variance, set vec(i) = mu + sqrt(sigma(i))*vec(i).
c Routine ranset should be called before the first subroutine call.
      integer n
      double precision vec(n),mean,var(n),
      * temp,p0,p1,p2,p3,p4,q0,q1,q2,q3,q4
      parameter (p0=-.322232431088d0, p1=-1d0, p2=-.342242088547d0,
      * p3=-.204231210245d-1, p4=-.453642210148d-4,
      * q0=.99348462606d-1, q1=.588581570495d0, q2=.531103462366d0,
      * q3=.10353775285d0, q4=.38560700634d-2)
      if (n .lt. 1) return
      call ranuv(n, vec)
      do 10 i = 1, n
         temp = vec(i)
         if (temp .gt. 0.5d0) vec(i) = 1d0 - vec(i)
         vec(i) = sqrt(log(1d0/vec(i)**2))
         vec(i) = vec(i) +
         * (((((vec(i) * p4 + p3) * vec(i) + p2) *
         * vec(i) + p1) * vec(i) + p0) /
         * (((((vec(i) * q4 + q3) * vec(i) + q2) *
         * vec(i) + q1) * vec(i) + q0)
         if (temp .lt. 0.5d0) vec(i) = -vec(i)
10 continue
      do 20 i = 1, n
         vec(i) = sqrt(var(i)) * vec(i) + mean
20 continue
      return
      end

```

```

C*****
      subroutine ranuv (n, vec)
c Generate a vector of n pseudorandom uniform variates
      integer n, a, m, q, r, seed
      double precision vec(n), rm
      parameter (a=16807, m=2147483647, q=127773, r=2836, rm=1d0/m)
      common /random/ seed
      save /random/
      if (n .lt. 1) return
      do 10 i = 1, n
         seed = a * mod(seed, q) - r * (seed/q)
         if (seed .le. 0) seed = seed + m
         vec(i) = seed * rm
10   continue
      return
      end
C*****

```

For example, to set the initial velocity vector according to the target temperature using the equipartition theorem (each degree of freedom has  $k_B T/2$  energy at thermal equilibrium), the routines above are used for the velocity vector  $V$  of  $3n$  components  $\{v_i\}$  with  $\mu = 0$  and  $\text{var}(i) = (k_B T)/m_i$ . For the Langevin random force vector, the variance for each vector coordinate  $i$  is:  $(2\gamma k_B m_i T)/\Delta t$ , where the delta function  $\delta$  in eq. (12.22) is discretized on the basis of the timestep  $\Delta t$  (see also Chapter 13 on molecular dynamics).

### 12.3.4 Box/Muller/Marsaglia Method

Another popular algorithm to form normal variates  $x_1$  and  $x_2$  is the Box/Muller/Marsaglia method [663, pp. 117–118]. It involves generating two *uniformly* distributed random variates  $u_1$  and  $u_2$ , setting  $v_1$  and  $v_2$  as uniform variates between  $-1$  and  $+1$  ( $v_1 \leftarrow 2u_1 - 1$ ,  $v_2 \leftarrow 2u_2 - 1$ ), checking that  $s = v_1^2 + v_2^2$  is less than  $1$  (if  $s \geq 1$ , the procedure is repeated), and then setting the two *normal* variates  $x_1$  and  $x_2$  as:

$$x_1 = v_1 \sqrt{-2 \ln s/s}, \quad x_2 = v_2 \sqrt{-2 \ln s/s}. \quad (12.24)$$

Essentially, we are using the polar-coordinate representation of  $x_1$  and  $x_2$  by  $v_1$  and  $v_2$  ( $x_1 = \tilde{r} \cos \theta$ ,  $x_2 = \tilde{r} \sin \theta$ ,  $\tilde{r} = \sqrt{-2 \ln s}$ ,  $\theta = \tan^{-1}(v_2/v_1)$ ) to construct the joint probability distribution of the two normal variates in polar coordinates:

$$\begin{aligned} & \left( \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_1} e^{-x^2/2} dx \right) \left( \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_2} e^{-y^2/2} dy \right) \\ &= \frac{1}{2\pi} \int_{\substack{\{(r, \theta) | \\ r \cos \theta \leq x_1 \\ r \sin \theta \leq x_2\}}} e^{-r^2/2} r dr d\theta. \end{aligned} \quad (12.25)$$

## 12.4 Means for Monte Carlo Sampling

### 12.4.1 Expected Values

Armed with a uniform random variate generator, we can now address the important task of estimating a mean property of interest. In molecular simulations, we might seek the average geometric and energetic properties associated with an equilibrium distribution of conformations.

#### MC Estimate

In its simplest form, we write such a mean, or expected value, as an integral

$$I = \int_{\mathcal{D}} A(x) dx = \langle A(x) \rangle_{\mathcal{D}} \quad (12.26)$$

where the average is computed over the uniformly distributed elements  $x \in \mathcal{D}$ . For example, assume that the function  $A(x)$  is defined on  $[0, 1]$ . Choose a sequence of  $N$  random variates for large  $N$ ,

$$x_1, x_2, \dots, x_N,$$

and generate corresponding function values  $y_i = A(x_i)$ :

$$y_1, y_2, \dots, y_N.$$

Then we compute the average, termed the Monte Carlo estimate of  $I$ , by:

$$\bar{y}_N = \frac{1}{N} \sum_{i=1}^N y_i. \quad (12.27)$$

#### Simple Example: Calculate $\pi$ by MC

As a simple example, consider calculating  $\pi$  by Monte-Carlo integration of the area of a quarter-circle of radius 1 circumscribed inside the unit square in the plane (with center at the origin of the plane). The integral to be evaluated is:

$$\int_0^1 \int_0^1 \rho(x, y) dx dy$$

where

$$\rho(x, y) = \begin{cases} 1 & x^2 + y^2 \leq 1 \\ 0 & \text{else} \end{cases}.$$

This integral's value is  $\pi/4$ . A simple Fortran program to perform this integration by Monte Carlo sampling consists of the following:

```
*****
subroutine monte(nstep)
implicit none
integer nstep, i, nin, iseed
double precision x, y, tmp, rand
```

```

nin = 0
iseed = 12345
call srand(iseed)
do 30 i = 1, nstep
    x = rand()
    y = rand()
    tmp = sqrt(x*x + y*y)
    if (tmp .lt. 1.d0) then
        nin = nin + 1
    endif
30 continue
print *, nstep, (4.d0 * nin)/nstep
return
end
C*****

```

Results as a function of the sample size (`nstep`) are presented in Table 12.1 and Figure 12.3 using the Unix `rand` and `drand48` random number generators and also more sophisticated methods. Since `drand48` is the fastest of the generators, we also record the estimated value corresponding to  $10^{12}$  steps (only up to  $10^9$  steps for the rest).

Note that, unfortunately, this procedure for calculating  $\pi$  is not very accurate. At best, the first six decimal places of  $\pi = 3.14159265358979323846\dots$  are obtained. The accuracy is limited not only by the sample size — statistical error — but also by any possible defects of the random number generator (e.g., lattice structure and limited coverage; see Figure 12.2). Here we see that the accuracy of the means starts to deteriorate after the number of steps exceeds the period length. We also learn from this example that the longer-period generators have greater resolution (another order of magnitude of two).

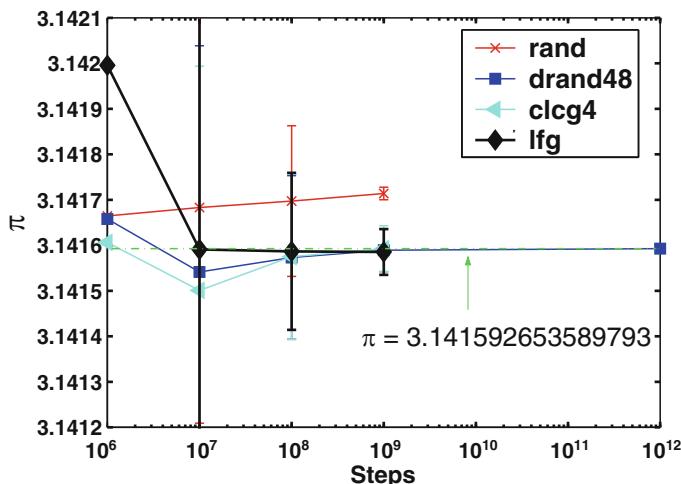


Figure 12.3. Results (means and error bars) of MC estimates for  $\pi$  based on different random number generators, as tabulated in Table 12.1.

Table 12.1. Results of computing  $\pi$  by MC integration with various random number generators and sample sizes by the procedure described in the text, subroutine `monte` (on an SGI R12K/300 MHz Octane processor). The value  $\sigma$  is computed from 100 runs for each NSTEP value, except for  $10^{12}$ , for which it is based on one run. `rand` has  $\tau \approx 2^{31}$ ;  $10^9$  calls require 771 seconds (13 min.). `drand48` has  $\tau \approx 2^{48}$ ;  $10^9$  calls require 656 seconds (11 min.). `clg4` has  $\tau \approx 2^{31}$ ;  $10^9$  calls require 5205 seconds (87 min.). `lfg` has  $\tau \approx 2^{64}$ ;  $10^9$  calls require 2106 seconds (35 min.). See also Figure 12.3.

Nstep	Estimate	Error	$\sigma$ , s.d.
rand, IRIX 6.5 system library			
$10^2$	3.1604000000000E+00	1.8807346410208E-02	1.59E-01
$10^3$	3.1346800000000E+00	-6.9126535897936E-03	4.75E-02
$10^4$	3.1383680000000E+00	-3.2246535897933E-03	1.79E-02
$10^5$	3.1416736000000E+00	8.0946410208504E-05	5.22E-03
$10^6$	3.1416646400000E+00	7.1986410206115E-05	1.61E-03
$10^7$	3.1416831800000E+00	9.0526410207570E-05	4.74E-04
$10^8$	3.1416971676000E+00	1.0451401020672E-04	1.65E-04
$10^9$	3.1417139545200E+00	1.2130093020657E-04	1.37E-05
drand48, IRIX 6.5 system library			
$10^2$	3.1408000000000E+00	-7.9265358979264E-04	1.72E-01
$10^3$	3.1456400000000E+00	4.0473464102098E-03	5.62E-02
$10^4$	3.1431880000000E+00	1.5953464102063E-03	1.53E-02
$10^5$	3.1418572000000E+00	2.6454641020690E-04	5.92E-03
$10^6$	3.1416580000000E+00	6.5346410206946E-05	1.61E-03
$10^7$	3.1415415040000E+00	-5.1149589793908E-05	4.97E-04
$10^8$	3.1415733104000E+00	-1.9343189793464E-05	1.79E-04
$10^9$	3.1415892614400E+00	-3.3921497935019E-06	4.70E-05
$10^{12}$	3.1415928451280E+00	1.9153820707274E-07	1.00E-06
clg4, based on four linear congruential generators [720]			
$10^2$	3.1600000000000E+00	1.8407346410206E-02	1.51E-01
$10^3$	3.1413200000000E+00	-2.7265358979189E-04	4.98E-02
$10^4$	3.1405440000000E+00	-1.0486535897933E-03	1.74E-02
$10^5$	3.1416476000000E+00	5.4946410205758E-05	4.94E-03
$10^6$	3.1416064800000E+00	1.3826410206530E-05	1.59E-03
$10^7$	3.1415008760000E+00	-9.1777589792841E-05	4.92E-04
$10^8$	3.1415751016000E+00	-1.7551989793141E-05	1.82E-04
$10^9$	3.1415925054000E+00	-1.4818979332532E-07	5.05E-05
lfg, SPRNG package, modified lagged-Fibonacci generator			
$10^2$	3.1372000000000E+00	-4.3926535897922E-03	1.70E-01
$10^3$	3.1414400000000E+00	-1.5265358979244E-04	5.20E-02
$10^4$	3.1442160000000E+00	2.6233464102083E-03	1.55E-02
$10^5$	3.1431008000000E+00	1.5081464102074E-03	5.23E-03
$10^6$	3.1419958000000E+00	4.0314641020611E-04	1.64E-03
$10^7$	3.1415909920000E+00	-1.6615897910910E-06	5.46E-04
$10^8$	3.1415866680000E+00	-5.9855897953653E-06	1.72E-04
$10^9$	3.1415854581200E+00	-7.1954697937748E-06	5.03E-05

### 12.4.2 Error Bars

#### Law of Large Numbers

According to the *Law of Large Numbers* in probability theory, the average of  $N$  sampled random variables converges (in probability) to its expected value. Stated more formally, if the uniform variates are independent and drawn from the same distribution so that the expected value of each  $y_i$  is  $\mu$ , then as  $N \rightarrow \infty$  the average value  $\bar{y}_N$  converges to  $\mu$  asymptotically:

$$P \left\{ \lim_{N \rightarrow \infty} (\bar{y}_N) = \mu \right\} = 1.$$

However, the rate of convergence to the expected value is a different matter and requires stronger assumptions.

#### Variance

As stressed in this chapter's introduction, it is essential to provide *error bars* when reporting an MC average. The variance of  $\bar{y}_N$  is defined as

$$\sigma_{\bar{y}}^2 = \text{var}(\bar{y}) = \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y}_N)^2. \quad (12.28)$$

The variance measures the distribution of  $\bar{y}$  about its mean  $\mu$ ; the larger  $N$  is, the narrower the interval about  $I$  where  $\bar{y}_N$  can be found.

#### Variance Relation to Central Limit Theorem

This interval can be determined as a probability of deviation in units of  $\sigma$  on the basis of the *Central Limit Theorem*. This beautiful and powerful result states that as  $N \rightarrow \infty$ , the *limiting distribution* for a *sum of random variates* is the *normal distribution*.

Specifically, if  $\{y_1, y_2, \dots\}$  is a sequence of independent, identically distributed random variates having mean  $\mu$  and finite nonzero variance  $\sigma^2$ , then the random variable

$$S_N = y_1 + y_2 + \dots + y_N$$

has the normal density with mean  $N\mu$  and variance  $N\sigma^2$ ,  $\mathcal{N}(N\mu, N\sigma^2)$ . In other words, the normalized random variable

$$\frac{S_N - N\mu}{\sqrt{\text{var}(S_N)}} = \frac{S_N - N\mu}{\sigma\sqrt{N}}$$

has the standard normal distribution:

$$\lim_{N \rightarrow \infty} P \left( \frac{S_N - N\mu}{\sigma\sqrt{N}} \leq x \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp[-t^2/2] dt.$$

Thus, in reporting an MC average, we say that we have estimated  $I$  within one standard error (i.e.,  $\sigma/\sqrt{N}$ ) of  $\bar{y}_N$  as:

$$I = \bar{y}_N \pm \sigma_{\bar{y}}/\sqrt{N}. \quad (12.29)$$

For example,  $N = 10,000$  yields a result that is at best roughly 1% accurate; for correlated data, such as from molecular dynamics simulations, a much larger  $N$  is required for that accuracy. Note that this  $1/\sqrt{N}$  scaling of errors is a general feature of MC methods and is independent of the space dimension involved; that is why MC is frequently the method of choice for multidimensional integrals.

Since we know that the limiting distribution for  $\bar{y}_N$  is the normal distribution  $\mathcal{N}$ , we say that 68.3% of the time this estimate is within one standard error of  $I$  [625]. The above integral estimates can be generalized to independent random samples from other probability densities, as described in the next section.

*Note that the above errors are statistical and can be controlled. The more serious errors in MC algorithms are the systematic errors, such as discussed above in connection with random-number-generator artifacts. Both errors should be monitored to the extent possible.*

### 12.4.3 Batch Means

When the MC data are highly correlated, the error bars may decrease much more slowly with  $N$  as in the above idealized case. The effective number of samples is then  $N/\tau$  where  $\tau$  is the decorrelation time (number of steps) for the data. This value can be determined by examining auto and cross-correlation functions for the most slowly-varying properties of the system or by using the method of *batch means*. Extensive mathematical/statistical tests for independence are also available to estimate *confidence intervals* of independent means [706, Chapter 4].

Essentially, we divide the sample size  $N$  into  $M$  batches  $\{B_1, B_2, \dots, B_M\}$  each of  $b = N/m$  elements where  $M$  should be significantly greater than  $\tau$ ; we then obtain a mean over each batch sample:

$$\bar{y}_{B_i} = \frac{1}{b} \sum_{y_i \in B_i} y_i, \quad i = 1, \dots, M, \quad (12.30)$$

and then set the  $\bar{y}_N$  estimate as the average over all these means:

$$\bar{y}_M = \frac{1}{M} \sum_{i=1}^M \bar{y}_{B_i}. \quad (12.31)$$

In reporting the estimator of form (12.29), the relevant sample size ( $M$  rather than  $N$ ) and variance to is determined from the above mean, that is:

$$\sigma_{\bar{y}_M}^2 = \text{var}(\bar{y}_M) = \frac{1}{M} \sum_{i=1}^M (\bar{y}_{B_i} - \bar{y}_M)^2. \quad (12.32)$$

It can be shown that if the batch size  $M$  is sufficiently large, the means of the batches are approximately uncorrelated. In practice, variations on the basic batch means method sketched above, and additional tests, are needed to yield good statistics [706, pages 528–530].

## 12.5 Monte Carlo Sampling

### 12.5.1 Density Function

The properties of many molecular systems can be described by a separable Hamiltonian of general form

$$H(q, p) = E_k(p) + E_p(q) = \frac{1}{2} p^T \mathbf{M}^{-1} p + E_p(q), \quad (12.33)$$

where  $E_k$  and  $E_p$  are the kinetic and potential energy components, respectively, and  $q$  and  $p$  are the collective position and momentum vectors of the system. This Hamiltonian function forms the basis for MC simulations applied to estimate various properties of large molecular systems, such as geometric and thermodynamic functions. However, the MC estimates must emulate a probability density function  $\rho(q, t)$  or  $\rho(q, p, t)$  appropriate for the statistical ensemble ( $t$  denotes time). This probability density  $\rho$  may or may not be known.

### 12.5.2 Dynamic and Equilibrium MC: Ergodicity, Detailed Balance

MC simulations can be used to mimic a *dynamic process* ( $\rho$  depends on time  $t$ ), as in Brownian dynamics. They can also generate an ensemble around a *statistical equilibrium*, as in some conformational sampling studies.

#### Dynamic Process

In the former case, a deterministic rule (such as based on Newton's equations of motion in the diffusive limit) is used to generate each configuration from the previous configuration given initial conditions  $\rho(q, p, t_0)$ , and that rule determines the resulting  $\rho(q, p, t)$  for  $t > t_0$ .

*Below we use the notation  $X$  to represent the collective phase-space vector  $(q, p)$ ; when discussing the Metropolis algorithm later, the momentum component drops out.* (This variable  $X$  should not be confused with the random variable  $X$  defined in Box 12.1).

#### Equilibrium Process

The equilibrium ensemble regime is appropriate when  $\rho(X, t) = \rho_0(X)$  for some  $t > t_0$ , as in the Metropolis algorithm (see below). The *ensemble average* is then

considered as an estimate for the *time average*, which may be much more complex to follow. This assumption, though very difficult to prove in practice, is known as the *ergodic hypothesis*.

In this statistical equilibrium case, the rule that generates  $X_{n+1}$  from  $X_n$  need not have a clear physical interpretation. However, to be useful for sampling, the rule must ensure that any starting distribution  $\rho(X, t)$  should tend to the stationary density  $\rho_0(X)$  and that the system be *ergodic* (i.e., as  $t \rightarrow \infty$ , the system spends equal times in equal volumes of phase space); see [547] for a rigorous definition. When the rule also obeys *detailed balance* (i.e., moving from state  $X$  to  $Y$  is as likely as returning to  $X$  from  $Y$ ), an equilibrium process is approached (though biased techniques may violate detailed balance and still approach the right answer through correcting for violations).

These criteria are crucial for constructing practical sampling algorithms for physical systems; efficient sampling of configurational space is another important aspect of computer simulations, especially for large systems where configuration space cannot be sampled exhaustively.

### 12.5.3 Statistical Ensembles

Common statistical ensembles used in biomolecular simulations are the *canonical* or constant–NVT ( $N$  = number of particles,  $V$  = volume,  $T$  = temperature), *microcanonical* or constant–NVE ( $E$  = energy), *isothermal-isobaric* or constant–NPT ( $P$  = pressure), and *grand canonical* or constant– $\mu$ VT ( $\mu$  = chemical potential) [22].

#### Canonical Ensemble and Boltzmann Factor

The probability density function for the *canonical* ensemble is proportional to the Boltzmann factor:

$$\rho_{\text{NVT}}(X) \propto \exp(-\beta E(X)), \quad (12.34)$$

where  $E$  is the total energy of the system and  $\beta = (1/k_B T)$ .

(In the Metropolis algorithm, it is sufficient to work with the *potential energy*, since the potential energy is independent of momenta; see below).

Hence for two system states  $X$  and  $X'$ , the corresponding probability ratio is:

$$\frac{\rho_{\text{NVT}}(X)}{\rho_{\text{NVT}}(X')} = \exp(-\beta \Delta E), \quad (12.35)$$

where

$$\Delta E = E(X) - E(X').$$

See the sketch of Figure 12.4. The normalizing factor in the proportionality relation (eq. (12.34)) is the total partition function for all of phase space. That is:

$$\rho_{\text{NVT}}(X) = \frac{1}{(h^{3N}) N!} \frac{\exp(-\beta E(X))}{Q_{\text{NVT}}} \quad (12.36)$$

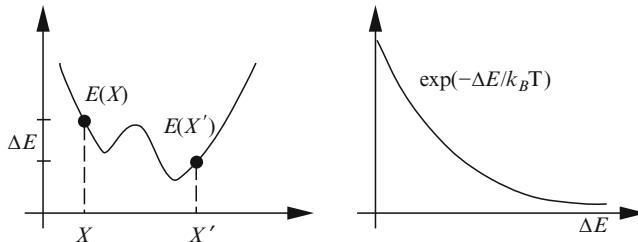


Figure 12.4. The Boltzmann probabilities for two system states  $X$  and  $X'$  with energies  $E(X)$  and  $E(X')$  are related at equilibrium by the probability ratio  $\exp[-\Delta E/k_B T]$ .

where  $h$  is Planck's constant, the factor  $N!$  accounts for the indistinguishability of the  $N$  particles, and  $Q_{\text{NVT}}$  is the canonical partition function:

$$Q_{\text{NVT}} = \frac{1}{(h^{3N}) N!} \int \exp(-\beta E(x)) dx, \quad (12.37)$$

where  $x$  is a point in phase space.

The corresponding means for the system for a function  $A$  can then be written as:

$$\langle A(x) \rangle_{\text{NVT}} = \int \rho_{\text{NVT}}(x) A(x) dx. \quad (12.38)$$

The Metropolis algorithm described below is used to generate an appropriate Markov chain (see below) from which the expected value of  $A$  is calculated as:

$$\langle A(x) \rangle_{\text{NVT}} = \lim_{M \rightarrow \infty} \frac{1}{M} \sum_{i=1}^M (A(x_i)). \quad (12.39)$$

In other words, the density function is already built into the generation algorithm through an appropriate acceptance probability.

#### 12.5.4 Importance Sampling: Metropolis Algorithm and Markov Chains

To obtain reliable statistical averages, it is essential to use computer time efficiently to concentrate calculations of the configurational-dependent functions in regions that make important contributions. This concept is known as *importance sampling*.

In many applications, it is possible to focus sampling on the *configurational* part of phase space (i.e.,  $E$  is the potential energy component) since the projection of the corresponding trajectory on the momentum subspace is essentially independent from that projection on the coordinate subspace. Hence  $X$  below is the *collective position vector only*.

### Markov Chain

Metropolis *et al.* [856] described such an efficient and elegantly simple procedure for the canonical ensemble. In mathematical terms, we generate a *Markov chain* of molecular states  $\{X_1, X_2, X_3, \dots\}$  constructed to have the limiting distribution  $\rho_{\text{NVT}}(X)$ . In a Markov chain, each state belongs to a finite set of states contained in the state space  $\mathcal{D}_0 \in \mathcal{D}$ , and the conditional distribution of each state relative to all the preceding states is equivalent to that conditional distribution relative to the last state:

$$P\{X_{n+1} \in \mathcal{D}_0 \mid X_0, \dots, X_n\} = P\{X_{n+1} \in \mathcal{D}_0 \mid X_n\};$$

in other words, the outcome  $X_{n+1}$  depends only on  $X_n$ . The Metropolis algorithm constructs a transition matrix for the Markov chain that is stochastic and ergodic so that the limiting distribution for each state  $X_i$  is  $\rho_i = \rho_{\text{NVT}}(X_i)$  and thereby generates a phase space trajectory in the canonical ensemble.

This transition matrix is defined by specifying a transitional probability  $\pi_{ij}$  for  $X_i$  to  $X_j$  so that *microscopic reversibility* is satisfied:

$$\rho_i \pi_{ij} = \rho_j \pi_{ji}. \quad (12.40)$$

In other words, the ratio of transitional probabilities depends only on the energy change between states  $i$  and  $j$ :

$$\frac{\rho_i}{\rho_j} = \frac{\pi_{ji}}{\pi_{ij}} \propto \exp(-\beta \Delta E_{ij}), \quad (12.41)$$

where  $\Delta E_{ij} = E(X_i) - E(X_j)$ . See subsection below on MC Moves with examples of biased sampling.

### Metropolis Algorithm

Briefly, the Metropolis algorithm generates a trial  $\tilde{X}_{i+1}$  from  $X_i$  by a system-appropriate random perturbation (satisfying detailed balance) and accepts that state if the corresponding energy is lower. If, however,  $E(\tilde{X}_{i+1}) > E(X_i)$ , then the new state is accepted with probability  $p = \exp(-\beta \Delta E)$ , where  $\Delta E = E(\tilde{X}_{i+1}) - E(X_i) > 0$ , by comparing  $p$  to a uniformly-generated number on (0,1): if  $p > \text{ran}$ , accept  $\tilde{X}_{i+1}$ , and if  $p \geq \text{ran}$ , generate another trial  $\tilde{X}_{i+1}$  but recount  $X_i$  in the Markov chain (see Fig. 12.4).

The result of this procedure is the acceptance probability at step  $i$  of:

$$\begin{aligned} p_{\text{acc, MC}} &= \min [1, \exp(-\beta \Delta E)] \\ &= \min \left[ 1, \frac{\rho_{\text{NVT}}(\tilde{X}_{i+1})}{\rho_{\text{NVT}}(X_i)} \right]. \end{aligned} \quad (12.42)$$

In this manner, states with lower energies are always accepted but states with higher energies have a nonzero probability of acceptance too. Consequently, the sequence tends to regions of configuration space with low energies, but the system can always escape to other energy basins.

### Simulated Annealing

An extension of the Metropolis algorithm is often employed as a global-minimization technique known as *simulated annealing* where the temperature is lowered as the simulation evolves in an attempt to locate the global energy basin without getting trapped in local wells.

Simulated annealing can be considered an extension of either MC or molecular/Langevin dynamics. Simulated annealing is often used for refinement of experimental models (NMR or crystallography) with added nonphysical, constraint terms that direct the search to target experimental quantities (e.g., interproton distances or crystallography R factors; see [676], for example, for a review of the application of simulated annealing in such contexts).

### Metropolis Algorithm Implementation

The Metropolis algorithm for the canonical ensemble can be implemented with the potential energy  $E_p$  rather than the total energy when the target measurement  $A$  for MC averaging is velocity independent. This is because the momentum integral can be factored and canceled. From eq. (12.38) combined with eq. (12.34), we expand the state variable  $x$  to represent both the momentum ( $p$ ) and position ( $q$ ) variables, both over which integration must be performed:

$$\begin{aligned}\langle A(x) \rangle &= \frac{\int \exp[-\beta E_k] dp \int A(q) \exp[-\beta E_p] dq}{\int \exp[-\beta E_k] dp \int \exp[-\beta E_p] dq} \\ &= \frac{\int A(q) \exp[-\beta E_p] dq}{\int \exp[-\beta E_p] dq}.\end{aligned}\quad (12.43)$$

The Metropolis algorithm is summarized below.

**Metropolis Algorithm (Canonical Ensemble)**

For  $i = 0, 1, 2, \dots$ , given  $X_0$ :

1. Generate  $\tilde{X}_{i+1}$  from  $X_i$  by a perturbation technique that satisfies *detailed balance* (i.e., the probability to obtain  $\tilde{X}_{i+1}$  from  $X_i$  is identical to that going to  $X_i$  from  $\tilde{X}_{i+1}$ ).
  2. Compute  $\Delta E = E(\tilde{X}_{i+1}) - E(X_i)$ .
  3. **If**  $\Delta E \leq 0$  (downhill move), accept  $X_{i+1}$  :  $X_{i+1} = \tilde{X}_{i+1}$ ;  
**Else**, set  $p = \exp(-\beta \Delta E)$ . Then
    - If**  $p > \text{ran}$ , accept  $\tilde{X}_{i+1}$  :  $X_{i+1} = \tilde{X}_{i+1}$ .
    - Else**, reject  $\tilde{X}_{i+1}$  :  $X_{i+1} = X_i$ .
  4. Continue the  $i$  loop.
-

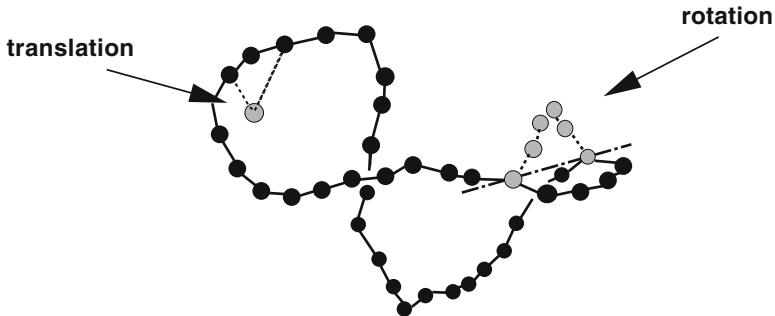


Figure 12.5. Translational and rotational MC moves for a bead model of DNA.

### MC Moves

Specifying appropriate MC moves for step 1 is an art by itself. Ideally, this could be done by perturbing all atoms by independent (symmetric) Gaussian variates with zero mean and variance  $\sigma^2$ , where  $\sigma^2$  is parameterized to yield a certain acceptance ratio (e.g., 50%). However, in biomolecular simulations, moving all atoms is highly inefficient (that is, leads to a large percentage of rejections) [428], and it has been more common to perturb one or few atoms at each step.

The type of perturbations depends on the system and the energy representation (e.g., rigid or nonrigid molecules, a biomolecule or pure liquid system, Cartesian or internal degrees of freedom). The perturbation can be set as translational, rotational, local, and/or global moves. For example, in the atomistic CHARMM molecular mechanics and dynamics program, protein moves are prescribed from a list of possibilities including rigid-residue translation/rotation, single-atom translation, and single or multiple torsional motions.

For MC simulations of a bead/wormlike chain model of long DNA, we use local translational moves of one bead at a time combined with a rotational move of a chain segment [607]; we must also ensure that no move changes the system's topology (e.g., linking number of a closed chain) for simulating the correct equilibrium ensemble.

Figure 12.5 illustrates such moves for long DNA. Figure 12.6 illustrates corresponding MC (versus Brownian dynamics) distributions of the DNA writhing number ( $Wr$ ) and the associated mean, as a function of length for two salt environments. Figure 12.7 demonstrates how a faulty move (like moving only a subset of the DNA beads instead of all) can corrupt the probability distributions of  $Wr$  and the radius of gyration ( $Rg$ ). Not only do we note a corruption of the distributions when incorrect MC protocols are used, but a large sensitivity to the initial configuration (sharp distributions around starting configurations).

The rule of thumb usually employed in MC simulations is to aim for a perturbation in Step 1 (e.g., displacement magnitude or the variance  $\sigma^2$  associated with the random Gaussian variate) that yields about 50% acceptance. Thus, we seek to balance too small a perturbation that moves the system in state space slowly

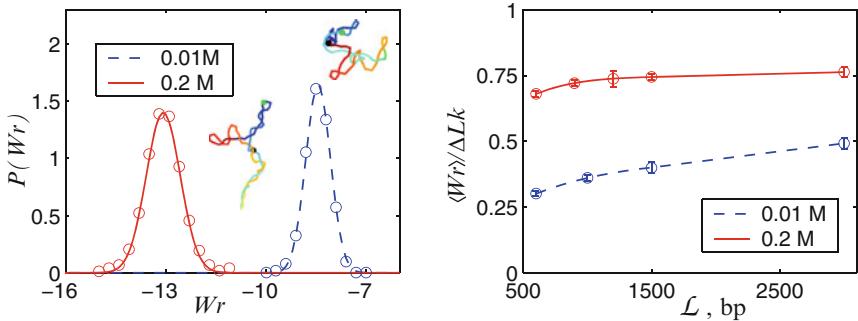


Figure 12.6. MC distributions (curves) versus Brownian dynamics data (circles) of the writhing number ( $Wr$ ) distribution of supercoiled DNA (left), and mean values of  $Wr$  (normalized by the linking number difference) as a function of the DNA length (right), at two salt concentrations. The DNA superhelical density is  $\sigma = -0.06$  in both panels, and the left panel involves DNA of length 3000 base pairs. Error bars for BD are shown only if they are larger than the circle symbol.

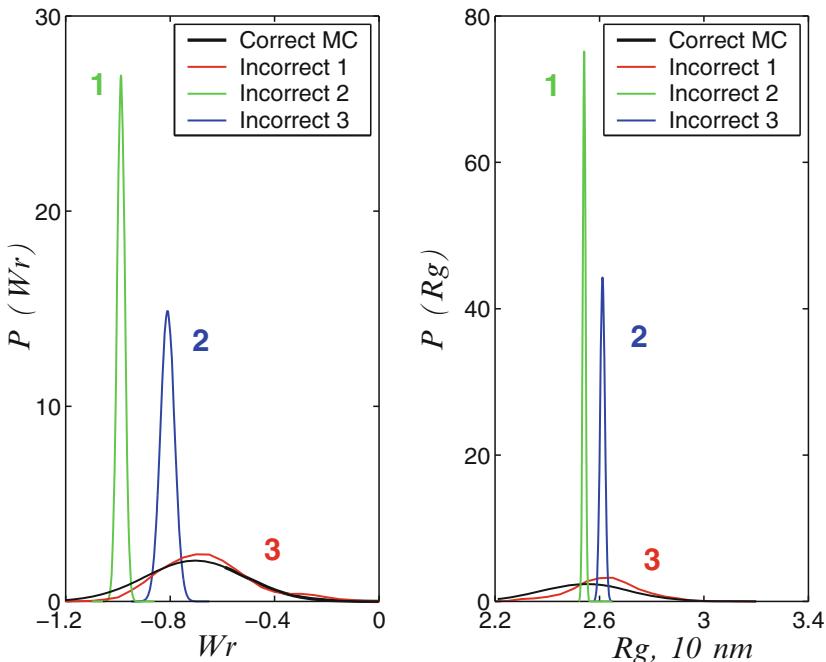


Figure 12.7. DNA writhing number  $Wr$  (left) and radius of gyration  $Rg$  distributions as generated by one correct versus three incorrect MC procedures. The incorrect MC protocols allow only a subset of the beads to move: 25 out of 30 (“Incorrect 1”), or 5 out of 30 (“Incorrect 2, 3”); the last two schemes are started from different initial points. The DNA modeled has 600 base pairs, with superhelical density of  $\sigma = -0.04$ , and the MC simulations consist of ten million steps.

with too large a perturbation that yields high trial-energy configurations, most of which are rejected. *However, the appropriate percentage depends on the application and is best determined by experimentation guided by known outcomes of the statistical means sought.* Much smaller acceptance probabilities may be perfectly adequate for some systems.

## 12.6 Monte Carlo Applications to Molecular Systems

### 12.6.1 Ease of Application

The simplicity and general applicability of Monte Carlo approaches has long been exploited for molecular applications, especially biomolecules, as reviewed in [129, 347, 781, 1117, 1457].

For example, MC methods can easily be applied to force fields with various constraints, to functions whose derivative routines are not available, or even to discontinuous potentials, like the square well potential for fluid or colloidal suspensions [155] or lattice and off-lattice protein models (e.g., [226]). MC methods also allow facile exploration of variable conditions such as the conformational dependencies on variable side-chain protonation states for proteins in the electrostatically driven MC method (EDMC) of Scheraga and co-workers [1051]. EDMC in combination with different dihedral angle constraints successfully folded a villin headpiece [1051].

As described above, MC sampling is used to generate a set of conformations under Boltzmann statistics. Thus, states that decrease the energy are always accepted and those that increase the energy are accepted with a probability  $p = \exp(-\beta\Delta E)$  where  $\beta = 1/k_B T$  and  $\Delta E$  is the energy difference between the internal energy of the new and old configurations (see eq. (12.42)). In this way, the molecular system can overcome barriers in the vast conformational space and escape from local minima.

Though in theory a good MC protocol would sample configuration space exhaustively, this becomes more difficult and inefficient in practice as the system size increases. When the cost of evaluating the energy function for large biomolecular systems is also a factor, millions of MC steps (with possibly many rejections) can become quite expensive. In addition, selecting the appropriate trial move set and movement magnitudes for a biomolecule without high rejection rates can be challenging in practice, and requires a thorough tested and implemented protocol. Thus, many MC variants have been developed to enhance sampling as well as efficiency.

*Simulated annealing* (SA), as described in the previous section, is a way to use MC for the purpose of global optimization. The idea is to lower the effective temperature gradually according to a specified cooling protocol to overcome barriers in the rugged landscape. SA can be used successfully as an extended form of MC, as well as structure optimization and molecular, Langevin or Brownian dynamics.

In fact, the Scheraga group has developed many stochastic global optimization methods based on MC [1000]. For example, an MC/Minimization (MCM) hybrid method [760] searches for the global energy minimum of a protein by an iterative procedure involving generating a large random conformational change followed by local minimization of the potential energy; this move is accepted or rejected using the Metropolis criterion. Friesner and coworkers have found this MC method to perform well in a variety of applications [372].

In this connection, see homework 13 for the *deterministic* global optimization approach based on the *diffusion equation* as suggested and implemented for molecular systems by Scheraga and colleagues [997].

Besides such simulated annealing and global optimization methods based on MC, many MC variants have been developed, such as biased MC, hybrid MC, and parallel tempering (or REM), as will be outlined below. See [224, 931], for example, for reviews and performance comparisons of several MC methods for molecular systems.

### 12.6.2 Biased MC

*Biased MC* variants have been devised with trial moves and hence the conformational deformations designed to move the system to more probable states. The Rosenbluth, instead of the Metropolis criterion, is therefore used to factor in the probability (Boltzmann weights) of all trial positions that were skipped in favor of the biased moves:

$$p_{\text{acc, biased MC}} = \min [1, W_{i+1}/W_i] \quad (12.44)$$

where the Rosenbluth factor  $W$  is equal to the product of the sum of the Boltzmann weights of trial positions for each segment  $i$  insertion:

$$W = \prod_{i=1}^N \sum_{k=1}^n \exp(-\beta E_k^i). \quad (12.45)$$

Here,  $N$  is the number of chain segments and  $E_k^i$  is the internal energy of the  $k$ th trial move for adjusting the  $i$ th segment. That is, in each step one trial move is selected for each segment  $i$  with a probability proportional to its Boltzmann weight, and this process is repeated for all segments until the entire chain is regrown. In this fashion, additional overhead is required in biased MC simulations to calculate that probability ratio.

*Configurational bias MC* (CB-MC) is a biased MC variant that helps “grow” a molecule toward particular states. Traditional CB-MC “re-grows” a deleted position of a polymer at the same end in variable orientations (instead of trying out all neighboring sites randomly). This results in an exponential scaling time with polymer length to re-grow a self-avoiding lattice chain due to the high probability of segment overlaps. In certain applications, much more effective variants

can be developed, as in the “end-transfer CB-MC” for chromatin, where one end of the polymer is grown at the other end. Dramatic efficiency can be achieved – quadratic vs. exponential scaling – in such applications [63].

A comparison of several biased MC schemes for small molecule design [224] demonstrated the excellent performance of such methods compared to genetic algorithms.

### 12.6.3 Hybrid MC

#### Exploiting Strengths of MC and MD

To enhance the efficiency of MC simulations, a simple idea emerged that attempts to combine the favorable properties of molecular dynamics (MD) simulations — sampling phase space in a directed manner guided by the shape of the energy gradient — with that of MC — sampling phase space more globally. Ideally, following conformation space by MD would generate a correct Boltzmann distribution of states, but the relatively short simulation lengths that are possible (see MD chapters) imply local rather than global sampling.

This MC/MD combination [344, 347, 688] — moving some particles by MC rules and others by MD, for example by combining global updates in position space via MD with reasonable acceptance criteria by MC — can be more effective for solvated biomolecular systems than either MC or MD alone. Such *hybrid MC* methods have been very effective for small systems [575].

#### Overall Idea

For example, the first step of a hybrid MC method can use a molecular dynamics framework to specify the system’s candidate move:  $X[i\Delta t] \longrightarrow X[(i+1)\Delta t]$  where  $\Delta t$  is the timestep. The MD algorithm must be *time reversible* and volume preserving so as to ensure detailed balance. The commonly used symplectic Verlet method satisfies this requirement (see Chapters 13 and 14). Since the recursive MD recipe relies on the velocities in addition to positions, the required velocity vectors  $V = M^{-1}P$  ( $P$  here designates momentum, not to be confused with the symbol used earlier for *probability*) are generated from a Gaussian distribution so as to obtain the target kinetic energy at the temperature  $T$  assuming energy equipartition. That is, the velocity components are drawn from a Gaussian distribution proportional to the Boltzmann factor applied to the kinetic energy:

$$\rho(P) \propto \exp[-\beta E_k(P)]. \quad (12.46)$$

Following this MD-guided step, the second step of the hybrid MC method applies the standard Metropolis acceptance criterion where the energy in the Boltzmann factor is the *Hamiltonian* (potential *plus* kinetic energy). Namely, the Metropolis criterion is applied to accept the new candidate  $\tilde{X}_{i+1}$  with probability

$$\begin{aligned}
p_{\text{acc, HMC}, \rho} &= \min \left[ 1, \exp \left[ -\beta \left( H(\tilde{X}_{i+1}, \tilde{P}_{i+1}) - H(X_i, P_i) \right) \right] \right] \\
&= \min \left[ 1, \frac{\exp[-\beta E_p(\tilde{X}_{i+1})] \exp[-\beta E_k(\tilde{P}_{i+1})]}{\exp[-\beta E_p(X_i)] \exp[-\beta E_k(P_i)]} \right] \\
&= \min \left[ 1, \frac{\rho_{\text{NVT}}(\tilde{X}_{i+1}) \exp[-\beta E_k(\tilde{P}_{i+1})]}{\rho_{\text{NVT}}(X_i) \exp[-\beta E_k(P_i)]} \right]. \quad (12.47)
\end{aligned}$$

Note that *generalized ensembles* can be emulated by HMC methods by applying weighting factors to the Metropolis-generated configurations. That is, to sample from a general ensemble with  $\mu(x)$  rather than  $\rho(x)$ , we adjust the acceptance criteria of eq. (12.47) to be:

$$p_{\text{acc, HMC}, \mu} = \min \left[ 1, \frac{\mu(\tilde{X}_{i+1}) \exp[-\beta E_k(\tilde{P}_{i+1})]}{\mu(X_i) \exp[-\beta E_k(P_i)]} \right]. \quad (12.48)$$

The weighting must be accomplished to maintain *detailed balance*.

HMC methods have been quite successful for biomolecular sampling, and many extensions to general ensembles and hybrid methods have been described. See [15], for example, for a comparison of several HMC methods with various detailed balance conditions, and [596] for a shadow HMC method which improves sampling while adding modest computational cost.

#### 12.6.4 Parallel Tempering and Other MC Variants

One idea that has gained popularity involves temperature jumps to accelerate sampling. This idea of *replica exchange* is similar to simulated annealing but here multiple simulations of non-interacting systems are involved [508, 1237]. Specifically, replicas (multiple copies) of the system that do not interact with one another are simulated at different temperatures, and systems from different ensembles are periodically exchanged with a transition probability that maintains each temperature's equilibrium ensemble distribution:

$$p_{\text{acc, MC-PT}} = \min [1, \exp((\beta_j - \beta_i)(E_j - E_i))] \quad (12.49)$$

where  $\beta_i = 1/(k_B T_i)$ .

REM, however, requires many parallel processors and a careful protocol of ensemble exchange and parameter selection to obtain rapid sampling and convergence. These issues have been addressed mostly in the dynamics analog of REM termed REMD. See a separate discussion of these issues in the MD chapters.

An energy-restricted multiple random walk MC protocol with similar advantages to REM tempering in escaping from local barriers was introduced by Wang & Landau [1328]. The method performs multiple random walks in energy space, each to sample a different range of energy; the resulting information is combined

to produce canonical averages for calculating thermodynamic quantities at any temperature. Both this method and REM perform similarly for protein conformational sampling and are faster by two orders of magnitude when compared to a canonical MC simulation at a low temperature. However, the Wang/Landau MC method was found to be easier to implement on single-processor systems, while parallel tempering is advantageous for multi-processor machines [623].

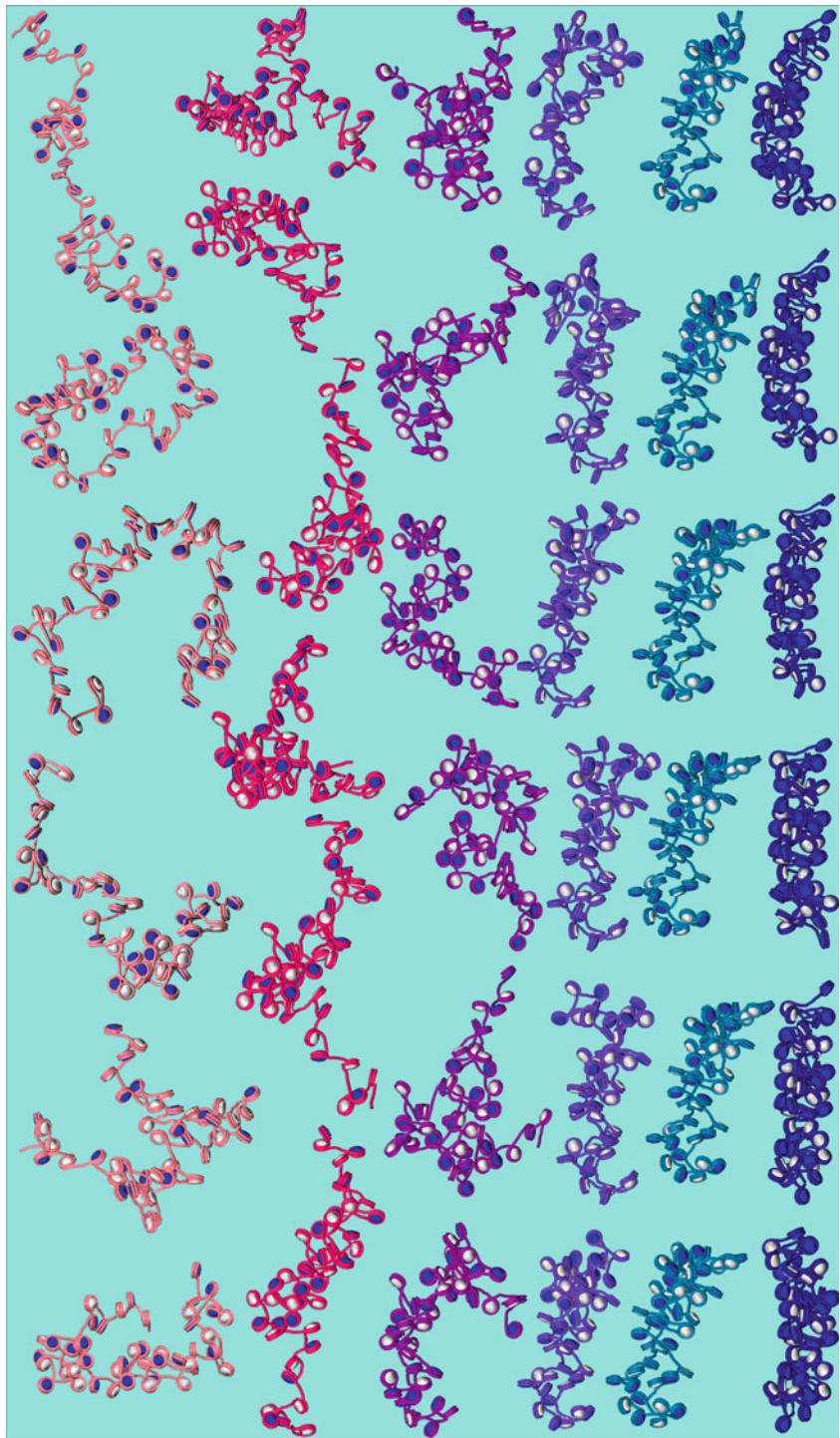
Additional MC variants include *Smart MC* [1162], stochastic dynamics, umbrella sampling, various methods that incorporate MC elements to enhance barrier crossing events (e.g., the transition path sampling method of Chandler and coworkers [147]), and potential-modification approaches such as smoothing and potential biasing techniques. See separate section for enhanced sampling and recent reviews [1117, 1117].

Given recent successes of MC methods for biomolecules [347, 781, 1457], some advocate that recent improvements in MC methodology and increased computer memory and speed lend support for the increased application of MC algorithms for folding small biomolecules. Indeed, canonical, multi-canonical, biased MC protocols that incorporate experimental information (e.g., knowledge-based dihedral angle distributions, and hybrids involving global optimization techniques and MD) can significantly enhance the sampling of low energy configurations and reveal folding ensembles of small proteins.

General and flexible MC modules have been built into standard programs like CHARMM [575], with automatic optimization of step sizes and efficient combinations with minimization or MD modules. These optimized MC methods were found to outperform standard Langevin dynamics simulations in reaching folded states of small proteins.

In general, MC methods can become inefficient for large systems but can certainly be effective for coarse grained methods (e.g., chromatin folding [64, 1240]) and as vital components of other methods (e.g., transition path sampling [147]). With increased usage and interest in coarse-grained models for larger biomolecular complexes, further development of MC methods, including the various extensions and hybrids described above, is warranted for biomolecular applications as a whole.

There has been continued discussion on the relative merits of MC and MD for biomolecules (e.g., [247, 619]). The limits of MD for simulating slow molecular events are clearly recognized, as is the high computational cost, though only MD can ultimately yield detailed dynamic information such as folding pathways and rates of conformational changes. Today, many enhanced sampling tools are available (discussed separately). Thus, a practitioner is well-advised to use a combination of MC, MD, local and global minimization algorithms, and enhanced sampling approaches as appropriate, for the problems at hand.



# 13

## Molecular Dynamics: Basics

### Chapter 13 Notation

SYMBOL	DEFINITION
<b>Matrices</b>	
<b>M</b>	mass matrix
<b>P</b>	pressure tensor (stress)
<b>Vectors</b>	
$g(X)$	constrained dynamics vector, components $g_i = r_{jk}^2 - \overline{r_{jk}}^2$
$\mathbf{r}_i$	position vector for atom $i$
$\mathbf{v}_i$	velocity vector component corresponding to atom $i$
$F$	force
$\ddot{F}$	acceleration ( $\ddot{X}$ ), or $\mathbf{M}^{-1} F(X(t))$
$F_i$	force on particle $i$ due to all particles
$V, \dot{V}, \ddot{V}$	velocity = $\dot{X}$ (components $\{v_i\}$ ), and its first and second time derivatives
$V_0$	initial velocity
$X, \dot{X}, \ddot{X}$	position (components $\{x_i\}$ ), and its first and second time derivatives
$X_0$	initial position
$\Lambda$	Lagrange multiplier vector for constrained dynamics (components $\lambda_i$ )
$\nabla E$	energy gradient
<b>Scalars &amp; Functions</b>	
$c$	speed of light
$c_p, c_t$	coordinate and velocity scaling parameters (for constant pressure and temperature simulations)
$h$	Planck's constant
$m$	particle mass

Chapter 13 Notation Table (continued)

SYMBOL	DEFINITION
$m_p, m_t$	masses for fictitious thermostat and barostat dynamic variables
$r_{ij}, \bar{r}_{ij}$	interatomic distance and associated equilibrium value
$v_{\text{cm}}$	velocity of the center of mass
$x_p, x_t$	dynamic variables for fictitious thermostat and barostat
$E_k$	kinetic energy
$E_p$	potential energy
$H, \tilde{H}$	Hamiltonian and nearby Hamiltonian
$N$	number of atoms
$N_F$	number of degrees of freedom
$P, P_0$	pressure, target pressure
$S_h$	harmonic bond stretching constant
$T, T_0$	temperature, target temperature
$\beta$	isothermal compressibility
$\gamma_t$	thermostat coupling parameter
$\delta(t)$	Lyapunov exponent
$\epsilon$	trajectory perturbation
$\zeta_p, \zeta_t$	thermodynamic coefficients for constant pressure and temperature simulations
$\lambda$	wavelength ( $1/\lambda$ is wave number)
$\nu, \omega$	characteristic frequencies
$\nu_l$	volume variable
$\tau$	thermostat coupling parameter
$\Delta t, \Delta t_m, \Delta \tau$	timesteps

Time is defined so that motion looks simple.

John Archibald Wheeler, American theoretical physicist (1911–2008).

Molecular dynamics is clearly on the way to being a universal tool, as if it were the differential calculus.

Sir John Royden Maddox [811], British science writer and *Nature* editor (1925–2009).

## 13.1 Introduction: Statistical Mechanics by Numbers

### 13.1.1 Why Molecular Dynamics?

Molecular dynamics (MD) simulations represent the computer approach to statistical mechanics. As a counterpart to experiment, MD simulations are used to estimate equilibrium and dynamic properties of complex systems that cannot be calculated analytically. Representing the exciting interface between theory and experiment, MD simulations occupy a venerable position at the crossroads of mathematics, biology, chemistry, physics, and computer science.

The static view of a biomolecule, as obtained from X-ray crystallography for example — while extremely valuable — is still insufficient for understanding a wide range of biological activity. It only provides an average, frozen view of a complex system. Certainly, molecules are live entities, with their constituent atoms continuously interacting among themselves and with their environment. Their dynamic motions can explain the wide range of thermally-accessible states of a system and thereby connect sequence to structure and function. Thus, by following the dynamics of a molecular system in space and time, we can obtain a rich amount of information concerning structural and dynamic properties. Though considered obvious today, relating motion to function of proteins through sampling rugged energy landscapes was innovative when first described by Frauenfelder and Wolynes [422, 1386].

Indeed, following the dynamics of molecular systems can provide valuable information concerning molecular geometries and energies; mean atomic fluctuations; local fluctuations (like formation/breakage of hydrogen bonds, water/solute/ion interaction patterns, or nucleic-acid backbone torsion motions); rates of configurational changes (ring flips, nucleic-acid sugar repuckering, diffusion), enzyme/substrate binding; free energies; and the nature of various types of concerted motions. Ultimately perhaps, large-scale deformations of macromolecules such as protein folding might be simulated, as discussed in the first chapter. This formidable aspect, however, is more likely to be an outgrowth of hand-in-hand advances in both experiment and theory, not to speak of high-end computing.

Though the MD approach remains popular because of its essential simplicity and physical appeal (see below), it complements many other computational tools for exploring molecular structures and properties, such as Monte Carlo simulations, Poisson-Boltzmann analyses, and energy minimization, as discussed in preceding chapters, and Brownian dynamics and enhanced sampling methods, as described in the next chapter. See examples in Table 13.1 and Figure 13.1. Each technique is appropriate for a different class of problems.

### 13.1.2 *Background*

A solid grounding in classical statistical mechanics, thermodynamic ensembles, time-correlation functions, and basic simulation protocols is essential for MD practitioners. Such a background can be found, for example, in the books by McQuarrie [853], Allen and Tildesley [22], and Frenkel and Smit [428]. Basic elements of simulation can be learned from the book of Ross [1067] and Bratley, Fox and Schrage [165]. MD fundamentals, as well as advanced topics, are also available in texts by Allen and Tildesley [22], Frenkel and Smit [428] and Berendsen [122]. Basic introductions and useful examples in Gould and Tobochnik [474], Rapaport [1038], Haile [494], and Field [398]. Some of these texts also describe analysis tools for MD trajectories, a topic not considered here.

Good advanced texts for Hamiltonian dynamics and integration schemes are those by Leimkuhler and Reich [731] and by Sanz-Serna and Calvo [1090]. Two

Table 13.1. Selected biomolecular sampling methods. Continuum solvation includes empirical constructs, generalized Born models, stochastic dynamics, or Poisson Boltzmann solutions, as discussed in Chapter 10. Targeted MD involves generating hypothetical paths between two conformational endpoints [389]. See Figure 13.1 for illustrations.

Method	Pros	Cons	CPU
• <i>Molecular Dynamics</i> (MD)	continuous motion, experimental bridge between structures and macroscopic kinetic data	expensive; short timespan	high
• <i>Targeted MD</i> (TMD)	connection between two states; useful for ruling out steric clashes and suggesting high barriers	not necessarily physical	moderate
• <i>Continuum Solvation</i>	mean-force potential approximates environment and reduces model's cost; useful information on ionic atmosphere and intermolecular associations	approximate	moderate
• <i>Brownian Dynamics</i> (BD)	large-scale and long-time motion	approximate hydrodynamics; limited to systems with small inertia	moderate
• <i>Monte Carlo</i> (MC)	large-scale sampling; useful statistics	move definitions are difficult; unphysical paths	low
• <i>Minimization</i>	valuable equilibria information; experimental constraints can be incorporated	no dynamic information	low

books written in the early days of biomolecular simulations are by McCammon and Harvey [846] and by Brooks, Pettitt and Karplus [178]. And there are always good reviews of MD from time to time (e.g., [4, 636, 658]).

Given these many sources for the theory and application of MD simulations, the focus in this chapter and the next is on providing a broad introduction into the computational difficulties of biomolecular MD simulations and some long-time integration approaches to these important problems. The reader is advised to complement these aspects with the more standard topics available elsewhere.

### 13.1.3 Outline of MD Chapters

We begin by describing in Section 13.2 the roots of molecular dynamics in Laplace's 19th-century vision. We then describe two basic limitations of Newtonian (classical) mechanics: determinism, and failure to capture quantum effects (or electronic motion). We elaborate on the latter by noting that the classical physics approximation is especially poor at lower temperatures and/or higher frequencies. Section 13.3 discusses general issues in molecular and biomolecular dynamics simulations, such as setting up the system (initial conditions, solvent and ions, equilibration), estimating temperature, and other simple aspects of the simulation protocol (equilibration, nonbonded interaction handling). We also

elaborate upon the inherent chaotic (but bounded) behavior of biomolecules first mentioned in Section 13.2 by demonstrating sensitivity to initial conditions. We motivate various methods for reducing the computational time of dynamic simulations by emphasizing the large computational times required for biomolecules, as well as extensive data-analysis and graphical requirements associated with the voluminous data that are generated.

Given the large computational requirements, we introduce in the same overview section the concepts of accuracy, stability, symplecticness, and multiple-timesteps (MTS) that are key to developing numerical integration techniques for biomolecular dynamics simulations. In this context, we also mention constrained dynamics and rigid-body approximations.

*Though we refer to several of the methods in this introductory section (Section 13.3) that we only define later, a detailed understanding is not needed for grasping the essentials that are communicated here. I recommend that students re-read Section 13.3 after completing the two MD chapters.*

Section 13.4 describes the symplectic Verlet/Störmer method and develops its variants known as leapfrog, velocity Verlet, and position Verlet. This material is followed by extension of Verlet to constrained dynamics formulations (Section 13.5) and a reference to various MD ensembles (Section 13.6).

The next chapter introduces more advanced topics on various integration approaches for biomolecular dynamics; the material is suitable for students with a good mathematical background. Specifically, we discuss symplectic integration methods, MTS (or force splitting) methods — via *extrapolation* and *impulse* splitting techniques — and introduce the notion of resonance artifacts. Enhancing our understanding of resonance artifacts in recent years has been important for developing large-timestep methods for MD. We continue to describe methods for Langevin dynamics, Brownian dynamics, implicit integration, and enhanced sampling.

## 13.2 Laplace's Vision of Newtonian Mechanics

### 13.2.1 The Dream Becomes Reality

Besides “statistical mechanics by numbers”, I like to term MD as “Laplace’s vision of Newtonian mechanics on supercomputers”.

Sir Isaac Newton (1642–1727) described in 1687 in his *Principia* masterpiece (*Philosophiae Naturalis Principia Mathematica*) the grand synthesis of basic concepts involving force and motion. This achievement was made possible by the large pillars of empirical data laid earlier by scientists who studied the motion of celestial and earthly bodies: Galileo Galilei (1564–1642), Nicolas Copernicus (1473–1543), Tycho Brahe (1546–1601), and Johannes Kepler (1571–1630). Newton’s second law of motion — that a body’s acceleration equals the net force divided by its mass — is the foundation for molecular dynamics.

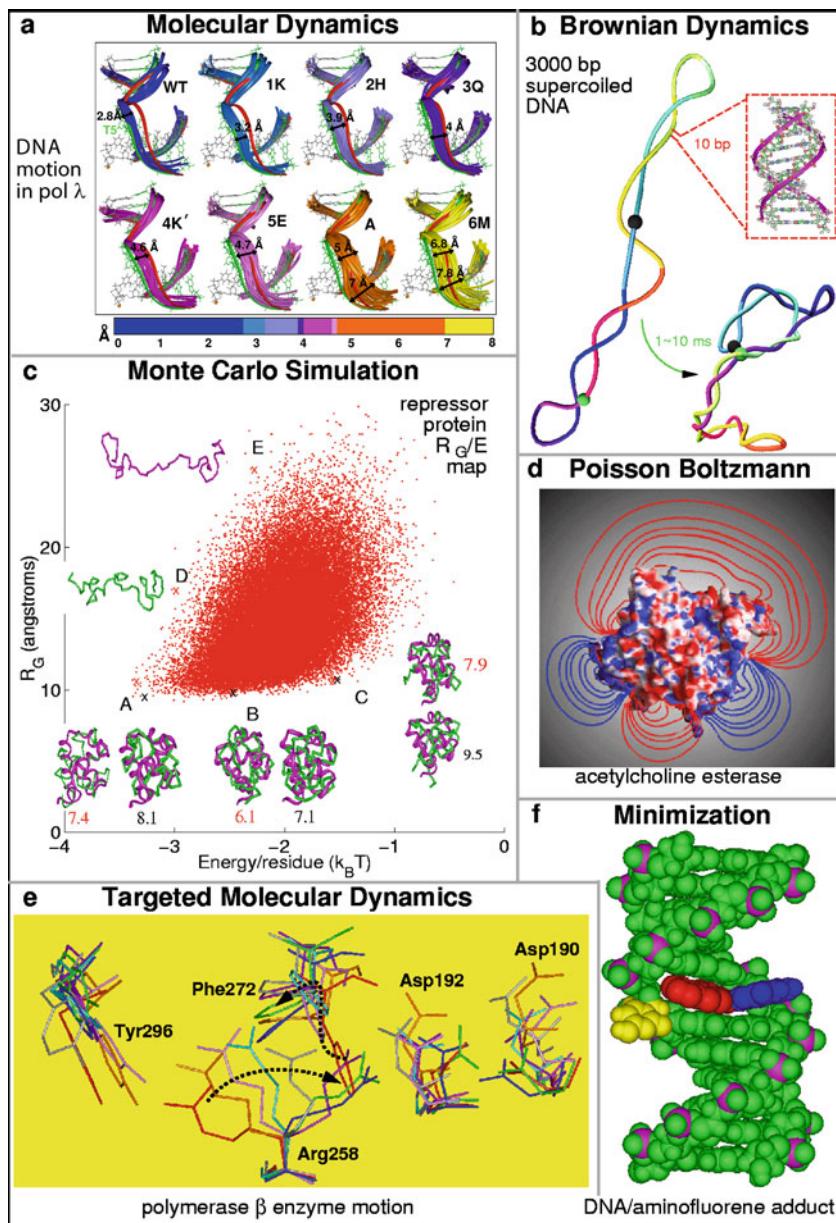


Figure 13.1. Illustrative images for selected biomolecular sampling methods. See caption on the following page.

**Caption to Figure 13.1:** (a) MD simulations of 7 single-residue variants at residue Arg517 of a polymerase  $\lambda$ /DNA complex illustrate this residue's importance in stabilizing the complex, as evident by the wide residue-dependent fluctuations captured in dynamics simulations [413]. The DNA snapshots are superimposed with respect to the active (green) and inactive (red) DNA positions. (b) BD snapshots of long DNA capture large-scale motions, such as site-juxtaposition kinetics (see Box 6.7 of Chapter 6) of two segments located 900-bp along the DNA contour, in the millisecond timeframe [577]. (c) MC configurations and ensemble radius of gyration/energy plot from a folding simulation of 434-repressor protein show RMS values for predicted structures, superimposed with native conformations [437]. (d) PB contours illustrate the electrostatic potential of mouse acetylcholinesterase (D. Sept, K. Tai, and J.A. McCammon, personal communication, and see [356, 1253]). (e) TMD snapshots of polymerase  $\beta$  capture the large-scale conformational change of the enzyme from closed (red) to open (green) forms [1408] shown here as reflected by motion of several residues near the active site. (f) The minimized adduct of a DNA duplex with a carcinogen (2-aminoanthracene) was computed with NMR constraints [969].

---

The celebrated French mathematician Pierre Simon de Laplace (1749–1827) recognized the far-reaching implications of Newtonian physics almost two centuries ago. In this now-classic piece, Laplace dreams about predicting the future, as well as reproducing the past, by animating Nature's forces (see Box 13.1) [290].

While stated in an era when the effect of high-speed computers on modern life and science could hardly be imagined, Laplace's vision already contains the essential ingredients of present-day biomolecular simulations: mathematical construction of the suitable force field (*forces*), design of appropriate numerical integration tools (*analysis*), and long-time propagation of the equations of motion (*far-reaching intelligence*).

One hundred years later, the British theoretical physicist Paul Dirac (1933 Nobel Laureate in Physics with Erwin Schrödinger for pioneering contributions in quantum physics), acknowledged that we now have formulations for Nature's forces but lamented that these equations are too complex to solve [320]:

The fundamental laws necessary for the mathematical treatment of a large part of physics and the whole of chemistry are thus known, and the difficulty lies only in the fact that application of these laws leads to equations that are too complex to be solved.

Indeed, the energy landscape is complex for biomolecules [422, 1386, 1387]. The various contacts — be they hydrogen bonds, disulfide bonds, or noncovalent interactions like stacking and favorable electrostatics — are difficult to predict *a priori*. Thus, the multidimensional potential energy surface that governs biomolecular structure has many maxima, minima, and saddle points. The distributions about each favorable or unfavorable state are highly anisotropic, with the width depending on the entropy associated with that state.

Biomolecules are also asymmetric in comparison to simple systems, such as homogeneous liquid clusters, which were successfully simulated much earlier. Certainly, there are symmetries in many aspects of protein and nucleic acid structure (e.g., many proteins are dimers, and the “ideal” DNA double helix has an axis of symmetry), but in realistic environments there are many sequence-specific motifs and binding interactions

However, modern computers have overcome the limitation noted by Dirac, allowing us to solve the complex equations!

### 13.2.2 Deterministic Mechanics

As Laplace stated, the capability to analyze and predict motion — be it of the solar system or a biological system — provides the link between the past and the future. Still, even Newtonian mechanics taken to its extreme cannot predict with certainty the future motion of all bodies. As became evident by the work of Poincaré less than a century after Laplace’s statement, the solar system is chaotic even though the underlying laws can be clearly expressed. This understanding, however, should not deter us from pursuing Laplace’s dream; rather, it should stimulate us to explore as deeply as possible the consequences of Newtonian physics.

### 13.2.3 Neglect of Electronic Motion

In addition to the limitation of deterministic mechanics, only the *nuclear* motion of many-body systems is typically followed in molecular dynamics. Thus, electronic motion is not considered, and quantum effects are generally ignored. The classical approximation is excellent for a wide range of systems and materials but is unsuitable for reactions involving electronic rearrangements such as bond formation and cleavage, polarization, and chemical bonding of metal ions. Quantum dynamical approaches are used for this purpose. They are, however, at a relatively early stage with respect to macromolecular applications and are not covered in this text.

#### **Box 13.1: Laplace’s ‘Far-Reaching Intelligence’**

Laplace, the son of a Normandy farmer, is famous for his masterpieces on Celestial Mechanics and the Theory of Probability [290]. The Laplace equation (though written by Euler in 1752 in a hydrodynamic context) is Laplace’s chief contribution to potential theory. In his 1820 *oeuvre* [290], Laplace states:

*Une intelligence qui, pour un instant donné, connaît toutes les forces dont la nature est animée et la situation respective des êtres qui la composent, si d’ailleurs elle était assez vaste pour soumettre ces données à l’Analyse, embrasserait dans la même formule les mouvements des plus grands corps de l’univers et ceux du plus léger atome: rien ne serait incertain pour elle,*

*et l'avenir, comme le passé, serait présent à ses yeux. L'esprit humain offre, dans la perfection qu'il a su donner à l'Astronomie, une faible esquisse de cette intelligence.*

*An intelligence which could, at any moment, comprehend all the forces by which nature is animated and the respective positions of the beings of which it is composed, and moreover, if this intelligence were far-reaching enough to subject these data to analysis, it would encompass in that formula both the movements of the largest bodies in the universe and those of the lightest atom: to it nothing would be uncertain, and the future, as well as the past, would be present to its eyes. The human mind offers us, in the perfection which it has given to astronomy, a faint sketch of this intelligence.*

---

### 13.2.4 Critical Frequencies

Classical MD simulations are also unsuitable for low temperatures, where the energy gaps among the discrete levels of energy dictated by quantum physics are much larger than thermal energy available to the system. This is because the system is confined to one or a few of the low-energy states under such conditions. This discrete description of energy states becomes less important as the temperature is increased and/or the frequencies associated with motion are decreased (i.e., have longer timescales). Under those conditions, more energy states become thermally accessible.

Rough estimates for the characteristic motions for which Newtonian physics is reasonable can be made on the basis of harmonic analysis. For a harmonic oscillator, the quantized energies are separated by  $h\nu$  where  $h$  is Planck's constant and  $\nu$  is the vibrational frequency.

Clearly, the classical approach is unsuitable for capturing motions with relatively high frequencies  $\nu$ , that is with  $\nu \gg k_B T/h$ , or

$$\frac{h\nu}{k_B T} \gg 1, \quad (13.1)$$

where  $k_B$  is Boltzmann's constant, and  $T$  is the temperature. This is because the probability of finding the system with this mode at the ground state energy is high. The larger this ratio, the greater this probability.

Conversely, classical behavior is approached for frequency/temperature combinations for which

$$\frac{h\nu}{k_B T} \ll 1. \quad (13.2)$$

Around the room temperature of 300 K,  $k_B T = 0.6$  kcal/mol. As we see from Table 13.2, the high-frequency vibrational modes present in biomolecules

Table 13.2. Ratios for some high-frequency vibrational modes at T = 300 K.

Vibrational mode	Wave number (1/λ) [cm <sup>-1</sup> ]	Frequency ν = c/λ [s <sup>-1</sup> ]	Ratio hν/(k <sub>B</sub> T)
O–H stretch	3600	1.1 × 10 <sup>14</sup>	17
C–H stretch	3000	9.0 × 10 <sup>13</sup>	14
O–C–O asym. stretch	2400	7.2 × 10 <sup>13</sup>	12
C=O (carbonyl) stretch	1700	5.1 × 10 <sup>13</sup>	8
C–N stretch (amines)	1250	3.8 × 10 <sup>13</sup>	6
O–C–O bend	700	2.1 × 10 <sup>13</sup>	3

have ratios larger than unity. They are thus not well treated by classical physics. Specifically, Newtonian physics, which distributes the energy equally among all vibrational modes according to the equipartition theorem, overestimates the energy/motions associated with these high-frequency modes.

The critical frequency  $\nu$  for which equality of the above ratio holds is around  $6.25 \times 10^{12}$  s<sup>-1</sup> or 6 ps<sup>-1</sup>, corresponding to an absorption wavelength of 208 cm<sup>-1</sup> (or period of 160 fs). Thus, modes with characteristic timescales in the picosecond and longer timeframes are reasonably treated by classical, Newtonian physics. The second and third classes of motions identified for biomolecules in Table 13.3 fall in this range.

Table 13.3. The broad spectrum of characteristic timescales in biomolecules.

Internal Motion	Timescale [seconds]
Light-atom bond stretch	10 <sup>-14</sup>
Double-bond stretch	2 × 10 <sup>-14</sup>
Light-atom angle bend	2 × 10 <sup>-14</sup>
Heavy-atom bond stretch	3 × 10 <sup>-14</sup>
Heavy-atom angle bend	5 × 10 <sup>-14</sup>
Global DNA twisting	10 <sup>-12</sup>
Sugar puckering (nucleic acids)	10 <sup>-12</sup> –10 <sup>-9</sup>
Collective subgroup motion (e.g., hinge bending, allosteric transitions)	10 <sup>-11</sup> –10 <sup>-7</sup>
Surface-sidechain rotation (proteins)	10 <sup>-11</sup> –10 <sup>-10</sup>
Global DNA bending	10 <sup>-10</sup> –10 <sup>-7</sup>
Site-juxtaposition (superhelical DNA)	10 <sup>-6</sup> –1
Interior-sidechain rotation (proteins)	10 <sup>-4</sup> –1
Protein folding	10 <sup>-5</sup> –10

### 13.2.5 Hybrid Quantum/Classical Mechanics Treatments

Electronic motions which have much higher characteristic frequencies must be treated by alternative approaches such as quantum-mechanical and hybrid quantum/classical approximations; excellent progress has been made in recent years on these techniques as discussed in a previous chapter [75, 270, 441, 452, 572–574, 628, 1163–1165, 1342, 1348, 1442]. Essentially, in these approaches, the reacting part of the system (e.g., active site of an enzyme) is treated quantum mechanically, while the other components (e.g., remaining amino acids and solvent) are modeled classically, by molecular mechanics, as first proposed in the late 1970s [1344]. Such treatments are critical for calculating reaction pathways and intermediates for bond-breaking events, reaction rates for formation of organic compounds in solution, or free energies of hydration. They are also necessary for describing localized enzymatic activity entailing charge transfer, or solvent polarization effects. Important recent progress in this area was mentioned in Chapter 8. Still, many technical details must be perfected regarding these more complex simulation protocols. For example, challenging are the proper definition of the quantum-mechanical treatment and the merging between the quantum and classical approximations and the proper sampling involved to compute accurate reaction rates.

Besides resorting to QM/MM techniques, many innovative sampling methods have now been developed that can estimate reaction mechanisms and rates for conformational pathways. These are mentioned in the end of the next chapter.

## 13.3 The Basics: An Overview

### 13.3.1 Following the Equations of Motion

The molecular dynamics approach is simple in principle. We simulate motion of a system under the influence of a specified force field by following molecular configurations in time according to Newton's equation of motion. We write these equations for a system of  $N$  atoms as the following pair of first-order differential equations:

$$\begin{aligned} \mathbf{M}\dot{\mathbf{V}}(t) &= \mathbf{F}(\mathbf{X}) = -\nabla E(\mathbf{X}(t)) + \dots, \\ \dot{\mathbf{X}}(t) &= \mathbf{V}(t). \end{aligned} \quad (13.3)$$

In these equations,  $\mathbf{X} \in \mathbf{R}^{3N}$  denotes the collective Cartesian vector of the system (i.e., the  $x$ ,  $y$ , and  $z$  components of each atom are listed in turn);  $\mathbf{V}$  is the corresponding collective velocity vector;  $\mathbf{M}$  is the diagonal mass matrix (i.e., the masses of each atom are repeated three times in the diagonal array of length  $3N$ ); and the dot superscripts denote differentiation with respect to time,  $t$ .

The total force  $\mathbf{F}$  in the right-hand-side of eq. (13.3) is composed of the systematic force, which is the negative gradient (vector of first partial derivatives) of the potential energy  $E$  and, possibly, additional terms that mimic the environment.

(See section 14.4 on stochastic dynamics for an example of these additional terms). Each gradient component  $i$ ,  $i = 1, \dots, 3N$ , is given by:

$$\nabla E(X)_i = \partial E(X) / \partial \alpha_i,$$

where  $\alpha_i$  denotes an  $x$ ,  $y$ , or  $z$  component of an atom. These equations must be integrated numerically since analytic (closed-form) solutions are only known for the simplest systems. Such numerical integration generates a sequence of positions and velocity pairs,  $\{X^n, V^n\}$ , for integers  $n$  that represent discrete times  $t = n\Delta t$  at intervals (timesteps)  $\Delta t$ .

### 13.3.2 Perspective on MD Trajectories

#### Force Field Dependency

Results of a molecular dynamics simulation can only be as good as the governing force field. Essentially, the mechanical representation of a system — particles connected by springs — assumes simple, pairwise-additive potentials. These express how the composite atoms stretch, vibrate, and rotate about the bonds in response to intramolecular and intermolecular forces. The resultant potential energy  $E$ , as described in Chapters 9 and 10, is still highly approximate for biomolecules and undergoes continuous improvements. Uncertainties are well recognized in the representation of solvent [1099], polarization effects [510], and electrostatic interactions [455, 668], and in the functional form of the local potentials (i.e., lack of anharmonic [826] and cross terms [300]).

#### Statics Vs. Dynamics

Minimization of this approximate energy function yields information on favorable regions in configuration space (this approach is termed molecular mechanics or statics). The numerical integration of the differential equations of motion reveals the intrinsic motions of the system under the influence of the associated force field. Thus, in principle, MD simulations can combine both the spatial and temporal aspects of conformational sampling. They are thus used in many cases as conformational search tools — to bypass the multiple-minimum problem — and as vehicles to refine low-resolution X-ray or NMR (nuclear magnetic resonance) data (e.g., [676]). However, for this purpose, special strategies, such as energy modifications or high-temperature settings, are required if substantial movements are desired. This is because MD simulations are severely limited by the very small timesteps that must be used relative to the timescales of major biological interest.

#### Range of Timescales

Indeed, the motion of biomolecules involves an *extraordinary* range of timescales (see Table 13.3). In general, the higher frequencies have smaller associated amplitude displacements. For example, while bond vibrations have characteristic

amplitudes of a tenth of an Ångstrom, global deformations can be in the order of 100 Å (see Table I in [178] for classes of timescales and associated amplitudes in biomolecular motions). The energies associated with these motions also span a large range.

Though the fastest, high-frequency modes have the smallest amplitudes, these motions affect other modes and thus their effect must be approximated in some way. Yet, the existence of high-frequency modes severely affects the timestep that can successfully be used in biomolecular simulations. Their timescale dictates a timestep of 1 fs or less in standard explicit schemes for acceptable resolution (a tenth or less of a period). This stepsize already implies one million steps to cover only a nanosecond and falls short, by more than *ten orders of magnitude*, of the slow and large-amplitude processes of major biological interest. Dealing with this severe timestep problem has been the focus of many research groups, and some results will be the subject of sections that follow.

### Challenges

Thus, in summary, although the basic idea is simple, the art of MD simulations is challenging in practice. The practical difficulties arise from various components that enter into biomolecular simulations: setting initial conditions, implementing various simulation protocols to ensure reliability, using suitable numerical integrators, considering the sensitivity of trajectories to initial conditions and other choices, meeting the large computational requirements, and visualizing and analyzing the voluminous data that are generated.

#### 13.3.3 Initial System Settings

A molecular dynamics trajectory consists of three essential parts: initialization, equilibration, and production. Initialization requires specifying the initial coordinates and velocities for the solute macromolecule, as well as for the solvent and ion atoms.

##### Structure

Even when initial coordinates are available from experiment (e.g., crystal structure), the starting vector may not correspond to a minimum in the potential energy function used, and hence minimization (further refinement) is needed to relax strained contacts. When an experimental structure is not available, a *build-up* technique may be used to construct a structure on the basis of the known building blocks, and minimization again is required.

##### Solvation

When water molecules and salt ions are also used, special care is needed to carve an appropriate solvation model around the biopolymer. The water coordinates

are typically taken from a pure water simulation (which reproduces the experimentally determined density), with initial ion placement guided by the biopolymer charge distribution and experimental measurements. Special iterative ion-placement algorithms are available in molecular dynamics programs that place positive ions in electronegative pockets and negative ions in positively-charged cavities so as to achieve a desired total ionic concentration. The final model is constructed following appropriate superimpositioning, removal of overlapping atoms, and energy relaxation. See [1419, 1421], for example, for detailed descriptions of such equilibration procedures for B-DNA and in [525, 526] for RNA.

### Velocity

The initial velocity vector is typically set pseudorandomly so that the total kinetic energy of the system,  $E_k$ , corresponds to the expected value at the target temperature  $T$ . According to the classical equipartition theorem, each normal mode has  $(k_B T)/2$  energy, on average, at thermal equilibrium. Thus

$$\langle E_k \rangle = \frac{1}{2} \sum_{i=1}^{3N} m_i v_i^2 \equiv \frac{1}{2} (V^0)^T \mathbf{M}(V^0) = (N_F k_B T)/2, \quad (13.4)$$

where  $N_F$  is the total number of degrees of freedom in the system ( $3N$  or  $3N - 3$ ; see below).

Such a velocity setting can be accomplished by assigning velocity components from a Gaussian distribution. Namely, each component  $v_i$  can be chosen from a Gaussian distribution with zero mean and a variance of  $(k_B T)/m_i$ . This implies that  $\langle v_i^2 \rangle = (k_B T)/m_i$ , where the bracket  $\langle \cdot \rangle$  denotes the expected value. Utilities to generate such distributions are based on uniform pseudorandom number generators and are available in many software packages (see discussion in the Monte Carlo chapter).

According to the *Central Limit theorem* (see Chapter 12), the sum of independent random variates, each chosen from a normal distribution, has itself a normal density with a mean and variance that are simply the sum of the individual means and variances. Thus, the expected value of  $E_k$  will correspond to the target temperature:

$$\langle E_k \rangle = \frac{1}{2} \left\langle \sum_{i=1}^{3N} m_i v_i^2 \right\rangle = 3N (k_B T)/2. \quad (13.5)$$

Often, the three translational degrees of freedom are removed by modifying the initial velocity vector so that the initial momentum is zero. This is done by subtracting from each component of  $V^0$  the center of mass velocity,  $v_{cm}$  ( $v_{cm} = (\sum_i m_i v_i) / \sum_i m_i$ , where  $i = 1, \dots, N$ ). This modification leaves the final temperature, measured with  $N_F = 3N - 3$ , unchanged.

## Equilibration

Once the system is set and initial coordinates and velocities are assigned, an initial round of *equilibration* is necessary before the production phase of the simulation can begin. In this equilibration period, there is an exchange between kinetic and potential energies. Stability is evident when both terms, and the total energy, appear to converge, with fluctuations occurring about the mean value. This is the expected behavior from a *microcanonical* ensemble (fixed total energy, volume, and particle number, or NVE for short); see Chapter 12, Subsection 12.5.3 and Section 13.6.

## Illustration

Figure 13.2 illustrates the convergence phase of a simulation, as obtained for a hydrated DNA dodecamer system (12389 atoms: 760 of DNA, 22 sodium, and 3869 water molecules). A heating phase of 15,000 steps (5000 steps, 3°K per every 50 steps) first brings the system to the target temperature. Note that during the equilibration phase the potential and kinetic energies fluctuate and the total energy remains nearly constant (see enlargement at bottom). Second, the leapfrog Verlet integrator (see later in chapter) with a timestep of 2 fs is applied, in combination with bond-length restraints (see below), to all water molecules. The water model consists of a periodic hexagonal-prism system of height 65 Å and side length of 26 Å. In general, the equilibration time is system and protocol dependent. A stochastic approach such as Langevin dynamics (as introduced in Subsection 10.6.3 of Chapter 10) can simplify this equilibration task because the random forces quickly lead the system to thermal equilibrium, even with a zero initial velocity vector.

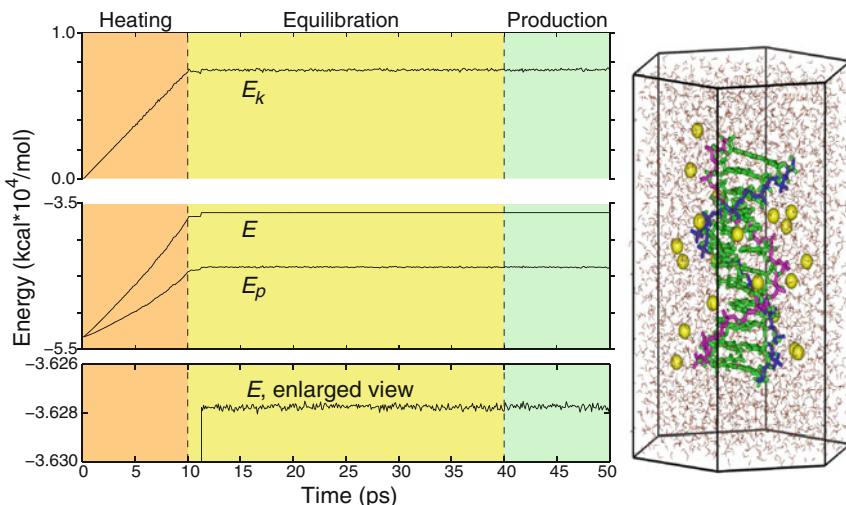


Figure 13.2. Heating and equilibration of a hydrated DNA dodecamer system of 12389 atoms in a hexagonal-prism periodic domain.

### 13.3.4 Sensitivity to Initial Conditions and Other Computational Choices

The chaotic nature of individual MD trajectories is well appreciated. It is especially important to keep this aspect in mind when analyzing data for complex, many-body systems. Roughly speaking, chaotic behavior means that a small change in initial conditions (e.g., a fraction of an Ångstrom difference in Cartesian coordinates) can lead to exponentially-diverging trajectories in a relatively short time. The larger the initial difference and/or the timestep, the more rapid this *Lyapunov instability* [849].<sup>1</sup> Though chaos is a feature of the analytical equations themselves, the finite arithmetic used by computers and the discretization errors of the integrators used to simulate the trajectory are other factors that produce divergence.

#### Chaos and Saturation

Figure 13.3 illustrates this sensitivity to initial conditions. Shown are the root-mean-square (RMS) differences in atomic coordinates as a function of time  $t$  for a water tetramer system:

$$e(t) = \left[ \sum_{i=1}^{3N} (x_i(t) - x'_i(t))^2 \right]^{1/2}; \quad (13.6)$$

here, the coordinate of the perturbed trajectory is distinguished from the reference trajectory by the prime superscript.

For each curve shown in the figure, the initial coordinates of the second trajectory were perturbed by  $\pm\epsilon$  from the reference trajectory ( $\epsilon = 10^{-8}, 10^{-6}, 10^{-4}$ ). The velocity Verlet algorithm (see below) used timesteps of 0.1 fs and 1 fs, and the difference  $e(t)$  is plotted every 20 fs, reflecting the average over that interval.

Several interesting features can be gleaned from this figure. A short period of rapid exponential divergence is evident, followed by linear oscillations around a threshold value. The ‘chaos’ often mentioned in MD texts refers to the *initial period* of exponential growth. This is in fact *local instability* [849]. It is well known from chaos theory that compact systems, like biomolecules, whose phase space is finite, reach *saturation* for the trajectory error when the magnitude of the deviations corresponds to a system-dependent value.

In other words, the phase trajectories cannot be found beyond the distance characteristic of the phase space. As the plots show, the smaller the initial perturbation ( $\epsilon$ ), the longer the time required to reach saturation. The same threshold is reached for both timestep values.

---

<sup>1</sup>Lyapunov exponents are related to the eigenvalues of the Jacobian of the transformation associated with a dynamical system; they describe the rate of growth of an instability. For example, a system that possesses local instability has a direction along which the distance between two points in phase space at time  $t$ ,  $|\delta(t)|$ , grows exponentially at the rate  $\lambda_0$ , the largest Lyapunov exponent:  $|\delta(t)| = |\delta(0)| \exp(\lambda_0 t)$ .

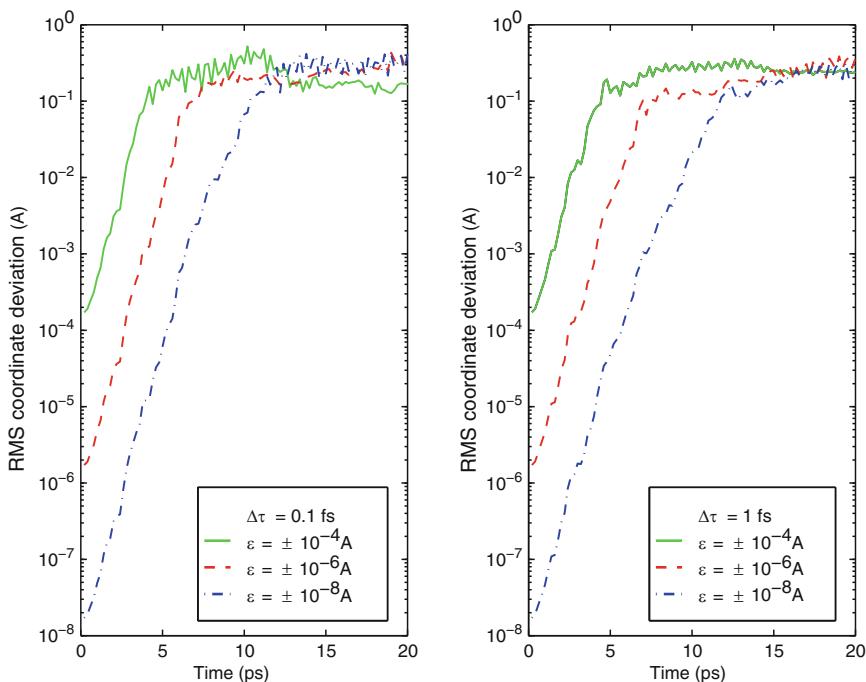


Figure 13.3. The rapid divergence of dynamic trajectories for four water molecules differing slightly in initial conditions.

This stabilization behavior realized in practice for MD simulations can also be understood by the different mathematical limit involved in the strict notion of chaos compared to that relevant for MD. Physicists define chaos rigorously for *finite times* but in the limit of *timesteps approaching zero*; in MD applications, we instead work with *finite timesteps* and simulation *times approaching infinity*.

### Statistical View

For many other reasons than inherent chaos, including the approximate nature of biomolecular simulations, MD trajectories are not intended to animate the life of a biomolecule faithfully, as if by an unbiased video camera. Rather, our experimental-by-nature computer ‘snapshots’ aim at gaining structural insights, which can be tested, and predicting meaningful *statistical properties* of a complex system in the ultimate goal of relating structure to function. The more reliable quantities from MD simulations are thus space and time averaged. This is unlike differential-equation applications in astrophysics, where precise trajectory pathways are sought (e.g., planetary orbits). Often, more useful data are obtained by averaging over, or calculating from, several trajectories rather than from one long trajectory (e.g., [158, 286, 360, 364, 1428]). This has become very suitable

now with readily available parallel processor architectures and loosely connected computer networks used for a common goal, as in the `folding@home` distributed computing initiative for simulating protein folding and related processes. Launched in 2001, `folding@home` has become an international enterprise, with clients like Sony PlayStations and numerous platforms. See also Skeel's recent analysis [1197] for a perturbation analysis on the effect of errors that aims to show why molecular dynamics simulations "work".

Let us consider the convergence of the end-to-end distance of a butane molecule, as first illustrated in [304]. Figure 13.4 shows the *time evolution* of the end-to-end-distance of the butane molecule for different choices of timesteps in the range of 0.02 to 2.0 fs. Clearly, different paths are realized even for this very small system. The associated *average end-to-end distance* values as a function of the timestep are the points corresponding to the dashed curve in Figure 13.5.

These data reveal convergence only for small timesteps, less than 0.2 fs. This value is typically *much smaller* than standard timesteps used in biomolecular simulations (0.5 fs or greater). Still, larger values can be quite reasonable depending on the specific questions being asked from a simulation.

### 13.3.5 Simulation Protocol

A careful simulation protocol is essential for MD simulations, to ensure not only the equilibration phases but also proper enforcement of the boundary conditions

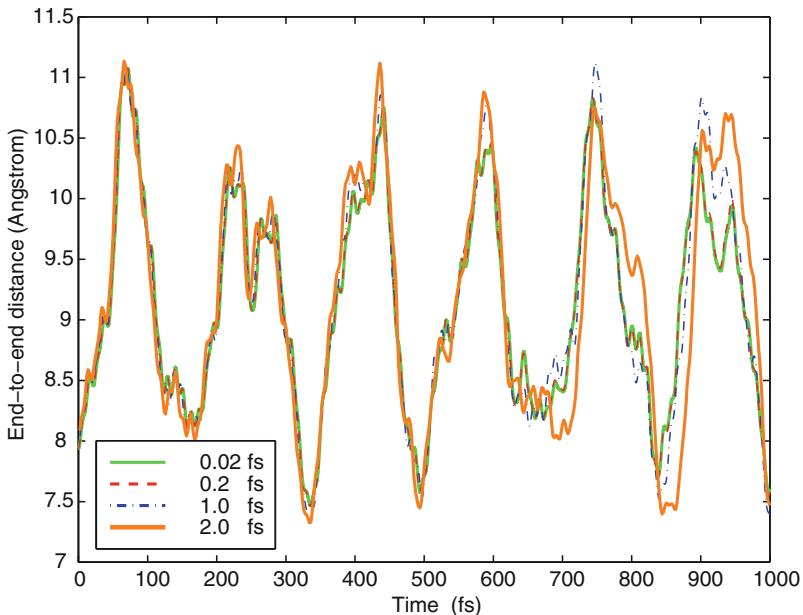


Figure 13.4. Time evolution of the end-to-end distance of butane for different timesteps.

(when, for example, a biomolecule is placed in a box of water molecules), positioning of solvent and salt molecules, or computation of the nonbonded terms and associated pairlist arrays. Also important is monitoring the kinetic temperature and the energy fluctuations, for the detection of systematic drifts or abrupt changes, both of which may indicate numerical problems.

For example, severe artifacts arise when the nonbonded energy interactions are truncated abruptly. This is because the associated force terms rise violently at the truncation boundary. Improved techniques for smoothly handling the nonbonded terms are necessary [1220], as described in Chapter 10. The alternative approach described also in that chapter is to forgo cutoffs and compute all the nonbonded terms by fast summation techniques such as multipole or Ewald [142, 486]. Fortunately, such schemes approach linear, rather than quadratic, dependence on system size.

### 13.3.6 High-Speed Implementations

High-speed computers are essential for performing the computationally-intensive MD simulations [143, 658]. The dynamics of condensed systems was simulated much earlier this century, but macromolecular applications only gained momentum in the mid-to-late 1980s with the advent of high-speed computing. There are many reasons why this lag occurred, and the issue at heart is best captured by the following statement by Frauenfelder and Wolynes [423]:

*Whatever complexity means, most people agree that biological systems have it.*

Indeed, the energy landscape is complex for biomolecules [422, 1386, 1387]. The various contacts — be they hydrogen bonds, disulfide bonds, or noncovalent interactions like stacking and favorable electrostatics — are difficult to predict *a priori*. Thus, the multidimensional potential energy surface that governs biomolecular structure has many maxima, minima, and saddle points. The distributions about each favorable or unfavorable state are highly anisotropic, with the width depending on the entropy associated with that state.

Biomolecules are also asymmetric in comparison to simple systems, such as homogeneous liquid clusters, which were successfully simulated much earlier. Certainly, there are symmetries in many aspects of protein and nucleic acid structure (e.g., many proteins are dimers, and the “ideal” DNA double helix has an axis of symmetry), but in realistic environments there are many sequence-specific motifs and binding interactions with other biomolecules in the environment that induce local structural variations in macromolecules. These local trends can produce profound global effects.

The motion of biomolecules is also more complex than that of small or homogeneous systems. The collective motion is a superposition of many fundamental motions that characterize the dynamics of a biomolecule: bond stretches, angle bends, torsions, and combinations of those “normal modes” (see Chapter 9

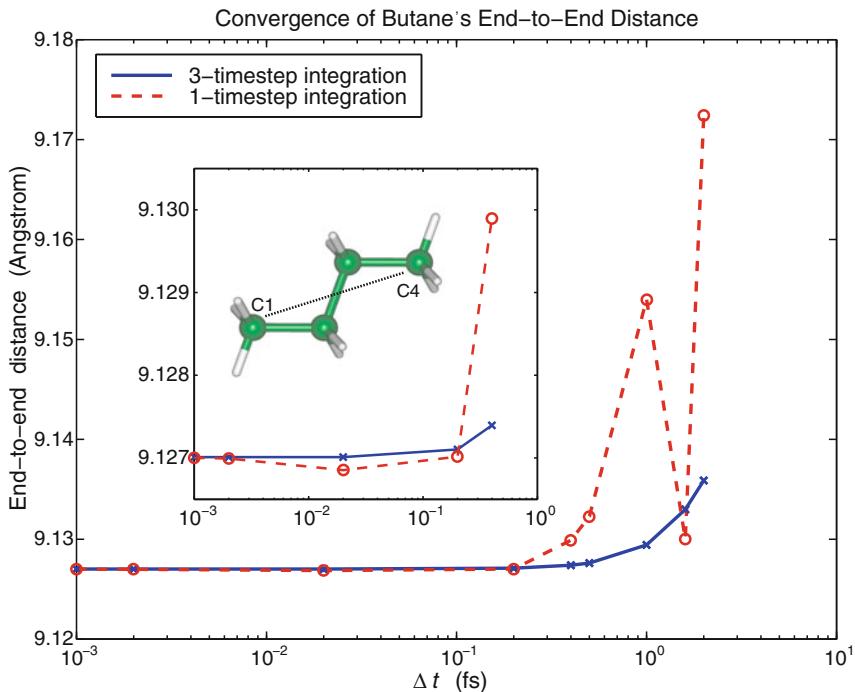


Figure 13.5. Average butane end-to-end distance for single (dashed line) and triple (solid line) timestep protocols (note logarithmic scale) as computed over 1 ps (see Figure 13.4). In the latter, the timestep used to define the data point is the outermost timestep  $\Delta t$  (the interval of updating the nonbonded forces), with the two smaller values used as  $\Delta t/2$  and  $\Delta t/4$  (for updating the dihedral-angle terms and the bond-length and angle terms, respectively).

for definitions and examples). Although these components have associated frequencies of disparate magnitudes, the overall motion of a biomolecule is highly cooperative. That is, small local fluctuations can trigger a chain of events that will lead to a large global rearrangement. Energy transfer among vibrational modes may also be involved.

In addition to the overall asymmetry and complexity (of energy landscape and motion), the solvent and ions in the environment influence the structure and dynamics of biomolecules profoundly. Solvent molecules surround and interpenetrate proteins and nucleic acids and damp many characteristic motions that might be present in vacuum. Similarly, ions (sodium, calcium, magnesium, etc.) influence structure and dynamics significantly, especially of the polyelectrolytic nucleic acids [850, for example]. Thus, for physical reliability, these long-range electrostatic interactions are essential to consider in MD simulations of macromolecules. This has proven especially important for stabilizing DNA simulations and accurately describing kinetic processes involving conformational transitions of DNA [219, 220].

### 13.3.7 Analysis and Visualization

Careful analysis of the results and visualization techniques are also essential components of biomolecular simulations today. *With the increasing ease and accessibility of generating computer trajectories, the challenge remains of carefully analyzing the voluminous data to distill the essential findings.* While many scalar and vector functions can be computed from a long series of configurations, understanding the dynamics behavior requires a combination of sophisticated analysis and chemical/biological intuition.

Molecular graphics techniques were more of a problem in the early days of molecular mechanics than they are now. Before the surge of graphics innovations, researchers relied more heavily on mechanical models and physical sense (see related comment in this book's Preface).

These days, molecular modeling software and analysis tools have become a large industry. Many sophisticated and useful tools are available to both experimentalists and computational/theoretical chemists. The dazzling capability of computer graphics today to render and animate a large, complex three-dimensional image — often so “real” in appearance that the source may be obscured — has certainly made biological interpretation much easier, but one still has to know exactly where, and for what, to look!

### 13.3.8 Reliable Numerical Integration

To increase the reliability of macromolecular simulations, special care is needed to formulate efficient numerical procedures for generating dynamic trajectories.

Mathematically, there are classes of methods for conservative Hamiltonian systems termed *symplectic* that possess favorable numerical properties in theory and practice [1090]. In particular, these schemes preserve volumes in phase space (as measured rigorously by the Jacobian of the transformation from one set of coordinates and momenta to the next). This preservation in turn implies certain physical invariants for the system.

Another view of symplectic integration is that the computed trajectory (with associated Hamiltonian  $H(X^n, V^n)$ ) remains close in time to the solution of a nearby Hamiltonian  $\tilde{H}$ . This proximity is rigorously defined as a trajectory whose energy is order  $\mathcal{O}(\Delta t)^p$  away from the initial value of the true Hamiltonian  $H$ :

$$H \equiv \frac{1}{2}(V^0)^T \mathbf{M}(V^0) + E(X^0),$$

where  $p$  is the order of the integrator, and  $X^n$  and  $V^n$  are the collective position and velocity vectors at time  $n\Delta t$ . This symplectic property translates to good long-time behavior in practice: small fluctuations about the initial (conserved in theory) value of  $H$ , and no systematic drift in energy, as might be realized by a nonsymplectic method.

Figure 13.6 illustrates this favorable symplectic behavior for a simple nonlinear system, a cluster of four water molecules, integrated by Verlet and by the classical

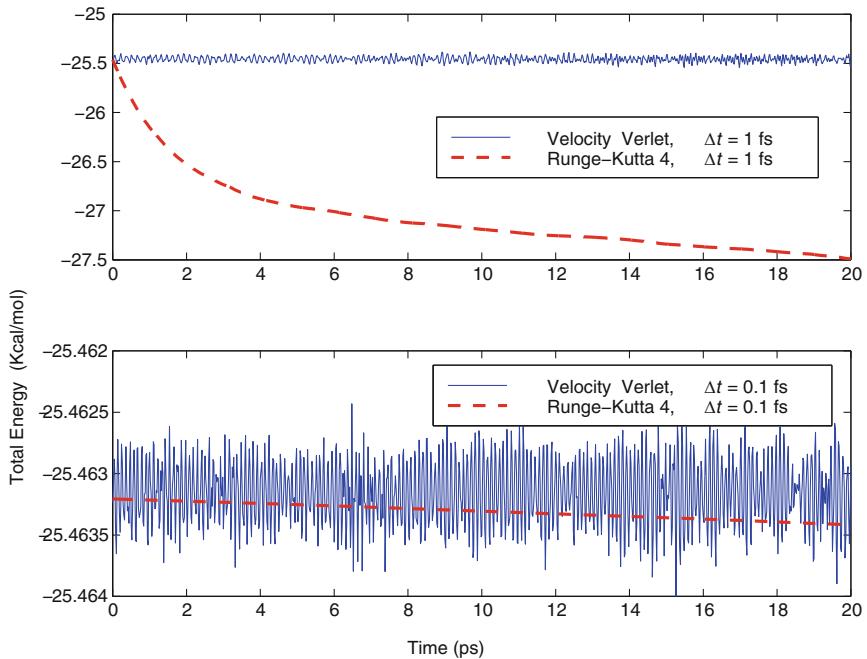


Figure 13.6. Energy evolution of a water tetramer simulated by the symplectic Verlet scheme (solid line) versus the nonsymplectic Runge-Kutta integrator (dashed line) at two timesteps  $\Delta t$  (0.1 and 1 fs). Note the different energy scales of the plots.

fourth-order (nonsymplectic) Runge-Kutta method. A clear damping trend (i.e., decrease of energy with time) is seen by the latter, non-symplectic integrator, especially at the larger timestep (top).

While symplectic methods can follow kinetics accurately (assuming the force field is adequate), other approaches as discussed in the next chapter — Langevin and Brownian dynamics — can be effective for thermodynamic sampling. Though additional terms are added (e.g., random thermal forces), enhanced sampling made possible by larger timesteps can yield more information on the thermally-accessible configuration space.

### 13.3.9 Computational Complexity

#### Intensive Requirements

Trimming the computational time of MD simulations remains a challenge [1110]. As seen from Table 1.2 of Chapter 1, the all-atom molecular models used for biomolecules are quite large (thousands of atoms). The energy and derivative computations required at each dynamic step are expensive and limit the total time that can be simulated: typical timesteps (1 fs) imply one million integration steps per nanosecond. Interestingly, Table 1.2 also shows our tendency to increase

system size rather than the simulation time. This can be explained, in part, by the desire to model more complex and realistic systems and by the timespan/sampling limitation of MD. Dynamic motions over substantially longer times, as well as larger spatial scales, can only be approximated through macroscopic models of proteins and DNA (see [1119] for example).

To reduce computational time of MD simulations, two basic routes can be taken and combined: reducing the work per timestep, and/or increasing the integration timestep,  $\Delta t$ . In the latter case, the parameter we aim to lengthen is that associated with long-range force calculations, that is, the ‘outermost timestep’ in multiple-timestep (MTS) methods.

### Less Work Per Step

Reducing the work per step can be achieved with various protocols used in practice: employing cutoff ranges for the nonbonded interactions, updating the nonbonded pairlist infrequently (i.e., less often than every outer timestep), and reducing the number of degrees of freedom by constraining certain molecular regions (e.g., protein residues revealed by experiment not to be important for a certain localized transition).

### Larger Timesteps: Accuracy vs. Stability

The maximum value for the outer timestep that can be successfully used is limited by both *accuracy* and *stability* considerations, as will be detailed below. Essentially, *instability* of a trajectory is easily detected when coordinates and energies grow uncontrollably. This often signals that the timestep is too large. However, unless the timestep is far too large, this instability may only be apparent from a trajectory of length of hundreds of picoseconds or longer. The guidelines for stability for a given integrator are often derived from linear analysis (i.e., theoretical analysis for a harmonic oscillator) [866, 967, 1437, for example]. In practice, the timestep must typically be smaller than this limit for complex nonlinear problems to ensure stability as well as reasonable accuracy.

Accuracy limits the timestep  $\Delta t$  in the sense that the resolution produced must be adequate for the processes being followed and the measurements targeted. Thus, a simulation may be stable but the timestep too large to capture high-frequency fluctuations accurately.

Closely related to stability and accuracy are resonance artifacts. This topic of recent interest will be discussed separately.

Techniques for increasing the timestep range from constraining the fastest degrees of freedom via algebraic constraints to using multiple-timestep (MTS) schemes to formulating rigid-body dynamics. For recent reviews, see [1110, 1118], for example.

### Constraining Fastest Motions

The technique of constrained dynamics has been used for a long time to enforce rigidity of bond lengths [868, 1079, 1291, 1292]. This approach helps increase the

timestep by a typical factor of two (e.g., from 1 to 2 fs), with similar gains in the computational time. Freezing bond angles, however, alters the overall motion substantially [1296]. See the last section of this chapter.

### Splitting Forces in MTS Schemes

In MTS schemes, we divide the force components into classes and update the associated forces at timesteps appropriate for the class. Three-class partitioning schemes are common in biomolecular simulations, though more classes can be used for added accuracy/efficiency.

For example, a simple partitioning based on potential-energy components [1349] can place bond-length and bond-angle terms into the fastest class, dihedral-angle terms into the medium class, and the remaining forces into the slow-force class. Incorporating distance classes (e.g., interactions within 6 Å into the medium-force class) involves more complex bookkeeping work for the nonbonded pairlist generation but often works better. The numerical details of how these different updates are merged affect the stability of the trajectory significantly [94, 134]; see the separate section on MTS methods in the next chapter for details.

To illustrate a simple MTS scheme, consider using three timesteps for the butane system analyzed above for computing the end-to-end distance. We set the three values as  $\Delta\tau$ ,  $2\Delta\tau$ , and  $4\Delta\tau \equiv \Delta t$ ; thus a 3-timestep integrator associated with the outer timestep of  $\Delta t = 2$  fs has an inner timestep of 0.5 fs and a medium timestep of 1 fs. Bond-length and angle terms are updated every  $\Delta\tau$ ; dihedral-angle terms are updated every  $2\Delta\tau$ ; and nonbonded interactions are recalculated every  $4\Delta\tau$  for this butane model. This partitioning should thus save work for this MTS protocol  $\{\Delta\tau, 2\Delta\tau, 4\Delta\tau\}$  relative to the single-timestep integrator at  $\Delta\tau$ .

The data shown in Figure 13.5 for the average end-to-end distance of butane compares values obtained from single-timestep methods to those obtained by Verlet-based MTS simulations. We see that the accuracy is determined by the innermost timestep. This is precisely the advantage exploited in MTS schemes: same accuracy at less computation (under the appropriate comparison) [579, 1448].

For the small butane system, the MTS force splitting is not computationally advantageous because the nonbonded interactions represent a small fraction of the total force-evaluation work. However, for biomolecules, non-negligible speedups (e.g., factor of 2 or more) can be reaped with respect to single-timestep simulations performed at the inner timestep ( $\Delta\tau$ ) used in the MTS protocol.

## 13.4 The Verlet Algorithm

One of the simplest and best family of integrators for biomolecular dynamics is the leapfrog/Verlet/Störmer group. The basic scheme was derived as a truncation of a higher-order method used by Störmer in 1907 [1227] to integrate trajectories of electrons. The scheme was adapted by Verlet 60 years later [1302]. A favorable

numerical property of this family is exceptional stability over long times (recall the example in Figure 13.6, where the Verlet trajectory shows favorable energy conservation compared to the Runge-Kutta trajectory).

Stability far more than accuracy limits the timestep in simulations of biomolecular motion governed by approximate force fields. Programming ease is a boon for weary simulators who must wade through thousands of lines of molecular mechanics and dynamics programs. Far from mere convenience, simplicity in the integration also facilitates the implementation of variations in the basic method (e.g., constrained dynamics, extensions to various statistical ensembles) and the merging of integration schemes with other program components (e.g., fast electrostatics by the particle-mesh Ewald method).

For both these practical and theoretical reasons, it has become very difficult for alternative schemes — such as higher-order and implicit discretizations — to compete with Verlet in popularity [1110]. For the same reason, the single-timestep Verlet method is still more popular in general than MTS variants (as introduced in Subsection 13.3.9 and discussed in detail in the next chapter), but this is likely to change as comprehensive and reliable MTS routines for various statistical ensembles enter into the popular biomolecular dynamics packages.

Not surprisingly, generalizations of the Verlet method have been applied to other formulations, such as constrained and stochastic dynamics. The Verlet scheme is also the basis of the common force-splitting MTS methods and for approaches using various thermodynamic ensembles.

Below we describe how the scheme is formulated and derive its leapfrog, velocity, and position variants.

### 13.4.1 Position and Velocity Propagation

#### Iterative Recipe

To derive the Verlet propagation scheme for Newtonian dynamics, we write the equation of motion, eq. (13.3), as

$$\mathbf{M} \ddot{\mathbf{X}}(t) = \mathbf{F}(\mathbf{X}(t)) \quad (13.7)$$

where

$$\mathbf{F}(\mathbf{X}(t)) = -\nabla E(\mathbf{X}(t)). \quad (13.8)$$

We also use the symbol  $\tilde{\mathbf{F}}$  (the acceleration, or the force scaled by the inverse-mass matrix) to simplify the formulas below, where

$$\tilde{\mathbf{F}}(\mathbf{X}(t)) = \mathbf{M}^{-1} \mathbf{F}(\mathbf{X}(t)) = -\mathbf{M}^{-1} \nabla E(\mathbf{X}(t)). \quad (13.9)$$

In a numerical discretization, a timestep  $\Delta t$  is chosen and the continuous variables  $\mathbf{X}(t)$  and  $\mathbf{V}(t)$  are approximated by values at discrete time intervals  $n\Delta t$  for  $n = 0, 1, 2, \dots$ , denoted as  $\mathbf{X}^n$  and  $\mathbf{V}^n$ . Below we also use the notation  $\mathbf{F}^n$  as short hand for the force at time  $n\Delta t$ , namely  $\mathbf{F}(\mathbf{X}^n)$ . Similarly,  $\tilde{\mathbf{F}}^n$  represents the numerical estimate to the acceleration at time  $n\Delta t$ :  $\tilde{\mathbf{F}}(\mathbf{X}^n)$ .

An iterative formula is then chosen to define recursively the positions and velocities of the molecular system. If the formula defines the solution at timestep  $n + 1$  in terms of quantities determined at previous iterates, as in

$$V^{n+1} = V^n + \Delta t \tilde{F}^n,$$

the discretization method is *explicit*. If instead the new solution is defined implicitly, in terms of known and unknown quantities, as in

$$V^{n+1} = V^n + \Delta t \tilde{F}^{n+1},$$

the method is *implicit*. The explicit versions generally involve simple algorithms that (for propagation only) use modest memory, while implicit methods involve more complex algorithms but may enjoy enhanced stability properties.

### Position Update

The original Verlet algorithm [1302] describes an update for trajectory positions, as follows:

$$X^{n+1} = 2X^n - X^{n-1} + \Delta t^2 \tilde{F}^n. \quad (13.10)$$

This is a two-step propagation scheme since positions from two prior steps ( $n$  and  $n - 1$ ) must be stored. A one-step scheme — preferable in terms of storage — can be obtained by using  $V^n$  instead of  $X^{n-1}$  as a variable on the left-hand side (see below).

The Verlet algorithm is fourth-order accurate in position. To see this, use the Taylor expansion around  $X(t)$  to obtain:

$$\begin{aligned} X(t + \Delta t) + X(t - \Delta t) \\ &= X(t) + \Delta t V(t) + \frac{\Delta t^2}{2} \tilde{F}(X(t)) + \frac{\Delta t^3}{6} \ddot{V}(t) + \mathcal{O}(\Delta t^4) \\ &\quad + X(t) - \Delta t V(t) + \frac{\Delta t^2}{2} \tilde{F}(X(t)) - \frac{\Delta t^3}{6} \ddot{V}(t) + \mathcal{O}(\Delta t^4) \\ &= 2X(t) + \Delta t^2 \tilde{F}(X(t)) + \mathcal{O}(\Delta t^4). \end{aligned} \quad (13.11)$$

### Velocity Update

The flexibility in defining compatible velocity formulas from the Verlet position propagation formula (eq. (13.10)) has led to Verlet variants. These variants rely on a velocity formula consistent with eq. (13.10).

We derive such a formula by subtracting the Taylor expansion for  $X(t - \Delta t)$  from that for  $X(t + \Delta t)$ , instead of adding as done above, to obtain:

$$\frac{X(t + \Delta t) - X(t - \Delta t)}{2\Delta t} = V(t) + \mathcal{O}(\Delta t^2). \quad (13.12)$$

This definition can be written as the *central difference approximation*

$$V^n = (X^{n+1} - X^{n-1})/(2\Delta t). \quad (13.13)$$

Though the position update formula of eq. (13.10) is fourth-order accurate, the Verlet method is overall *second-order accurate*. This is because eq. (13.10) divided by  $\Delta t^2$  represents an  $\mathcal{O}(\Delta t^2)$  approximation to  $\dot{X} = \tilde{F}(X)$ . Though the same trajectory positions are followed in theory, differences are realized in practice among Verlet variants due to finite computer arithmetic. Local accuracy together with stability considerations affect global measurements, like the kinetic temperature (as computed from a simulation by the relation of eq. (13.4)). This is because such global properties must be interpreted statistically.

### 13.4.2 Leapfrog, Velocity Verlet, and Position Verlet

Three popular variants of Verlet are the *leapfrog* scheme [553], *velocity Verlet* [1244], and *position Verlet* [95, 99, 1277]. All originate from the same leapfrog/Störmer/Verlet method [1227, 1302] and have favorable numerical properties like symplecticness (e.g., [1199]) and time-reversibility [1277].

The leapfrog scheme is so called because the velocity is defined at half steps ( $V^{n+1/2}$ ) while positions are defined at whole steps ( $X^n$ ) of  $\Delta t$ . The propagation formula of leapfrog thus transforms one point in phase space to the next as  $\{V^{n-1/2}, X^n\} \Rightarrow \{V^{n+1/2}, X^{n+1}\}$ . The velocity Verlet scheme transforms instead  $\{X^n, V^n\} \Rightarrow \{X^{n+1}, V^{n+1}\}$ .

We first present these three methods using a symmetric *propagation triplet*. Both leapfrog and velocity Verlet can be written as:

$$\begin{aligned} V^{n+1/2} &= V^n + \frac{\Delta t}{2} \tilde{F}^n \\ X^{n+1} &= X^n + \Delta t V^{n+1/2} \\ V^{n+1} &= V^{n+1/2} + \frac{\Delta t}{2} \tilde{F}^{n+1}. \end{aligned} \tag{13.14}$$

The position Verlet scheme can be written as:

$$\begin{aligned} X^{n+1/2} &= X^n + \frac{\Delta t}{2} V^n \\ V^{n+1} &= V^n + \Delta t \tilde{F}^{n+1/2} \\ X^{n+1} &= X^{n+1/2} + \frac{\Delta t}{2} V^{n+1}. \end{aligned} \tag{13.15}$$

Note that the force is evaluated at the endpoints in leapfrog and velocity Verlet but at the midpoint of position Verlet.

#### Leapfrog

The leapfrog Verlet method as described above can also be written as

$$V^{n+1/2} = V^{n-1/2} + \Delta t \tilde{F}^n, \tag{13.16}$$

$$X^{n+1} = X^n + \Delta t V^{n+1/2}. \tag{13.17}$$

It can be derived from the position propagation formula of eq. (13.10) in combination with the velocity formula (13.13) approximated at half steps:

$$V^{n-1/2} = (X^n - X^{n-1})/\Delta t. \quad (13.18)$$

It is transparent to see the equivalence of the position update formula of leapfrog (eq. (13.17)) to that used in the triplet scheme (eq. (13.14)). To see that the velocity updates are equivalent as well, replace the  $n+1$  superscript by  $n$  in the third equation of the triplet (13.14) to get:

$$V^n = V^{n-1/2} + \frac{\Delta t}{2} \tilde{F}^n, \quad (13.19)$$

and add this equation to the first equation (for  $V^{n+1/2}$ ) of the triplet (13.14). The result  $V^{n+1/2} = V^{n-1/2} + \Delta t \tilde{F}^n$  is the velocity formula of leapfrog (eq. (13.16)).

### Velocity Verlet

Besides the triplet equation (13.14), the velocity Verlet scheme can also be written as:

$$X^{n+1} = X^n + \Delta t V^n + \frac{\Delta t^2}{2} \tilde{F}^n, \quad (13.20)$$

$$V^{n+1} = V^n + \frac{\Delta t}{2} (\tilde{F}^n + \tilde{F}^{n+1}). \quad (13.21)$$

The first equation can be derived by using eq. (13.13) to express  $X^{n-1}$  as  $(X^{n+1} - 2\Delta t V^n)$  and plugging this substitution for  $X^{n-1}$  in eq. (13.10). The second equation of velocity Verlet is obtained by applying the velocity equation (13.13) to express  $V^{n+1}$  in terms of positions and then using eq. (13.20) to write  $X^{n+2}$  in terms of  $X^{n+1}$ ,  $V^{n+1}$ , and  $\tilde{F}^{n+1}$ , and  $X^n$  in terms of  $X^{n+1}$ ,  $V^n$ , and  $\tilde{F}^n$ . That is, we write  $V^{n+1}$  as

$$\begin{aligned} V^{n+1} &= \frac{X^{n+2} - X^n}{2\Delta t} \\ &= \left[ \frac{X^{n+1} + \Delta t V^{n+1} + \frac{\Delta t^2}{2} \tilde{F}^{n+1}}{2\Delta t} \right] - \left[ \frac{X^{n+1} - \Delta t V^n - \frac{\Delta t^2}{2} \tilde{F}^n}{2\Delta t} \right] \\ &= \frac{V^{n+1} + V^n}{2} + \frac{\Delta t}{4} (\tilde{F}^n + \tilde{F}^{n+1}). \end{aligned} \quad (13.22)$$

Collecting the two  $V^{n+1}$  terms yields

$$\frac{V^{n+1}}{2} = \frac{V^n}{2} + \frac{\Delta t}{4} (\tilde{F}^n + \tilde{F}^{n+1}), \quad (13.23)$$

leading immediately to the position update formula of eq. (13.21).

The equivalence of the triplet version (eq. (13.14)) of velocity Verlet to the form in eqs. (13.20), (13.21) is straightforward.

### Position Verlet

Besides the triplet (13.15), we can express position Verlet as:

$$X^{n+1} = X^n + \Delta t V^n + \frac{\Delta t^2}{2} \tilde{F}^{n+1/2}, \quad (13.24)$$

$$V^{n+1} = V^n + \Delta t \tilde{F}^{n+1/2}. \quad (13.25)$$

### MTS Preference

The velocity Verlet scheme has been customarily used in MTS schemes because the forces are evaluated at the ends of the interval and this is more amenable to force splitting. Moreover, it only requires one constrained dynamics application per inner timestep cycle (when position is updated) when constrained dynamics techniques are used; two such constrained-dynamics iterations are required for position Verlet. However, position Verlet — first suggested to be preferable in large-timestep methods [95] — was recently shown to have advantages in connection with moderate to large timesteps [97, 99].

## 13.5 Constrained Dynamics

In constrained dynamics, the equations of motion are augmented by algebraic constraints to freeze the highest-frequency interactions. To illustrate, consider a typical modeling of bond-length potentials as the harmonic form

$$E_{\text{bond}} = \frac{S_h}{2} (r_{ij} - \bar{r}_{ij})^2,$$

where  $r_{ij}$  is an interatomic distance with equilibrium value  $\bar{r}_{ij}$ , and  $S_h$  is a force constant. The constraint implemented in the context of MD simulations is thus

$$g_k = r_{ij}^2 - \bar{r}_{ij}^2 = 0.$$

Using the formalism of Lagrange multipliers, we have then in place of eq. (13.3) the following system to solve:

$$\begin{aligned} \dot{\mathbf{M}}V(t) &= -\nabla E(X(t)) - g'(X(t))^T \Lambda, \\ \dot{X}(t) &= V(t), \\ g(X(t)) &= 0. \end{aligned} \quad (13.26)$$

Here  $g(X(t))$  is a vector with entries  $g_i$  containing the individual constraints, and  $\Lambda$  is the vector of Lagrange multipliers.

## SHAKE

In 1977, the SHAKE algorithm was introduced on the basis of the leapfrog Verlet scheme of eqs. (13.16, 13.17) [1079]:

$$\begin{aligned} V^{n+1/2} &= V^{n-1/2} - \Delta t \nabla E(X^n) - \Delta t g'(X^n)^T \Lambda^n, \\ X^{n+1} &= X^n + \Delta t V^{n+1/2}, \\ g(X^{n+1}) &= 0. \end{aligned} \quad (13.27)$$

Like the Verlet method, the forces  $-\nabla E$  must be computed only once per step. However, the  $m$  constraints require at each timestep, in addition, the solution of a nonlinear system of  $m$  equations in the  $m$  unknowns  $\{\lambda_i\}$ . Thus, the SHAKE method is *semi-explicit*.

## RATTLE

A variant termed RATTLE was later introduced [44]. Similar techniques have been proposed for constraining other internal degrees of freedom [1264], and direct methods have been developed for the special case of rigid water molecules [868]. Leimkuhler and Skeel [733] have shown that the RATTLE method is symplectic and that SHAKE, while not technically symplectic, yields solutions identical to those of RATTLE for the positions and only slightly perturbed velocities.

### Computational Advantage

With the fastest vibrations removed from the model, the integration timestep can be lengthened, resulting in a typical force calculation per time ratio of 0.5 per fs. This computational advantage must be balanced against the additional work per step required to solve the nonlinear equations of constraints. Each proposed constrained model is thus accompanied by a practical numerical algorithm.

For the time discretization of eq. (13.27), an iterative scheme for solution of the nonlinear equations was presented in [1079], where the individual constraints are imposed sequentially to reset the coordinates that result from an unconstrained Verlet step. This iterative scheme is widely used due to its simplicity and frugal consumption of computer memory. However, the SHAKE iteration can converge very slowly, or even fail to converge, for complex bond geometries [92]. Improved convergence was later reported by Barth *et al.* [92], who proposed enhancements based on the relationship between the SHAKE iteration process and the iterative Gauss-Seidel/Successive-Over-Relaxation techniques for solving nonlinear systems. This was also independently developed by Xie and Scott, as reported later by Xie *et al.* [1401].

### Limitations

The computational advantage of constrained models is clear, but the verity of the constrained model for reproducing the dynamics of unconstrained systems is a separate issue. Van Gunsteren and Karplus [1296] showed through simulations of the protein BPTI in vacuum that the use of fixed bond lengths does not

significantly alter the dynamical properties of the system, whereas fixing bond angles does. Similar conclusions were reported by Toxvaerd [1269] for decane molecules.

Though it is possible that the former study was also influenced by poor convergence of SHAKE for bond-angle constraints (which can lead to overdetermination and singularities in the constraints equations), the intricate vibrational coupling in biomolecules argues against the general use of angle constraints. As shown in [93], from the point of view of increasing the timestep in biomolecular simulations, only constraints of light-atom angle terms in addition to all bond lengths might be beneficial. Furthermore, it is common practice not to constrain motion beyond that associated with light-atom bonds because the overall motion can be disturbed due to vibrational coupling.

## 13.6 Various MD Ensembles

### 13.6.1 Need for Other Ensembles

The algorithmic framework discussed thus far is appropriate for the microcanonical (constant NVE) ensemble (see the equilibration illustration in Figure 13.2, for example), where the total energy  $E$  is a constant of the motion. This assumes that the time averages are equivalent to the ensemble averages.

The study of molecular properties as a function of temperature and pressure — rather than volume and energy — is of general importance. Thus, microcanonical ensembles are inappropriate for simulating certain systems which require constant pressure and/or temperature conditions and allow the energy and volume to fluctuate. This is the case for a micelle or lipid bilayer complex (proteins and lipids in an ionic solution) under constant pressure (isotropic pressure or the pressure tensor), and for crystalline solids under constant stress (i.e., constant pressure tensor). For such systems, other ensembles may be more appropriate, such as *canonical* (constant temperature and volume, NVT) *isothermal-isobaric* (constant temperature and pressure, NPT), or constant pressure and enthalpy (NPH). Special techniques have been developed for these ensembles, and most rely on the Verlet framework just described, including the MTS variants discussed in the next chapter [836].

For pioneering works, see the articles by Andersen [43] and the extension by Parrinello and Rahman [964] for NPH, and that of Nosé [925] and Hoover [566] for NVT and NPT ensembles (started in 1984). For details on algorithms for these extended ensembles, see the texts [22, 428, 731] and current literature.

*In this section we only give a flavor of some of these methods.* Methods for handling these thermodynamic constraints or generating the various ensembles can be grouped into simple *constrained formulations*, including approaches that involve stochastic-motivated models, and more sophisticated techniques that involve additional degrees of freedom (*extended system methods*). The former group, while easy to implement, does not generally generate the desired ensemble in rigorous terms.

### 13.6.2 Simple Algorithms

Constraint-based methods can involve Lagrange-multiplier coupling, or simple scaling of phase-space variables. For example, the velocity vector might be simply scaled at each timestep to fix the desired kinetic temperature  $T$  at the target value  $T_0$  [22]. This can be achieved by the scaling

$$V^{\text{new}} \leftarrow c_t V^{\text{old}}, \quad (13.28)$$

where

$$c_t = \sqrt{T_0/T}. \quad (13.29)$$

This drastic approach, however, implies rapid energy transfer to, from, and among the various degrees of freedom in the system. In particular, it can be shown that velocity rescaling leads to an artifactual pumping of energy into low-frequency modes [524]; occasional resetting of velocities can mitigate this problem.

#### Weak Coupling Thermostat for Constant T

Similar algorithms that are motivated by stochastic approaches are also easy to implement but can steer the system toward the desired constant (e.g., target temperature or pressure) more gently. This can be accomplished by mimicking a diffusive process governed by a fictitious frictional coefficient  $\gamma$ , whose value controls the relaxation rate of this coupling [123].

For example, consider the weak coupling thermostat proposed by Berendsen and coworkers [123], widely used for constant-temperature MD. In this approach, the equations of motion (eq. (13.3)) are modified as:

$$\begin{aligned} \dot{X}(t) &= V(t), \\ \mathbf{M}\dot{V}(t) &= -\nabla E(X(t)) - \gamma_t \mathbf{M}V(t), \\ \gamma_t &= \frac{1}{2\tau} \left( 1 - \frac{T_0}{T} \right), \end{aligned} \quad (13.30)$$

where  $\gamma_t$  has units of inverse time,  $T$  is the instantaneous kinetic temperature, and  $\tau$  is the time constant of the coupling to the heat bath. (Compare to the Langevin formalism described in the next chapter, Section 14.4).

This augmentation effectively scales the velocity vector via eq. (13.28) by the factor:

$$c_t = \sqrt{1 - \frac{\Delta t}{\tau} \left( 1 - \frac{T_0}{T} \right)}, \quad (13.31)$$

where  $\Delta t$  is the (Verlet) timestep. This scaling produces a least-squares local disturbance to the velocity vector, so as to satisfy the global temperature constraint.

The advantage of this weak bath coupling approach is the introduction of added thermostat parameters (not dynamic variables, as below). The factor  $\tau$  controls the characteristic decay time — and hence the strength — of the coupling to the heat bath. When  $\tau$  is large,  $\gamma_t$  is small, and the scaling factor  $c_t$  approaches unity (the microcanonical ensemble is approached); when  $\tau$  is small,  $\gamma_t$  is large — that is, the coupling to the heat bath is strong — and the energy exchange between the molecular system and the thermal reservoir is significant.

In particular, when  $\tau = \Delta t$ ,  $c_t = \sqrt{T_0/T}$ , producing the simplest scaling approach for constant-T simulations, but also the most artificial, as mentioned above. The advantage of the flexible coupling  $\tau$  parameter (or  $\gamma_t$ ) is the control of the *rate* at which the target temperature is reached.

Still, this approach, while convenient, does not generate the true canonical ensemble; it can also lead to artifacts as simple velocity scaling [524]. See below for a more rigorous approach that reproduces the desired thermal ensemble properly.

### Weak Coupling Barostat for Constant P

In the same spirit, for the purpose of constant pressure simulations for isotropic systems modeled in a finite volume (generally via periodic boundary conditions), Berendsen and coworkers suggest a simple scaling to the Cartesian positions,

$$X^{\text{new}} \leftarrow c_p X^{\text{old}}, \quad (13.32)$$

as well as a volume scaling (see below), where the scale factor is:

$$c_p = \left[ 1 - \frac{\beta \Delta t}{\tau} (P_0 - P) \right]^{1/3}. \quad (13.33)$$

Here  $\beta$  is the *isothermal compressibility*,  $P$  is the *instantaneous (or internal) pressure*,  $P_0$  is the target pressure (or external pressure  $P_{\text{ex}}$ ), and  $\tau$  is the pressure coupling time (see Box 13.2). The goal of this formulation is to allow the volume of the macromolecular system ( $\nu_l$ ) to fluctuate as the instantaneous (internal) pressure ( $P$ ) approaches the applied (external) pressure,  $P_0$ .

This procedure is applied to the coordinates of particles under periodic boundary conditions, and to the box length  $L$  for an isotropic system in a cubic box ( $L^{\text{new}} \leftarrow c_p L^{\text{old}}$ ). From the point of view of the guiding dynamic equations, this coordinate and volume scaling protocol effectively solves the augmented equation of motion for the time derivative of position ( $\dot{X}(t) = V(t)$ ), where the additional term is proportional to  $X$  as follows:

$$\begin{aligned} \dot{X}(t) &= V(t) - \tilde{\beta} X(t), \\ \tilde{\beta} &= \frac{\beta (P_0 - P(t))}{3\tau}. \end{aligned} \quad (13.34)$$

This equation can also be written as a pair of differential equations describing the evolution of the position ( $X$ ) and volume ( $\nu_l$ ) by using eq. (13.36) in Box 13.2 to produce the following equations (we suppress the time dependency for clarity):

$$\begin{aligned}\dot{X} &= V + \frac{1}{3} \frac{\dot{\nu}_l}{\nu_l}, \\ \dot{\nu}_l &= \frac{\beta(P - P_0)}{\tau} \nu_l.\end{aligned}\tag{13.35}$$

Thus, this approach to constant pressure simulations allows the volume of the system to fluctuate as the pressure is held constant by changing the cell size uniformly but not its shape, as appropriate for an isotropic system. For an anisotropic system, instead of the scalar pressure variable, the  $3 \times 3$  pressure matrix (see Box 13.2) is relevant [853].

As for the weak coupling thermostat method, the above approach does not generally produce trajectories from any known ensemble. A rigorous approach using an extended system method which includes dynamic thermostat and barostat variables is outlined below. See also [382] for an improved constant-pressure simulation protocol using a Langevin piston that is straightforward to implement. The Langevin piston method attempts to eliminate the nonphysical vibrations of the volume associated with the approach above and to generate in theory the correct NPT ensemble.

### Box 13.2: Pressure Variables Definitions

The *isothermal compressibility* describes the change in volume ( $\nu_l$ ) of a substance under external pressure  $P$  via:

$$\frac{dP}{dt} = \frac{-1}{\beta \nu_l} \left( \frac{d\nu_l}{dt} \right)\tag{13.36}$$

or

$$\beta \dot{P} = -\dot{\nu}_l / \nu_l.\tag{13.37}$$

The *instantaneous pressure*  $P$  is calculated from the thermodynamic relation between the pressure, temperature, volume, and internal virial (*vir*) of a system by [22]:

$$P = \frac{2}{3 \nu_l} (E_k - vir).\tag{13.38}$$

The internal virial is proportional to the inner product of the each atom's position vector ( $\mathbf{r}_i$ ) with the corresponding force component acting on atom  $i$  due to all particles ( $F_i$ ):

$$vir = - \sum_i (\mathbf{r}_i^T F_i).\tag{13.39}$$

(Intramolecular forces are thus considered in addition to intermolecular forces). Note that in the expression for the pressure (eq. (13.38)) we have used the relation between the temperature  $T$  and the kinetic energy  $E_k$  ( $\langle E_k \rangle = \frac{N_F}{2} k_B T$ ), where the number of degrees of freedom,  $N_F$ , is three times the number of atoms in the system.

For an anisotropic system, the pressure becomes a  $3 \times 3$  matrix called the *pressure tensor*  $\mathbf{P}$  [853]. This tensor is defined as:

$$\mathbf{P} = \frac{1}{\nu_l} \left[ \sum_i (m_i \mathbf{v}_i \mathbf{v}_i^T) + \sum_i (\mathbf{r}_i F_i^T) \right], \quad (13.40)$$

where  $m_i$  and  $\mathbf{v}_i$  are the scalar mass and 3-component velocity vector corresponding to atom  $i$ , respectively; the result of an outer product between a  $3 \times 1$  vector and its transpose is a  $3 \times 3$  matrix.

---

### 13.6.3 Extended System Methods

Extended-system methods introduce additional degrees of freedom to represent the environment of the macromolecular system (e.g., pressure piston, thermostat, etc.). These degrees of freedom have positions, momenta, and/or other associated variables. Examples include an external volume variable  $\nu_l$  associated with a piston of given mass, whose potential energy is expressed in terms of the desired pressure [924], or effective thermodynamic variables that are functions of the positions and momenta of the added dynamic variables.

The equations of motion for the extended system are expressed as augmented versions of the standard equations of motions, to represent the evolution of both the internal and external variables in the desired ensemble. The solutions for the time evolution of the extended system by standard numerical methods (for the *microcanonical* ensemble) then produce the phase-space variables appropriate for the desired ensemble.

#### Canonical Ensemble

For example, to produce the true canonical (NVT) ensemble — which the simple scaling and thermostat approaches above do not accomplish — the method termed Nosé-Hoover [566,924] adds a fictitious degree of freedom to the physical system with ‘coordinate’ parameter  $x_t$  (effectively a scaling parameter [428]), mass  $m_t$ , and thermodynamic friction coefficient  $\zeta_t$ . (This friction coefficient is related to  $x_t$  and the corresponding momentum  $\dot{x}_t$ ).

After appropriate scaling of variables (see [428], for example), the equations of motion for the real and artificial system combine to yield:

$$\begin{aligned} \dot{X}(t) &= V(t), \\ \mathbf{M}\dot{V}(t) &= -\nabla E(X(t)) - \zeta_t \mathbf{M}V(t), \\ m_t \dot{\zeta}_t(t) &= 2V^T \mathbf{M}V - g k_B T_0, \end{aligned} \quad (13.41)$$

where  $g$  is the number of degrees of freedom in the system. The conserved quantity under these augmented equations of motion has two energy terms in addition to the internal kinetic and potential energies:

$$\hat{H}^{\text{NVT}} = \frac{1}{2}(V^T \mathbf{M}V) + E(X) + \frac{1}{2}(m_t \zeta_t^2) + g k_B T_0 x_t. \quad (13.42)$$

The choice of the fictitious mass  $m_t$  should ensure that the thermalization process is efficient. Too small a value implies large harmonic motion and rapid thermal fluctuations for the extended degree of freedom, and this in turn restricts the timestep too severely. Too large a value, on the other hand, makes the thermalization process slow. In general,  $m_t$  is chosen to be proportional to  $g k_B T$ , and the integration of the above system is performed fairly accurately so that the energy is well conserved. See the original literature and [428,835] for algorithmic details.

### Isothermal-Isobaric Ensemble

Controlling the pressure in MD simulations is more involved. Instead of fixing the volume, as in microcanonical simulations, the volume of the system is considered as a dynamic variable, and it is allowed to fluctuate while the pressure is held constant. An approach analogous to the Nosé-Hoover NVT form above has been described by Andersen [43] and Hoover [566] for the isothermal-isobaric, or NPT ensembles.

Essentially, in addition to the effective coordinate, mass, and friction set  $(x_t, m_t, \zeta_t)$  associated with the fictitious thermostat variable, a set  $(x_p, m_p, \zeta_p)$  is introduced and associated with a virtual pressure piston (“barostat”).

The effective equations of motion for a 3-dimensional system become:

$$\begin{aligned}\dot{X}(t) &= V(t) + \zeta_p X, \\ \mathbf{M}\dot{V}(t) &= F(X(t)) - \mathbf{M}\dot{V}(t) \left[ \left(1 + \frac{3}{g}\right) \zeta_p + \zeta_t \right], \\ \dot{\nu}_l &= 3 \nu_l \zeta_p, \\ m_p \dot{\zeta}_p(t) &= 3 \nu_l (P_{\text{in}} - P_0) + \frac{3}{g} (2V^T \mathbf{M}V) - m_p \zeta_t \zeta_p, \\ m_t \dot{\zeta}_t(t) &= 2V^T \mathbf{M}V + \frac{\zeta_p^2}{m_p} - (g+1) k_B T_0,\end{aligned}\tag{13.43}$$

where  $g$  is the number of degrees of freedom in the system,  $P_0$  is the external applied pressure, and  $P_{\text{in}}$  is the internal pressure, defined as:

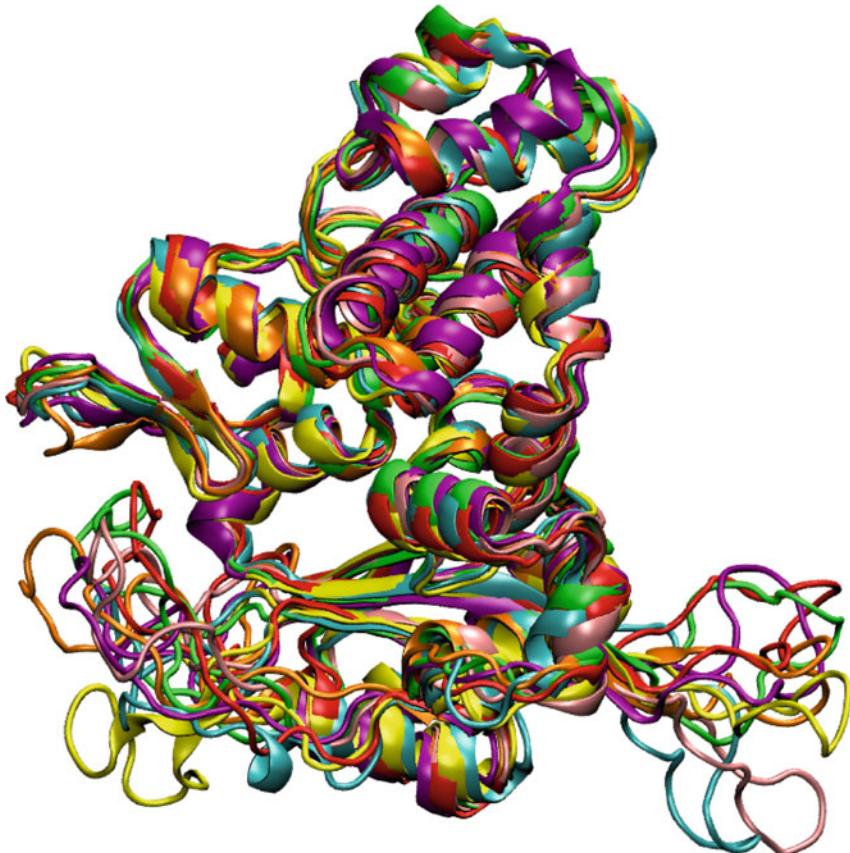
$$P_{\text{in}} = \frac{2}{3 \nu_l} \left[ E_k - vir - \left( \frac{3 \nu_l}{2} \right) \frac{\partial E(X, \nu_l)}{\partial \nu_l} \right];\tag{13.44}$$

the virial  $vir$  is defined in equation (13.39). The conserved quantity under these augmented equations of motion is:

$$\hat{H}^{\text{NPT}} = \frac{1}{2} (V^T \mathbf{M}V) + E(X, \nu_l) + \frac{1}{2} (m_t \zeta_t^2 + m_p \zeta_p^2) + (g+1) k_B T_0 x_t + P_0 \nu_l.\tag{13.45}$$

By this approach, the volume of the system fluctuates under the applied thermostat and barostats so that the system is driven to the steady state at which the average internal pressure  $P$  is equal to the external applied force  $P_0$ .

Algorithmic extensions to anisotropic systems, to allow changes in cell shape in addition to size, have also been developed using the pressure tensor and matrix analogs of the above approaches [123, 835, for example].



# 14

## Molecular Dynamics: Further Topics

### Chapter 14 Notation

SYMBOL	DEFINITION
<b>Matrices</b>	
<b>C</b>	covariance matrix formulated for PCA
<b>D</b>	diffusion tensor (with sub-block matrices $\bar{\mathbf{D}}_{ij}$ )
$\tilde{\mathbf{H}}$	local Hessian approximation
<b>I</b>	identity matrix
<b>J</b>	constant matrix used in defining symplectic transformation
<b>A</b>	diagonal matrix with eigenvalues $\lambda_1, \lambda_2, \dots, 3N$
<b>M</b>	mass matrix
<b>S</b>	phase space transformation ( $\mathbf{S} = \mathbf{DQD}^{-1}$ ; also Cholesky factor of $\mathbf{D}$ in BD)
<b>T</b>	hydrodynamic tensor
<b>V</b>	eigenvector matrix
<b>Z</b>	friction tensor (related to <b>D</b> via $k_B \mathbf{T} \mathbf{D}^{-1}$ )
$\Psi_J$	Jacobian matrix of <b>J</b>
<b>Vectors</b>	
<b>F</b>	force
$\tilde{\mathbf{F}}$	acceleration ( $\ddot{\mathbf{X}}$ ), or $\mathbf{M}^{-1} F(X(t))$
$F_{\text{fast}}, F_{\text{med}}, F_{\text{slow}}$	forces of ‘fast’, ‘medium’ and ‘slow’ components
$P, \dot{P}$	momentum and its first time derivative
$R, R_B$	random forces
<b>V</b>	velocity ( $\dot{\mathbf{X}}$ ), components $\{v_i\}$ (also used for eigenvector)
$X, \dot{X}, \ddot{X}$	position (components $\{x_i\}$ ), and its first and second time derivatives
$X_r$	reference position for MTS extrapolation

Chapter 14 Notation Table (continued)

SYMBOL	DEFINITION
$\vec{Y}, \dot{\vec{Y}}$	vector and its first time derivative
$\nabla E$	energy gradient
<b>Scalars &amp; Functions</b>	
$a$	hydrodynamic bead radius
$k_1, k_2$	integers (used in MTS to relate timesteps; $r = k_1 k_2$ )
$\lambda_n$	eigenvalue $n$
$n, m$	integers (defining resonance condition); also $m$ = mass
$D_t$	translational diffusion constant
$E_k$	kinetic energy
$H$	Hamiltonian
$\mathcal{L}$	Liouville operator
$N$	number of atoms
$N_b$	number of beads (Brownian dynamics)
$T$	temperature
$T_p$	period (of characteristic frequency)
$\gamma$	Langevin damping constant
$\zeta$	frictional coefficient ( $m\gamma$ for example)
$\eta$	solvent viscosity
$\theta$	phase space rotation defined by integrator
$\theta_{\text{eff}}$	effective phase space rotation
$\omega$	natural frequency ( $\epsilon = \omega \Delta t$ )
$\omega_{\text{eff}}$	effective (scheme-dependent) frequency
$\Delta\tau, \Delta t_m, \Delta t$	timesteps (inner, medium, and outer)
$\Phi(X)$	'dynamics function' for implicit integration

Nature laughs at the difficulties of integration.

Pierre-Simon de Laplace (1749–1827).

Only by taking an infinitesimally small unit for observation (the differential of history ...) and attaining to the art of integrating them (that is, finding the sum of these infinitesimals) can we hope to arrive at the laws of history.

Lev Tolstoy, in *War and Peace* (1828–1910).

## 14.1 Introduction

In this chapter we survey more advanced aspects of integration approaches for biomolecular dynamics that are suitable for students interested in the mathematical issues of numerical schemes. We begin in Section 14.2 by introducing symplectic integrators in terms of an effective rotation in phase space and illustrate basic concepts by simple harmonic analysis.

The multiple-timestep (MTS), or force splitting, methods are presented next (Section 14.3), and the extreme variants of splitting by extrapolation versus splitting by impulses are contrasted with regard to resonance artifacts. Understanding favorable and unfavorable features of MTS schemes with respect to stability, accuracy, and resonance artifacts has led to important developments in recent years of efficient, long-timestep methods for biomolecular dynamics, as well as to new frameworks for method design and interpretation.

In this connection, we introduce in Sections 14.4 and 14.5, in turn, the stochastic dynamics approaches of Langevin and Brownian dynamics; both are in fact closely related, and the separation of terms may be arbitrary. These stochastic formulations may be viewed as constructs that enhance sampling, or that, through the introduction of random forces, make possible large timesteps by masking mild instabilities resulting from Newtonian integration.

Stochastic approaches also constitute approximate models for following large-scale motions in systems where random fluctuations (e.g., introduced by the bulk solvent) are at least as important as the systematic forces. This is the case, for example, in long DNA polymers that move with agility in solution, sampling many equilibrium configurations, rather than remaining near a single state. Studies of DNA supercoiling and of diffusion-controlled ligand gating events in enzyme catalysis, for example, have relied on such Brownian dynamics approaches.

We also mention implicit integration schemes in Section 14.6. This class of generally-more-expensive schemes has been explored as a possible way to increase the timestep in biomolecular dynamics simulations. Section 14.7 describes various enhanced sampling approaches for biomolecules, such as harmonic-analysis based methods like essential dynamics, force biasing, altered protocols (like replica-exchange MD), and various innovative methods for exploring conformational space, deducing mechanisms and computing reaction rates. We conclude in Section 14.8 with future perspectives on MD algorithm developments and with a description of some promising integration alternatives.

*The reader should have read Chapter 13, where basic aspects of molecular dynamics and associated notation have been presented.*

## 14.2 Symplectic Integrators

As introduced in the last chapter, symplectic or canonical integrators *preserve* special properties associated with the Hamiltonian system of differential equations [731, 1090]. These properties include *volume elements in phase space* and the Hamiltonian value (*energy*). In practice, the total energy is not preserved exactly, but the energy error remains constant over long times. This is different from nonsymplectic methods, which typically display a systematic energy drift in time (usually damping). See Figure 13.6 for such an example of integration by the symplectic Verlet versus the nonsymplectic Runge-Kutta method.

Such favorable properties can be explained by *backward error analysis*,<sup>1</sup> which shows that the numerical trajectory of a symplectic integrator is the *exact* solution of a nearby Hamiltonian. The proximity measure depends on the timestep of the integrator,  $\Delta t$ , and on the scheme's order of accuracy,  $p$ . That is, it can be shown that the Hamiltonian corresponding to the numerical trajectory of a symplectic method remains order  $\mathcal{O}(\Delta t)^p$  away from the initial value of the true Hamiltonian.

A comprehensive treatment of symplectic integration can be found in [1090]. Here only a few key concepts will be introduced, many of which are relevant to future sections.

### 14.2.1 Symplectic Transformation

A Hamiltonian system is generally described by the motion of the collective position and momenta vectors  $X, P \in \mathbb{R}^{3N}$  in time (here  $P = MV$  where  $M$  is the diagonal mass matrix; see last chapter). The Hamiltonian  $H(X, P)$  is given as the sum of the kinetic and potential energy of the system:

$$H(X, P) = \frac{1}{2}(P^T M^{-1} P) + E(X). \quad (14.1)$$

The evolution of this system is governed by the equations:

$$\dot{P} = -\partial H / \partial X, \quad \dot{X} = \partial H / \partial P, \quad (14.2)$$

or, equivalently, in vector form as:

$$\begin{pmatrix} \dot{X} \\ \dot{P} \end{pmatrix} = \begin{bmatrix} 0 & \mathbf{I} \\ -\mathbf{I} & 0 \end{bmatrix} \begin{pmatrix} \partial H / \partial X \\ \partial H / \partial P \end{pmatrix} \equiv \mathbf{J} \cdot \nabla H, \quad (14.3)$$

where  $\mathbf{I}$  is the  $3N \times 3N$  identity matrix.

An integrator defines a mapping  $\Psi$  that transforms coordinates and momenta at time  $t$ ,  $\{X^n, P^n\}$ , to coordinates and momenta at time  $t + \Delta t$ ,  $\{X^{n+1}, P^{n+1}\}$ . This transformation is *symplectic* if and only if its Jacobian matrix,  $\Psi_J$ , satisfies:

$$\Psi_J^T \mathbf{J} \Psi_J = \mathbf{J}, \quad (14.4)$$

where  $\Psi_J$  is the matrix

$$\Psi_J \equiv \begin{bmatrix} \partial X^{n+1} / \partial X^n & \partial X^{n+1} / \partial P^n \\ \partial P^{n+1} / \partial X^n & \partial P^{n+1} / \partial P^n \end{bmatrix}. \quad (14.5)$$

---

<sup>1</sup>See [280, Section 2.4.1] or [357, Section 2.7], for example. Roughly speaking, backward error analysis transforms the problem of estimating computational errors (in our case due to the finite-difference approximation) back to the problem of estimating the effect of changing the input data slightly.

It can be shown that a *composition* of symplectic transformations is also symplectic and that the inverse of a symplectic mapping is symplectic. These properties are often used in practice to prove that an integrator is symplectic.

For a proof that the Verlet scheme is symplectic, see [1078]. Below we only illustrate how the Verlet transformation can be interpreted for a linear system as a rotation in phase space. Analysis of a harmonic oscillator is often a first step in analyzing behavior of a numerical scheme and already sheds considerable insight [821, 866, 967, 1126, 1199, 1437, for example].

### 14.2.2 Harmonic Oscillator Example

Consider the Verlet scheme of eq. (13.15) for linear forces  $F(X) = -\omega^2 X$ , where  $\omega$  is the natural frequency of the oscillator. The transformation  $\mathbf{S}$  can be used to relate one phase point to the next [1199]. That is,

$$\begin{pmatrix} \omega X^{n+1} \\ V^{n+1} \end{pmatrix} = \mathbf{S} \begin{pmatrix} \omega X^n \\ V^n \end{pmatrix}, \quad (14.6)$$

where  $\mathbf{S}$  is defined as

$$\mathbf{S} = \begin{bmatrix} 1 - \epsilon^2/2 & \epsilon(1 - \epsilon^2/4) \\ -\epsilon & 1 - \epsilon^2/2 \end{bmatrix}, \quad (14.7)$$

with the parameter  $\epsilon$  as the frequency times the timestep:

$$\epsilon = \omega \Delta t. \quad (14.8)$$

It is simple to show<sup>2</sup> that the product of the two eigenvalues is equal to 1 ( $\lambda_1 \lambda_2 = 1$ ) and that the absolute value of their sum is less than or equal to 2 ( $|\lambda_1 + \lambda_2| \leq 2$ ). Thus, powers of  $\mathbf{S}$  (which are matrices) are bounded if and only if

$$\epsilon^2 < 4, \quad (14.9)$$

or equivalently

$$\Delta t < 2/\omega. \quad (14.10)$$

### 14.2.3 Linear Stability

The above restriction on the timestep size is the *stability condition* for Verlet. Thus, if the period of the oscillator is  $T_p = 2\pi/\omega$ , this *linear stability* condition becomes

$$\Delta t < T_p/\pi. \quad (14.11)$$

---

<sup>2</sup>Hint: For a  $2 \times 2$  matrix, the matrix trace (sum of diagonals) is the sum of the eigenvalues, here  $2 - \epsilon^2$ , and the matrix determinant is their product, here 1.

Table 14.1 gives corresponding linear stability limits on the timestep for the high-frequency end of biomolecular vibrational modes. Clearly, in general, we have:

$$T_p \text{ (stretch)} < T_p \text{ (bend)}, \text{ and}$$

$$T_p \text{ (light atoms)} < T_p \text{ (heavy atoms)}.$$

We also note that the light-atom bends (e.g., H–O–H) and the heavy-atom bond stretches (e.g., C=C) have very similar periods.

In general, the table emphasizes the *lack of clear gaps in timescales* among these modes. This explains why constraining some high-frequency motions in a dynamic formulation typically affects other vibrational modes and why the computational benefit (larger permitted timestep) is relatively modest.

For example, if we constrain all bonds for a water model, thereby eliminating the motion and stability limit corresponding to the O–H stretch (period  $\sim 10$  fs), the next relevant period that limits the timestep is the H–O–H bend (period  $\sim 21$  fs), followed by water libration ( $\sim 42$  fs). Thus, the effective timestep can at most be doubled by constrained dynamics.

Table 14.1. The timestep limit for Verlet based on a harmonic oscillator analysis.

Vibrational Mode <sup>a</sup>	Wave number ( $1/\lambda$ ) [ $\text{cm}^{-1}$ ]	Period $T_p$ ( $\lambda/c$ ) [fs] <sup>b</sup>	$T_p/\pi$ [fs]
O–H, N–H stretch	3200–3600	9.8	3.1
C–H stretch	3000	11.1	3.5
O–C–O asymm. stretch	2400	13.9	4.5
C≡C, C≡N stretch	2100	15.9	5.1
C=O (carbonyl) stretch	1700	19.6	6.2
C=C stretch			
H–O–H bend	1600	20.8	6.4
C–N–H, H–N–H bend	1500	22.2	7.1
C=C (aromatic) stretch			
C–N stretch (amines)	1250	26.2	8.4
Water Libration (rocking)	800	41.7	13
O–C–O bending	700	47.6	15
C=C–H bending (alkenes)			
C=C–H bending (aromatic)			

<sup>a</sup>All values are approximate.

<sup>b</sup>The value of the speed of light is taken as  $c = 3.00 \times 10^{10} \text{ cm s}^{-1}$ .

#### 14.2.4 Timestep-Dependent Rotation in Phase Space

Under the linear stability assumption, the matrix  $\mathbf{S}$  of eq. (14.7) has the two eigenvalues  $\exp(\pm i\theta)$  ( $i = \sqrt{-1}$ ), where

$$\theta = 2 \sin^{-1}(\omega \Delta t / 2) \quad (14.12)$$

$$= \omega \Delta t + \frac{1}{24} (\omega \Delta t)^3 + \mathcal{O}(\omega \Delta t)^5. \quad (14.13)$$

Thus, the angle  $\theta$  depends on the timestep (and on  $\omega$ ). To see that this transformation defines a rotation in phase space, we decompose the phase-space transforming matrix  $\mathbf{S}$  as

$$\mathbf{S} = \mathbf{D} \mathbf{Q} \mathbf{D}^{-1}. \quad (14.14)$$

In this definition, the matrix

$$\mathbf{Q} = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \quad (14.15)$$

defines a rotation of  $-\theta$  radians in phase space, and  $\mathbf{D}$  is the diagonal matrix

$$\mathbf{D} = \begin{bmatrix} 1 & 0 \\ 0 & 1 + \tan^2 \frac{\theta}{2} \end{bmatrix}. \quad (14.16)$$

Behavior of the integrator in time can be interpreted through analysis of the powers of  $\mathbf{S}$  given by

$$\mathbf{S}^n = \mathbf{D} \mathbf{Q}^n \mathbf{D}^{-1},$$

where

$$\mathbf{Q}^n = \begin{bmatrix} \cos n\theta & \sin n\theta \\ -\sin n\theta & \cos n\theta \end{bmatrix}. \quad (14.17)$$

The timestep-dependent behavior of the transformation  $\mathbf{S}$  can be interpreted as follows. Equations (14.12) and (14.13) show that the integrator is using  $\theta$  as an approximation to the exact rotation  $\omega \Delta t$ . The smaller the timestep, the closer the approximation. This is shown in Figure 14.1, where rapid divergence from the target value (dashed diagonal line) is evident as  $\Delta t$  is increased.

The *effective rotation*  $\theta_{\text{eff}}$  can thus be defined as:

$$\theta_{\text{eff}} = \omega_{\text{eff}} \Delta t. \quad (14.18)$$

For the Verlet method, the effective rotation is given by equation (14.12):

$$\theta_{\text{eff}}^{\text{verlet}} = 2 \sin^{-1}(\omega \Delta t / 2). \quad (14.19)$$

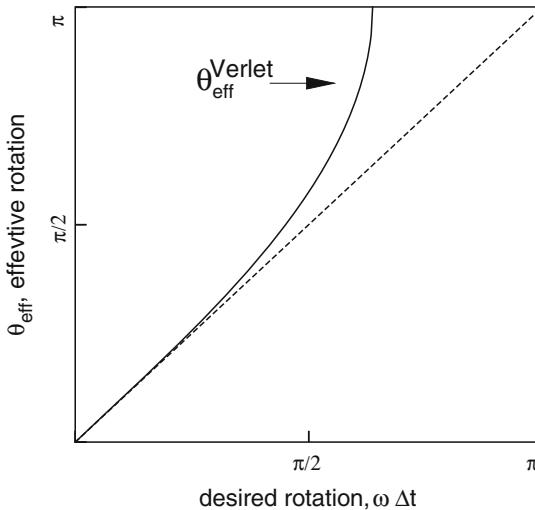


Figure 14.1. The effective rotation  $\theta_{\text{eff}}^{\text{verlet}}$  (in radians), eq. (14.19), plotted against the desired rotation,  $\omega\Delta t$ , for the Verlet scheme.

#### 14.2.5 Resonance Condition for Periodic Motion

For periodic motion with natural frequency  $\omega$ , nonphysical resonance — an artifact of the symplectic integrator — can occur when  $\omega$  is related by integers  $n$  and  $m$  to the forcing frequency  $(2\pi/\Delta t)$ :

$$\frac{n}{m}\omega = \frac{2\pi}{\Delta t}. \quad (14.20)$$

Here  $n$  is the resonance order, and the integers  $n$  and  $m$  are relatively prime. Now recall that above we have shown that the Verlet method has the timestep-dependent frequency  $\omega_{\text{eff}}$  given by  $\theta_{\text{eff}}^{\text{verlet}}/\Delta t$  where  $\theta_{\text{eff}}^{\text{verlet}}$  is defined in eq. (14.19); this frequency thus depends on the timestep in a nonlinear way. Given this frequency, the integrator-dependent resonance condition becomes

$$\frac{n}{m}\omega_{\text{eff}} = \frac{2\pi}{\Delta t}. \quad (14.21)$$

A “resonance of order  $n : m$ ” means that  $n$  phase space points are sampled in  $m$  revolutions:

$$n\theta_{\text{eff}} = n\Delta t\omega_{\text{eff}} = 2\pi m.$$

This special, finite-coverage of phase space can lead to incorrect, limited sampling of configuration space. See Figure 14.2 for an illustration.

As was shown in [821], equation (14.19) can be used to formulate *a condition for a resonant timestep* for the harmonic oscillator system. That is, using

$$\omega_{\text{eff}}^{\text{verlet}} = \frac{2\sin(\omega\Delta t/2)}{\Delta t} \quad (14.22)$$

with the resonance condition of eq. (14.21), we have

$$\frac{\omega\Delta t}{2} = \sin\left(\frac{m\pi}{n}\right). \quad (14.23)$$

Equivalently,

$$\Delta t_{n:m}^{\text{verlet}} = \frac{2}{\omega} \sin\left(\frac{m\pi}{n}\right) = \frac{T_p}{\pi} \sin\left(\frac{m\pi}{n}\right), \quad (14.24)$$

where  $T_p$  is the natural period of the oscillator.

Table 14.2 lists the values for low-order resonances (which are the most severe) for  $m = 1$  based on a period  $T_p = 10$  fs (or  $\omega = 0.63$  fs $^{-1}$ ). This period corresponds to the fastest period in biomolecular simulations, such as an O–H bond stretch (see Table 14.1). For  $n = 2$ , we have *linear stability*.

Clearly, since the limiting timesteps  $\Delta t_{n:1}$  for resonance orders  $n > 2$  are *smaller* than the linear stability limit ( $\Delta t_{2:1}$ ), resonance limits the timestep to values *lower* than classical stability. Since the third-order resonance leads to instability and the fourth-order resonance often leads to instability, in practice we require  $\Delta t < \Delta t_{4:1}$ . This implies for Verlet a stricter restriction than eq. (14.10), which comes from the second-order resonance (or linear stability), to

$$\Delta t < \sqrt{2}/\omega, \quad (14.25)$$

which corresponds to the fourth-order resonance ( $\Delta t_{2:1}$  and  $\Delta t_{4:1}$  are listed in Table 14.2).

### 14.2.6 Resonance Artifacts

For nonlinear systems, the effective forcing frequency for Verlet is not known, but for relatively small energy values, it is reasonable to approximate this effective frequency by expressions known for harmonic oscillators, as derived above. Such approximations can be appropriate, as shown in [821] and Figure 14.2.

This figure from [821] shows the resonance artifacts of the Verlet scheme in terms of peculiar phase-space diagrams (position versus momentum plot) obtained for a Morse oscillator<sup>3</sup> integrated at  $\Delta t = 2$  fs for increasing initial energies (or corresponding temperatures). Around this timestep, resonances of order 6 and higher are captured (see the reference resonant timesteps based on linear analysis in Table 14.2). The 6:1 resonance shows the 6 islands (each a closed orbit; see inset) covered in one revolution. As the initial energy increases, so does the resonance order: see the 13:2 and 20:3 resonances near the box periphery. Analysis of stability and resonance for MTS schemes can be similarly performed by extending the model to more than one frequency [94, 1089].

---

<sup>3</sup>The Morse bond potential and its relation to the harmonic potential is discussed in the Bond Length Potentials section of Chapter 9 (see eq. (9.7)).

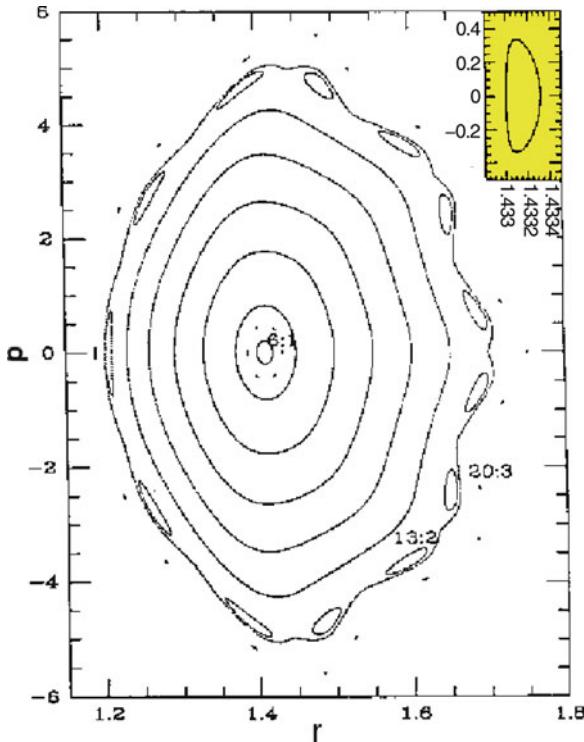


Figure 14.2. Morse oscillator phase diagrams obtained from the leapfrog Verlet integrator with  $\Delta t = 2$  fs at increasing initial energies. Resonances of order 6 and higher can be noted. One of the inner-most islands (near the 6:1 label) is enlarged in the inset to show that each corresponds to a closed orbit. See [821] for details.

Section 14.6 presents a resonance analysis for the implicit-midpoint (IM) scheme, and Box 14.5 derives IM's resonance condition (based on its phase-space rotation  $\omega_{\text{eff}}^{\text{IM}}$ ). Figure 14.16 illustrates resonance for a nonlinear system for both Verlet and IM.

## 14.3 Multiple-Timestep (MTS) Methods

### 14.3.1 Basic Idea

MTS schemes were introduced in the 1970s [1232, 1243] in an effort to reduce the computational cost of molecular simulations. Savings can be realized if the forces due to distant interactions are held constant over longer intervals than the short-range forces. Standard integration procedures can then be modified by evaluating the long-range forces less often than the short-range terms. Between updates, the slow forces can be incorporated into the discretization as a piecewise constant

Table 14.2. Stability limit and resonant timestep formulas for the Verlet scheme, with numerical values for a vibrational mode with period  $T_p = 10$  fs.

$n$ , res. order <sup>a</sup>	$\Delta t_{n:1}(\omega)$	$\Delta t_{n:1}(T_p)$	$\Delta t_{n:1}$ for MD <sup>b</sup>
2	$2/\omega$ [= $T_p/\pi$ ]	$0.318 T_p$	3.2 fs
3	$\sqrt{3}/\omega$	$0.276 T_p$	2.8 fs
4	$\sqrt{2}/\omega$	$0.225 T_p$	2.3 fs
5	$1.176/\omega$ [= $2 \sin(\pi/5)/\omega$ ]	$0.188 T_p$	1.9 fs
6	$1/\omega$ [= $T_p/2\pi$ ]	$0.159 T_p$	1.6 fs

<sup>a</sup>For  $n = 2$ , we have the linear stability condition.

<sup>b</sup>To compare resonance in Verlet to implicit-midpoint integration (see end of chapter), consider the analogous resonant timesteps  $\Delta t_{n:1}$  for  $n = 3, 4, 5, 6$  of 5.5, 3.2, 2.3, and 1.8 fs, respectively, for the implicit-midpoint scheme.

function (via *extrapolation*) or as a sum of delta functions (via *impulses*). See Figure 14.4 for a schematic illustration of these two approaches.

To illustrate, in the discussion below we consider below a force splitting scheme that involves three components for updating the fast, medium, and slow forces:

$$F_{\text{fast}}, F_{\text{med}}, \text{ and } F_{\text{slow}},$$

with corresponding timesteps

$$\Delta\tau \leq \Delta t_m \leq \Delta t.$$

The timesteps are related via the ratios

$$\begin{aligned} k_1 &= \Delta t_m / \Delta\tau, \\ k_2 &= \Delta t / \Delta t_m, \\ r &= k_1 k_2 = \Delta t / \Delta\tau, \end{aligned} \tag{14.26}$$

where  $k_1$  and  $k_2$  are integers. Note that when  $k_1 = k_2 = 1$  we have a single-timestep (STS) method.

### 14.3.2 Extrapolation

A simple, Verlet-based, extrapolative force-splitting approach is best formulated on the basis of *position Verlet* (eq. (13.15)) since the fast and medium forces are evaluated at the *middle* of the corresponding interval rather than at the beginning (and end). In programming style, where new iterates overwrite the old (i.e.,  $X \leftarrow X + \dots$ ), this extrapolative force-splitting scheme can be written as the doubly-nested loop for covering each  $\Delta t$  sweep shown in Figure 14.3, left-hand side.

Extrapolative MTS based on Position Verlet	Impulse MTS based on Velocity Verlet
$X_r^0 \equiv X + \frac{\Delta t_m}{2} V$ $\tilde{F}_{\text{slow}} \equiv -\mathbf{M}^{-1} \nabla E_{\text{slow}}(X_r)$	$\tilde{F}_{\text{slow}} \equiv -\mathbf{M}^{-1} \nabla E_{\text{slow}}(X)$ $V \leftarrow V + \frac{\Delta t}{2} \tilde{F}_{\text{slow}}$
<b>For</b> $j = 1$ to $k_2$ $X_r \equiv X_r^j \leftarrow X + \frac{\Delta t_m}{2} V$ $\tilde{F}_{\text{med}} \equiv -\mathbf{M}^{-1} \nabla E_{\text{med}}(X_r)$ $\tilde{F} \leftarrow \tilde{F}_{\text{med}} + \tilde{F}_{\text{slow}}$	<b>For</b> $j = 0$ to $k_2 - 1$ $\tilde{F}_{\text{med}} \equiv -\mathbf{M}^{-1} \nabla E_{\text{med}}(X)$ $V \leftarrow V + \frac{\Delta t_m}{2} \tilde{F}_{\text{med}}$
<b>For</b> $i = 1$ to $k_1$ $X \leftarrow X + \frac{\Delta \tau}{2} V$ $V \leftarrow V + \Delta \tau (\tilde{F} + \tilde{F}_{\text{fast}})$ $X \leftarrow X + \frac{\Delta \tau}{2} V$ <b>End</b>	<b>For</b> $i = 0$ to $k_1 - 1$ $V \leftarrow V + \frac{\Delta \tau}{2} \tilde{F}_{\text{fast}}$ $X \leftarrow X + \Delta \tau V$ $V \leftarrow V + \frac{\Delta \tau}{2} \tilde{F}_{\text{fast}}$ <b>End</b>
<b>End</b>	<b>End</b> $V \leftarrow V + \frac{\Delta t_m}{2} \tilde{F}_{\text{med}}$ $V \leftarrow V + \frac{\Delta t}{2} \tilde{F}_{\text{slow}}$

Figure 14.3. Algorithmic sketches of molecular dynamics integration by two force-splitting variants: *extrapolation* (left) versus *impulses* (right), based on position Verlet (extrapolation) and velocity Verlet (impulses).

Note that in this loop the *slow force* is evaluated *once* (at the point  $X_r^0$ ), the *medium force* is evaluated  $k_2$  *times* (for each  $X_r^j$ ), and the *fast force* is evaluated  $k_1 k_2$  (or  $r$ ) *times* at a corresponding midpoint. If the *slow force* calculations take the majority of the CPU time, the MTS approach will result in significant computational savings.

In practical implementations, the force distance classes are best treated by a smooth force-switching approach; see Chapter 10, Spherical-Cutoff section.

Simple extrapolation formulations, as above, were first tried for molecular dynamics [486, 1231, 1232, 1243, 1279]. However, these variants exhibited systematic energy drifts, a result of their nonsymplecticness.

### 14.3.3 Impulses

Work continued in the 1980s on MTS methods in a variety of contexts [1276, 1278], leading to the introduction in 1991/1992 (by Schulten and coworkers [486] and independently by Berne and coworkers [1277]) to an MTS method which

is symplectic and time-reversible. This similar MTS variant, based on impulses rather than extrapolation, was termed Verlet-I by the former group [486] and r-Respa [1277] by the latter.

The reversible Respa method was derived from a general Trotter factorization associated with the Liouville operator  $\mathcal{L}$ . Liouville operators are fundamental tools in statistical mechanics for the description of the canonical equations of motion of Hamiltonian systems. The Liouville operator can be decomposed into parts corresponding to different components of the energy using the reversible Trotter factorization. Here, we write the Liouville operator  $\mathcal{L}$  as a sum of three operators that characterize the scales of motions associated with different potential components:

$$\mathcal{L} = \mathcal{L}_{\text{fast}} + \mathcal{L}_{\text{med}} + \mathcal{L}_{\text{slow}}.$$

We then use a symmetric factorization of the components to arrive at (recall that  $k_1\Delta\tau = \Delta t_m$  and  $k_2\Delta t_m = \Delta t$ ):

$$\begin{aligned} \exp[i\Delta t \mathcal{L}] &= \exp[i\Delta t(\mathcal{L}_{\text{fast}} + \mathcal{L}_{\text{med}} + \mathcal{L}_{\text{slow}})] \\ &= \exp\left[i\left(\frac{\Delta t}{2}\right)\mathcal{L}_{\text{slow}}\right] \times \\ &\quad \left(\exp\left[i\left(\frac{\Delta t_m}{2}\right)\mathcal{L}_{\text{med}}\right] (\exp[i\Delta\tau\mathcal{L}_{\text{fast}}])^{k_1} \exp\left[i\left(\frac{\Delta t_m}{2}\right)\mathcal{L}_{\text{med}}\right]\right)^{k_2} \\ &\quad \times \exp\left[i\left(\frac{\Delta t}{2}\right)\mathcal{L}_{\text{slow}}\right] + \mathcal{O}(\Delta t^3). \end{aligned}$$

This factorization effectively shows that the propagation of the solution can be approximated by a combination of terms corresponding to several force components, each of which is resolved on a suitable timescale (e.g.,  $\Delta\tau/\Delta t_m/\Delta t$  for fast/medium/slow components). The middle term, corresponding to the fast components of the motion, is discretized with the Verlet method at a small timestep ( $\Delta\tau$ ).

A sweep over one  $\Delta t$  interval by an impulse-MTS Verlet approach, based on the leapfrog or velocity Verlet triplet (eq. 13.14), can be written as the doubly-nested iteration process shown in Figure 14.3, right-hand side.

#### 14.3.4 Vulnerability of Impulse Splitting to Resonance Artifacts

Note that the application of the slow force results in an *impulse*. The velocities are modified only outside of the inner loop (i.e., at the onset and at the end of a sweep covering  $\Delta t$ ) by a term proportional to  $r\Delta\tau$ , thus  $r$  times larger than the changes made to  $X$  and  $V$  in the inner loop. This is shown schematically in Figure 14.4 for a dual-timestep method with  $r = 4$  (see tall spikes at bottom).

Thus, as the time interval between slow-force updates increases, the size of these “impulses” grows. This causes a *resonance artifact* when the impulse frequency, or the MTS outer timestep  $\Delta t$ , occurs near a natural frequency of the

**MTS Protocol  $\{\Delta\tau, \Delta t\}$  where  $\Delta t = k\Delta\tau$**

### EXTRAPOLATION:

$$M\dot{V}(t) = \Delta\tau \sum_i \delta(t - i\Delta\tau)[F_{\text{fast}}(X(t)) + F_{\text{slow}}(X(t_i))],$$

$t_i = \Delta t \cdot [\text{largest integer}] < i/k]$



### IMPULSE:

$$M\dot{V}(t) = \Delta\tau \sum_i \delta(t - i\Delta\tau)[F_{\text{fast}}(X(t))] + \Delta t \sum_j \delta(t - j\Delta t)[F_{\text{slow}}(X(t))]$$

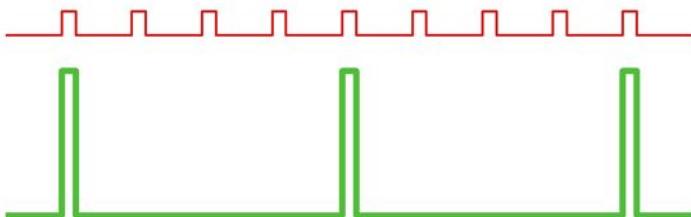


Figure 14.4. Schematic illustration of extrapolative vs. impulse force splitting integration for a dual-timestep protocol with inner timestep  $\Delta\tau$  and outer timestep  $\Delta t = k\Delta\tau$  ( $k = 4$  used). In extrapolative splitting, a slow-force contribution (green) is made each time the fast force (red) is evaluated (i.e., every  $\Delta\tau$  interval). In impulse splitting, in contrast, contributions of the slow forces (tall spikes) are considered only at the time of their evaluation (e.g., every four fast-force evaluations).

system. This artifact results because the long-range forces add energy at pulses that correspond to the natural frequency of the system. Of course, the true physical system experiences continuous variation of the long-range forces. Next we analyze resonance artifacts of MTS methods in more detail for a harmonic model.

#### 14.3.5 Resonance Artifacts in MTS

Analysis of a simple linear system is instructive to illustrate resonance in MTS schemes [94, 446]. Recent works [94, 1024, 1089] have shown that impulse methods are *generally stable* except at integer multiples of half the period of the fastest motion, with the severity of the *instability worsening* with the timestep.

Extrapolation methods are *generally unstable* for the Newtonian model problem, but the *instability is bounded* for increasing timesteps. Similar results hold for stochastic extensions of MTS [94].

### Simple Example

Figure 14.5 from [1089] illustrates this behavior as analyzed on a one-dimensional linear oscillator obeying the equations

$$\dot{X} = V, \quad \dot{V} = -(\lambda_1 + \lambda_2)X,$$

where  $X$  and  $V$  denote the scalar position and velocity, respectively, and a unit mass for the particle is assumed. The scalars  $\lambda_1 > \lambda_2$  represent two motion components differing in timescales. The characteristic angular frequencies associated with these components and respective periods are

$$\omega_i = \sqrt{\lambda_i}, \quad T_{p_i} = 2\pi/\omega_i, \quad i = 1, 2.$$

For the analysis of Figure 14.5, we set  $\lambda_1 = \pi^2$  and  $\lambda_2 = (\pi/5)^2$  to produce  $T_{p_2} = 10 = 5 T_{p_1}$  time units. The slow force component is defined as  $-\lambda_1 X$  and is updated at timesteps  $\Delta t$  that are  $k$  times larger than those ( $\Delta\tau$ ) used for the fast components,  $-\lambda_2 X$ .

In the linear analysis of [1089], eigenvalue magnitudes derived from the propagation matrices associated with the impulse and extrapolation force splitting schemes are plotted against the outer timestep. A scheme is *unstable* if the eigenvalue magnitude exceeds unity. Figure 14.5 shows results for both Newtonian (left) and Langevin (right) dynamics; the latter, stochastic dynamics approach is described in more detail below. The inner timestep was set to  $\Delta\tau = 0.001$  and the outer timestep to  $k\Delta\tau$  (compare to  $T_{p_1} = 2$  and  $T_{p_2} = 10$ ).

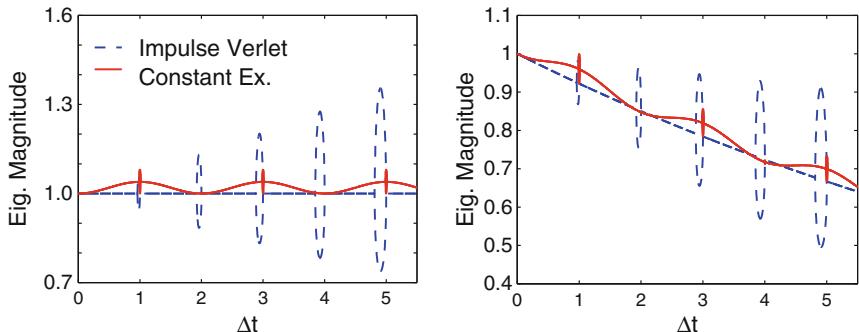


Figure 14.5. Propagation-matrix eigenvalue magnitudes for Newtonian (left) and Langevin (right,  $\gamma = 0.162$ ) *impulse* and *extrapolative* force-splitting methods as a function of the outer timestep as calculated for a one-dimensional linear test system governed by  $\ddot{X} = -(\lambda_1 + \lambda_2)X$  with  $\lambda_1 = \pi^2$ ,  $\lambda_2 = (\pi/5)^2$ , where  $X$  denotes the scalar position of the particle with unit mass. These settings produce associated characteristic periods for the fast and slow components of  $T_{p_1} = 2\pi/\sqrt{\lambda_1} = 2$  and  $T_{p_2} = 2\pi/\sqrt{\lambda_2} = 10$  time units. The inner timestep is  $\Delta\tau = 0.001$  [1089].

We see that the impulse method (dashed line) is unstable at integral multiples ( $m$ ) of half the fast period ( $m T_{p_1}/2$ , where  $T_{p_1}/2 = 1$  here). Furthermore, the severity of these corruptions increases with  $k$ , with the amplitudes of the resonant spikes increasing linearly with  $\Delta t$  and becoming wider.

While the use of impulses clearly leads to serious artifacts at certain timesteps, for other timesteps the impulse method is stable for this one-dimensional linear model. It might seem that avoiding the resonant timesteps is possible in practice. Unfortunately, experiments on a large nonlinear system show that once a resonant timestep is reached, good numerical behavior cannot be restored by increasing the timestep [1126]. See Figure 14.7 for an illustration of resonance on a nonlinear system.

For the extrapolation method, in contrast, we note in Figure 14.5 (solid curves) deviations from unit eigenvalues except around isolated outer timesteps. Since both eigenvalues exceed one in magnitude, the method is not volume-preserving or symplectic, explaining the energy drift observed in practice for force splitting by extrapolation. Still, the resonant spikes for constant extrapolation have magnitudes that are *independent of the outer timestep*, of values approximately  $1 + \lambda_2/\lambda_1$  ( $\lambda_2/\lambda_1 = 0.04$  here). Therefore, the larger the separation of frequencies, the more benign the numerical artifacts in practice.

### Extensions to Stochasticity and Nonlinearity

These resonance patterns generalize to stochastic dynamics, where friction and random terms are added (see next section), as seen in the same figure (the eigenvalue magnitudes decrease due to frictional damping). The value of  $\gamma$  used here ensures that the first spike for the impulse method has magnitude less than unity [1089]. Behavior is more complicated for a linear three-dimensional test problem [1089]. Still, applications to a nonlinear potential function for a protein, as shown in Figure 14.7 (discussed below), exhibit similar overall patterns for impulse versus extrapolative force splitting treatments.

#### 14.3.6 Limitations of Resonance Artifacts on Speedup; Possible Cures

The original work of Schulten and coworkers [486] expressed reservations regarding force impulses. A subsequent study by Biesiadecki and Skeel [134] discussed resonance as well as the systematic drift of extrapolative treatment in simple oscillator systems, though it conjectured that resonance artifacts were not likely to be so clear in nonlinear systems. This, unfortunately, turned out not to be true. Other articles noted a rapid energy growth [1349, 1448] when the interval between slow-force updates approaches and exceeds 5 fs, half the period of the fastest oscillations in biomolecules.

The presence of these resonances in impulse splitting limits the outermost timestep,  $\Delta t$  (or the interval of slow-force update), in MTS schemes. In turn, the overall speedup over a standard Verlet trajectory is capped. The achievable

speedup factor has been reported to be around 4–6 [1277, 1349, 1448] for biomolecules. However, this number depends critically on the reference STS (single-timestep) method to which MTS performance is compared. Since the inner timestep in MTS schemes is often small (e.g., 0.5 or even 0.25 fs), such speedup factors around 5 are obtained with respect to STS trajectories at 0.5 fs (10 with respect to 0.25 fs!). While accuracy is comparable when this small timestep is used, typically STS methods use larger timesteps, such as 1 or 2 fs (often with SHAKE). This means that the actual computational benefit from the use of MTS schemes in terms of CPU time per nanosecond, for example, is much less. Still, a careful MTS implementation with inner timestep  $\Delta\tau = 0.5$  fs can yield better accuracy than a STS method using 1 or 2 fs.

Given this resonance limitation on speedup, it is thus of great interest to revise these methods to yield larger speedups.

The mollified impulse method of Skeel and coworkers [446, 594, 595] has extended the outer timestep by roughly a factor of 1.5 (e.g., to 8 fs). With additional Langevin coupling, following the LN approach [95], the Langevin mollified method (termed LM) can compensate for the inaccuracies due to the use of impulses to approximate a slowly varying force. This is accomplished by substituting  $\mathcal{A}(X)^T F_{\text{slow}}(\mathcal{A}(X))$  for the slow force term where  $\mathcal{A}(X)$  is a time-averaging function.

Another approach altogether is to use extrapolation in the context of a stochastic formulation, as in the LN method [93–95] (see next section). This combination avoids the systematic energy drift, alleviates severe resonance, and allows much longer outer timesteps; see also discussion of masking resonances via the introduction of stochasticity in an editorial overview [665] and review [328].

Though the Newtonian description is naturally altered, the stochastic formulation may be useful for enhanced sampling. The contribution of the stochastic terms can also be made as small as possible, just to ensure stability [99, 1109]. For example, unlike the predictions in the above review [328], a smaller stochastic contribution (damping constant of 5 to 10 ps<sup>-1</sup>) has been used in the LN scheme without reducing the outer timestep, and hence without compromising the speedup [1230].

## 14.4 Langevin Dynamics

### 14.4.1 Many Uses

A stochastic alternative to Newtonian dynamics, namely Langevin dynamics, has been used in a variety of biomolecular simulation contexts for various numerical and physical reasons. The Langevin model has been employed to eliminate explicit representation of water molecules [966], treat droplet surface effects [180, 1189], represent hydration shell models in large systems [112–114], enhance sampling [298, 328, 513, 659, 783, 1036, 1428], and counteract numerical damping while masking mild instabilities of certain long-timestep approaches [95, 992,

1132, 1435, 1436]. See Pastor's comprehensive review on the use of the Langevin equation [966]. The Langevin equation is also discussed in Subsection 10.6.3 of Chapter 10 in connection with continuum solvation representations.

### 14.4.2 Phenomenological Heat Bath

The Langevin model is phenomenological [853] — adding friction and random forces to the systematic forces — but with the physical motivation to represent a simple heat bath for the macromolecule by accounting for molecular collisions. The continuous form of the simplest Langevin equation is given by:

$$\mathbf{M}\ddot{\mathbf{X}}(t) = -\nabla E(\mathbf{X}(t)) - \gamma\mathbf{M}\dot{\mathbf{X}}(t) + \mathbf{R}(t), \quad (14.27)$$

where  $\gamma$  is the collision parameter (in reciprocal units of time), also known as the damping constant. The random-force vector  $R$  is a stationary Gaussian process with statistical properties given by:

$$\langle \mathbf{R}(t) \rangle = 0, \quad \langle \mathbf{R}(t)\mathbf{R}(t')^T \rangle = 2\gamma k_B T \mathbf{M} \delta(t - t'), \quad (14.28)$$

where  $k_B$  is the Boltzmann constant,  $T$  is the target temperature, and  $\delta$  is the usual Dirac symbol.

### 14.4.3 The Effect of $\gamma$

The magnitude of  $\gamma$  determines the relative strength of the *inertial* forces with respect to the random (external) forces. Thus, as  $\gamma$  increases, we span the inertial to the *diffusive* (Brownian) regime. The Brownian range is used (with suitable algorithms) to explore configuration spaces of floppy systems efficiently. See [607, 1321], for instance, for applications to the large-scale opening/closing lid motion of the enzyme triosephosphate isomerase (TIM), and to the juxtaposition of linearly-distant segments in long DNA systems, respectively.

Figure 14.6 illustrates the effects of increasing  $\gamma$  on the trajectories and phase diagrams of a harmonic oscillator. The systematic harmonic motion and the closed, circular trajectories characteristic at zero viscosity change as the relative contribution of the random to systematic forces increases.

Since the stochastic Langevin forces mimic collisions between solvent molecules and the biomolecule (the solute), we see that the characteristic vibrational frequencies of a molecule in vacuum are damped. In particular, the low-frequency vibrational modes are overdamped, and various correlation functions are smoothed (see Case [202] for a review and further references). The magnitude of such disturbances with respect to Newtonian behavior depends on  $\gamma$  [95].

A physical value for  $\gamma$  for each particle can be chosen according to Stokes' law for a hydrodynamic particle of radius  $a$ :

$$\gamma = 6\pi\eta a/m, \quad (14.29)$$

where  $m$  is the particle's mass (not to be confused with the integer  $m$  used to define resonance), and  $\eta$  is the solvent viscosity. For example,  $\gamma = 50 \text{ ps}^{-1}$  is a

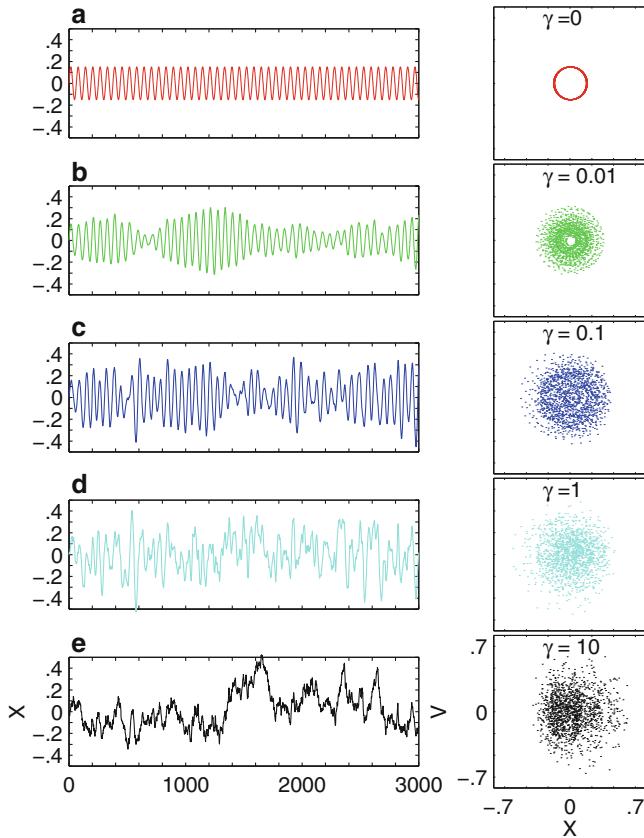


Figure 14.6. Langevin trajectories for a harmonic oscillator of angular frequency  $\omega = 1$  and unit mass simulated by a Verlet extension to Langevin dynamics at a timestep of 0.1 (about 1/60 the period) for various  $\gamma$ . Shown for each  $\gamma$  are plots for position versus time (left panels) and phase-space diagrams (right squares). Note that points in the phase-space diagrams are not connected and only appear connected for  $\gamma = 0$  because of the closed orbit.

typical collision frequency for protein atoms exposed to solvent having a viscosity of 1 cp at room temperature [967]; this value is also in the range of that estimated for water ( $\gamma = 54.9 \text{ ps}^{-1}$ ).

It is also possible to choose an appropriate value for  $\gamma$  for a system modeled by the simple Langevin equation so as to reproduce observed experimental translation diffusion constants,  $D_t$ . Namely, in the *diffusive limit*,  $D_t$  is related to  $\gamma$  by

$$D_t = k_B T / \sum m\gamma .$$

See [966, 1037] for example, for the use of these relations in studies of supercoiled DNA.

#### 14.4.4 Generalized Verlet for Langevin Dynamics

The Verlet algorithm can easily be generalized to include the friction and stochastic terms above. A common discretization is that described by Brooks, Brünger and Karplus, known as BBK [180, 967]:

##### Generalized Verlet Algorithm for Langevin Dynamics

$$\begin{aligned} V^{n+1/2} &= V^n + \mathbf{M}^{-1} \frac{\Delta t}{2} [-\nabla E(X^n) - \gamma \mathbf{M} V^n + R^n] \\ X^{n+1} &= X^n + \Delta t V^{n+1/2} \\ V^{n+1} &= V^{n+1/2} + \mathbf{M}^{-1} \frac{\Delta t}{2} [-\nabla E(X^{n+1}) - \gamma \mathbf{M} V^{n+1} + R^{n+1}] . \end{aligned} \quad (14.30)$$


---

This Langevin scheme reduces to velocity Verlet (triplet eq. (13.14)) when  $\gamma$  and hence  $R^n$  are zero.

Note that the third equation above defines  $V^{n+1}$  implicitly; the linear dependency, however, allows solution for  $V^{n+1}$  in closed form (i.e., explicitly). The superscript used for  $R$  has little significance, as the random force is chosen independent at each step.

When the Dirac delta function of eq. (14.28) is discretized,  $\delta(t - t')$  is replaced by  $\delta_{nm}/\Delta t$ .

The BBK method is only appropriate for the small- $\gamma$  regime. A more versatile algorithm was derived in [1293]. See also [1200] for a related Langevin discretization.

#### 14.4.5 The LN Method

The idea of combining force splitting via extrapolation with Langevin dynamics can alleviate severe resonance effects — as discussed in the MTS section (and shown in Figure 14.5 for a simple harmonic model) — and allow larger outer timesteps to be used than impulse-based MTS schemes [93–95].

The LN algorithm based on this combination (see background in Box 14.1) is sketched in Figure 14.8 in the same notation used for the MTS schemes. The direct-force algorithms on the left forms the basis for the algorithms implemented in the CHARMM [95] and AMBER [89, 97, 1025] programs.

Note that LN is based on position Verlet rather than velocity Verlet. If a constrained dynamics formulation is used (e.g., SHAKE), this splitting version requires on average two SHAKE iterations per inner loop rather than the one required by velocity Verlet. However, we have found stability advantageous in practice for the position Verlet scheme over velocity Verlet in the unusual limit of moderate to large timesteps and large timescale separations [99].

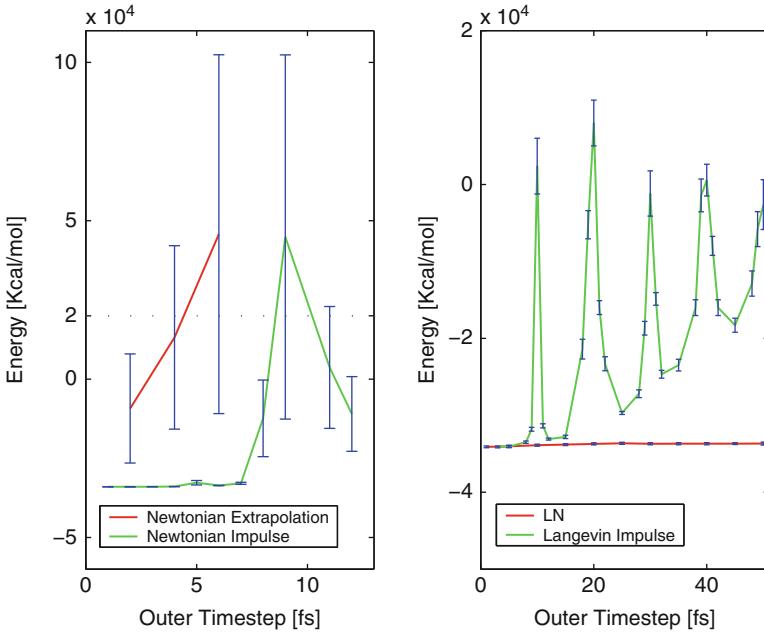


Figure 14.7. Energy means and deviations as computed over 5-ps Newtonian and Langevin impulse and extrapolative force splitting schemes for the protein BPTI with the damping constant  $\gamma = 20 \text{ ps}^{-1}$ , as functions of the outer timestep for a fixed inner timestep of 0.5 fs and medium timestep of 1 fs [1089].

### Resonance Alleviation

Applications to the solvated model of the protein BPTI in Figure 14.7 show how the Langevin extrapolative treatment alleviates severe resonances present in impulse treatments for both Newtonian and Langevin dynamics. Newtonian extrapolation yields large energy increases as the timestep increases. Both Newtonian and Langevin impulse splitting exhibit resonant spikes: beyond half the fastest period for Newtonian impulse, and beyond the fastest period for Langevin impulse splitting. Unfortunately, once such resonance is encountered, good behavior cannot be restored after the first resonant timestep. The LN combination, in contrast, does not exhibit resonant spikes for this value of damping constant,  $20 \text{ ps}^{-1}$ . It stabilizes the energy growth seen for Newtonian extrapolation. See also [1134] for detailed performance evaluations for a large protein/DNA complex.

The exchange of Hamiltonian dynamics for stochastic dynamics can guarantee better numerical behavior, but the resulting dynamics are altered. Stochastic dynamics are, however, suitable for many thermodynamic and sampling questions. Small-timestep dynamics simulations can always be performed in tandem once an interesting region of conformation space is identified.

(Direct Force Version)	LN Algorithm (Linearization Variant changes)
$X_r^0 \equiv X$	$X_r \equiv X + \frac{\Delta t_m}{2} V$
$F_{\text{slow}} \equiv -\nabla E_{\text{slow}}(X_r)$	
<b>For</b> $j = 1$ to $k_2$	
$X_r \equiv X_r^j \leftarrow X + \frac{\Delta t_m}{2} V$	$X_r \leftarrow X + \frac{\Delta t_m}{2} V; \quad \tilde{\mathbf{H}}_r \equiv \tilde{\mathbf{H}}(X_r)$
$F_{\text{med}} \equiv -\nabla E_{\text{med}}(X_r)$	$F_{\text{med}} \equiv -\nabla E_{\text{med}}(X_r) - \nabla E_{\text{fast}}(X_r)$
$F \leftarrow F_{\text{med}} + F_{\text{slow}}$	
<b>For</b> $i = 1$ to $k_1$	
Evaluate Gaussian force $R$	
$X \leftarrow X + \frac{\Delta \tau}{2} V$	$F_{\text{tot}} \leftarrow F - \tilde{\mathbf{H}}_r \cdot (X - X_r) + R$
$F_{\text{tot}} \leftarrow F + F_{\text{fast}}(X) + R$	
$V \leftarrow (V + \mathbf{M}^{-1} \Delta \tau F_{\text{tot}}) / \tilde{\gamma}$	
$X \leftarrow X + \frac{\Delta \tau}{2} V$	
<b>End</b>	
<b>End</b>	$[\tilde{\gamma} = 1 + \gamma \Delta \tau]$
	$[\tilde{\mathbf{H}}_r = \text{local Hessian approx.}]$

Figure 14.8. Algorithmic sketch of the LN scheme [95] for Langevin dynamics by extrapolative force-splitting based on position Verlet. The version on the left uses direct fast-force evaluations, while the reference version on the right (only alternative statements shown) uses linearization to approximate the fast forces over a  $\Delta t_m$  interval. The latter requires a local Hessian formulation,  $\tilde{\mathbf{H}}_r = \tilde{\mathbf{H}}(X_r)$ , at point  $X_r$  every time the medium forces are evaluated. See Box 14.1 for background details.

Skeel, Izaguirre, and coworkers have also adopted the stochastic coupling [594, 595] in the context of the mollified impulse method [1133] to extend the timestep beyond half the fast period. Because their  $\gamma$  is yet smaller, the agreement with Hamiltonian dynamics is better, but the largest possible outer timesteps and hence speedup factors are reduced.

An interesting recent variation called “NML” [1242] is an extension of a LIN [1435, 1436] developed as a way to increase the timestep in standard MD integration. The idea in LIN was to use implicit integration for the low-frequency modes and normal-mode analysis (NMA) for the high-frequency modes; Langevin rather than Newtonian dynamics was also used to dampen resonance effects, which limit the timestep to around 3.3 fs when all light-atom motions are considered. In NML, the costly NMA of LIN is replaced by Brownian dynamics for the high-frequency modes to keep the fast oscillations around their equilibrium values; the low-frequency modes are propagated by Langevin dynamics as in LIN.

---

### Box 14.1: LN Background and Assessment

**Background.** LN arose fortuitously [93] upon analysis of the range of harmonic validity of the Langevin/Normal-mode method LIN [1435, 1436]. Essentially, in LIN the equations of motion are linearized using an approximate Hessian and solved for the harmonic component of the motion; an implicit integration step with a large timestep then resolves the residual motion (see subsection on implicit methods). Approaches based on linearization of the equations of motion have been attempted for MD [67, 603, 1209, 1280], but computational issues ruled out general macromolecular applications. Indeed, the LIN method is stable over large timesteps such as 15 fs but the speedup is modest due to the cost of the minimization subproblem involved in the implicit discretization. The discarding of LIN’s implicit-discretization phase — while reducing the frequency of the linearization — in combination with a force splitting strategy forms the basis of the LN approach [95].

**Performance: Energetics and Speedup.** Performance of MTS schemes can be analyzed by ‘Manhattan plots’, as shown in Figure 14.9 for a large polymerase/DNA system of 41,973 atoms (illustrated in Figure 14.10) [1408]; that is, differences of mean energy components are reported as a function of the outer timestep  $\Delta t$  relative to STS Langevin simulations. For three LN protocols — using different combinations of  $\Delta\tau$ ,  $\Delta t_m$ ,  $\gamma$ , and bond constraints (SHAKE on or off) — these plots, along with corresponding CPU times and speedup, show that the first protocol has the optimal combination of low relative error in all energy components (below 3%) and low CPU time per physical time unit. The computational speedup factor is 4 or more in all cases.

**Performance: Dynamics.** The assignment of the Langevin parameter  $\gamma$  in the LN scheme ensures numerical stability on one hand and minimizes the perturbations to Hamiltonian dynamics on the other; we have used  $\gamma = 10 \text{ ps}^{-1}$  or smaller in biomolecular simulations. To assess the effect of  $\gamma$  of dynamic properties, the protocol-sensitive spectral density functions computed from various trajectories can be analyzed (see Figure 14.11 caption). The densities for solvated BPTI in Figure 14.11 show how the characteristic frequencies can be more closely approximated as  $\gamma$  is decreased; the densities for the large polymerase system in Figure 14.12 show, in addition, the good agreement between the STS Langevin and LN-computed frequencies for the same  $\gamma$ . This emphasizes the success of MTS integrators as long as the inner timestep is small. (Recall another illustration of this point for butane in Figure 13.5, where the average butane end-to-end distance is shown for STS versus MTS protocols).

Detailed comparisons of the evolution of various geometric variables (Figure 14.13) reflect the agreement between LN and the reference Langevin simulation as well [1408]. As expected, individual trajectories diverge, but the angular fluctuations are all in reasonable ranges. The flexibility of the DNA backbone angles is expected at the base pair near the kink induced by the polymerase [1408].

---

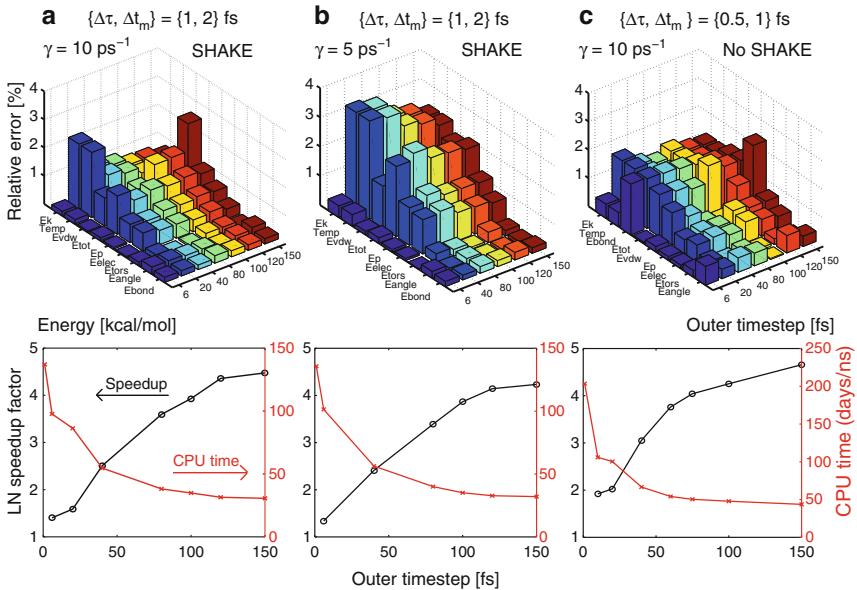


Figure 14.9. ‘Manhattan plots’ for a solvated polymerase/DNA system showing the LN relative errors in different energy components (bond, bond angle, torsion, electrostatic, potential, total, van der Waals) and the kinetic temperature as a function of the outer timestep  $\Delta t$  compared to single-timestep Langevin trajectories at the same  $\Delta\tau$ , as well as corresponding speedups (scale at left) and CPU times per nanosecond (scale at right). ‘SHAKE’ refers to constraining all bonds involving hydrogen. The data for three parameter combination in (a–c) are based on simulations of length 12 ps started from the intermediate polymerase/DNA complex [1408], pictured in Figure 14.10. See also Box 14.1.

### Testing and Application

Results have shown that good parameter choices for a 3-class LN scheme are  $\Delta\tau = 0.5$  fs,  $\Delta t_m = 1$  fs, and  $\Delta t$  up to 150 fs. If constrained dynamics for the light atom bonds are used, the inner timestep can be increased to 1 fs and the medium timestep to around 2 fs. Various biomolecular applications have shown good agreement of LN trajectories to small-timestep Langevin simulations and significant speedup factors [95, 1230, 1408]. See Box 14.1 for performance assessment of energetics, dynamics, and speedup on a biomolecule.

Recent work has tailored elements of the LN integrator to particle-mesh Ewald protocols [97, 99, 1024, 1025]. Challenges remain regarding the most effective MTS splitting procedures for Ewald formulations, given the fast terms present in the Ewald reciprocal component [89, 97, 98, 1025, 1110, 1236, 1449] and the numerical artifact stemming from subtraction of the term accounting for excluded-nonbonded atoms pairs [89, 1019, 1025].

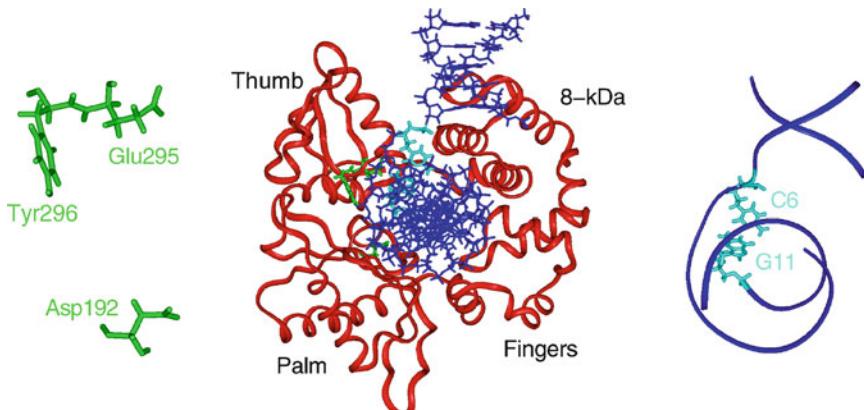


Figure 14.10. Solvated polymerase/DNA system used for evaluating LN [1408].

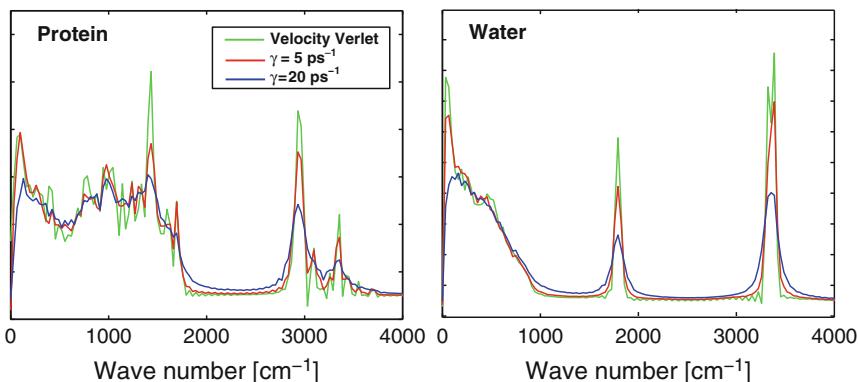


Figure 14.11. Spectral density functions calculated by Verlet and LN over 5 ps runs for a solvated system of the protein BPTI (protein and water frequencies shown separately) at three  $\gamma$  values: 0 (by the velocity Verlet scheme), 5 and  $20 \text{ ps}^{-1}$  by the LN scheme [1089]. The functions are computed by Fourier transforming the velocity autocorrelation time series for each atom in the system to obtain a power spectrum for each atom. These spectra are averaged over the water and biomolecule atoms separately for global characterization of the motion. See [1089] for the detailed protocol. Note that the characteristic frequencies obtained by this procedure reflect the force field constants rather than physical frequencies *per se*.

## 14.5 Brownian Dynamics (BD)

### 14.5.1 Brownian Motion

The mathematical theory of Brownian motion is rich and subtle, involving high-level physics and mathematics. Important contributors to the theory include Einstein (who explained Brownian motion in 1905) and Planck. Here we present

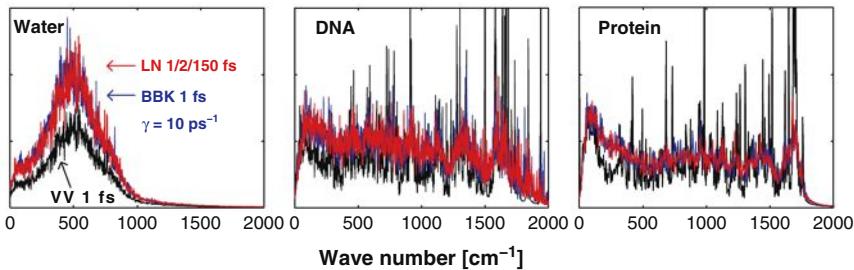


Figure 14.12. Spectral density functions calculated by the STS Newtonian (velocity Verlet) and STS Langevin (BBK,  $\gamma = 10 \text{ ps}^{-1}$ ) schemes versus the MTS LN scheme at  $\gamma = 10 \text{ ps}^{-1}$  (triple timestep protocol 1/2/150 fs) for a solvated polymerase/DNA system simulated over 12 ps and sampled every 2 fs. Spectral densities for the protein, DNA, and water atoms are shown separately. See also caption to Figure 14.11.

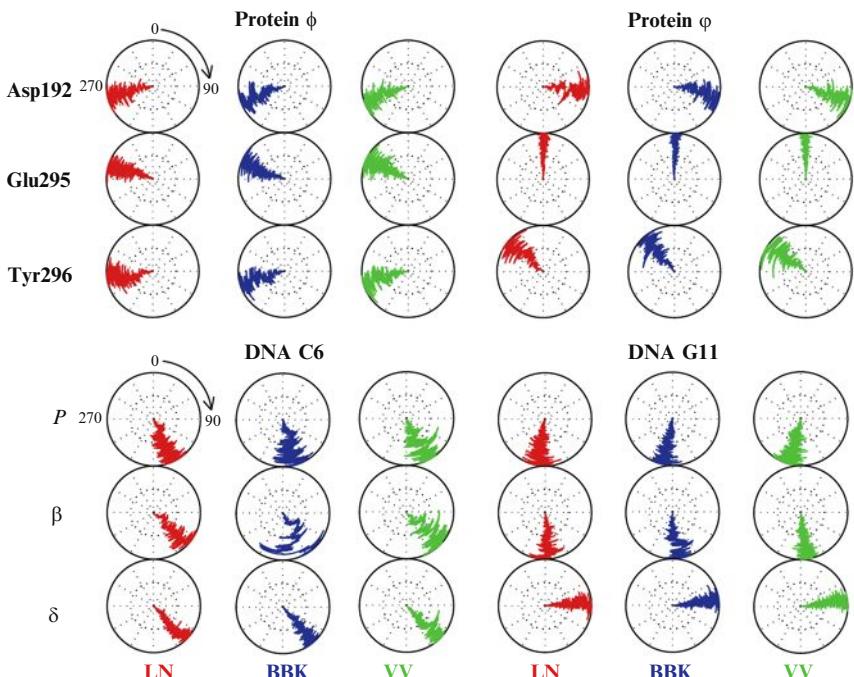


Figure 14.13. Comparisons of the evolution of representative dihedral angles (from protein  $\phi$  and  $\psi$  and DNA sugar and backbone near one CG base pair) over 12 ps for the STS velocity Verlet, STS Langevin (BBK), and the MTS LN scheme for a solvated polymerase/DNA system. The LN protocol used is  $\Delta\tau = 1 \text{ fs}$ ,  $\Delta t_m = 2 \text{ fs}$ , and  $\Delta t = 150 \text{ fs}$ , with SHAKE. The polar vector coordinates correspond to time and angle. See Figure 14.10 for location of residues.

the BD framework as an extension to the Langevin model presented above and focus on numerical algorithms for Brownian dynamics. See statistical mechanics texts, such as [853], for more comprehensive presentations.

The term *Brownian* is credited to the botanist Robert Brown, who in 1827 observed that fine particles — like pollen grains, dust, and soot — immersed in a fluid undergo a continuous irregular motion due to collisions of the particles with the solvent molecules. Dutch physician Jan Ingenhausz was actually the first, in 1785, to report this motion for powdered charcoal on an alcohol surface. The effective force on such particles originates from friction, as governed by Stokes' law, and a fluctuating random force.

### 14.5.2 Brownian Framework

#### Generalized Friction

Generalized frictional interactions among the particles can be incorporated into the Langevin equation introduced in Section 14.4 using a friction tensor  $\mathbf{Z}$ . This matrix replaces the single parameter  $\gamma$ , so as to describe the action of the solvent by:

$$\mathbf{M}\ddot{\mathbf{X}}(t) = -\nabla E(\mathbf{X}(t)) - \mathbf{Z}\dot{\mathbf{X}}(t) + \mathbf{R}(t), \quad (14.31)$$

where the mean and covariance of the random force  $R$  are given by:

$$\langle \mathbf{R}(t) \rangle = 0, \quad \langle \mathbf{R}(t)\mathbf{R}(t')^T \rangle = 2k_B T \mathbf{Z} \delta(t-t'). \quad (14.32)$$

The above relation is based on the fluctuation/dissipation theorem, a fundamental result showing how friction is related to fluctuations of the random force, assuming the Brownian particle is randomly moving about thermal equilibrium. This description ensures that the ensembles of the trajectories generated from eq. (14.31) are governed by the Fokker-Planck equation, a partial-differential equation describing the density function in phase space of a particle undergoing diffusive motion.

The random force  $R$  is white noise and has no natural timescale. Thus, the inertial relaxation times given by the inverses of the eigenvalues of the matrix  $\mathbf{M}^{-1} \mathbf{Z}$  define the characteristic timescale of the thermal motion in eq. (14.31).

#### Neglect of Inertia

When the inertial relaxation times are short compared to the timescale of interest, it is often possible to ignore inertia in the governing equation, that is, discard the momentum variables, assuming  $\mathbf{M}\ddot{\mathbf{X}}(t) = 0$ . From eq. (14.31), we have:

$$\dot{\mathbf{X}}(t) = -\mathbf{Z}^{-1} \nabla E(\mathbf{X}(t)) + \mathbf{Z}^{-1} \mathbf{R}(t), \quad (14.33)$$

and this can be written in the *Brownian dynamics* form:

$$\dot{\mathbf{X}}(t) = -\frac{\mathbf{D}}{k_B T} \nabla E(\mathbf{X}(t)) + \mathbf{R}_B(t); \quad (14.34)$$

here  $\mathbf{D}$  is the diffusion tensor

$$\mathbf{D} = k_B T \mathbf{Z}^{-1}, \quad (14.35)$$

and the mean and covariance of the random force  $R_B$  depend on  $\mathbf{D}$  as:

$$\langle R_B(t) \rangle = 0, \quad \langle R_B(t) R_B(t')^T \rangle = 2\mathbf{D} \delta(t - t'). \quad (14.36)$$

Thus, solvent effects are sufficiently large to make inertial forces negligible, and the motion is overall *Brownian* and random in character. This description is effective for very large, dense systems whose conformations in solution are continuously and significantly altered by the fluid flow in their environment.

### Transport Properties

Brownian theory allows us to determine average behavior, such as transport properties, for systems governed by such diffusional motion. For example, a molecule modeled as a free Brownian particle that moves a net distance  $x$  over a time interval  $\Delta t$  has an expected mean square distance  $\langle x^2 \rangle$  (analogous to the mean square end-to-end distance in a polymer chain) proportional to  $\Delta t$ :  $\langle x^2 \rangle \propto \Delta t$ . Einstein showed that the proportionality constant is  $2D$ , where  $D$  is the diffusion coefficient:

$$\langle x^2 \rangle = 2D \Delta t.$$

### Algorithms

An appropriate simulation algorithm for following Brownian motion can be defined by assuming that the timescales for momentum and position relaxation are well separated, with the former occurring much faster. BD algorithms then prescribe recursion recipes that displace each current position by a random force — similar in flavor to the Langevin random force — and additional terms that depend on the diffusion tensor.

It is not always possible to neglect inertial contributions. For example, it was shown that inertial contributions can affect long-time processes in long DNA due to mode coupling [107]. An “inertial BD” algorithm termed IBD has been developed, tested on simple systems [106], and applied to bead models of long DNA [107]. IBD more accurately approximates long-time kinetic processes that occur within equilibrium ensembles, as long as the timestep is not too small.

In practice, the IBD scheme adds a mass-dependent correction term to the usual BD propagation scheme (see below). Though IBD has an additional computational cost of a factor of two over BD, the computational complexity is the same as for the usual BD scheme based on the Cholesky factorization (consult a numerical methods textbook like [280] for the Cholesky factorization, and see Box 14.4 and below for IBD details).

### 14.5.3 General Propagation Framework

In general, larger timesteps are used for Brownian dynamics simulations than for molecular and Langevin dynamics simulations. As a first example, consider a free particle in one dimension whose diffusion constant  $D$  is (by definition) the mean square displacement divided by  $2t$ :

$$2tD \approx \langle |x(t) - x(0)|^2 \rangle$$

over sufficiently long times. This diffusional motion can be simulated by the simple scheme:

$$x^{n+1} = x^n + R^n, \quad (14.37)$$

where the random force  $R$  is related to  $D$  by:

$$\langle R^n \rangle = 0, \quad \langle (R^n)^2 \rangle = 2D\Delta t. \quad (14.38)$$

This propagation scheme reproduces  $D$  over long times  $t$ ; see homework assignment 12.

#### Ermak/McCammon

Ermak and McCammon [367] derived a basic BD propagation scheme from the generalized Langevin equation in the high-friction limit, where it is assumed that momentum relaxation occurs much faster than position relaxation. For a three-dimensional particle diffusing relative to other particles and subject to a force  $F(X)$ , the derived BD scheme becomes:

$$X^{n+1} = X^n + \frac{\Delta t}{k_B T} \mathbf{D}^n F(X^n) + R^n, \quad (14.39)$$

$$\langle R^n \rangle = 0, \quad \langle R^n (R^m)^T \rangle = 2\mathbf{D}^n \Delta t \delta_{nm}. \quad (14.40)$$

For reference, the IBD algorithm developed in [106] has the following mass-dependent correction term:

$$X^{n+1} = X^n + \frac{\Delta t}{k_B T} \mathbf{D}^n F(X^n) + \frac{1}{(k_B T)^2} \mathbf{D}^n M \mathbf{D}^n [F(X^{n-1}) - F(X^n)] + R^n, \quad (14.41)$$

with the same random-force properties for  $R$  as above (eq. (14.40)).

### 14.5.4 Hydrodynamic Interactions

Various approaches have been developed to define the diffusion tensor  $\mathbf{D}$  in these equations. Recall from Subsection 14.2.5 that the simple Langevin formulation offers only a simple isotropic description of viscous effects: frictional effects are taken to be isotropic, so  $\gamma$  is a scalar. Furthermore, the random force  $R$  acts independently on each particle, thereby ignoring changes in force due to the solvent-mediated dynamic interparticle interactions.

### Tensor $\mathbf{T}$

Account of these modified particle interactions due to solvent structure requires formulation of a configuration and momenta-dependent *hydrodynamic tensor*  $\mathbf{T}$  to express the instantaneous effective solute force. This is because each atom's force changes the solvent flow, and this in turn affects forces on other atoms through the frictional forces affecting them. The tensor  $\mathbf{T}$  is related to  $\mathbf{D}$  by the  $k_B T$  factor (eq. (14.43) below).

Various expressions for hydrodynamic tensors are derived from hydrodynamic theories (due to Kirkwood and Riseman, Oseen, Burgers, and others) so as to describe the effective velocity perturbation term,  $\Delta V^n$ , as the product of the tensor and force,  $\mathbf{T}(X^n) F(X^n)$ . Typically, the dependence of the tensor on momenta is ignored, and only configuration-dependent effects are considered. The governing BD equations that include hydrodynamic interactions then become:

$$\dot{X}(t) = \mathbf{T}(X) F(X(t)) + k_B T \nabla_X \cdot \mathbf{T}(X) + R_B(t), \quad (14.42)$$

where the gradient term  $\nabla_X \cdot \mathbf{T}(X)$  is the vector with components  $i$  given by  $\sum_j \partial \mathbf{T}_{ij}(r_{ij}) / \partial X_j$ .

The Oseen and Rotne-Prager tensors expressed below have the favorable property that the derivative of the tensor with respect to the displacement vector is zero, a term omitted in the BD propagation scheme of eq. (14.39) [367]. These tensors have been derived for polymer systems modeled as hydrodynamic (spherical) beads of radius  $a$  immersed in a fluid of viscosity  $\eta$  [1407]. For a polymer system of  $N$  beads, the diffusion tensor  $\mathbf{D}$  is then a configuration-dependent, symmetric  $3N \times 3N$  matrix in which each  $3 \times 3$  subblock  $\bar{\mathbf{D}}_{ij}$  is defined as:

$$\mathbf{D} = \begin{pmatrix} \bar{\mathbf{D}}_{11} & \bar{\mathbf{D}}_{12} & \dots & \bar{\mathbf{D}}_{1N} \\ \bar{\mathbf{D}}_{21} & \bar{\mathbf{D}}_{22} & \dots & \bar{\mathbf{D}}_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{\mathbf{D}}_{N1} & \bar{\mathbf{D}}_{N2} & \dots & \bar{\mathbf{D}}_{NN} \end{pmatrix}; \quad \bar{\mathbf{D}}_{ij} = (k_B T) \mathbf{T}_{ij}. \quad (14.43)$$

Box 14.2 outlines the principles on the basis of which the tensor approximations are derived.

The frictional coefficient  $\zeta$  introduced in the derivation of Box 14.3 can be related to the Langevin parameter  $\gamma$  by:  $\zeta = \gamma m$  if all beads have the same mass  $m$ . In fact, we can approximate roughly the reduction of the effective total friction for a polymer due to hydrodynamic effects using the Kirkwood-Riseman (KR) approximation for bead hydrodynamics based on the Oseen tensor [966]. See Box 14.3 for this approximation.

#### Box 14.2: Derivation of Tensor Expressions

For macromolecules in solution, the velocity of the solvent at a polymer segment differs from the velocity of the bulk solvent since the polymer alters the local flow. The solvent motion can thus be considered as an applied force. Let  $u_i$  be the velocity of a polymer

particle (e.g., bead),  $v_i$  be the velocity of the solvent at that particle, and  $v^0$  be the velocity of the bulk solvent. Then  $v_i - u_i$  represents the net velocity, proportional to the effective force on the solvent at particle  $i$ :

$$F_i = -\zeta(v_i - u_i), \quad (14.44)$$

where  $\zeta$  is the translational friction coefficient of a particle. For example,  $\zeta = 6\pi\eta a$  for a bead of radius  $a$  and solvent viscosity  $\eta$ ; assume for simplicity that  $\zeta$  is the same for all beads. The velocity at the  $i$ th bead can be written as  $v^0$  plus some perturbation due to all other segments:

$$v_i = v^0 + \sum_{j \neq i} \mathbf{T}_{ij} F_j. \quad (14.45)$$

By inserting eq. (14.45) into eq. (14.44), we have:

$$F_i = -\zeta \left( v^0 + \sum_{j \neq i} \mathbf{T}_{ij} F_j - u_i \right) = -\zeta(v^0 - u_i) - \zeta \sum_{j \neq i} \mathbf{T}_{ij} F_j. \quad (14.46)$$

It follows that

$$F_i + \zeta \sum_{j \neq i} \mathbf{T}_{ij} F_j = -\zeta(u_i - v^0). \quad (14.47)$$

This equation is the basis for deriving the hydrodynamic tensor  $\mathbf{T}$ , an equation that must be solved to find the net force for a given system.

---

### Oseen Tensor

One of the simplest hydrodynamic tensors is the Oseen tensor, developed by Oseen and Burgers (circa 1930) for the hydrodynamic interaction between a point of friction and a fluid, defined as:

$$\mathbf{T}_{ij} = \begin{cases} \frac{1}{6\pi\eta a} \mathbf{I}_{3 \times 3} & i = j \\ \frac{1}{8\pi\eta r_{ij}} \left[ \mathbf{I}_{3 \times 3} + \frac{\mathbf{r}_{ij} \mathbf{r}_{ij}^T}{r_{ij}^2} \right] & i \neq j \end{cases}. \quad (14.48)$$

Here  $\mathbf{r}_{ij}$  denotes an interbead distance vector, and  $r_{ij}$  is the corresponding scalar distance. The expression for  $i \neq j$  is the first term in an expansion corresponding to the pair diffusion for an incompressible fluid in inverse powers of  $r_{ij}$ .

### Rotne-Prager Tensor

Another term in the expansion produces the Rotne-Prager hydrodynamic tensor  $\mathbf{T}_{ij}$ . For nonoverlapping beads ( $r_{ij} > 2a$ ), this matrix is defined as [1071]:

$$\mathbf{T}_{ij} = \begin{cases} \frac{1}{6\pi\eta a} \mathbf{I}_{3 \times 3} & i = j \\ \frac{1}{8\pi\eta r_{ij}} \left[ \left( \mathbf{I}_{3 \times 3} + \frac{\mathbf{r}_{ij} \mathbf{r}_{ij}^T}{r_{ij}^2} \right) + \frac{2a^2}{r_{ij}^2} \left( \frac{\mathbf{I}_{3 \times 3}}{3} - \frac{\mathbf{r}_{ij} \mathbf{r}_{ij}^T}{r_{ij}^2} \right) \right] & i \neq j \end{cases} \quad (14.49)$$

Note that both tensors are a generalization of the scalar Langevin friction term expressed in terms of the  $\gamma$  set according to Stokes' law.

---

### Box 14.3: Approximation of The Effective BD Friction Coefficient

In the absence of hydrodynamic interactions (the “free-draining” limit), effective friction constant  $f_T$  is the sum of the friction constants of the  $N_b$  individual beads, i.e.,  $f_T = N_b\zeta = N_b(\gamma m)$ . The Kirkwood-Riseman (KR) equation for  $f_T$ , which approximately incorporates hydrodynamic interactions, can be written as:

$$f_T = \frac{N_b\zeta}{1 + \frac{\zeta}{6\pi\eta N_b} \sum_{i \neq j} \langle \frac{1}{r_{ij}} \rangle}, \quad (14.50)$$

where  $\langle 1/r_{ij} \rangle$  is the mean inverse distance between beads  $i$  and  $j$  averaged over an ensemble of configurations. Thus, the KR approximation is based on a preaveraged Oseen configuration tensor [966]. If we use Stokes' law to set  $\zeta$  as  $6\pi\eta a$ , we obtain:

$$f_T = \frac{N_b\zeta}{1 + \frac{a}{N_b} \sum_{i \neq j} \langle \frac{1}{r_{ij}} \rangle}. \quad (14.51)$$

Thus, the denominator of this expression reflects the reduction by hydrodynamics of the effective friction from the reference value of  $N_b\zeta$ . This approximation has been used to estimate hydrodynamic effects for long DNA [1037].

---

### 14.5.5 BD Propagation Scheme: Cholesky vs. Chebyshev Approximation

Once the matrix  $\mathbf{D}$  is formulated at each step  $n$  of the BD algorithm from the hydrodynamic tensor (eq. (14.43)) the random force  $R$  must be set to satisfy eq. (14.40). *This is actually the computationally intensive part of the BD propagation scheme when hydrodynamics interactions are considered.*

The traditional way, based on a Cholesky decomposition of  $\mathbf{D}$  (see [280] and Box 14.4) increases in computational time as the cube of the system size, since a Cholesky factorization of  $\mathbf{D}$  is required at every step. The alternative approach based on Chebyshev polynomials proposed by Marshall Fixman [404] only increases in complexity roughly with the square of the number of variables. See Box 14.4 for details.

Essentially, both methods for determining the random force  $R$  first compute a  $3N$ -vector  $Z$  from a Gaussian distribution so that

$$\langle Z_i \rangle = 0, \quad \langle Z_i Z_j \rangle = 2\Delta t \delta_{ij};$$

(the indices run from 1 to  $3N$ ). The second step is different. In the Cholesky-based approach,  $R$  is computed as  $R^n = \mathbf{S}^n Z$ , where  $\mathbf{S}$  is the Cholesky factor of  $\mathbf{D}$ . In the Chebyshev procedure, we compute the random force vector instead as

$R^n = \tilde{\mathbf{S}}^n Z$ , where  $\tilde{\mathbf{S}}$  is the square root matrix of  $\mathbf{D}$ , and the product is computed as a series of Chebyshev polynomials (see Box 14.4).

A recent application of the Chebyshev approach for computing the Brownian random force in simulations of long DNA demonstrates computational savings for large systems [1119]. Figure 14.14 compares the percentage CPU work required for the systematic versus hydrodynamic forces in BD simulations of DNA modeled as macroscopic hydrodynamic beads using the Cholesky versus Chebyshev approaches. The figure also shows the total CPU time required to simulate 10 ms in both cases.

We see that for the largest, 12,000 base-pair DNA system studied, the Chebyshev approach is twice as fast. The overall speedup is not more dramatic since the system size (in terms of beads) is relatively small.

Perhaps more significantly, the Chebyshev alternative to the Cholesky factorization also opens the door to other BD protocols (such as the recent inertial BD idea [106, 107]) and is crucial to BD studies of finer macroscopic models of DNA, such as residue-based rather than bead-based.

Finally, note that once the BD computational bottleneck is reduced to electrostatics and hydrodynamics (roughly  $\mathcal{O}(N^2)$  for both), fast electrostatic methods, such as described in Chapter 10, will help accelerate such BD computations further.

#### Box 14.4: BD Implementation: Cholesky vs. Chebyshev Approach

In the Cholesky approach to computing the random force to satisfy the properties of eq. (14.40), the Cholesky decomposition of the diffusion tensor  $\mathbf{D}$  is determined:

$$\mathbf{D} = \mathbf{S}\mathbf{S}^T,$$

where  $\mathbf{S}$  is a lower triangular matrix. The desired vector  $R^n$  is then computed from the following matrix/vector product:

$$R^n = \mathbf{S}^n Z. \quad (14.52)$$

It can be easily shown that this  $R^n$  satisfies  $\langle R^n (R^m)^T \rangle = 2\mathbf{D}^n \Delta t \delta_{nm}$ , as desired.

Note that the Cholesky decomposition of  $\mathbf{D}$  can be written in closed form as the following procedure for determining the elements of  $\mathbf{S}$ ,  $s_{ij}$ , row by row (i.e.,  $s_{11}, s_{21}, s_{22}, s_{31}, s_{32}, s_{33}, \dots, s_{3N} s_{NN}$ ):

$$s_{ij} = \begin{cases} \left( D_{ii} - \sum_{k=1}^{i-1} s_{ik}^2 \right)^{1/2} & i = j \\ \left( D_{ij} - \sum_{k=1}^{j-1} s_{ik} s_{jk} \right) / s_{jj} & i > j \end{cases}. \quad (14.53)$$

An advantage of the Cholesky approach is that the factors can be reused. Thus, it is possible to reduce the overall cost of the BD simulation by less-frequent updating of the hydrodynamic tensor; parallelization of some of the numerical linear algebra tasks can also further accelerate the total computational time [577, 607].

The Chebyshev alternative was proposed over a decade ago by Marshall Fixman [404] and recently applied [605, 681, 1119]. Instead of a Cholesky decomposition  $\mathbf{D} = \mathbf{S}\mathbf{S}^T$ ,

Fixman suggests to calculate  $R$  from the relation

$$R^n = \tilde{\mathbf{S}}^n Z, \quad (14.54)$$

rather than eq. (14.52) where  $\tilde{\mathbf{S}}$  is the *square root matrix* of  $\mathbf{D}$ :

$$\mathbf{D} = \tilde{\mathbf{S}}^2.$$

This idea is based on expanding the matrix/vector product  $\tilde{\mathbf{S}}Z$  as a series of Chebyshev polynomials which approximate the function  $\sqrt{x}$  on some given interval believed to contain eigenvalues. This calculation requires about  $\mathcal{O}(N^2)$  operations, thus reflecting substantial savings with respect to the Cholesky approach.

The details of computing the expansion of  $R$  in this Chebyshev approximation were recently given in the appendix of [1119] in an application of BD to supercoiled DNA. See also [605,681] for other applications. The Chebyshev implementation requires determining bounds on the maximum and minimum eigenvalues of  $\mathbf{D}^n$  and then computing an expansion of desired order (according to some error criterion) of  $R$  in terms of polynomials, with coefficients determined for the square-root function.

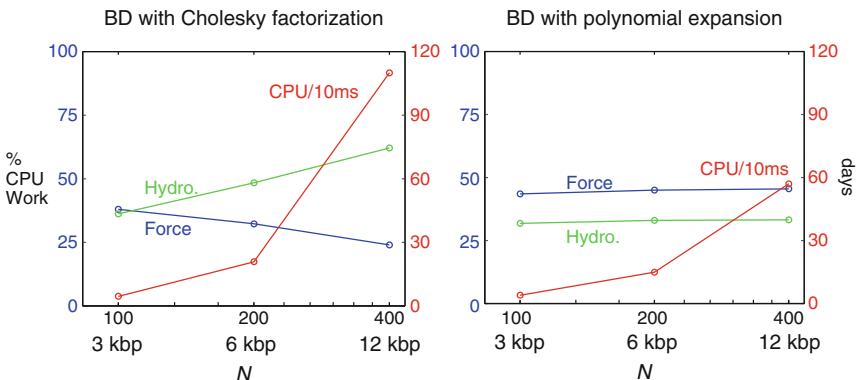


Figure 14.14. Computational complexity for BD schemes with hydrodynamics as obtained for a bead model of DNA using Cholesky and Chebyshev approaches for computing the Brownian random force. The fraction of CPU associated with the hydrodynamics and systematic force calculations is shown in each case (% scales at left) as a function of system size, with corresponding total number of days required to compute a 10 ms trajectory (shown with the scale at right). Computations are reported on an SGI Origin 2000 machine with 300 MHz R12000 processor [1119].

## 14.6 Implicit Integration

A reasonable approach for allowing long timesteps is the use of *implicit* integration schemes [447]. These methods are designed specifically for problems with disparate timescales where explicit methods do not usually perform well, such

as chemical reactions [495]. The integration formulas of implicit methods are designed to increase the range of stability for the difference equation.

Typically, implicit methods work well when the fast components are rapidly decaying and largely decoupled from the slower components. This is not the case for biomolecules, where vibrational modes are intricately coupled and the fast motions are oscillatory.

### 14.6.1 Implicit vs. Explicit Euler

To illustrate the use of implicit methods, we consider a simple example for the solution of the general differential equation

$$\dot{Y}(t) = \mathcal{F}[Y(t)], \quad (14.55)$$

where  $Y = (X, V)$  is a vector and  $\mathcal{F}$  is a vector function. The Newtonian system of equations (13.7, 13.8) can be written in this form with the composite vector  $Y = (X, \dot{X})$  and the function

$$\mathcal{F}[Y(t)] = [V(t), -\mathbf{M}^{-1}\nabla E(X(t))].$$

The *implicit-Euler* (IE) scheme discretizes eq. (14.55) as

$$(Y^{n+1} - Y^n)/\Delta t = \mathcal{F}(Y^{n+1}), \quad (14.56)$$

while the *explicit Euler* (EE) analog has the different right-hand-side:

$$(Y^{n+1} - Y^n)/\Delta t = \mathcal{F}(Y^n). \quad (14.57)$$

In the former case, the solution  $Y^{n+1}$  is derived implicitly, since  $\mathcal{F}(Y^{n+1})$  is not known and must be solved by some procedure. In the latter, the solution can be explicitly formulated in terms of quantities known from previous steps.

Though the explicit approach is simpler to solve, it imposes a severe restriction on the timestep. Implicit schemes can yield much better stability behavior for ‘stiff’ problems [495]. This can be seen in Figure 14.15, where we solve the one-dimensional problem  $y' = -ay, a > 0$ , whose exact solution is  $y = \exp(-at)$  by the implicit and explicit Euler methods.

It can be shown that the former gives the recursion relation

$$y^n = y^0/(1 + a\Delta t)^n,$$

while the latter gives

$$y^n = y^0(1 - a\Delta t)^n.$$

The IE scheme is always stable since  $a\Delta t > 0$ , but EE requires that  $a\Delta t < 2$ .

The experience with implicit methods in the context of biomolecular dynamics has been limited and rather disappointing (e.g., [604, 1437]). The disappointment stems from three basic reasons: intrinsic damping, large computational demands, and resonance artifacts.

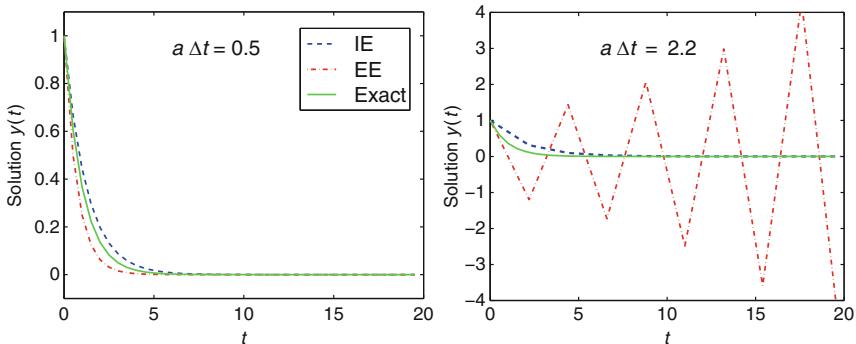


Figure 14.15. Implicit (IE) and explicit (EE) Euler solutions to the one-dimensional differential equation  $y' = -ay$  with  $a > 0$  are shown. They indicate stability of IE at small and large timesteps but growing instability of EE with time if the timestep is not sufficiently small. Here we use  $y^0 = 1$  and two ratios for  $a\Delta t$  (0.5, left, and 2.2, right). The corresponding IE and EE solutions (see text) yield, respectively,  $y^n = 1/(1.5)^n$  and  $y^n = (0.5)^n$  for the smaller timestep, and  $y^n = 1/(3.2)^n$  and  $y^n = (-1.2)^n$  for the larger timestep.

### 14.6.2 Intrinsic Damping

Despite the high stability of the IE discretization, the IE scheme is nonconservative and exhibits intrinsic damping. For system (14.27), the IE discretization is:

$$\begin{aligned} \mathbf{M}(V^{n+1} - V^n)/\Delta t &= F(X^{n+1}) - \gamma \mathbf{M}V^{n+1} + R^{n+1}, \\ (X^{n+1} - X^n)/\Delta t &= V^{n+1}. \end{aligned} \quad (14.58)$$

This scheme was introduced to molecular dynamics with the Langevin restoring force over a decade ago [992, 1132], but its  $\gamma$  and timestep-dependent damping [992, 1107] ruled out general applications to biomolecules. This was concluded from the tendency of the scheme to preserve ‘local’ structure (as measured by ice-like features of liquid water [1120], the much slower decay of autocorrelation functions [298, 927], or lower effective temperature [298]).

Thus, at large timesteps, the IE method was numerically stable but behaved like an energy minimizer. Still, the IE scheme was incorporated as an ingredient in dynamic simulations of macroscopic separable models in which high-frequency modes are not relevant [1105, 1125, 1127, 1128], and in enhanced sampling schemes with additional mechanisms that counteract numerical damping [298, 299, 513].

### 14.6.3 Computational Time

In addition to the general damping problem, there is often no net computational advantage in implicit schemes. Each timestep, albeit longer, is much more expensive than an explicit timestep because a subproblem must be solved at each step.

This subproblem arises because the solution for  $X^{n+1}$  and  $V^{n+1}$  cannot be solved in closed form (i.e., in terms of previous iterates) as in explicit schemes. Note that the  $(n + 1)$  iterates of  $V$  or  $X$  appear on both the right and left-hand sides of eq. (14.58). Solving for  $X^{n+1}$  requires solution of a nonlinear system.

It has been shown that this solution can be formulated as a nonlinear optimization problem and solved using efficient Newton schemes [992, 1437], such as the truncated-Newton method introduced in Chapter 11. As fast as the minimization algorithm might be, the added cost of the minimization makes implicit schemes competitive with explicit, small-timestep integrators only at very large timesteps, a regime where reliability is questionable given the damping and resonance effects.

#### 14.6.4 Resonance Artifacts

Resonance, the third problem associated with implicit methods, is relevant to implicit schemes since large timesteps may yield stable methods [821].

A detailed examination of resonance artifacts with the implicit midpoint (IM) was described in [821]. IM differs from IE in that it is symmetric and symplectic. It is also special in the sense that the transformation matrix for the model linear problem is unitary, partitioning kinetic and potential-energy components identically. For these reasons, several researchers have explored the application of IM to long-timestep dynamics [468, 632, 1186, 1187].

##### IM Analysis

The IM discretization applied to system (14.27) is:

$$\begin{aligned} \mathbf{M}(V^{n+1} - V^n)/\Delta t &= \frac{-\nabla E}{2}(X^n + X^{n+1}) - \frac{\gamma \mathbf{M}}{2}(V^n + V^{n+1}) + R^n, \\ (X^{n+1} - X^n)/\Delta t &= (V^n + V^{n+1})/2. \end{aligned} \quad (14.59)$$

This nonlinear system can be solved, following [992], by obtaining  $X^{n+1}$  as a minimum of the “dynamics function”  $\Phi(X)$ :

$$\Phi(X) = \frac{\tilde{\gamma}}{2}(X - X_0^n)^T \mathbf{M}(X - X_0^n) + (\Delta t)^2 E\left(\frac{X + X^n}{2}\right), \quad (14.60)$$

where

$$X_0^n = X^n + (\Delta t V^n)/\tilde{\gamma} + \mathbf{M}^{-1}(\Delta t)^2 R^n/(2\tilde{\gamma}),$$

and

$$\tilde{\gamma} = 1 + (\gamma \Delta t)/2.$$

Hence, for IM applied to Newtonian dynamics  $\tilde{\gamma} = 1$ , and the  $R^n$  term in  $X_0^n$  is absent. Following minimization of the IM dynamics function to obtain  $X^{n+1}$ , the new velocity,  $V^{n+1}$ , is obtained from the second equation of system (14.59).

An examination of the application of IM to MD showed good numerical properties (e.g., energy conservation and stability) for moderate timesteps, larger than

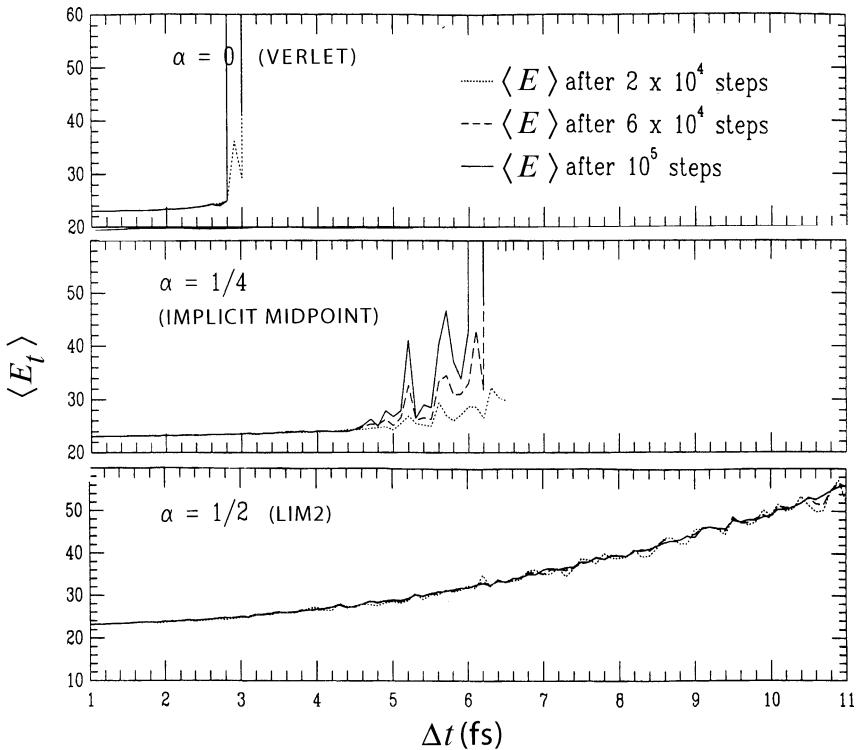


Figure 14.16. Mean total energy for a blocked alanine model as a function of the timestep, as obtained by Verlet, implicit-midpoint, and the symplectic implicit scheme LIM2 [1437]; see [1126].

Verlet [821, 1437]. However integrator-induced *resonance* artifacts were found to be severe as the timestep approaches half the characteristic period [821] (see Box 14.5).

Figure 14.16 shows the mean total energy for Verlet and IM for a blocked alanine model [1126] (misnamed ‘dipeptide’ in some works). Note that Verlet becomes unstable around 2.8 fs (recall Table 14.2 for reference linear-analysis values), and IM exhibits resonance peaks around 5.2 and 5.7 fs and instability around 6 fs, all near half the period of the fast motion.

#### Box 14.5: Analysis of Implicit-Midpoint (IM) Resonance

The linear analysis of [821] has shown that the *effective angular rotation*  $\theta_{\text{eff}}$  of IM is:

$$\theta_{\text{eff}}^{\text{IM}} = \frac{2}{\Delta t} \tan^{-1} (\omega \Delta t / 2). \quad (14.61)$$

Recall that the corresponding value for the Verlet scheme (eq. (14.19)) is:

$$\theta_{\text{eff}}^{\text{verlet}} = 2 \sin^{-1} (\omega \Delta t / 2).$$

As derived above, the resonant timestep of order  $n:m$  can be obtained for the model linear problem by using the function for  $\omega_{\text{eff}}$  in the expression

$$n \Delta t \omega_{\text{eff}} = 2\pi m,$$

since those timesteps correspond to a sampling of  $n$  phase-space points in  $m$  revolutions.

This substitution yields for the IM scheme

$$\Delta t_{n:m}^{\text{IM}} = \frac{2}{\omega} \tan \left( \frac{\pi m}{n} \right) = \frac{T_p}{\pi} \tan \left( \frac{\pi m}{n} \right). \quad (14.62)$$

For  $n = 3, 4$  and  $m = 1$ , we have, for example:  $\Delta t_{3:1} = 2\sqrt{3}/\omega$  and  $\Delta t_{4:1} = 2/\omega$  for IM. These resonance times are larger than corresponding values for Verlet:  $\sqrt{3}/\omega$  and  $\sqrt{2}/\omega$ , respectively. Hence IM confers better stability than Verlet.

If we consider a fast period of 10 fs (see Table 14.2), the IM resonant timesteps for orders  $n = 3, 4, 5, 6$  are 5.5, 3.2, 2.3, and 1.8 fs, respectively, compared to 2.8, 2.3, 1.9, and 1.6 fs for Verlet. This analysis explains the results shown in Figure 14.16.

---

### A Family of Symplectic Implicit Schemes

The IE and IM methods described above turn out to be quite special in that IE's damping is extreme and IM's resonance patterns are quite severe relative to related symplectic methods. However, success was not much greater with a symplectic implicit Runge-Kutta integrator examined by Janežič and coworkers [604].

A family of implicit symplectic methods including IM and Verlet (as a special case) was also explored in subsequent works by Skeel, Schlick, and coworkers [446, 1126]. The family of methods can be parameterized via  $\alpha$  (see Box 14.6). The values  $\alpha = 0$ ,  $\alpha = 1/2$ , and  $1/4$  correspond to the Verlet, 'LIM2', and IM methods, respectively. The LIM2 method appeared attractive based on the effective rotation analysis (see Box 14.6), since it should exhibit no erratic resonance patterns; this was indeed observed in practice (Figure 14.16.) Unfortunately, this resonance removal comes at the price of large increases in energy with the timestep, as also seen in Figure 14.16.

These scheme-dependent resonances were concluded by the analysis in [1126] on the basis of the effect of the parameter  $\alpha$  on the numerical frequency of the integrator for the system simulated. Specifically, the maximum possible phase angle change per timestep *decreases* as the parameter  $\alpha$  *increases*. Hence, the angle change can be limited by selecting a suitable  $\alpha$ .

The choice  $\alpha \geq \frac{1}{2}$ , as in the LIM2 method, restricts the phase angle change to less than one quarter of a period and thus is expected to eliminate notable disturbances due to fourth-order resonance.

The requirement that  $\alpha \geq \frac{1}{3}$  guarantees that the phase angle change per timestep is less than one third of a period and therefore should also avoid third-order resonance for the model problem.

This was verified in an application to a representative nonlinear system, a blocked alanine model [1126] in Figure 14.16. This figure contrasts the early (i.e., with small timestep) instability of Verlet with the resonance of IM around one half the fastest period, and the absence of resonances in LIM2 ( $\alpha = 1/2$ ).

Unfortunately, the increases of the average energy in LIM2 with the timestep are unacceptable: they reflect values that are approximately 30% and 100% larger than the small-timestep value, for  $\Delta t = 5$  and 9 fs, respectively, for this system. Part of this behavior is also due to an error constant for LIM2 that is greater than that of leapfrog/Verlet.

### Perspective

In sum, it is difficult to expect reasonable resolution by implicit methods in the large-timestep regime given the stability limit and resonance problems mentioned above. Perhaps semi-implicit [847] or cheaper implementations of implicit schemes [1440] will better handle this problem: it might be possible to treat the local terms implicitly and the nonlocal terms explicitly. Exploitation of parallel machine architecture has potential for further speedup but, if experience to date on parallelization of linear algebra codes can be considered representative, parallel computers tend to favor explicit methods.

#### Box 14.6: Implicit Symplectic Integrator Family

The family of implicit symplectic methods parameterized by  $\alpha$  can be formulated as follows [446, 1126]:

$$\begin{aligned} X^{n+1/2} &= X^n + \frac{\Delta t}{2} V^n \\ V^{n+1/2} &= V^n + \frac{\Delta t}{2} \mathbf{M}^{-1} F^{n+1/2} \\ V^{n+1} &= V^{n+1/2} + \frac{\Delta t}{2} \mathbf{M}^{-1} F^{n+1/2} \\ X^{n+1} &= X^{n+1/2} + \frac{\Delta t}{2} V^{n+1}, \end{aligned} \quad (14.63)$$

where

$$F^{n+1/2} = F(X^{n+1/2} + \alpha \Delta t^2 \mathbf{M}^{-1} F^{n+1/2}). \quad (14.64)$$

The values  $\alpha = 0$  and  $1/4$  correspond to the Verlet and IM methods, respectively, and  $\alpha = 1/2$  corresponds to the LIM2 method introduced in the text.

The above implicit system can be solved by minimizing the dynamics function  $\Phi$  to obtain  $X$  and then evaluating  $F$  at this minimum point. Here  $\Phi$  is defined as

$$\Phi(X) = \frac{1}{\gamma \Delta t^2} (X - X_0^n)^T \mathbf{M} (X - X_0^n) + \alpha E(X), \quad (14.65)$$

with

$$X_0^n = X^n + \frac{\Delta t}{2} V^n.$$

The effective rotation angle for this  $\alpha$ -family is [1126]:

$$\begin{aligned}\theta_{\text{eff}}^{\alpha \text{ family}} &= 2 \sin^{-1} \left( \frac{\omega \Delta t}{2} \sqrt{\phi} \right) \\ &= \omega \Delta t + \left( \frac{1}{24} - \frac{\alpha}{2} \right) (\omega \Delta t)^3 + \mathcal{O}((\omega \Delta t)^5)\end{aligned}\quad (14.66)$$

where

$$\phi = 1 / [1 + \alpha(\omega \Delta t)^2]. \quad (14.67)$$

Thus the maximal effective rotation can be controlled by the choice of  $\alpha$  (see text).

---

## 14.7 Enhanced Sampling Methods

### 14.7.1 Overview

The well known shortcomings of traditional MD have stimulated many innovative methods for capturing large-scale, long-time configurational changes of biomolecules. Indeed, an enormous range of methods has been developed to enhance coverage of the conformational space; see [729, 781, 1116, 1117], for example, for recent reviews. Each method depends on the computation goals and available computing resources. While many of these methods are suitable for probing the thermally-accessible configuration space at the cost of altered kinetics, more sophisticated approaches like transition path sampling or Markov chain models can yield valuable mechanistic insights, reaction pathways, and reaction rates.

Approaches mentioned already for addressing the sampling problem include simulating multiple trajectories for improved statistics [69], especially effective when applied to small systems, as in folding studies of peptides [285] or protein segments [364, 365]. Other approaches mentioned are Monte Carlo techniques [509] and variants like simulated annealing, parallel tempering/replica exchange MC [508, 1237], and hybrid MC methods (Monte Carlo chapter). Implicit solvation (Chapter 10) and Langevin and Brownian dynamics approaches (this chapter) can also be viewed as approaches for enhancing the sampling because they can cover longer times and/or reduce computational cost.

Below, additional approaches are described, including methods based on harmonic analysis and coordinate transformations, coarse-grained models, biasing approaches and altered MD protocols, and various novel methods for computing reaction mechanisms and reaction rates.

### 14.7.2 Harmonic-Analysis Based Techniques

The fundamental oscillations of each molecule about its equilibrium state, termed *normal modes*, were described in Chapter 9 in connection to force-field parameterization. Each such fundamental mode has a frequency associated with it that

is related to a force constant and energy potential that can be propagated to describe molecular displacements about equilibrium. The same idea based on harmonic theory has been extended more generally to describe collective molecular motions, including for complex molecular systems.

In their purest forms, *normal mode analysis* (NMA) and the related *principal component analysis* (PCA) involve propagating the normal modes based on a spectral decomposition (diagonalization) of a mass-weighted Hessian at thermal equilibrium [535]. Elastic networks [77, 463, 1262] are modern extensions of these techniques which forgo the computationally-demanding diagonalization because the simplified bead/spring coarse-grained models are assumed by construction to reflect minimum states of the molecular system. Clearly, this harmonic approximation to describe small-amplitude, high-frequency motions is far from accurate at ambient temperatures when significant biomolecular fluctuations between minimum-energy regions, as well as rearrangements, occur. Still, these techniques have provided valuable information on collective motions of biomolecules.

Besides elastic networks, another successful extension of these harmonic-analysis-based techniques is called *essential dynamics* (ED) [36, 37, 650, 1289]. ED attempts to characterize low-frequency, high-amplitude motions modes by expressing the dynamics in terms of alternative coordinates (principal components corresponding to low-frequency motion) which filter out other modes. This low-dimensional space is constructed from the variance/co-variance matrix of positional fluctuations obtained from a time series of coordinates, by projecting the original configurations onto each of the principal components and then following the principal components of interest in time. ED makes no explicit assumption of thermal equilibrium.

Specifically, to describe the collective motions, a covariance matrix  $\mathbf{C}$  of atomic fluctuations is constructed from a series of  $M$  coordinates  $X_1, X_2, \dots, X_M$  where  $X_k$  is the collective coordinate vector of the system at the  $k$ th configuration, with respect to an average structure  $\langle X \rangle$ :

$$\langle X \rangle = \frac{1}{M} \sum_{k=1,M} X_k, \quad (14.68)$$

from the expression:

$$\mathbf{C} = \frac{1}{M} \sum_{k=1,M} (X_k - \langle X \rangle) (X_k - \langle X \rangle)^T. \quad (14.69)$$

Diagonalization of this covariance matrix  $\mathbf{C}$  produces the eigenvectors  $\{V_n\}$  and eigenvalues  $\{\lambda_n\}$  as entries of the diagonal matrix  $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_{3N})$  from the spectral decomposition:

$$V^T \mathbf{C} V = \Lambda. \quad (14.70)$$

Thus, the projection

$$\mathbf{C} V_n = \lambda_n V_n, \quad n = 1, 2, \dots, 3N, \quad (14.71)$$

can be used to visualize and propagate selected motions. In this description, each eigenvector  $V_n$  defines the direction of motion of the molecular system as a displacement about the average structure. The normalized magnitude of the corresponding eigenvalue  $\lambda_n$  is a measure of the amplitudes of motion along the eigenvector  $V_n$ .

The literature is vast with applications of PCA, NMA, and ED with both all-atom and coarse-grained models in combination with various algorithms, including molecular, Langevin and Brownian dynamics to biomolecular conformational flexibility and dynamics. These approaches have provided valuable insights into biomolecular flexibility and functional activity. However, the results depend strongly on the level of convergence of the sampling, which influences the results and hence the interpretations.

Some examples of using PCA and ED include analyses of protein subdomain motion correlations in DNA pol  $\beta$  [60] and Dpo4 [1336] and sequence/flexibility relationships for a series of single-based DNA TATA elements [1026, 1229] (Fig. 14.17). Other examples include folding dynamics of a  $\beta$ -protein WW domain using the coarse-grained protein model UNRES [816], and an investigation of protein flexibility in water [1075]. Network models have been particularly effective for applications to molecular machines like the ribosome modeled by coarse-grained formulations [1248].

### 14.7.3 Other Coordinate Transformations

This idea in ED of projecting the dynamics onto principal modes is related to various variable/coordinate transformation methods in classical statistical mechanical using configuration partition functions. For example, *metadynamics* rewrites the equations of motion in terms of a few collective variables so that key regions of space are identified and explored [963], and the *reference potential spatial warping algorithm* (REPSWA) [860] introduces a variable transformation in the classical partition function that increases attraction to basins.

Internal or torsion-angle dynamics has long been attempted for the goal of enhanced sampling, with the rationale that the fewer degrees of freedom compared to Cartesian coordinates will allow longer integration timesteps and hence greater sampling. Indeed, peptide folding and refinement with dihedral-angle MD demonstrated a computational advantage of several orders of magnitude over Cartesian analogues [585] as well as the capturing of folding pathways of helical peptides and local side-chain and domain dynamics [1426]. Dihedral-space MD has also been combined with PCA in a clever way to systematically construct the low-dimensional free energy landscape from a classical MD simulation [34]. Though this analysis is interpretive, it shows that major conformational states, barriers, and reaction pathways for solvated peptides can be visualized from the constructed energy landscape.

In general, dihedral-angle MD approaches for propagating biomolecular motion have not caught on at large, perhaps both due to the added cost of the transformation involved in the Newtonian laws of motion and the fact that

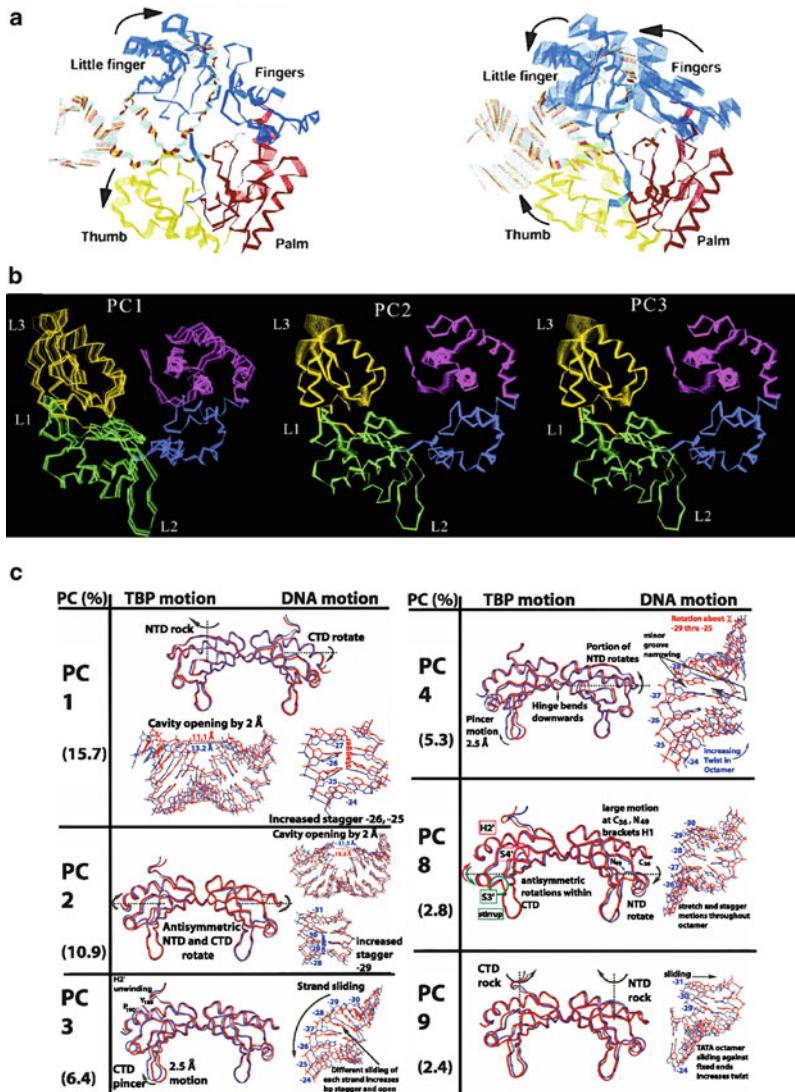


Figure 14.17. Examples of PCA and ED applications. (a) A study of the conformational changes of Dpo4 after chemistry revealed that the little finger and fingers domains may help translocate the DNA for the next cycle of nucleotide insertion [1336]. (b) A study of the closing conformational change of DNA polymerase  $\beta$  upon binding the nucleotide substrate revealed from the top three principal components correlations between the thumb subdomain and other regions of the protein (palm, 8-kDa) [59]. (c) A study of the motions of single-base variants of TATA-box DNA sequences bound to the TPB protein helped explain why these variants revealed a wide range of transcriptional efficiency despite remarkably similar structures: high-efficiency variants favored complexation motions while low-efficiency variants tended toward dissociation deformations. The dominant motions common to all complexes are shown, dissected for the protein and bound TATA-box DNA separately [1026].

biomolecular vibrational modes are intricately coupled and hence dynamics can be critically altered by neglecting the high-frequency bond-length and bond-angle modes. However, the increased interest in coarse-graining models argues for their resurgence.

#### 14.7.4 Coarse Graining Models

System specific coarse-grained methods are attractive because they drastically reduce the number of degrees of freedom. However, their formulations are highly system dependent and require as much art as science in constructing, testing/validating, and applying them to appropriately formulated questions. Coarse graining can involve bead models, implicit solvent approximations, discrete lattice models, and general multiscale formulations.

The simplest type of coarse graining involves bead models, long used for proteins (e.g., Warshel & Levitt's united residue model [1344]). Bead models have also been widely used for supercoiled DNA (wormlike chain model of Allison and co-workers [32], see Chapter 6) and more recently developed for RNA (e.g., [315]). Such methods can lead to meaningful insights into large-scale rearrangements, including folding, not typically amenable to all-atom simulations. However, the neglect of many details (e.g., solvent/solute interactions) must be considered in the biological interpretations.

Lattice models also reduce the conformational degrees of freedom to a discrete set, therefore allowing in theory exhaustive sampling of the conformational space. Lattice models of proteins, such as developed by Gō [463], and by Miyazawa & Jernigan [869], are associated with ideal funnel energy landscapes: a protein chain is modeled by attractive interactions between pairs of residues that interact in the native structures and repulsive interactions of the other pairs, based on statistical data.

General coarse-grained or multiscale models are most challenging to formulate and validate because the various components need to be resolved by different approaches and combined effectively. For example, the simplified models of the chromatin fiber developed by the Langowski [1312], Schiessel [883], and Schlick [64, 1114] groups and others necessarily select the molecular parts to resolve in detail and those that can be effectively approximated (i.e., coarse grained). For example, in studies aimed at deducing the architecture of the 30nm chromatin fiber [483], the nucleosome core, histone tails, linker DNA, and linker histones are each modeled differently in a mesoscale model (see Fig. 6.11) and sampled by MC; parameterization for the core is done by a Poisson Boltzmann electric field approximation and for the proteins by comparison to atomistic MD [65, 109, 1441].

Other examples of general coarse-grained models are innovative models developed for membrane systems to investigate: the reshaping of the electrostatics-dominated surface [57, 655], pore formation [461], membrane architecture [1150], and protein/membrane binding interactions [647]. Complex processes involving

virus capsid stability [56], GroEL/GroES chaperonin-guided folding [260], and an ATP-driven molecular motor [667] have also been studied by coarse-graining approaches.

It is well recognized that rigorous coarse-graining approaches are needed to ultimately address many levels of biomolecular pathways and reactions to allow a “telescoping” of views from one level of resolution to another. Some notable work in that direction has already been reported [506, 794, 921]. The appropriateness of the coarse graining has been assessed quantitatively from the collective motions (ED) compared to atomistic motion [794], by computing solvation free energies [506], and by a formal statistical mechanical framework [921].

### 14.7.5 *Biasing Approaches*

If altering the model — through biasing forces or guiding restraints — is considered fair game, much room for ingenuity remains, as described by the various biasing techniques.

The simplest and crudest approach for enhancing the sampling is to manipulate the energy function used as a basis for MD, by using restrained potentials, as in targeted MD (TMD) or umbrella sampling, and by using various experimentally-based biases or guides.

TMD can be used for generating pathways between known endpoints, such as unfolded and folded state of a peptide [389]. In TMD, an artificial restraining potential is added like a Lagrange multiplier with parameter  $\lambda$  that is 0 at the initial state and 1 at the target state. This forces the system to evolve toward a target state in a specified number of steps [389, 1418]. Despite the fictitious trajectory that results from such targeted MD, insights into disallowed configurational states can be obtained, as well as conclusions regarding common pathway themes, though individual trajectories diverge from one another.

For example, TMD has been used to suggest possible opening pathways between substrate-free and substrate-bound enzyme states (see Figure 13.1) [1408] and between closed and open states of a membrane channel [669]. Importantly, besides suggesting pathways, intermediate pathway configurations generated in TMD can serve as initial states for regular MD to explore pathways and conformational transitions.

Many other methods for simulating transitions between defined conformations have been reported (e.g., [355, 521, 955]). For example, the simplified multiple-basin Hamiltonian funnel-based potential model [932] was developed for very large molecular complexes with known endpoint structures to simulate transitions between ligand-bound and unbound states.

Umbrella sampling (US) allows sampling of specified regions of phase space by using a similar restraining potential, typically by restraining key conformational variables to specific regions (e.g., a certain range of sugar puckering for DNA [59, 1030]). It is also possible to employ experimental information to define various restraints to guide or restrict the pathways. Such experimental information can

involve distances determined by FRET experiments [859] or the ratio of native contacts present in protein intermediates relative to those in the native state [956]. US has also been useful in computation of free energies where extensive sampling is required.

Besides various constraints or restraints, the energy can be alternatively modified by using other types of biasing terms to facilitate barrier crossing events. These include a *conformation flooding* technique in which the bias is determined based on a coarse-grained simulation for the conformational space density [485], *hyper-MD* in which the biasing potential is constructed based on the smallest eigenvalue of the Hessian matrix [1315], simple *local boost* method based on the total potential energy [1332], a *bond boost* method based on bond-length deviations [862], and accelerated MD by a potential-energy term “boost” [504].

Umbrella sampling methods can also be very effective in combination with such biased trajectories (e.g., [130]). For example, diffusion-limited processes associated with high-energy or entropy barrier can be ‘accelerated’ through biasing forces in Brownian dynamics simulations, with rate calculations adjusted by associating lower weights with movements along high biases and vice versa [1458]. Biomolecular systems can be ‘steered’ [592, 593] or ‘guided’ [1393, 1394] — subjected to time-dependent external forces along certain degrees of freedom or along a local free-energy gradient — to probe molecular details of certain experiments, such as atomic force microscopy and optical tweezer manipulations (see citations in [592, 593]), or to study folding/unfolding events [1395]. Interactive MD [475, 1111] allows such exploration of pathways manually, by a combination of advanced computer graphics and simulation (VMD and NAMD): the user steers the system to the desired state by “feeling” the potential force and applying any desired magnitude (to “pull” the system).

#### 14.7.6 Variations in MD Algorithm and Protocol

The MD simulation protocol rather than the energy can also be manipulated directly. In this subclass, we have the popular approach LES (locally enhanced sampling) in which several configurations are generated in a single energy evaluation [1063], and Replica exchange MD (REMD) [1237], based on parallel tempering MC [508] (described in the MC chapter), which itself is based on Metropolis-coupled Markov chain MC.

The idea in REMD is to simulate the dynamics of multiple, non-interacting copies (replicas) of identical systems at different temperatures. This requires integration by using canonical-ensemble algorithms as described in Chapter 13, such as the Berendsen weak coupling thermostat (or heat-bath) [123] or the extended-simple Nosé-Hoover thermostat approach [566, 924]. Periodically, the configurations of the different replicas (e.g., replica  $i$  at temperature  $T_i$  with replica  $j$  at temperature  $T_j$ ) are exchanged with a transition probability  $p$  that maintains each temperature’s equilibrium ensemble distribution. Thus, the

thermodynamic or temperature states in the canonical ensemble are exchanged, with the goal of retaining those systems that are making better progress:

$$p_{i \rightarrow j} = \min [1, \exp [(\beta_j - \beta_i)(E_j - E_i)]], \quad (14.72)$$

where  $\beta_i$  is the Boltzmann factor for temperature  $i$ :  $\beta_i = 1/(k_B T_i)$ .

These temperature-ensemble exchanges are intended to accelerate barrier crossings. Though obeying detailed balance for an extended ensemble of the canonical states, the state-exchange probability of eq. (14.72) destroys real kinetic properties. However, REMD has been used at large to simulate folding/unfolding equilibria of biomolecules. Examples include the folding of a small solvated RNA hairpin from an extended state [444], folding of a solvated protein A [443], folding of villin within 1.78 Å of the native state [730], and determination of equilibrium ensembles of proteins in a coarse-grained implicit-solvation model, OPEP [222].

Though REMD has been in wide usage for applications ranging from small peptides to complex biological systems, its success has largely been empirical and only recently have some technical issues come to the surface. Indeed, practitioners of REMD have emphasized the need to formulate careful configurational swapping protocols (temperature ladders and exchange/acceptance ratios) and other ways to increase the conformational sampling efficiency (see [1064, 1192] for example) and to use a large number of concurrent processors/replicas for efficient sampling. In fact, a variant distributed replica sampling has been proposed for serial implementations when a large number of processors is not available [1060], and a variant for enhanced sampling based on a Tsallis biasing potential was also proposed [627]. It has also been remarked that REMD may only be computationally advantageous for systems with relatively high energy barriers [1459].

More fundamentally, there are issues with REMD because on the inherent canonical-ensemble integrators that are used. In their thorough analysis of REMD, Cooke & Schmidler explain that the formulation of REMD by extension of parallel tempering MC to MD has introduced sampling and ergodicity problems stemming from the failure of the underlying constant-temperature MD integrators to preserve certain variants of REMD [264]. This is because while the MC analog is based on Markov chains, REMD algorithms cannot use the symplectic leap-frog integrator popular for microcanonical (constant energy) MD and resort to isothermal integrators as mentioned above. Unfortunately, these methods are not rigorously ergodic, and this can affect the dynamics of even small systems, as practitioners noted. As a remedy, combining REMD with hybrid MC to ensure ergodicity was suggested [264], thereby returning to the original advantages of parallel tempering MC. Other cures, such as using recently developed entropy-preserving constant-temperature integrators for sampling the canonical distribution based on a Nosé-Hoover formulation in the context of Markov chain Monte Carlo can also solve the ergodicity problem in REMD [732]; systematic discretization errors are also eliminated by this hybrid MC approach for canonical sampling.

### 14.7.7 Other Rigorous Approaches for Deducing Mechanisms, Free Energies, and Reaction Rates

Many other innovative methods have been developed by using various constructs and ideas from physics, mathematics, and engineering such as domain decomposition, clustering techniques, and Markov models to sample and/or survey conformational space and dynamics.

For example, for exploring the free energy landscape, the canonical adiabatic free energy sampling (*CAFES*) method [1299] propagates the dynamics of decoupled solute and solvent systems so that the former follows adiabatic dynamics of the free energy surface created by the latter at increased temperatures. *Metadynamics* explores the free energy surface by following non-Markovian dynamics determined by rewriting the equations of motion in terms of a few collective variables so that key regions of space are identified as simulation time increases [963]. And a *sweep method* combines temperature-accelerated MD (TAMD) with a free-energy reconstruction from the mean force using radial basis functions via optimization [828].

Methods for deducing reaction mechanisms include *transition path sampling* (TPS) by Chandler and co-workers which follows a Monte Carlo protocol in MD-trajectory space to locate key transition states ([147] and recently reviewed in [294]); the *nudged elastic band method* (NEB) [543] that optimizes minimum energy pathways; the *max-flux* approach [583], which constructs variationally optimized reaction pathways based on the max flux for diffusive dynamics; and the *string method* [858], which prunes MD trajectories to retain reactive segments. A recent extension of TPS connects more than two intermediate states in phase space [1062].

A TPS application to a biomolecular complex, a DNA polymerase  $\beta$ /DNA complex, combined TPS with an efficient free-energy protocol termed “BOLAS” [1030] and network models to deduce the mechanism, compute the free energy profile, and estimate the rate for pol  $\beta$ ’s closing conformational change [1031]. The study revealed a complex landscape where sequential, subtle-side chain rearrangements lead the enzyme from open to the closed state and where Arg258 is a “gate keeper” for the reaction (Fig. 14.18); the overall computed rate of 10 per second corresponds to the  $27k_B T$  barrier and agrees well with the experimental value of 3–10 per second for the overall reaction. A similar TPS protocol applied to the mismatch (G:A instead of G:C) suggested that the higher free energy barrier for the mismatch comes from the instability of the closed mismatched state compared to the matched base pair system [1032] (Figure 14.19).

More recently, innovative divide and conquer methods have been developed to compute reaction rates. These include *forward flux* simulations (FFS) [152], *milestoning* [354], and *Markov State Models* (MSM) [919].

FFS [152] enhances sampling of rare events in stochastic non-equilibrium systems in which the phase space distribution is not known *a priori* by using interfaces to partition phase space. Following this partitioning, an adaptive procedure

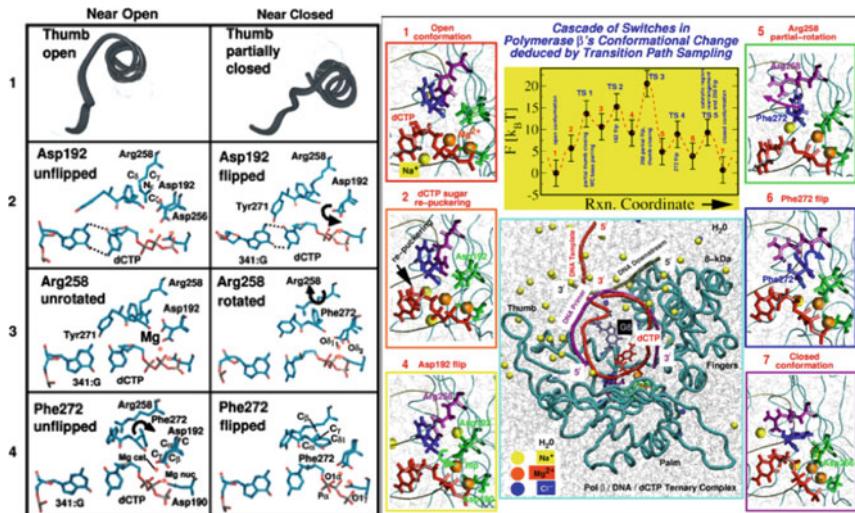


Figure 14.18. Left: Molecular snapshots near open (left column) and closed (right column) states of  $\text{pol} \beta$  for four transition state regions [1031]: (1) Partial thumb closing. (2) Asp-192 flip. (3) Arg-258 partial rotation. (4) Phe-272 flip. Right: Overall captured reaction kinetics profile (from TPS) for the conformational transition of  $\text{pol} \beta$  (for G:C) from open (state 1) to closed (state 7) forms showing free energies (in  $k_B T$ ) associated with the different transition state regions. The meta-stable basins (in red) along the reaction coordinate are numbered 1–7.

is used to find kinetic “bottleneck regions” by estimating rate constants associated with reaching subsequent interfaces. Then, FFS concentrates on sampling those bottleneck regions only. FFS is appropriate for discretely defined surfaces, like protein folding on a lattice [23].

Milestoning [354] coarse grains temporal and spatial descriptors of the system by sequential transitions and then recovers reaction kinetics by integral equations and global path optimization strategies. Similarly, Markov State Models (MSM) [919] compute reaction rates from different interfaces by using transition networks to describe biomolecular kinetics and thermodynamics. See also [920]. Combinations of MSM with the single-sweep method [828] can be successful for proteins [957] because phase space is surveyed to map dynamically important regions, from which free simulations are initiated, and then the transition matrix is constructed by piecing the information together. Clustering techniques for partitioning space to locate key regions can also be used in this connection [1357].

Milestoning was compared to MSM, FFS, and transition interface sampling in [1298], where the advantages of milestoning compared to the other methods were suggested in terms of accuracy, efficiency, and parallel-machine implementations.

Various combinations of these methods, such as REMD with TIS (transition interface sampling) to compute free energy profiles and rate constants when the barriers are high [146], emphasize how different sampling approaches can be combined in clever ways to suit the problem. Often, coarse-grained models are

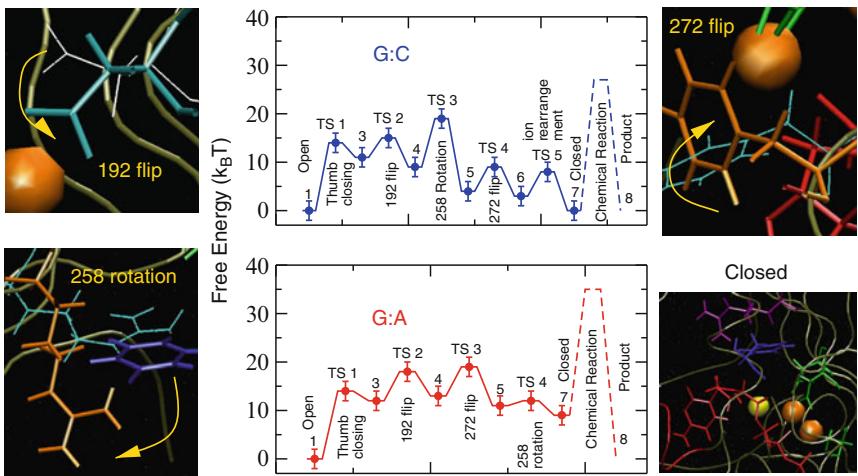


Figure 14.19. Overall captured reaction kinetics profile for pol  $\beta$ 's closing transition followed by chemical incorporation of dNTP for G:C and G:A systems [1032]. The barriers to chemistry (dashed peaks) are derived from experimentally measured  $k_{\text{pol}}$  values. The profiles were constructed by employing reaction coordinate characterizing order parameters in conjunction with transition path sampling. The potential of mean force along each reaction coordinate is computed for each conformational event.

used in combination with implicit solvation and Langevin or Brownian dynamics or MC for added benefits. Still, applications of these methods to biomolecular complexes at large remain a challenge.

## 14.8 Future Outlook

### 14.8.1 Integration Ingenuity

#### Many Approaches

In this chapter we have discussed various numerical integration techniques for Newtonian and Langevin/Brownian dynamics and focused on variations that can increase the timestep and lead to overall computational gains. For very detailed dynamic pathways, the standard Newtonian approach is necessary, but for certain systems (like long DNA supercoils or diffusion-controlled mechanisms in enzyme catalysis), stochastic models are appropriate and computationally advantageous.

The approaches discussed to ameliorate the severe timestep restriction in biomolecular dynamics involve force splitting or multiple timesteps, various harmonic approximations, implicit integration, and other separating frameworks. Each such avenue has encountered different problems along the way, but these obstacles serve to increase our understanding of the intriguing numerical, computational, and accuracy issues involved in propagating the motion of complex nonlinear systems.

The most successful integrators that address stability and resonance limitations of MD integrators due to the high-frequency vibrational modes and the intricate coupling among the vibrational modes of a biomolecule consist of multiple-timestep integrators combined with stochastic dynamics. Langevin and Brownian dynamics (LD and BD), where solvent collisions and thermal fluctuations are incorporated in an average sense, have long been used as a way to allow larger timesteps and hence larger timespans in dynamics simulations. For stochastic dynamics methods, it is also easier to prove ergodicity.

Certainly, long MD simulations are being reported more regularly now compared to several years ago, made possible by efficient and parallel simulation packages like NAMD, DESMOND, GROMACS and new computer systems like Anton, hard-wired for MD simulations [1169, 1462]. For example, a  $1\ \mu\text{s}$  atomic-level simulation of the voltage-modulated potassium channel Kv1.2 (120,000 atoms) recently described the opening/closing mechanism involved [136]. Thus, many problems thought to be intractable several decades ago are now being solved.

But faster computer technology alone is insufficient to address challenging biomolecular problems. Algorithmic developments for enhanced sampling such as described here, especially those divide and conquer methods suitable for loosely coupled processor architectures that are more readily available to the average user, are needed to complement long atomistic simulations.

## Resonance

The heightened appreciation of resonance problems [134, 821], in particular, contrasts with the more systematic error associated with numerical stability that grows monotonically with the discretization size. Ironically, resonance artifacts are worse in the modern impulse multiple-timestep methods, formulated to be symplectic and reversible; the earlier extrapolative variants were abandoned due to energy drifts. Stochasticity and slow-force averaging, as in the LM method [594, 595], or stochasticity and extrapolation, as in the LN method [95], can dampen and/or remove resonance artifacts and allow larger timesteps, but this requires added forces to standard Newtonian dynamics.

*Ultimately, the compromise between the realized speedup and the accuracy obtained for the governing dynamic model should depend on the applications for which the dynamic simulations are used (e.g., configurational sampling versus detailed dynamics).*

### 14.8.2 Current Challenges

#### PME Protocols

A problem that remains unsolved in part in connection with MTS methods is their optimal integration with Ewald and particle-mesh Ewald protocols.

The problem is due to the presence of fast terms in the reciprocal Ewald component; this limits the outer MTS timestep and hence the speedup. For some discussion and approaches, see [89, 97, 98, 1019, 1025, 1236, 1449]. Improving such algorithms will allow us to use larger outer timesteps and hence simulate systems over longer times.

There is also a related parallelization problem in PME implementations: memory requirements create a bottleneck in typical MD simulations longer than a microsecond. This is because the contribution of the long-range electrostatic forces imposes a global data dependency on all the system charges. In PME, this dependency appears as a convolution between the meshed charge distribution and a chosen kernel evaluated using 3D Fast Fourier Transforms. These 3D FFT computations impose communication problems and define the rate limiting step for parallel implementations [403]. In practice, parameters such as mesh sizes, precision, size of real space and k-space have been optimized to delay the communication bottleneck as possible (e.g., [1462]), but overall errors in long simulations are far from trivial [1204]. These issues will likely receive more attention in the future. In addition, these communication requirements and hence limitations on parallel implementations suggest that exploring alternatives to PME, like fast multipole (Chapter 10) or multigrid [1082] methods are needed. These and other approaches may be necessary as long-time MD is implemented on massively parallel computer architectures.

### Technology’s Role

Not to be downplayed as a factor in the increasing of simulation scope is technology improvement. The steady increase in computer power, the declining cost of fast and highly-modular processors, and the rise of efficient parallelization of MD codes surely are helping to bridge the timescale gap between simulation range and experimental durations. The triumphant report of Duan and Kollman’s 1  $\mu\text{s}$  trajectory approaching the folding state of a 36-residue villin headpiece, from a disordered configuration (4 dedicated months of 256 processors on the Cray T3D/E), is a case in point [338]. However, as discussed in Chapter 1 (see Table 1.2 and Figures 1.3 and 1.4), microsecond simulations are possible today and simulating milliseconds in the life of a biomolecule is on the horizon [658, 1462]. But many experts consider that certain technical bottlenecks must be overcome, such as mentioned above, concerning force-field accuracy, long-range force calculations, integration accuracy, and parallelization implementations. Indeed, designing reliable, efficient, and general software for large-scale MD applications on multiprocessors remains a challenge.

### Sampling Issues

The scope of dynamics simulations is certainly being enhanced when used in combination with various methods to analyze biomolecular motion, sample the large conformation space, and obtain relevant information on reaction mechanisms, pathways, and rates. Because coarse-grained models are vital for addressing

problems related to large macromolecular complexes or ensemble properties of smaller systems, Monte Carlo methods deserve further consideration and development in general. In addition, exciting new approaches for rigorous frameworks for general multiscale models are on the horizon.

While harmonic approximation methods like PCA, NMA, ED and elastic networks continue to add valuable insights into biomolecular flexibility and function, they are also participating in more applications with the growth of network models for molecular machines that help dissect and distill complex functional motions.

Simple potential modifications to MD like TMD can provide conformational insights, and REMD approaches can enhance sampling, but success has been rather empirical rather than rigorous. Because the actual kinetics of the systems are altered, caution is warranted in biological interpretations. However, enhanced sampling protocols can be useful in specific contexts, as recently demonstrated by using accelerated MD [504] to reproduce residual dipolar coupling measurements concerning the slow modes from NMR data on a domain of the protein GB3 and pinpoint slow motions [831].

REMD has recently received some scrutiny concerning effective ensemble exchange protocols, general computational efficiency, and ergodicity questions. This has led to variants with stochastic elements — motivated by REMD’s origin in Monte Carlo parallel tempering methods — that can improve sampling (e.g., [264]). Because REMD applications require concurrent processors (one per replica), the technique is not always practical, especially for large atomistic systems. The distributed replica variant [1060] that performs stochastic moves of independent replicas instead of pairwise exchanges of replicas offers an alternative.

TPS, Markov state models, milestoning and approaches that aim to compute reaction rates have been more rigorously grounded in theory, and successful biomolecular applications have been reported. Still, their application to biomolecules in general remains far from routine. In this goal, various combinations of methods that deduce mechanisms and compute reaction rates by divide and conquer approaches will undoubtedly be effective on today’s readily available distributed computing resources of cluster networks, especially by enhancing them with coarse-grained models and Monte Carlo elements.

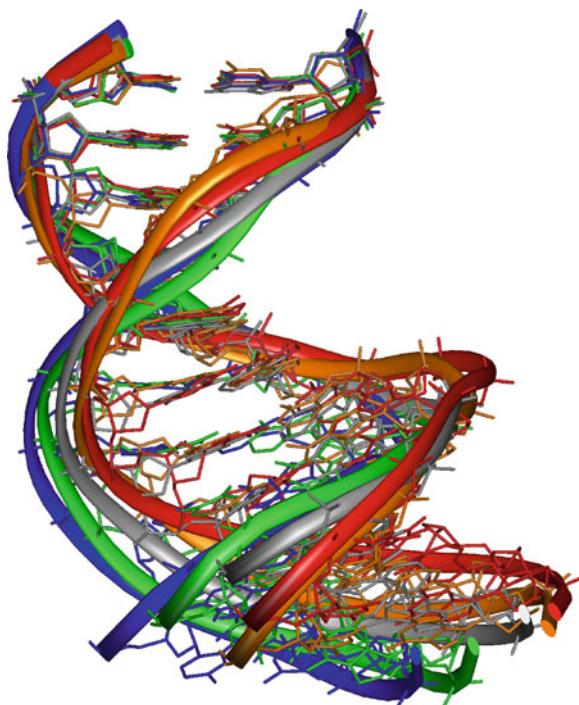
Significantly, all these approaches for enhanced sampling can be combined for cumulative and significant computational advantages.

### Tip of the Iceberg

These long-time and greater-sampling approaches represent only the tip of the iceberg of those possible. Undoubtedly, motivated by fundamental biological problems like protein folding, the zealous computational biologists and the scientists they enlist in these quests will continue to combine algorithmic ingenuity with state-of-the-art computing to overcome, or circumvent, one of the fundamental challenges in molecular dynamics simulations. All these advances are collectively opening the way to exciting applications of a rich variety of

biomolecular systems regarding large-scale conformational changes and functional dynamics on millisecond and longer timescales that are helping close the gap between experimental and theoretical timeframes.

Ultimately, the most successful applications reflect a combination of novel and rigorous mathematical ideas with physical intuition. And such applications can be successful not only by reproducing experimental data concerning slow motions but also in helping suggest mechanistic and energetic details. Clearly, the gap between experimental and theoretical timeframes is steadily narrowing, leading to renewed interest in MD-based modeling by theorists and experimentalists alike.



# 15

## Similarity and Diversity in Chemical Design

### Chapter 15 Notation

SYMBOL	DEFINITION
<b>Matrices</b>	
$A$	Dataset matrix ( $n \times m$ )
$A_k$	Rank $k$ approximation to $A$
$C$	covariance matrix ( $m \times m$ ), elements $c_{jj'}$
$P_k$	projection matrix
$U$	SVD factor of $A$ ( $n \times n$ ), contains left singular values
$V$	SVD factor of $A$ ( $m \times m$ ), contains right singular values; also eigenvector matrix of $C$
$V_k$	low-rank approximation to eigenvector matrix ( $m \times k$ )
$\Sigma$	SVD factor ( $n \times m$ ), contains singular values
$\Sigma_k$	low-rank approximation to $\Sigma$
<b>Vectors</b>	
$u_j$	left singular value
$v_j$	right singular value
$X_i$	vector of compound $i$ (components $X_{i1}, X_{i2}, \dots, X_{im}$ )
$\hat{X}_i$	scaled version of $X_i$
$Y_i$	projection of $X_i$ ; also principal component of $C$
<b>Scalars &amp; Functions</b>	
$d_{ij}$	intercompound distance $ij$ in the projected representation
$f, E$	target optimization functions
$l_{ij}$	lower bounds on intercompound distance $ij$
$u_{ij}$	upper bounds on intercompound distance $ij$
$m$	number of dataset descriptors
$n$	number of dataset components

Chapter 15 Notation Table (continued)

SYMBOL	DEFINITION
$N$	number of variables
$T_d$	total number of distance segments satisfying a given deviation from target
$\alpha, \beta$	scaling factors
$\delta$	Euclidean distance (with upper/lower bounds $u, l$ )
$\lambda$	eigenvalues
$\mu$	mean value
$\omega$	weights used in target optimization function
$\sigma$	singular values

Every sentence I utter must be understood not as an affirmation but as a question.

Niels Bohr (1885–1962).

## 15.1 Introduction to Drug Design

Following a simple introduction to drug discovery research, this chapter presents some mathematical formulations and approaches to problems involved in chemical database analysis that might interest mathematical/physical scientists. With continued advances in structure determination, genomics, and high-throughput screening and related (more focused) techniques, *in silico* drug design is playing an important role as never before. Thus, traditional structure-directed library design methods in combination with newer approaches like fragment-based drug design [496, 1447], virtual screening [453, 1179], and system-scale approaches to drug design [236, 278, 649] will form important areas of research.

For a historical perspective of drug discovery, see [7, 159, 335, 507, 589, 727, 772], for example, and for specialized treatments in drug design modeling consult the texts by Leach [709] and Cohen [254].

### 15.1.1 Chemical Libraries

The field of combinatorial chemistry was recognized by *Science* in 1997 as one of nine “discoveries that transform our ideas about the natural world and also offer potential benefits to society”. Indeed, the systematic assembly of chemical building blocks to form potential biologically-active compounds and their rapid testing for bioactivity has experienced a rapid growth in both experimental and theoretical approaches (e.g., [640, 692, 1241]); see the editorial overview on combinatorial chemistry [207] and the associated group of articles. Two combinatorial chemistry journals were launched in 1997, with new journals since then, and a Gordon Research conference on Combinatorial Chemistry was created. The number of new-drug candidates reaching the clinical-trial stage is greater than ever.

Indeed, it was stated in 1999: “Recent advances in solid-phase synthesis, informatics, and high-throughput screening suggest combinatorial chemistry is coming of age” [151].

Accelerated (automated and parallel) synthesis techniques combined with screening by molecular modeling and database analysis are the tools of combinatorial chemists. These tools can be applied to propose candidate molecules that resemble antibiotics, to find novel catalysts for certain reactions, to design inhibitors for the HIV protease, or to construct molecular sieves for the chemical industries based on zeolites. Thus, combinatorial technology is used to develop not only new drugs but also new materials, such as for electronic devices. Indeed, as electronic instruments become smaller, thin insulating materials for integrated circuit technology are needed. For example, the design of a new thin-film insulator at Bell Labs of Lucent Technologies [333] combined an optimal mixture of the metals zirconium (Zr), tin (Sn), and titanium (Ti) with oxygen.

As such experimental synthesis techniques are becoming cheaper and faster, huge chemical databases are becoming available for computer-aided [159] and structure-based [41, 453, 1179, 1447] drug design; the development of reliable computational tools for the study of these database compounds is thus becoming more important than ever. The term *cheminformatics* (*chemical informatics*, also called *chemoinformatics*), has been coined to describe this emerging discipline that aims at transforming such data into information, and that information into knowledge useful for faster identification and optimization of lead drugs.

### 15.1.2 Early Drug Development Work

Before the 1970s, proposals for new drug candidates came mostly from laboratory syntheses or extractions from Nature. A notable example of the latter is Carl Djerassi’s use of locally grown yams near his laboratory in Mexico City to synthesize cortisone; a year later, this led to his creation of the first steroid effective as a birth control pill [323]. Synthetic technology has certainly risen, but natural products have been and remain vital as pharmaceuticals (see [666, 1006] and Box 15.1 for a historical perspective).

A pioneer in the systematic development of therapeutic substances is James W. Black, who won the Nobel Prize in Physiology or Medicine in 1988 for his research on drugs beginning in 1964, including histamine H<sub>2</sub>-receptor antagonists. Black’s team at Smith Kline & French in England synthesized and tested systematically compounds to block histamine, a natural component produced in the stomach that stimulates secretion of gastric juices. Their work led to development of a classic ‘rationally-designed’ drug in 1972 known as *Tagamet* (cimetidine). This drug effectively inhibits gastric-acid production and has revolutionized the treatment of peptic ulcers.

Later, the term *rational drug design* was introduced as our understanding of biochemical processes increased, as computer technology improved, and as the field of molecular modeling gained wider acceptance. ‘Rational drug design’ refers to the systematic study of correlations between compound composition and its bioactive properties.

---

**Box 15.1: Natural Pharmaceuticals**

Though burdened by political, environmental, and economic issues, pharmaceutical industries have long explored unusual venues for disease remedies, many in remote parts of the world and involving indigenous cures. Micro-organisms and fungi, in particular, are globally available and can be reproduced readily. For example, among the world's 25 top-selling drugs in 1997, seven were derived from natural sources. Some notable examples of products derived from Nature are listed below.

- A fungus found on a Japanese golf course is being used by Merck to make the cholesterol lowering drug mevacor, one of the 25 top-sellers of 1997.
- A fungus found on a Norwegian mountain is the basis for another 1997 top-seller, the transplant drug *Cyclosporin*, made by Novartis.
- A fungus from a Pacific yew tree is also the source of the anticancer agent paclitaxel (taxol).
- The rosy periwinkle of Madagascar is the source of Eli Lilly's two cancer drugs vincristine and vinblastine, which have helped fight testicular cancer and childhood leukemia since the 1960s.
- A microbe discovered in a Yellowstone hot spring is the source of a heat-resistant enzyme now key in DNA amplification processes.
- Ocean salmon is a source for osteoporosis drugs (*Calcimar* and *Miacalcin*), and coral extracts are used for bone replacement.
- The versatile polymer chitosan, extracted from crab and shrimp shells, is a well known fat-binding weight-loss aid, in addition to its usage in paper additives, pool cleaners, cosmetics, and hair gels.
- The *Artemisia annua* plant (also known as sweet wormwood), which grows in China, Vietnam, and some parts of the United States, provides the raw material for a malaria drug, artemisinin.
- Frog-skin secretions serve as models for development of painkillers with fewer side effects than morphine. This chemical secret, long exploited by Amazon rain forest tribesmen, is now being pursued with frogs from Ecuador by Abbott Labs.
- Marine organisms from the Philippines are being investigated as sources of chemicals toxic to cancer cells.
- The venomous lizard termed Gila monster inhabiting Phoenix, Arizona, may provide a powerful peptide, exendin, for treating diabetes, because it stimulates insulin secretion and aids digestion in lizards that gorge thrice-yearly.
- A compound isolated from a flowering plant in a Malaysian rainforest, calanolide A, is a promising drug candidate for AIDS therapy, in the class of non-nucleoside reverse transcriptase inhibitors.
- A protein from a West African berry was identified by University of Wisconsin scientists as 2000 times sweeter than sugar; sweeteners are being developed from this source to make possible sweeter food products by gene insertion.

- A natural marine product (ecteinascidin 743) derived from the Caribbean sea squirt *Ecteinascidia turbinata* was found to be an active inhibitor of cell proliferation in the late 1960s, but only recently purified, synthesized, and tested in clinical trials against certain cancers.
- A Caribbean marine fungus extract (developed as halimide) shows early promise against cancer, including some breast cancers resistant to other drugs.

One of the most challenging aspects of using natural products as pharmaceutical agents is a sourcing problem, namely extracting and purifying adequate supplies of the target chemicals. For example, biochemical variations within species combined with international laws restricting collection (e.g., of frogs from Ecuador whose skins contain an alkaloid compound with powerful painkilling effects) limit available natural sources. In the case of the frog skin chemical, this sourcing problem prompted the synthetic design of a new type of analgesic that is potentially nonaddictive [1006].

---

### 15.1.3 Molecular Modeling in Rational Drug Design

Since the 1980s, further improvements in modeling methodology, computer technology, as well as X-ray crystallography and NMR spectroscopy for biomolecules, have increased the participation of molecular modeling in this lucrative field. Molecular modeling is playing a more significant role in drug development [453,496,666,772,1179,1301,1376] as more disease targets are being identified and solved at atomic resolution (e.g., HIV-1 protease, HIV integrase, adenovirus receptor, protein kinases), as our understanding of the molecular and cellular aspects of disease is enhanced (e.g., regarding pain signaling mechanisms, or the immune invasion mechanism of the HIV virus), and as viral genomes are sequenced [529]. Indeed, in analogy to genomics and proteomics — which broadly define the enterprises of identifying and classifying the genes and the proteins in the genome — the discipline of *chemogenomics* [198] has been associated with the delineation of drugs for all possible drug targets.

As described in the first chapter, examples of drugs made famous by molecular modeling include HIV-protease inhibitors (AIDS treatments), SARS virus inhibitor, thrombin inhibitors (for blood coagulation and clotting diseases), neuropeptide inhibitors (for blocking the pain signals resulting from migraines), PDE-5 inhibitors (for treating impotence by blocking a chemical reaction which controls muscle relaxation and resulting blood flow rate), various antibacterial agents, and protein kinase inhibitors for metastatic lung cancer and other tumors [913]. See Figure 15.1 for illustrations of popular drugs for migraine, HIV/AIDS, and blood-flow related diseases.

Such computer modeling and analysis — rather than using trial and error and exhaustive database studies — was thought to lead to dramatic progress in the design of drugs. However, some believe that the field of rational drug design has not lived up to its expectations.

One reason for the restrained success is the limited reliability of modeling molecular interactions between drugs and target molecules; such interactions must be described very accurately energetically to be useful in predictions. Newer approaches consider multiple targets [278] and work in system-oriented approaches [649] to improve success.

Another reason for the limited success of drug modeling is that the design of compounds with the correct binding properties (e.g., dissociation constants in the micromolar range and higher) is only a first step in the complex process of drug design; many other considerations and long-term studies are needed to determine the drug's bioactivity and its effects on the human body [1364]. For example, a compound may bind well to the intended target but be inactive biologically if the reaction that the drug targets is influenced by other components (see Box 15.2 for an example). Even when a drug binds well to an appropriate target, an optimal therapeutic agent must be delivered precisely to its target [999], screened for undesirable drug/drug interactions [1061], lack toxicity and carcinogenicity (likewise for its metabolites), be stable, and have a long shelf life.

The problems of viability and efficacy are even more important now with the increased development and usage of *biologics* or *biotherapeutics* — biological molecules like proteins derived from living cells and used as drugs — rather than small-molecule drugs. Such biologics, which include various vaccines, are typically administered by injection or infusion. Successful recent examples are Wyeth's *Enbrel* for rheumatoid arthritis, Genetech's *Avastin* for cancer, and Amgen's *EpoGen* for anemia. Many large pharmaceutical companies are increasing their work on biologics because such drugs are more complex and expensive to replicate and hence much less vulnerable to the usual patent expiration which allows introduction of generics and thereby restricts the profits of the original manufacturers. However, the big challenge in biologics is dealing with the characteristic heterogeneity of such biological molecules and better understanding their mechanism of action related to the disease target and long-term effects.

#### 15.1.4 The Competition: Automated Technology

Even accepting those limitations of computer-based approaches, rational drug design has avid competition from automated technology: new synthesis techniques, such as robotic systems that can run hundreds of concurrent synthetic reactions, have emerged, thereby enhancing synthesis productivity enormously. With "high-throughput screening", these candidates can be screened rapidly to analyze binding affinities, determine transport properties, and assess conformational flexibility.

Many believe that such a production *en masse* is the key to establishing diverse databases of drug candidates. Thus, at this time, it might be viewed that *drug design need not be 'rational' if it can be exhaustive*. Still, others advocate a more focused design approach, based on structures of ligands or receptors [453], fragment-based drug design [1447], or virtual screening approaches applied to smaller subsets of compounds [453, 1179].

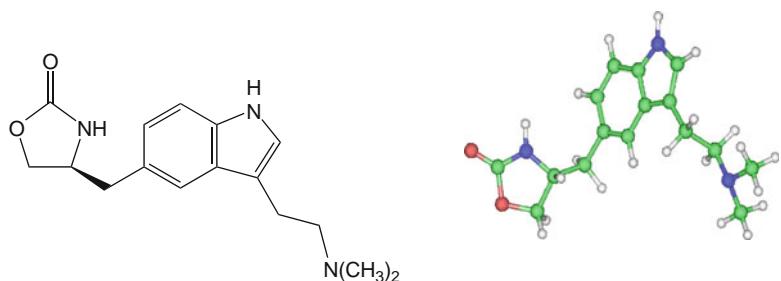
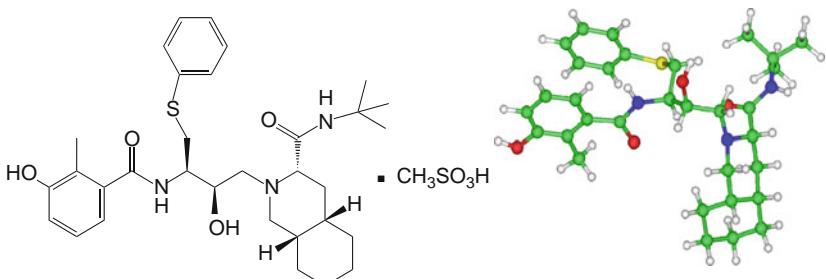
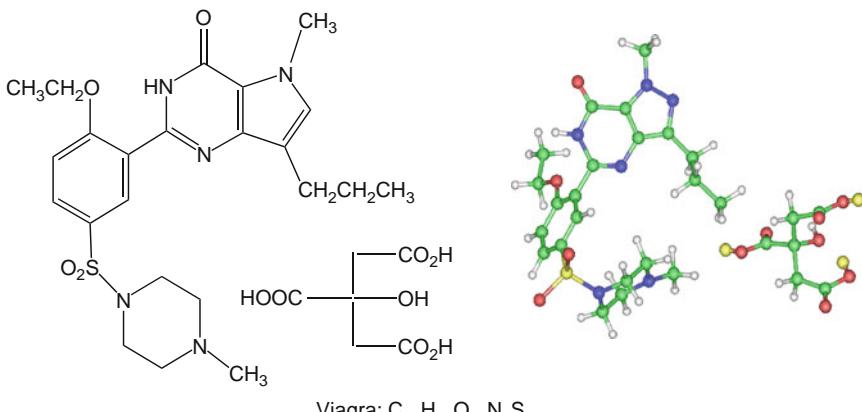
Zomig:  $\text{C}_{16}\text{H}_{21}\text{N}_3\text{O}_2$ Viracept:  $\text{C}_{33}\text{H}_{49}\text{O}_7\text{N}_3\text{S}_2$ Viagra:  $\text{C}_{28}\text{H}_{37}\text{O}_{11}\text{N}_6\text{S}$ 

Figure 15.1. Popular drug examples. Top: *Zolmitriptan* (zomig) for migraines, a 5-HT<sub>1</sub> receptor agonist that enhances the action of serotonin. Middle: *Nelfinavir Mesylate* (viracept), a protease inhibitor for AIDS treatment. Bottom: *Sildenafil Citrate* (viagra) for penile dysfunctions, a temporary inhibitor of phosphodiesterase-5, which regulates associated muscle relaxation and blood flow by converting cyclic guanosine monophosphate to guanosine monophosphate. See other household examples in Figure 15.3.

Another convincing argument for the focused design approach is that the amount of synthesized compounds is so vast (and rapidly generated) that computers will be essential to sort through the huge databases for compound management and applications. Such applications involve clustering analysis and *similarity* and *diversity sampling* (see below), preliminary steps in generating drug candidates or optimizing bioactive compounds.

This information explosion explains the resurrection of computer-aided drug design and its enhancement in scope under the new title **combinatorial chemistry**, affectionately endorsed as ‘the darling of chemistry’ [1376].

### 15.1.5 Chapter Overview

In this chapter, a brief introduction into some mathematical questions involved in this discipline of chemical library design is presented, namely *similarity* and *diversity sampling* for ligand-based drug design. Some ideas on cluster analysis and database searching are also described. This chapter is only intended to whet the appetite for chemical design and to invite mathematical scientists to work on related problems.

Because medicinal chemistry applications are an important subfield of chemical design, this last chapter also provides some perspectives on current developments in drug design, as well as mentioning emerging areas such as pharmacogenomics of personalized medicine and biochips (see Boxes 15.3 and 15.4).

## 15.2 Problems in Chemical Libraries

Chemical libraries consist of compounds (known chemical formulas) with potential and/or demonstrated therapeutic activities. Most libraries are proprietary, residing in pharmaceutical houses, but public sources also exist, like the National Cancer Institute’s (NCI’s) 3D structure database.

Both target-independent and target-specific libraries exist. The name ‘combinatorial libraries’ stems from the important combinatorial problems associated with the experimental design of compounds in chemical libraries, as well as computational searches for potential leads using concepts of **similarity** and **diversity** as introduced below.

### 15.2.1 Database Analysis

In broad terms, two general problem categories can be defined in chemical library analysis and design:

**Database systematics:** analysis and compound grouping, compound classification, elimination of redundancy in compound representation (dimensionality reduction), data visualization, etc., and

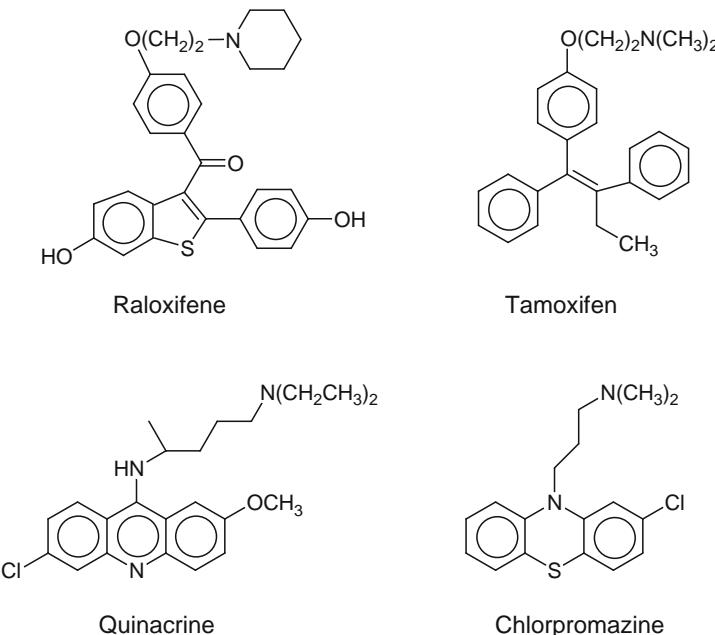


Figure 15.2. Related pairs of drugs: the antiestrogens raloxifene and tamoxifen, and the tricyclic compounds with aliphatic side-chains at the middle ring quinacrine and chlorpromazine.

**Database applications:** efficient formulation of quantitative links between compound properties and biological activity for compound selection and design optimization experiments.

Both of these general database problems involved in chemical libraries are associated with several mathematical disciplines. Those disciplines include multivariate statistical analysis and numerical linear algebra, multivariate nonlinear optimization (for continuous formulations), combinatorial optimization (for discrete formulations), distance geometry techniques, and configurational sampling.

### 15.2.2 Similarity and Diversity Sampling

Two specific problems, described formally in the next section after the introduction of chemical descriptors, are the similarity and diversity problems.

The *similarity* problem in drug design involves finding molecules that are ‘similar’ in physical, chemical, and/or biological characteristics to a known target compound. Deducing compound similarity is important, for example, when one drug is known and others are sought with similar physiochemical and biological properties, and perhaps with reduced side effects.

One example is the target bone-building drug *raloxifene*, whose chemical structure is somewhat related to the breast cancer drug *tamoxifen* (see Figure 15.2) (e.g., [1093]). Both are members of the family of *selective estrogen receptor modulators* (SERMs) that bind to estrogen receptors in the breast cancer cells and exert a profound influence on cell replication. It is hoped that raloxifene will be as effective for treating breast tumors but will reduce the increased risk of endometrial cancer noted for tamoxifen. Perhaps raloxifene will also not lose its effectiveness after five years like tamoxifen.

Another example of a related pair of drugs is *chlorpromazine* (for treating schizophrenia) and *quinacrine* (antimalarial drug). These tricyclic compounds with aliphatic side chains at the middle ring group (see Figure 15.2) were suggested as candidates for treating Creutzfeldt-Jakob and other prion diseases [677].

Because similarity in structure might serve as a first criterion for similarity in activity/function, similarity searching can be performed using 3D structural and energetic searches (e.g., induced fit or ‘docking’ [41, 818]) or using the concept of molecular descriptors introduced in the next section, possibly in combination with other discriminatory criteria.

The *diversity* problem in drug design involves delineating the most diverse subset of compounds within a given library. Diversity sampling is important for practical reasons. The smaller, representative subsets of chemical libraries (in the sense of being most ‘diverse’) might be searched first for lead compounds, thereby reducing the search time; representative databases might also be used to prioritize the choice of compounds to be purchased and/or synthesized, similarly resulting in an accelerated discovery process, not to speak of economic savings.

---

### Box 15.2: Treatments for Chronic Pain

Amazing breakthroughs have been achieved recently in the treatment of chronic pain. Such advances were made possible by an increased understanding of the distinct cellular mechanisms that cause pain due to different triggers, from sprained backs to arthritis to cancer.

**What is Pain?** Pain signals start when nerve fibers known as nociceptors, found throughout the human body, react to some disturbances in nearby tissues. The nerve fibers send chemical pain messengers that collect in the dorsal horn of the spinal cord. Their release depends on the opening of certain pain gates. Only when these messengers are released into the brain is pain felt in the body.

**Natural Ammunition.** Fortunately, the body has a battery of natural painkillers that can close those pain gates or send signals from the brain. These compensatory agents include endorphins, adrenaline, and serotonin (a peptide similar to opium). Many painkillers enhance or mimic the action of these natural aids (e.g., opium-bases drugs such as *morphine*, *codeine*, and *methadone*). However, these opiates have many undesirable side effects.

**Painkiller Targets.** To address the problem of pain, new treatments are targeting *specific* opiate receptors. For example, Actiq, developed by Anesta Corp. for intense cancer pain, is a lozenge placed in the cheek that is absorbed quickly into the bloodstream, avoiding the gut. Other pain relievers include a class of drugs known as COX-2 inhibitors, like Monsanto's *Celebrex* (celecoxib) and Merck's *Vioxx*, which relieve aches and inflammation with fewer stomach-damaging effects. They do so by targeting only one (COX-2) of two enzymes called cyclo-oxegenases (COX), which are believed to cause inflammation and thereby trigger pain.

While regular non-steroidal anti-inflammatory drugs (NSAIDs, like *Aspirin*, *Ibuprofen*, and *Naproxen*) and others available by prescription attack both COX-1 and COX-2, COX-1 is also known to protect the stomach lining; this explains the stomach pain that many people experience with NSAIDs and the pain relief without the side effects that COX-2 inhibitors can offer.

Modern pain treatment also involves compounds that stop pain signals before the brain gets the message, either by intercepting the signals in the spinal cord or by blocking their route to the spine. Evidence is emerging that a powerful chemical called 'substance P' can be used as an agent to deliver pain blockers to receptors found throughout the body; an experimental drug based on this idea (marketed by Pfizer) has proven effective at easing tooth pain.

---

### 15.2.3 Bioactivity Relationships

Besides database systematics, such as similarity and diversity sampling, the establishment of clear links between compound properties and bioactivity is, of course, the heart of drug design. In many respects, this association is not unlike the protein prediction problem in which we seek some target energy function that upon global minimization will produce the biologically relevant, or native, structure of a protein.

In our context, formulating that 'function' to relate sequence and structure while not ignoring the environment might even be more difficult, since we are studying small molecules for which the evolutionary relationships are not clear as they might be for proteins. Further, the bioactive properties of a drug depend on much more than its chemical composition, three-dimensional (3D) structure, and energetic properties. A complex orchestration of cellular machinery is often involved in a particular human ailment or symptom, and this network must be understood to alleviate the condition safely and successfully.

A successful drug has usually passed many rounds of chemical modifications that enhanced its potency, optimized its selectivity, and reduced its toxicity. An example involves obesity treatments by the hormone *leptin*. Limited clinical studies have shown that leptin injections do not lead to clear trends of weight loss in people, despite demonstrating dramatic slimming of mice. Though not a quick panacea in humans, leptin has nonetheless opened the door to pharmacological manipulations of body weight, a dream with many medical — not to speak of

monetary — benefits. Therapeutic manipulations will require an understanding of the complex mechanism associated with leptin regulation of our appetite, such as its signaling the brain on the status of body fat.

Box 15.2 contains another illustration of the need to understand such complex networks in connection with drug development for chronic pain. These examples clearly show that *lead generation*, the first step in drug development, is followed by *lead optimization*, the challenging, slower phase.

In fact, this complexity of the molecular machinery that underlies disease has given rise to the subdisciplines of *molecular medicine* and *personalized medicine* (see Boxes 15.3 and 15.4), where DNA technology plays an important role. Specifically, DNA chips — small glass wafers like computer chips studded with bits of DNA instead of transistors — can analyze the activities of thousands of genes at a time, helping to predict disease susceptibility in individuals, classify certain cancers, and to design treatments [400].

For example, DNA chips can study expression patterns in the tumor suppressor gene *p53* (the gene with the single most common mutations in human cancers), and such patterns can be useful for understanding and predicting response to chemotherapy and other drugs. DNA microarrays have also been used to identify genes that selectively stimulate metastasis (the spread of tumor cells from the original growth to other sites) in melanoma cells.

Besides developments on more personalized medicine, which will also be enhanced by a better understanding of the human body and its ailments, new advances in drug delivery systems may be important for improving the rate and period of drug delivery in general [1304].

---

### Box 15.3: Molecular and Personalized Medicine

**Pauling's Groundwork.** Molecular medicine seeks to enhance our therapeutic solutions by understanding the molecular basis of disease. Linus Pauling lay the groundwork for this field in his seminal 1949 paper [977] which demonstrated that the hemoglobin from sickle cell anemia sufferers has a different electric charge than that from healthy people. This difference was later explained by Vernon Ingram as arising from a single amino acid difference [590]. These pioneering works relied on electrophoretic mobility measurements and fingerprinting techniques (electrophoresis combined with paper chromatography) for peptides.

**Disease Simulations.** A modern incarnation of molecular medicine involves conducting virtual experiments by computer simulation with the goal of developing new hypotheses regarding disease mechanisms and prevention. For example, scientists at Entelos Inc. (Menlo Park, California) are simulating cell inflammation caused by asthma to try to learn how blocking certain inflammation factors might affect cellular receptors and then to identify targets for steroid inhalers.

**From SNPs to Tailored Drugs.** Another significant current trend in medicine is personalized medicine, the tailoring of drugs to individual genetic makeup. User-specific drugs

have great potential to be more potent and to eliminate adverse side effects experienced by some individuals. *Pharmacogenetics* is the field of studying how genetic factors influence drug response. Its newer sibling *pharmacogenomics* involves using genomics to describe individual responses to drugs. Pharmacogenomics (also abbreviated as Pgx or pgx) has become possible with the advent of microarray technology (e.g., [544, 1174]): these make possible large-scale genome-wide analyses to test thousands of genes for related activity with a specific drug. Developing tailored diets and vitamins based on individual responses to diet (determined in part by one's genes) is another growing field called *nutritional genomics* or *nutrigenomics*.

Specifically, the drug tailoring idea is based on identifying the small variations in people's DNA where a single nucleotide differs from the standard sequence. These mutations, or individual variations in genome sequence that occur once every couple of hundred of base pairs, are called single-nucleotide polymorphisms known as SNPs (pronounced "snips"). The presence of SNPs can be signaled visually using DNA chips or biochips, instruments of fancy of the biotechnology industry (See [400] and Box 1.4 of Chapter 1). Other genomic factors besides SNPs also serve as distinguishing factors in pharmacogenomics studies.

Pharmacogenetics gained momentum in April 1999 when eleven pharmaceutical and technology companies and the Wellcome Trust announced a genome *mapping* consortium for SNPs. The consortium's goal is to construct a fine-grained map of order 300,000 SNPs to permit searching for SNP patterns that correlate with particular drug responses. Efforts are ongoing, and many companies have specialized in this area. Pharmacogenomics now receives considerable attention both from the professional medical circles and the popular press. It has potential to markedly improve medical intervention, reduce hospitalization costs, and alleviate human suffering by increasing the efficacy and decreasing adverse effects in the drug treatment of various human diseases.

Some notable examples of success of pharmacogenomics include the genotype-based dosaging of the blood thinning drug *Warfarin*; administration of *Abacavir* (an RT inhibitor) to HIV patients; *Herceptin* treatment for HER2-positive breast cancer patients; and *Gleevac* and other cancer drugs for individual cancer patients. See Box 15.4 for details of some of these drugs.

Directed drugs are also under development to treat or diagnose diabetes, neurological diseases like Alzheimer's, prostate cancer, and ailments requiring antibiotics. Though there are many hurdles to this new field, not to mention possible financial drawbacks of genotyping, it is hoped that some benefits of cost savings in prescriptions and hospitalizations for adverse drug effects could be realized in the not-too-distant future [588].

---

---

#### Box 15.4: Examples of Successes in Pharmacogenomics

**Warfarin.** *Warfarin* is the "darling" of pharmacogenomics because international collaborations by the International Warfarin Pharmacogenetics Consortium and the Pharmacogenetics Research Network have led to development of a dosing algorithm [591].

This is a milestone in the evolution of drug prescription from trial and error to exact science [591]. Warfarin is an anti-coagulation agent given to patients with risk of heart disease. However, adverse effects can be catastrophic since the patient may bleed to death. Practical experience has shown that reactions to the drug vary widely from person to person. But why? Pharmacogenomics analyses revealed that a patient's response to Warfarin depends on the presence of two genes encoding two proteins: CYP2C9 which metabolizes warfarin, and VKORC1 which recycles vitamin K and affects clotting factors. Certain genotypes make reaction much more sensitive. In 2007, FDA modified Warfarin labels to highlight the potential relevance of genetic information to prescribing decisions.

**Abacavir.** *Abacavir* is a guanosine reverse-transcriptase inhibitor used as an anti-retroviral treatment against infections of HIV. However, 5 to 8% of the white population develops adverse side effects, namely toxic skin reaction. In 2002, it was discovered that the HLAB\*5701 gene variant is highly associated with this hyper sensitivity. Genotyping has thus been used to effectively reduce the number of such adverse reactions. Genetic testing for sensitivity to abacavir is now widely used.

**Herceptin.** *Herceptin* (Trastuzumab) is an antibody used to treat breast cancer. Studies have shown that Herceptin is effective for patients with over-expression of the human epidermal growth factor receptor HER2, which occurs in invasive breast carcinomas. Herceptin has now been approved by the FDA for patients with invasive breast cancer that over expresses HER2.

**Codeine for Breast-Feeding Mothers.** *Codeine* is a painkiller often prescribed to help women with post-delivery pain. Codeine is metabolized into morphine, but it was generally considered to be safe for breast-feeding mothers. In 2005, a 13-day old male baby in Toronto who was breastfed by a codeine-treated mother died of a morphine overdose [673]. Investigations revealed that the mother was an “ultra-metabolizer” of codeine, and this led to an unusually high amount of morphine in the baby. Studies have shown that the metabolism of codeine is related to the CYP2D6 gene. Subsequently, genetic testing for this variant has been suggested for mothers who want to breastfeed and receive codeine for post-delivery pain [1375]. Alternatively, breast feeding can be avoided, reduced, and/or the level of morphine in the neonate monitored carefully to prevent unnecessary deaths.

---

## 15.3 General Problem Definitions

### 15.3.1 *The Dataset*

Our given dataset of size  $n$  contains information on compounds with potential *biological activity* (drugs, herbicides, pesticides, etc.). A schematic illustration is presented in Figure 15.3. The value of  $n$  is large, say one million or more. Because of the enormous dataset size, the problems described below are simple to solve in principle but extremely challenging in practice because of the large associated computational times. Any systematic schemes to reduce this computing time can thus be valuable.

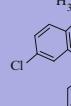
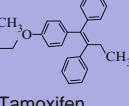
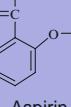
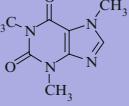
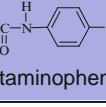
Compound ( $i = 1, \dots, n$ )		Vectorial Descriptors ( $k = 1, \dots, m$ )							
		$X_{i_1}$	$X_{i_2}$	$\dots$	$X_{i_m}$	$B_{i_1}$	$B_{i_2}$	$\dots$	$B_{i_{m_B}}$
1		0.873	0.763	$\dots$	0.531	0	1	$\dots$	0
	Valium								
2		0.912	0.131	$\dots$	0.834	0	0	$\dots$	1
	Tamoxifen								
3		0.763	0.214	$\dots$	0.533	0	0	$\dots$	0
	Aspirin								
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$i$		0.925	0.237	$\dots$	0.742	1	0	$\dots$	1
	Caffeine								
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n$		0.347	0.279	$\dots$	0.846	1	1	$\dots$	0
	Acetaminophen								

Figure 15.3. A chemical library can be represented by  $n$  compounds  $i$  (known or potential drugs), each associated with  $m$  characteristic descriptors ( $\{X_{ik}\}$ ) and activities  $\{B_{ij}\}$  with respect to  $m_B$  biological targets (known or potential).

### 15.3.2 The Compound Descriptors

Each compound in the database is characterized by a vector (the *descriptor*). The vector can have real or binary elements. There are many ways to formulate these descriptors so as to reduce the database search time and maximize success in generation of lead compounds.

Conventionally, each compound  $i$  is described by a list of **chemical descriptors**, which may reflect *molecular composition*, such as atom number, atom connectivity, or number of functional groups (like aromatic or heterocyclic rings, tertiary aliphatic amines, alcohols, and carboxamides), *molecular geometry*, such as number of rotatable bonds, *electrostatic properties*, such as charge distribution, and various *physiochemical measurements* that are important for bioactivity.

These descriptors are currently available from many commercial packages like Molconn-X and Molconn-Z (Hall Associates Consulting, Quincy, MD). Descriptors fall into many classes. Examples include:

*2D descriptors* — also called molecular connectivity or topological indices — reflecting molecular connectivity and other topological invariants;

*binary descriptors* — simpler encoded representations indicating the presence or absence of a property, such as whether or not the compound contains at least three nitrogen atoms, doubly-bonded nitrogens, or alcohol functional groups;

*3D descriptors* — reflecting geometric structural factors like van der Waals volume and surface area; and

*electronic descriptors* — characterizing the ionization potential, partial atomic charges, or electron densities.

See also [8] for further examples.

Binary descriptors allow rapid database analysis using Boolean algebra operations. The MolConn-X and MolConn-Z programs, for example, generate topological descriptors based on molecular connectivity indices (e.g., number of atoms, number of rings, molecular branching paths, atoms types, bond types, etc.). Such descriptors have been found to be a convenient and reasonably successful approximation to quantify molecular structure and relate structure to biological activity (see review in [6]). These descriptors can be used to characterize compounds in conjunction with other selectivity criteria based on activity data for a training set (e.g., [322, 582]). The search for the most appropriate descriptors is an ongoing enterprise, not unlike force-field development for macromolecules.

The number of these descriptors,  $m$ , is roughly on the order of 1000, thus much smaller than  $n$  (the number of compounds) but too large to permit standard systematic comparisons for the problems that arise.

Let us define the vector  $X_i$  associated with compound  $i$  to be the row  $m$ -vector

$$\{X_{i1}, X_{i2}, \dots, X_{im}\}.$$

Our dataset  $\mathcal{S}$  can then be described as the collection of  $n$  vectors

$$\mathcal{S} = \{X1, X2, X3, \dots, Xn\},$$

or expressed as a rectangular matrix  $A_{n \times m}$  by listing, in rows, the  $m$  chemical descriptors of the  $n$  database compounds:

$$A = \begin{pmatrix} X1_1 & X1_2 & \dots & \dots & \dots & X1_m \\ X2_1 & X2_2 & \dots & \dots & \dots & X2_m \\ \vdots & & & \ddots & & \vdots \\ \vdots & & & \ddots & & \vdots \\ \vdots & & & \ddots & & \vdots \\ \vdots & & & \ddots & & \vdots \\ Xn_1 & Xn_2 & \dots & \dots & \dots & Xn_m \end{pmatrix}. \quad (15.1)$$

In practice, this rectangular  $n \times m$  matrix has  $n \gg m$  (i.e., the matrix is long and narrow), where  $n$  is on the order of millions and  $m$  is several hundreds.

The compound descriptors are generally *highly redundant*. Yet, it is far from trivial how to select the “principal descriptors”. Thus, various statistical techniques (principal component analysis, classic multivariate regression; see below) have been used to assess the degree of correlation among variables so as to eliminate highly-correlated descriptors and reduce the dimension of the problems involved.

### 15.3.3 Characterizing Biological Activity

Another aspect of each compound in such databases is its *biological activity*. Pharmaceutical scientists might describe this property by associating a simple *affirmative* or *negative* score with each compound to indicate various areas of activity (e.g., with respect to various ailments or targets, which may include categories like headache, diabetes, protease inhibitors, etc.).

Drugs may enhance/activate (e.g., *agonists*) or inhibit (e.g., *antagonists*, *inhibitors*) certain biochemical processes. This bioactivity aspect of database problems is far less quantitative than the simple chemical descriptors. Of course, it also requires synthesis and biological testing for activity determination. Studies of several drug databases have suggested that active compounds can be associated with certain ranges of physiochemical properties like molecular weight and occurrence of functional groups [451].

For the purpose of the problems outlined here, it suffices to think of such an additional set of descriptors associated with each compound. For example, a matrix  $B_{n \times m_B}$  may complement the  $n \times m$  database matrix  $A$ ; see Figure 15.3. Each

row  $i$  of  $B$  may correspond to measures of activity of compound  $i$  with respect to specific targets (e.g., binary variables for active/nonactive target response).

*The ultimate goal in drug design is to find a compound that yields the desired pharmacological effect.* This quest has led to the broad area termed SAR, an acronym for Structure/Activity Relationship [709]. This discipline applies various statistical, modeling, or optimization techniques to relate compound properties to associated pharmacological activity. A simple linear model, for example, might attempt to solve for variables in the form of a matrix  $X_{m \times m_B}$ , satisfying

$$AX = B. \quad (15.2)$$

Explained more intuitively, SAR formulations attempt to relate the given compound descriptors to experimentally-determined bioactivity markers. While earlier models for ‘quantitative SAR’ (QSAR) involved simple linear formulations for fitting properties and various statistical techniques (e.g., multivariate regression, principal component analysis), nonlinear optimization techniques combined with other visual and computational techniques are more common today [448]. The problem remains very challenging, with rigorous frameworks continuously being sought.

### 15.3.4 The Target Function

To compare compounds in the database to each other and to new targets, a quantitative assessment can be based on common structural features. Whether characterized by topological (chemical-formula based) or 3D features, this assessment can be broadly based on the vectorial chemical descriptors provided by various computer packages. A target function  $f$  is defined, typically based on the *Euclidean distance* function between vector pairs,  $\delta$ , where

$$f(X_i, X_j) = \delta_{ij} \equiv \|X_i - X_j\| = \sqrt{\sum_{k=1}^m (X_{ik} - X_{jk})^2}. \quad (15.3)$$

Thus, to measure the similarity or diversity for each pair of compounds  $X_i$  and  $X_j$ , the function  $f(X_i, X_j)$  is often set to the simple distance function  $\delta_{ij}$ . Other functions of distance are also appropriate depending upon the objectives of the optimization task.

### 15.3.5 Scaling Descriptors

Scaling the descriptor components is important for proper assessment of the score function [1372]. This is because the individual chemical descriptors can vary drastically in their magnitudes as well as the variance within the dataset. Subsequently, a few large descriptors can overwhelm the similarity or diversity measures. For example, actual descriptor components of a database compound may look like the following:

11.0000	0.6433	4.5000	0.0833	150.2200	8.4831	0.0159	-1.0000	113.2239	..
1.000	0.2917	0.5000	0.0000	40.0000	7.2566	0.0801	1.0000	782.7121	..
-8.0000	0.2081	0.5000	0.0186	80.0000	0.0000	0.0017	1.0000	62.2016	..
2.0000	0.0000	2.5000	-0.9010	0.0000	1.3867	0.2500	1.0000	120.0030	..
0.0000	0.0000	3.0000	0.0326	0.0000	-4.3984	0.1759	1.0000	11.2189	..
80.0000	-0.0442	6.0000	0.7002	210.0000	-1.9784	0.0026	-1.0000	370.3473	..
-5.0000	-0.1491	0.0000	0.0000	10.0000	9.0909	0.1641	1.0000	98.2782	..
-1.0000	0.5427	4.5000	0.8963	35.0000	2.0061	0.0720	1.0000	119.8090	..
17.0000	-0.3209	0.5000	0.0803	0.0000	9.4765	0.0000	-1.0000	11.7011	..
19.0000	0.2690	1.0000	-0.3420	90.0000	0.0000	0.0000	-1.0000	201.0180	..
0.0000	0.0000	0.0000	0.2000	40.0000	9.1702	0.0429	-1.0000	23.2423	..
4.0000	0.3061	0.5000	0.6670	10.0000	2.3820	0.0023	1.0000	0.0000	..
4.0000	0.7702	1.5000	0.1870	0.0000	0.0000	0.7290	1.0000	0.0000	..
1.0000	-0.1134	1.5000	0.3356	40.0000	0.0000	0.7782	-1.0000	314.6658	..
0.0000	0.0000	0.0000	0.7842	0.0000	-6.1659	0.0000	1.0000	85.2285	..
3.0000	0.0000	0.0000	0.2382	75.0000	4.2276	0.1260	1.0000	7.2854	..
15.0000	0.3479	4.0000	0.0034	0.0000	0.5152	0.3018	1.0000	280.8721	..
7.0000	0.6945	3.5000	0.4552	0.0000	3.5315	0.3065	-1.0000	0.0000	..
.	.	.	.	.	.	.	.	.	..

Clearly, the ranges of individual descriptors vary (e.g., 0 to 1 versus 0 to 1000). Thus, given no chemical/physical guidance, it is customary to scale the vector entries before analysis. In practice, however, it is very difficult to determine the appropriate scaling and displacement factors for the specific application problem [1372]. A general scaling of each  $X_{ik}$  to produce  $\hat{X}_{ik}$  can be defined using two real numbers  $\alpha_k$  and  $\beta_k$ , for  $k = 1, 2, \dots, m$ , termed the *scaling* and *displacement* factors, respectively, where  $\alpha_k > 0$ . Namely, for  $k = 1, 2, \dots, m$ , we define the scaled components as

$$\hat{X}_{ik} = \alpha_k (X_{ik} - \beta_k), \quad 1 \leq i \leq n. \quad (15.4)$$

The following two scaling procedures are often used. The first makes each column in the range  $[0, 1]$ : each column of the matrix  $A$  is modified using eq. (15.4) by setting the factors as

$$\begin{aligned} \beta_k &= \min_{1 \leq i \leq n} X_{ik}, \\ \alpha_k &= 1 / \left( \max_{1 \leq i \leq n} X_{ik} - \beta_k \right). \end{aligned} \quad (15.5)$$

This scaling procedure is also termed “standardization of descriptors”.

The second scaling produces a new matrix  $A$  where each column has a mean of zero and a standard deviation of one. It does so by setting the factors (for  $k = 1, 2, \dots, m$ ) as

$$\begin{aligned} \beta_k &= \frac{1}{n} \sum_{i=1}^n X_{ik}, \\ \alpha_k &= 1 / \sqrt{\frac{1}{n} \sum_{i=1}^n (X_{ik} - \beta_k)^2}. \end{aligned} \quad (15.6)$$

Both scaling procedures defined by eqs. (15.5) and (15.6) are based on the assumption that no one descriptor dominates the overall distance measures.

### 15.3.6 The Similarity and Diversity Problems

The Euclidean distance function  $f(X_i, X_j) = \delta_{ij}$  based on the chemical descriptors can be used in performing similarity searches among the database compounds and between these compounds and a particular target. This involves optimization of the distance function over  $i = 1, \dots, n$ , for a fixed  $j$ :

$$\text{Minimize}_{X_i \in \mathcal{S}} \{f(\delta_{ij})\}. \quad (15.7)$$

More difficult and computationally-demanding is the diversity problem. Namely, we seek to reduce the database of the  $n$  compounds by selecting a “representative subset” of the compounds contained in  $\mathcal{S}$ , that is one that is “the most diverse” in terms of potential chemical activity. We can formulate the diversity problem as follows:

$$\text{Maximize} \sum_{X_i, X_j \in \mathcal{S}_0} \{f(\delta_{ij})\} \quad (15.8)$$

for a given subset  $\mathcal{S}_0$  of size  $n_0$ .

The molecular diversity problem naturally arises since pharmaceutical companies must scan huge databases each time they search for a specific pharmacological activity. Thus reducing the set of  $n$  compounds to  $n_0$  representative elements of the set  $\mathcal{S}_0$  is likely to accelerate such searches. ‘Combinatorial library design’ corresponds to this attempt to choose the best set of constituents for combinatorial synthetic schemes so as to maximize the likelihood of identifying lead compounds.

The molecular diversity problem involves maximizing the volume spanned by the elements of  $\mathcal{S}_0$  as well as the separation between those elements. Geometrically, we seek a well separated, uniform-like distribution of points in the high-dimensional compound space in which each chemical cluster has a ‘representative’.

A simple, heuristic formulation of this problem might be based on the similarity problem above: successively minimize  $f(\delta_{ij})$  over all  $i$ , for a fixed (target)  $j$ , so as to eliminate a subset  $\{X_i\}$  of compounds that are similar to  $X_j$ . This approach thus identifies groupings that *maximize intracluster similarity* as well as *maximize intercluster diversity*.

The *combinatorial optimization* problem, an example of a very difficult computational task, has *non-polynomial computational complexity* (‘NP-complete’) (see footnote in Chapter 11, Section 11.2). This is because an exhaustive calculation of the above distance-sum function over a *fixed set*  $\mathcal{S}_0$  of  $n_0$  elements requires a total of  $\mathcal{O}(n_0^2 m)$  operations. However, there are many possible subsets of  $\mathcal{S}$  of size  $n_0$ , namely  $C_n^{n_0}$  of them, where

$$\begin{aligned} C_n^{n_0} &= \frac{n!}{n_0! (n - n_0)!} \\ &= \frac{n(n-1)(n-2)\cdots(n-n_0+1)}{n_0!}. \end{aligned} \quad (15.9)$$

As a simple example, for  $n = 4$ , we have  $C_4^1 = 4/1 = 4$  subsets of one element;  $C_4^2 = (4 \times 3)/2 = 6$  different subsets of two elements,  $C_4^3 = (4 \times 3 \times 2)/(3!) = 4$  subsets of three elements, and  $C_4^4 = (4 \times 3 \times 2 \times 1)/(4!) = 1$  subset of four elements.

Typically, these combinatorial optimization problems are solved by stochastic and heuristic approaches. These include genetic algorithms, simulated annealing, and tabu-search variants. (See Agrafiotis [5], for example, for a review).

As in other applications, the efficiency of simulated annealing depends strongly on the choice of cooling schedule and other parameters. Several potentially valuable annealing algorithms such as deterministic annealing, multiscale annealing, and adaptive simulated annealing, as well as other variants, have been extensively studied.

Various formulations of the diversity problem have been used in practice. Examples include the maximin function — to maximize the minimum intermolecular similarity:

$$\text{Maximize}_{i, X_i \in \mathcal{S}_0} \left\{ \min_{\substack{j \neq i \\ X_j \in \mathcal{S}_0}} (\delta_{ij}) \right\} \quad (15.10)$$

or its variant — maximizing the sum of these distances:

$$\text{Maximize}_{X_i, X_j \in \mathcal{S}_0} \sum_i \left\{ \min_{j \neq i} (\delta_{ij}) \right\}. \quad (15.11)$$

The maximization problem above can be formulated as a minimization problem by standard techniques if  $f(x)$  is normalized so it is monotonic with range  $[0, 1]$ , since we can often write

$$\max[f(x)] \Leftrightarrow \min[-f(x)] \text{ or } \min[1 - f(x)].$$

In special cases, combinatorial optimization problems can be formulated as integer programming and mixed-integer programming problems. In this approach, linear programming techniques such as interior methods can be applied to the solution of combinatorial optimization problems, leading to branch and bound algorithms, cutting plane algorithms, and dynamic programming algorithms. Parallel implementation of combinatorial optimization algorithms is also important in practice to improve the performance.

Other important research areas in combinatorial optimization include the study of various algebraic structures (such as matroids and greedoids) within which some combinatorial optimization problems can more easily be solved [263].

Currently, practical algorithms for addressing the diversity problem in drug design are relatively simple heuristic schemes that have computational complexity of at most  $\mathcal{O}(n^2)$ , already a huge number for large  $n$ .

## 15.4 Data Compression and Cluster Analysis

Dimensionality reduction and data visualization are important aids in handling the similarity and diversity problems outlined above. Principal component analysis (PCA) is a classic technique for data compression (or dimensionality reduction). It has already shown to be useful in analyzing microarray data (e.g., [1009]), as discussed in Chapter 1. The singular value decomposition (SVD) is another closely related approach. Data visualization for cluster analysis requires dimensionality reduction in the form of a projection from a high-dimensional space to 2D or 3D so that the dataset can be easily visualized. Cluster analysis is heuristic in nature.

In this section we outline the PCA and SVD approaches for dimensionality reduction in turn, continue with the distance refinement that can follow such analyses, and illustrate projection and clustering results with some examples.

### 15.4.1 Data Compression Based on Principal Component Analysis (PCA)

PCA transforms the input system (our database matrix  $A$ ) into a smaller matrix described by a few uncorrelated variables called the **principal components** (PCs). These PCs are related to the eigenvectors of the covariance matrix defined by the component variables. The basic idea is to choose the orthogonal components so that the original data variance is well approximated. That is, the relations of similarity/dissimilarity among the compounds can be well approximated in the reduced description. This is done by performing eigenvalue analysis on the covariance matrix that describes the statistical relations among the descriptor variables.

#### Covariance Matrix and PCs

Let  $a_{ij}$  be an element of our  $n \times m$  database matrix  $A$ . The covariance matrix  $C_{m \times m}$  is formed by elements  $c_{jj'}$  where each entry is obtained from the sum

$$c_{jj'} = \frac{1}{n-1} \sum_{i=1}^n (a_{ij} - \mu_j)(a_{ij'} - \mu_{j'}) . \quad (15.12)$$

Here  $\mu_j$  is the mean of the column associated with descriptor  $j$ :

$$\mu_j = \frac{1}{n} \sum_{i=1}^n a_{ij} . \quad (15.13)$$

$C$  is a symmetric semi-definite matrix and thus has the spectral decomposition

$$C = V\Sigma V^T , \quad (15.14)$$

where the superscript  $T$  denotes the matrix transpose, and the matrix  $V$  ( $m \times m$ ) is the orthogonal eigenvector matrix satisfying  $VV^T = I_{m \times m}$  with  $m$  component

vectors  $\{v_i\}$ . The diagonal matrix  $\Sigma$  of dimension  $m$  contains the  $m$  ordered eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0.$$

We then define the  $m$  PCs  $Yj$  for  $j = 1, 2, \dots, m$  as the product of the original matrix  $A$  and the eigenvectors  $v_j$ :

$$Yj = Av_j, \quad j = 1, 2, \dots, m. \quad (15.15)$$

We also define the  $m \times m$  matrix  $Y$  corresponding to eq. (15.15), related to  $V$ , as the matrix that holds the  $m$  PCs  $Y1, Y2, \dots, YM$ ; this allows us to write eq. (15.15) in the matrix form  $Y = AV$ . Since  $VV^T = I$ , we then obtain an expression for the dataset matrix  $A$  in terms of the PCs:

$$A = YV^T. \quad (15.16)$$

### Dimensionality Reduction

The problem dimensionality can be reduced based on eq. (15.16). First note that eq. (15.16) can be written as:

$$A = \sum_{j=1}^m Yj \cdot v_j^T. \quad (15.17)$$

Second, note that  $Xi$ , the vector of compound  $i$ , is the transpose of the  $i$ th row vector of  $A$ :

$$Xi = A^T e_i, \quad (15.18)$$

where  $e_i$  is an  $n \times 1$  unit vector with 1 in the  $i$ th component and 0 elsewhere. Thus, compound  $Xi$  is expressed as the linear combination of the orthonormal set of eigenvectors  $\{v_j\}$  of the covariance matrix  $C$  derived from  $A$ :

$$Xi = \sum_{j=1}^m (Yj_i) v_j, \quad i = 1, 2, \dots, n, \quad (15.19)$$

where  $Yj_i$  is the  $i$ th component of the column vector  $Yj$ .

Based on eq. (15.19), the problem dimensionality  $m$  can be reduced by constructing a  $k$ -dimensional approximation to  $Xi$ ,  $Xi^k$ , in terms of the first  $k$  PCs:

$$Xi^k = \sum_{j=1}^k (Yj_i) v_j, \quad i = 1, 2, \dots, n. \quad (15.20)$$

The index  $k$  of the approximation can be chosen according a criterion involving the threshold variance  $\gamma$ , where

$$\left( \sum_{i=1}^k \lambda_i \right) / \left( \sum_{i=1}^m \lambda_i \right) \geq \gamma. \quad (15.21)$$

The eigenvalues of  $C$  represent the variances of the PCs. Thus, the measure  $\gamma = 1$  for  $k = m$  reflects a 100% variance representation. In practice, good approximations to the overall variance (e.g.,  $\gamma > 0.7$ ) can be obtained for  $k \ll m$  for large databases.

For such a suitably chosen  $k$ , the smaller database represented by components  $\{Xi^k\}$  for  $i = 1, 2, \dots, n$  approximates the variance of the original database  $A$  reasonably, making it valuable for cluster analysis.

As we show below, the singular value decomposition can be used to compute the factorization of the covariance matrix  $C$  when the ‘natural scaling’ of eq. (15.6) is used.

### 15.4.2 Data Compression Based on the Singular Value Decomposition (SVD)

SVD is a procedure for data compression used in many practical applications like image processing and cryptanalysis (code deciphering) [296, for example]. Essentially, it is a factorization for rectangular matrices that is a generalization of the eigenvalue decomposition for square matrices. Image processing techniques are common tools for managing large datasets, such as digital encyclopedias, or images transmitted to earth from space shuttles on limited-speed modems.

SVD defines two appropriate *orthogonal coordinate systems* for the domain and range of the mapping defined by a rectangular  $n \times m$  matrix  $A$ . This matrix maps a vector  $x \in \mathcal{R}^n$  to a vector  $y = Ax \in \mathcal{R}^m$ . The SVD determines the orthonormal coordinate system of  $\mathcal{R}^n$  (the columns of an  $n \times n$  matrix  $U$ ) and the orthonormal coordinate system of  $\mathcal{R}^m$  (the columns of an  $m \times m$  matrix  $V$ ) so that  $A$  is diagonal.

The SVD is used routinely for storing computer-generated images. If, a photograph is stored as a matrix where each entry corresponds to a pixel in the photo, fine resolution requires storage of a huge matrix. The SVD can factor this matrix and determine its *best rank- $k$  approximation*. This approximation is computed not as an explicit matrix but rather as a sum of  $k$  outer products, each term of which requires the storage of two vectors, one of dimension of  $n$  and another of dimension  $m$  ( $m+n$  storage for the pair). Hence, the total storage required for the image is reduced from  $nm$  to  $(m+n)k$ .

The SVD also provides the *rank of  $A$*  (the number of independent columns), thus specifying how the data may be stored more compactly via the best rank- $k$  approximation. This reformulation can reduce the computational work required for evaluation of the distance function used for similarity or diversity sampling.

#### SVD Factorization

The SVD decomposes the real matrix  $A$  as:

$$A = U\Sigma V^T, \quad (15.22)$$

where the matrices  $U$  ( $n \times n$ ) and  $V$  ( $m \times m$ ) are orthogonal, i.e.,  $UU^T = I_{n \times n}$  and  $VV^T = I_{m \times m}$ . The matrix  $\Sigma$  ( $n \times m$ ) contains at most  $m$  nonzero entries ( $\sigma_i, i = 1, \dots, m$ ), known as the *singular values*, in the first  $m$  diagonal elements:

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \sigma_2 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & \cdots & \sigma_r & \cdots & \cdots & \cdots & 0 \\ \vdots & & & & & & & \\ 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & \sigma_m \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & & & & & & & \\ 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix} \quad (15.23)$$

where

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r \dots \geq \sigma_m \geq 0.$$

The columns of  $U$ , namely  $u_1, \dots, u_n$ , are the *left singular vectors*; the columns of  $V$ , namely  $v_1, \dots, v_m$ , are the *right singular vectors*. In addition,  $r = \text{rank of } A = \text{number of nonzero singular values}$ . Thus if  $r \ll m$ , a rank- $r$  approximation of  $A$  is natural. Otherwise, we can set  $k$  to be smaller than  $r$  by neglecting the singular values beyond a certain threshold.

### Low-Rank Approximation

The rank- $k$  approximation to  $A$  can be obtained by noting that  $A$  can be written as the sum of rank-1 matrices:

$$A = \sum_{j=1}^r \sigma_j u_j v_j^T. \quad (15.24)$$

The rank- $k$  approximation,  $A_k$ , is simply formed by extending the summation in eq. (15.24) from 1 to  $k$  instead of 1 to  $r$ . In practice, this means storing  $k$  left singular vectors and  $k$  right singular vectors. This matrix  $A_k$  can also be written as

$$A_k = \sum_{j=1}^k \sigma_j u_j v_j^T = U \Sigma_k V^T \quad (15.25)$$

where

$$\Sigma_k = \text{diag} (\sigma_1, \dots, \sigma_k, 0, \dots, 0).$$

This matrix is closest to  $A$  in the sense that

$$\|A - A_k\| = \sigma_{k+1}$$

for the standard Euclidean norm.

Recall that we can express each  $X_i$  as:

$$\text{Row } i \text{ of } (A) = (A^T e_i)^T,$$

where  $e_i$  is an  $n \times 1$  unit vector with 1 in the  $i$ th component and 0 elsewhere. Using the decomposition of eq. (15.24), we have:

$$A^T e_i = \sum_{j=1}^r \sigma_j v_j u_j^T e_i = \sum_{j=1}^r (\sigma_j u_{j_i}) v_j.$$

The SVD transforms this row vector to  $[(A_k)^T e_i]^T$ , where:

$$(A_k)^T e_i = \sum_{j=1}^k (\sigma_j u_{j_i}) v_j. \quad (15.26)$$

### Projection

This transformation can be used to project a vector onto the first  $k$  principal components. That is, the projection matrix  $P_k = \sum_{j=1}^k [v_j v_j^T]$  maps a vector from  $m$  to  $k$  dimensions. For example, for  $k = 2$ , we have:

$$\begin{aligned} P_2 A^T e_i &= \sum_{j=1}^r (v_1 v_1^T + v_2 v_2^T) (\sigma_j u_{j_i}) v_j \\ &= (\sigma_1 u_{1_i}) v_1 + (\sigma_2 u_{2_i}) v_2. \end{aligned} \quad (15.27)$$

Thus, this projection maps the  $m$ -dimensional row vector  $X_i$  onto the two-dimensional (2D) vector  $Y_i$  with components  $\sigma_1 u_{1_i}$  and  $\sigma_2 u_{2_i}$ . This mapping generalizes to a projection onto the  $k$ -dimensional space where  $k \ll m$ :

$$Y_i^k = (\sigma_1 u_{1_i}, \sigma_2 u_{2_i}, \dots, \sigma_k u_{k_i}). \quad (15.28)$$

### 15.4.3 Relation Between PCA and SVD

It can be shown that the eigenvectors  $\{v_i\}$  of the covariance matrix (eq. (15.14)) coincide with the right eigenvectors  $\{v_i\}$  defined above when the second scaling (eq. (15.6)) is applied to the database matrix. Recall that this scaling makes all columns have zero means and a variance of unity.

Moreover, the left SVD vectors  $\{u_i\}$  can be related to the singular values  $\{\sigma_i\}$  and PC vectors  $\{Y_i\}$  of eq. (15.15) by

$$u_i = A v_i / \sigma_i = Y_i / \sigma_i. \quad (15.29)$$

Therefore, we can use the SVD factorization as defined above (eq. (15.22)) to compute the PCs  $\{Y_i\}$  of the covariance matrix  $C$ . The SVD approach is more efficient since formulation of the covariance matrix is not required.

The algorithm **ARPACK** [728] can compute the first  $k$  PCs, saving significant storage. It requires an order  $\mathcal{O}(nk)$  memory and  $\mathcal{O}(nm^2)$  floating point operations.

#### 15.4.4 Data Analysis via PCA or SVD and Distance Refinement

The SVD or the PCA projection is a first step in database visualization. The second step refines this projection so that the original Euclidean distances  $\{\delta_{ij}\}$  in the  $m$ -dimensional space are closely related to the corresponding distances  $\{d_{ij}\}$  in the reduced,  $k$ -D space. Here,

$$\delta_{ij} \equiv \|Xi - Xj\|$$

and

$$d_{ij} \equiv \|Yi - Yj\|$$

for all  $i, j$ , where the vectors  $\{Y_i\}$  are the  $k$ -D vectors produced by SVD defined by eq. (15.28).

#### Projection Refinement

This distance refinement is a common task in distance geometry refinement of NMR models. In the NMR context, a set of interatomic distances is given and the objective is to find the 3D coordinate vector (the molecular structure) that best fits the data. Since such a problem is typically overdetermined — there are  $\mathcal{O}(n^2)$  distances but only  $\mathcal{O}(n)$  Cartesian coordinates for a system of  $n$  atoms — an optimal *approximate solution* is sought.

For example, optimization work on evolutionary trees [1001] solved an identical mathematical problem in an unusual context that is closely related to the molecular similarity problem here. Specifically, the experimental distance-data in evolutionary studies reflect complex factors rather than simple spatial distances (e.g., interspecies data arise from immunological studies which compare the genetic material among taxa and assign similarity scores). Finding a 3D evolutionary tree by the distance-geometry approach, rather than the conventional 2D tree which conveys evolutionary linkages, helps identify subgroup similarities.

#### Distance Geometry

The distance-geometry problem in our evolutionary context can be formulated as follows. We are given a set of pairwise distances with associated lower and upper bounds:

$$\{l_{ij} \leq \delta_{ij} \leq u_{ij}\}, \quad \text{for } i, j = 1, 2, \dots, n,$$

where each  $\delta_{ij}$  is a target interspecies distance with associated lower and upper bounds  $l_{ij}$  and  $u_{ij}$ , respectively, and  $n$  is the number of species. Our goal is to compute a 3D “tree” for those species based on the measured distance/similarity data.

This distance geometry problem can be reduced to finding a coordinate vector  $Y$  that minimizes the objective function

$$E(Y) = \sum_{i < j} \omega_{ij} (d_{ij}^2(Y) - \delta_{ij}^2)^2, \quad (15.30)$$

where  $d_{ij}(Y)$  is Euclidean distance between points  $i$  and  $j$  in the vector  $Y$ , and the  $\{\omega_{ij}\}$  are appropriately-chosen weights.

In the combinatorial chemistry context, we use the same function  $E(Y)$  where  $Y$  is the vector of  $2n$  components, listing the 2D projections of each compound in turn. Details of this data clustering approach are described in [1399, 1402]. Minimization can be performed so that the high-dimensional distance relationships are approximated.

Besides the value of the objective function (eq. (15.30)), a useful measure of the distance approximation in the low-dimensional space is the percentage of intercompound distances  $\{i, j\}$  (out of  $n(n-1)/2$ ) that are within a certain threshold of the original distances. We first define the deviations from the targets by a percentage  $\eta$  so that

$$\begin{aligned} |d(Yi, Yj) - \delta_{ij}| &\leq \eta \delta_{ij} & \text{when } \delta_{ij} > d_{\min}, \\ d(Yi, Yj) &\leq \tilde{\epsilon} & \text{when } \delta_{ij} \leq d_{\min}, \end{aligned} \quad (15.31)$$

where  $\eta$ ,  $\tilde{\epsilon}$ , and  $d_{\min}$  are given small positive numbers less than one. For example,  $\eta = 0.1$  specifies a 10% accuracy; the other values may be set to small positive numbers such as  $d_{\min} = 10^{-12}$  and  $\tilde{\epsilon} = 10^{-8}$ . The second case above (very small original distance) may occur when two compounds in the datasets are highly similar.

With this definition, the total number  $T_d$  of the distance segments  $d(Yi, Yj)$  satisfying eq. (15.31) can be used to assess the degree of distance preservation of our mapping. We define the percentage  $\rho$  of the distance segments satisfying eq. (15.31) as

$$\rho = \frac{T_d}{n(n-1)/2} \times 100. \quad (15.32)$$

The greater the  $\rho$  value (the maximum is 100), the better the mapping and the more information that can be inferred from the projected views of the database compounds.

This minimization procedure (projection refinement) is quite difficult for scaled datasets. Experiments with several chemical datasets of size 58 to 27255 compounds show that the percentage of distances satisfying a threshold deviation  $\rho$  of 10% (eq. (15.31)) is in the range of 40% [1399, 1402]. Nonetheless, these low values can be made close to 100% with projections onto 10-dimensional space. This is illustrated in Figure 15.4, which shows the percentage of distances satisfying eq. (15.31) for  $\eta = 0.1$  as a function of the projection dimension for a database ARTF.

A similar improvement can be achieved with larger tolerances  $\eta$  (e.g., distances that are within 25% of the original values rather than 10%) [1399, 1402].

### 15.4.5 Projection, Refinement, and Clustering Example

As an illustration, consider the model database ARTF of 402 compounds and  $m = 312$  descriptors containing eight chemical subgroups. We have analyzed this

database by performing 2D and 3D projections based on the SVD factorization followed by minimization refinement by TNPACK [1121, 1122, 1397] for performance assessment in terms of accuracy as well as visual analysis of the compound interrelationships.

From Figure 15.4 we note that the refinement stage that follows the SVD projection is important for increasing the accuracy in every dimension. Namely, the accuracy is increased by 25–40% in this example.

The 2D and 3D projection patterns obtained for ARTF in Figure 15.5 show the utility of such a projection approach. The resemblance between the 2D and 3D views is evident, and the various 3D views offer different perspectives of the intercompound relationships.

We note that clusters corresponding to individual pharmacological subsets appear very close to each other, though partial overlap of clusters is evident. The *ecdysteroid* group forms a diverse but separate set of points. The *estrogen* class is also clustered and somewhat separate from the others. The strong overlap of the three clusters corresponding to *D1 agonists*, *D1 antagonists*, and *H1 receptor ligands* is reasonable given the relative chemical similarity of these compounds: all act at receptors of the same pharmacological class (i.e., G-protein coupled receptors). Thus, such data compression and visualization techniques can be used as a quick analysis tool of the database structure.

The chemical structures in Figure 15.6 reveal that compounds that are nearer in the projection are more closely related than those that are distant; this is seen when compounds are compared both within the same subgroup and within different subgroup. For example, the two labeled estrogen representatives that are

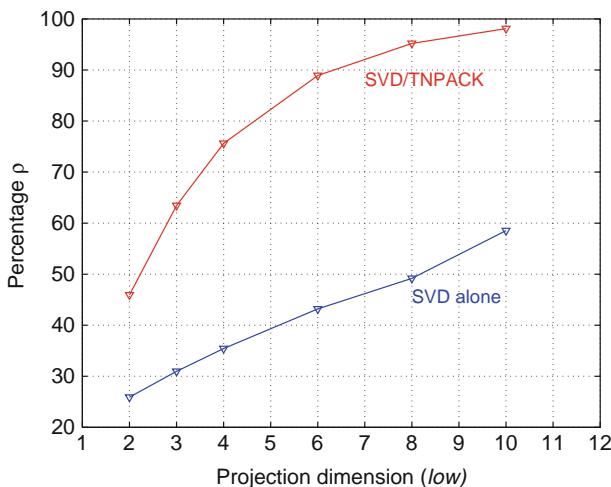


Figure 15.4. Performance of the SVD and SVD/minimization protocols for the ARTF chemical database in terms of the percentage of distances satisfying eq. (15.31) for  $\eta = 0.1$  (reflecting 10% distance deviations) as a function of the projection dimension [1399, 1402].

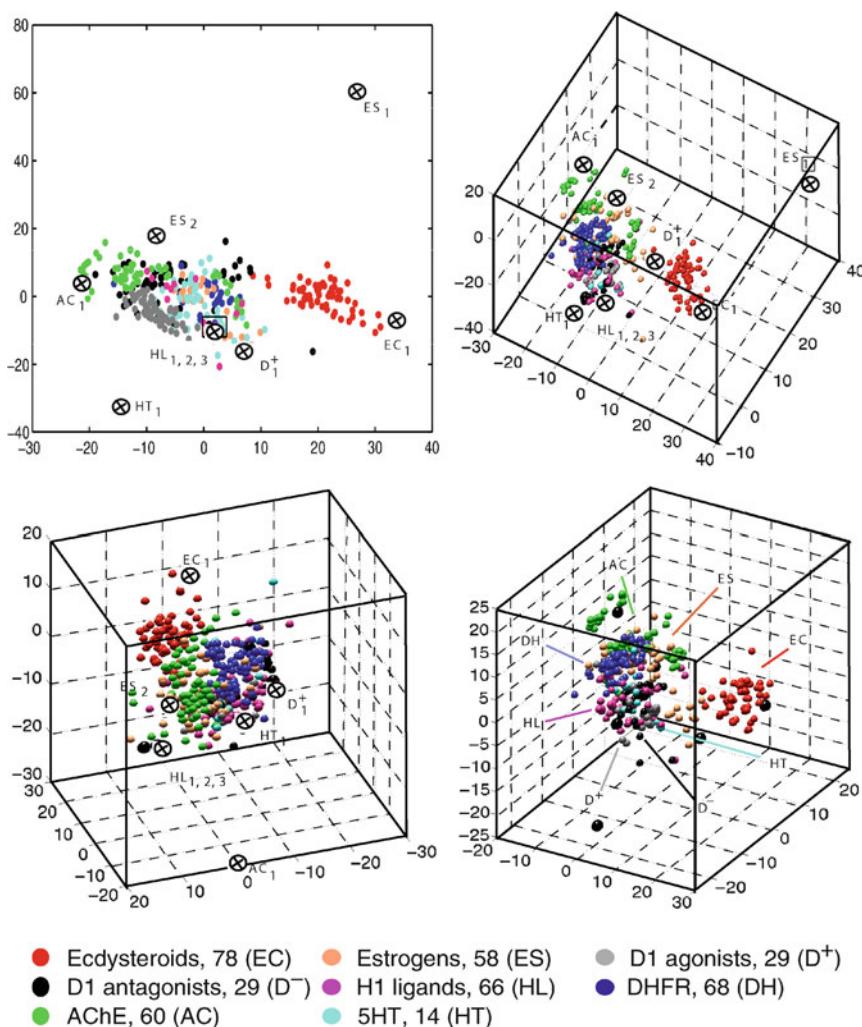


Figure 15.5. Two and three-dimensional projections of the chemical database ARTF of 402 compounds composed of the eight chemical subgroups ecdysteroids (EC), estrogens (ES), D1 agonists ( $D^+$ ), D1 antagonists ( $D^-$ ), H1 ligands (HL), DHFR inhibitors (DH), AchE inhibitors (AC), and 5HT ligands (HT) using the projection/refinement SVD/TNPACK approach [1399, 1402]. Three views are shown for the 3D projection. The accuracy of the 2D projection is about 46% and that of the 3D is 63% (with  $\eta = 0.1$ ); see eq. (15.31). The 2D projection was obtained by refining the 3D projection. The nine chemical structures labeled in the projections are drawn in Figure 15.6.

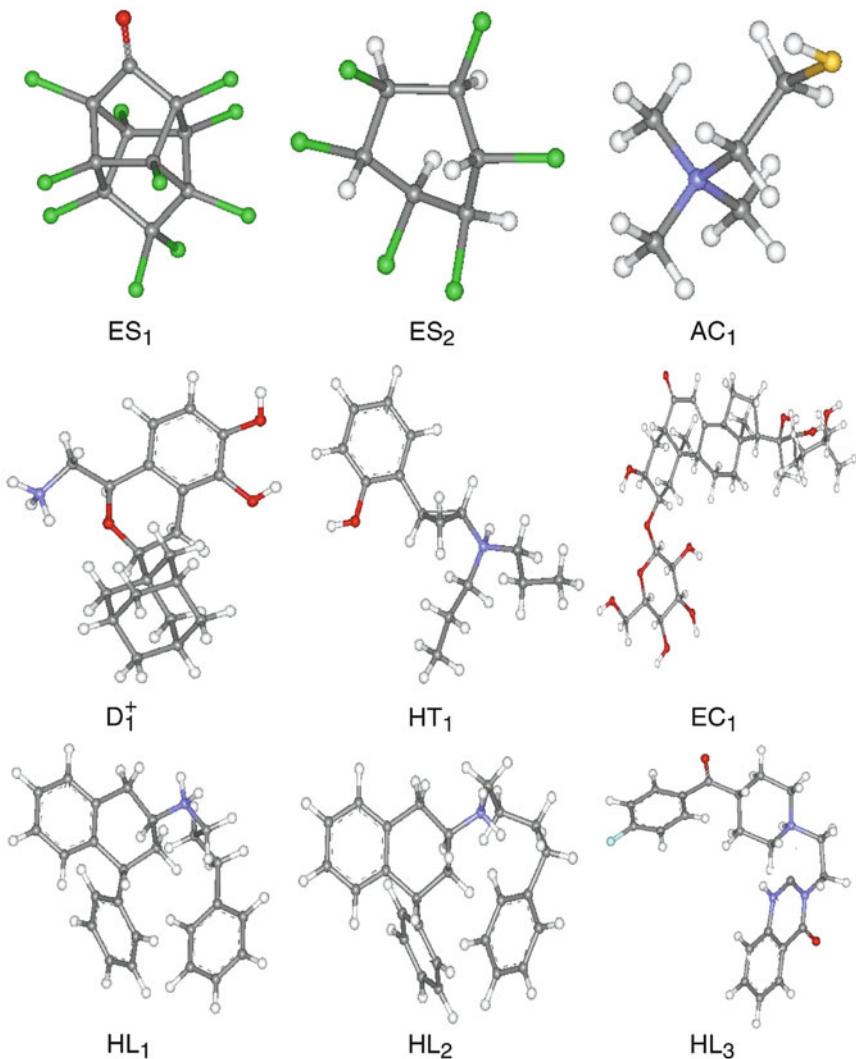


Figure 15.6. Selected chemical structures from the ARTF projection shown in Figure 15.5 reveal similarity of nearby structures and dissimilarity of distant compounds.

distant in the projection appear chemically quite different, while the three clustered H1 ligands appear similar to each other and perhaps to the nearby D1 agonist representative.

An example of a database projection in 2D by the alternative PCA approach followed by distance refinement is shown in Figures 15.7 and 15.8 for 832 compounds from the MDL Drug Data Report (MDDR) database using topological indices. (This work was performed in collaboration with Merck

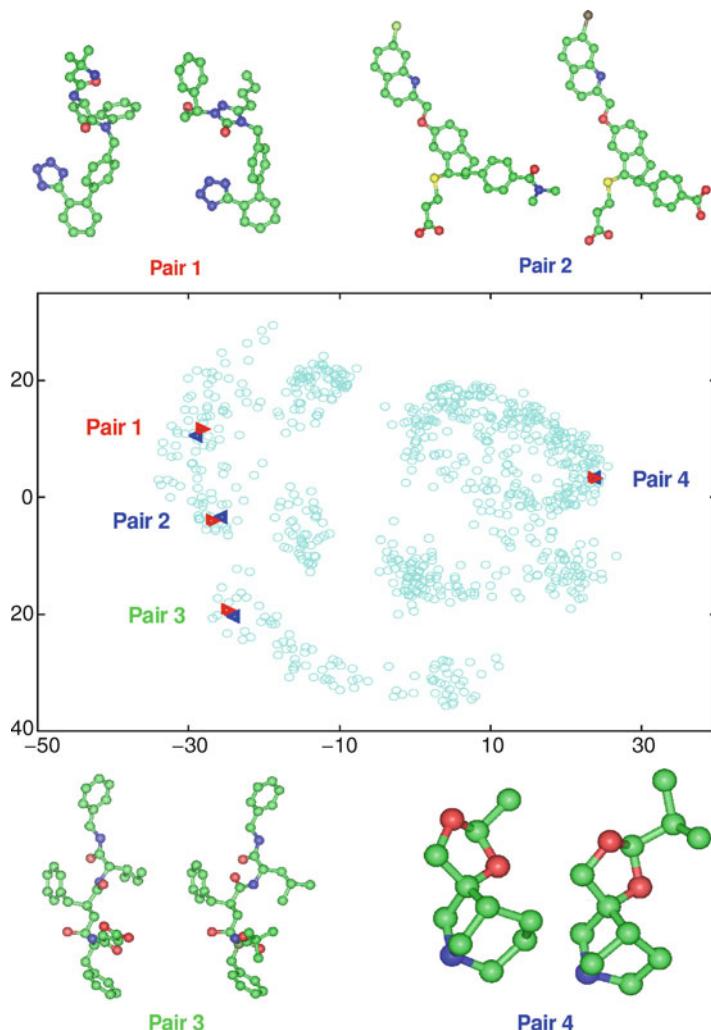


Figure 15.7. 2D projection using PCA for 832 compounds in the MDDR database showing the similarity of four compound pairs that are near in the projection.

Research Laboratories). The accuracy of this projection (the percentage of distances satisfying eq. (15.31) for  $\eta = 0.1$ ) is only 0.2% after PCA and 24.8% after PCA/TNPACK. Figure 15.7 shows that compounds close in the projection appear similar, and Figure 15.8 shows that more distantly related compounds tend to be different. Without knowing the grouping of these compounds according to bioactivity, the clusters identified in Figure 15.8 suggest a ‘diversity subset’ consisting of a few members from each cluster.

The approach described here appears promising, but further work is required to make the technique viable for very large databases.

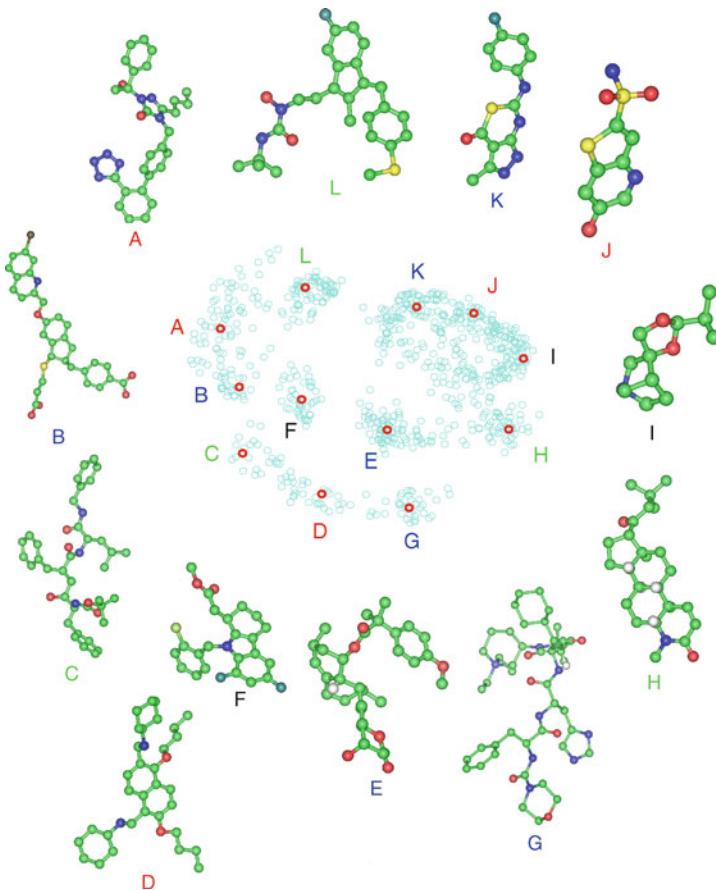


Figure 15.8. 2D projection using PCA for 832 compounds in the MDDR database showing the diversity of compounds that represent different clusters in the projection (distinguished by letters). A representative subset may thus consist of one or only a few members from each cluster.

## 15.5 Future Perspectives

Similarity and diversity sampling of combinatorial chemistry libraries is a field in its infancy. The choice of descriptors as well as metrics used to define similarity and diversity are empirical and perhaps application dependent. Thus, many challenges remain for future developments in the field, and the added involvement of mathematical scientists and new approaches borrowed from allied disciplines might be fruitful.

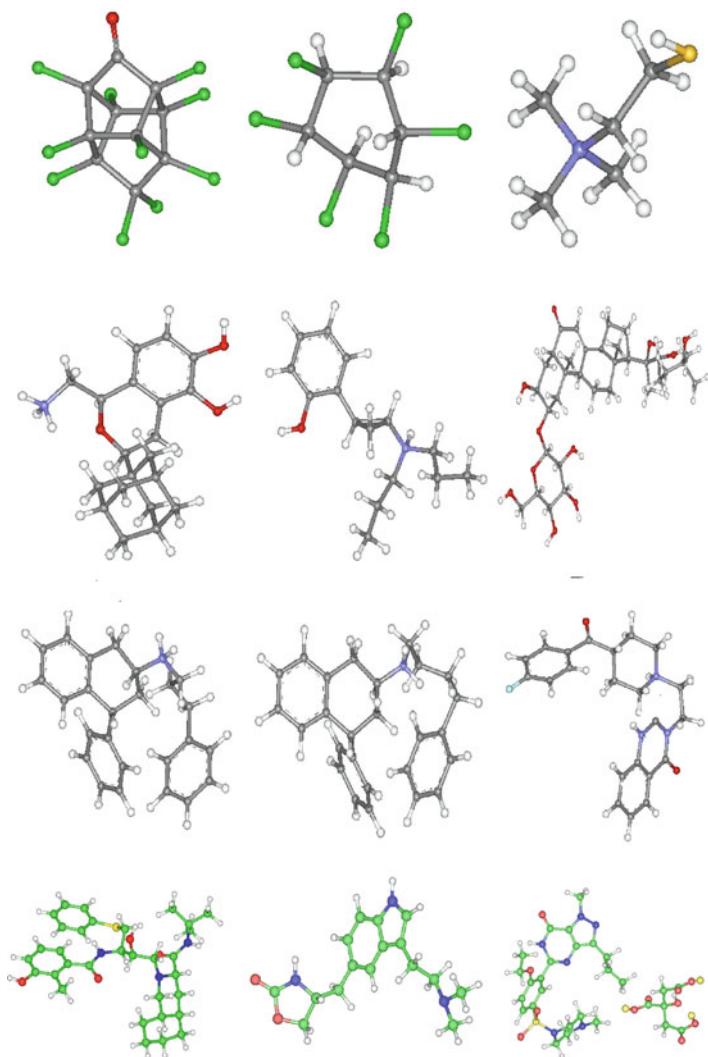
Developments are needed for formulation of descriptor sets, rigorous mathematical frameworks for their analysis, and efficient algorithms for very large-scale problems based on statistics, cluster analysis, and optimization. The algorithmic

challenge of manipulating large datasets might also explain the tendency toward smaller and focused libraries [555]; still, as argued in [621], this assumed defeat is premature!

The central assumption of structure/activity relationships of course remains a challenge to validate, develop, and further apply.

More broadly, structure-based drug design is likely to increase in importance as many more protein targets are identified and synthesized [1301], and as modeling programs improve in their ability to predict binding affinities of certain ligands (e.g., peptide-like) that share chemical groups with macromolecules, the focus of many biomodeling packages. The difficulty in determining membrane protein structures continues to be a limitation since membrane receptors are important pharmacological targets.

While perhaps not the dominant technique, it is clear that structure-based drug design will be an important component of drug modification and optimization after available leads have been generated. The search for the needle in the haystack (i.e., a successful drug) will likely be guided by the steady light generated by computer modeling. And, with additional genetic and genomic screening, disease treatment is likely to move forward to a new phase of greater scientific precision and success.



## Epilogue

The need for accuracy must be weighed against the need for finality.

Comments by the Justices of the U.S. Supreme Court regarding bringing to a close the uncertain outcome, one month after the elections, of the U.S. Presidential elections in 2000 between George W. Bush and Albert Gore (2000).

# Appendix A

## Molecular Modeling Sample Syllabus

Please Note: Article numbers in the Homework column refer to items in the course reading list (Appendix B).

Class	Subject	Homework
1	Course and Field Overview: <ul style="list-style-type: none"> <li>• What is molecular modeling and how has it evolved?</li> <li>• What are the practical applications and important questions?</li> </ul> <b>(Preface and Chapter 1)</b>	<b>1:</b> Introduction to sequence and structural databases and to the early molecular modeling literature. <i>Read papers 1,2,3,20,33,42, 44,47,52,54,58.</i>
2	<ul style="list-style-type: none"> <li>• Continuation of the overview on biomolecular modeling and simulation, from drug design to new materials.</li> <li>• Discussion of the 1959 paper of Alder &amp; Wainwright and 1971 work of Rahman &amp; Stillinger: difficulties then and now.</li> <li>• Introduction to interesting biomolecular modeling problems: protein folding, protein misfolding, nucleic acid/protein interactions, and RNA folding.</li> </ul> <b>(Chapter 2)</b>	<b>2:</b> Retrieval of structural information from the Protein Data Bank (PDB), and the display, manipulation, and analysis of three-dimensional biomolecular structures with the Insight II molecular graphics package. Explore kinemage tutorials. <i>Read papers 5,8,30,31,34,37</i>
3	<ul style="list-style-type: none"> <li>• Minitutorial on protein structure: amino acid repertoire, primary to quaternary structure, protein structure classification.</li> <li>• Kinemage tutorial demonstration: folding motifs and major protein classes.</li> </ul> <b>(Chapters 3 &amp; 4)</b>	<b>3:</b> Construction and analysis of the pentapeptide Met-enkephalin with the Insight II program. <i>Read papers 2,4,6.</i>
4	<ul style="list-style-type: none"> <li>• Discuss homework assignments 1 and 2.</li> <li>• Minitutorial on nucleic acid structure: building blocks, backbone conformational flexibility, helical parameters, and DNA supercoiling.</li> </ul> <b>(Chapters 5–7)</b>	<b>4:</b> Generation and analysis of Ramachandran plots for proteins and introduction to the NDB. <i>Read papers 23,29,47,50.</i>
5	Guest Lecturer: <i>The Nucleic Acid Database and the ‘New Protein Databank’ (RCSB)</i> , Prof. Helen Berman (Rutgers University, Department of Chemistry), Director of NDB and RCSB.	<b>5:</b> Analysis of Protein/DNA Complexes with Insight and NDB. <i>Read papers 7,21.</i>
6	<ul style="list-style-type: none"> <li>• Discuss homework assignment 4.</li> <li>• Computational and theoretical approaches to structure prediction (from the quantum-mechanical to the molecular mechanical description).</li> </ul> <b>(Chapters 8 &amp; 9)</b>	<b>6 (MIDTERM):</b> Sequence/Structure/Function Relationships in Proteins, A Contest! <i>Read papers 22,36,38,56.</i>

Table A.1: (continued)

Week	Subject	Homework
7	Guest Lecturer: <i>Protein Structure Modeling</i> , Dr. Andrej Sali, expert in protein modeling.	<b>7:</b> Molecular mechanics force fields: approximations, variations, and the assessment of results with respect to experiment and other simulations ( <i>papers 10,11,15,16</i> ). <i>Read papers 13,14,17,24.</i>
8	Amer. Chem. Soc. 1990 videotapes: <i>Molecular Modeling in Biological Systems</i> : 1 – Peter Kollman, “Methods in Molecular Modeling”, 4 – Panel Discussion.	
9	Guest Lecturer: <i>Ab Initio Calculation of Protein Structure by Global Optimization of Potential Energy</i> , Prof. Harold Scheraga (Cornell University, Department of Chemistry), pioneer of protein force fields and computation of protein structure.	<b>8:</b> A bit of programming: nonbonded versus bonded energy computations.
10	MIDTERM class presentations	
11	• Continue MIDTERM presentations. • <b>Force Field Debate!</b>	<b>9 (TERM PROJECT):</b> The Successes (Failures?) of Molecular Modeling. <i>Read papers 9,18,19.</i>
12	• Molecular mechanics force fields — origin, variations, and parameterization. • Special topics — molecular topology: book-keeping and data structures, potential energy differentiation. • Special issues in nonbonded energy computations — spherical cutoffs, fast electrostatics by the multipole method, periodic boundary conditions, and the Ewald summation. <b>(Chapter 10)</b>	<b>10:</b> Experiments in Geometry Structure Optimization: Minimization of Biphenyl with Insight II/Discover. <i>Read paper 12.</i>
13	• Go over Assignment 8, including general discussion of programming and timing strategies. • Optimization techniques for multivariate functions in computational chemistry. <b>(Chapter 11)</b>	<b>11:</b> A global optimization contest for a pentapeptide! <i>Read papers 25,26.</i>
14	Monte Carlo Simulations. <b>(Chapter 12)</b>	<b>12:</b> An exercise in Monte Carlo. <b>13 (Optional):</b> Advanced exercises in minimization and MC. <i>Read papers 27,28,32,41,53,57,59.</i>
15	Molecular dynamics simulations — theory and practice. <b>(Chapters 13 &amp; 14)</b>	<b>14 (Optional):</b> Advanced exercises in molecular dynamics. <b>15 (Optional):</b> Scaling of protein conformations and folding simulations.

# Appendix B

## Article Reading List

### Before 1970

1. B. J. Alder and T. E. Wainwright, “Studies in Molecular Dynamics. I. General Method”, *J. Chem. Phys.* **31**, 459–466 (1959).
2. G. Némethy and H. A. Scheraga, “Theoretical Determination of Sterically Allowed Conformations of a Polypeptide Chain by a Computer Method”, *Biopolymers* **3**, 155–184 (1965).

### 1970s

3. A. Rahman and F. H. Stillinger, “Molecular Dynamics Study of Liquid Water”, *J. Chem. Phys.* **55**, 3336–3359 (1971).
4. P. Y. Chou and G. D. Fasman, “Prediction of Protein Conformation”, *Biochemistry* **13**, 222–245 (1974).
5. M. Levitt and A. Warshel, “Computer Simulation of Protein Folding”, *Nature* **253**, 694–698 (1975).
6. M. Levitt and C. Chothia, “Structural Patterns in Globular Proteins”, *Nature* **261**, 552–558 (1976).

### 1980s

7. S. Lifson, “Potential Energy Functions for Structural Molecular Biology”, in *Methods in Structural Molecular Biology*, pp. 359–385, D. B. Davies, W. Saenger, and S. S. Danyluk, Eds., Plenum Press, London (1981).
8. M. Karplus and J. A. McCammon, “The Dynamics of Proteins”, *Sci. Amer.* **254**, 42–51 (1986).
9. M. S. Friedrichs and P. G. Wolynes, “Toward Protein Tertiary Structure Recognition by Means of Associative Memory Hamiltonians”, *Science* **246**, 371–373 (1989).
10. I. K. Roterman, M. H. Lambert, K. D. Gibson, and H. A. Scheraga, “Comparison of the CHARMM, AMBER and ECEPP Potentials for Peptides. I. Conformational Predictions for the Tandemly Repeated Peptide (Asn-Ala-Asn-Pro)<sub>9</sub>”, *J. Biomol. Struct. Dyn.* **7**, 391–419 (1989a).
11. I. K. Roterman, M. H. Lambert, K. D. Gibson, and H. A. Scheraga, “Comparison of the CHARMM, AMBER and ECEPP Potentials for Peptides. II.  $\phi$ – $\psi$  Maps for N-Methyl Amide: Comparisons, Contrasts and Simple Experimental Tests”, *J. Biomol. Struct. Dyn.* **7**, 421–453 (1989b).

**1990–1992**

12. M. Karplus and G. A. Petsko, “Molecular Dynamics Simulations in Biology”, *Nature* **347**, 631–639 (1990).
13. J. Skolnick and A. Kolinski, “Simulations of the Folding of a Globular Protein”, *Science* **250**, 1121–1125 (1990).
14. F. M. Richards, “The Protein Folding Problem” *Sci. Amer.* **264**, 54–63 (1991).
15. P. A. Kollman and K. A. Dill, “Decisions in Force Field Development: An Alternative to Those Described by Roterman *et al.*”, *J. Biomol. Struct. Dyn.* **8**, 1103–1107 (1991).
16. K. B. Gibson and H. A. Scheraga”, “Decisions in Force Field Development: Reply to Kollman and Dill”, *J. Biomol. Struct. Dyn.* **8**, 1109–1111 (1991).
17. H. A. Scheraga, “Predicting Three-Dimensional Structures of Oligopeptides”, in Reviews in Computational Chemistry, K. B. Lipkowitz and D. B. Boyd, Editors, Vol. 3, pp. 73–142, VCH Publishers, New York (1992).
18. T. Schlick, “Optimization Methods in Computational Chemistry”, in Reviews in Computational Chemistry, K. B. Lipkowitz and D. B. Boyd, Editors, Vol. 3, pp. 1–71, VCH Publishers, New York (1992). See also T. Schlick, “Geometry Optimization”, in the Encyclopedia of Computational Chemistry, P. von Ragué Schleyer (Editor-in-Chief) and N. L. Allinger and T. Clark and J. Gasteiger and P. A. Kollman and Schaefer, III, H. F., Editors, Vol. 3, pp. 1136–1157, John Wiley & Sons, West Sussex, England (1998).

**1993–1995**

19. R. A. Abagyan and M. M. Totrov, “Biased Probability Monte Carlo Conformational Searches and Electrostatic Calculations for Peptides and Proteins”, *J. Mol. Biol.* **235**, 983–1002 (1994).
20. J. A. Board, Jr., L. V. Kalé, K. Schulten, R. D. Skeel, and T. Schlick, “Modeling Biomolecules: Larger Scales, Longer Durations”, *IEEE Comp. Sci. Eng.* **1**, 19–30 (Winter 1994).
21. K. B. Lipkowitz, “Abuses of Molecular Mechanics. Pitfalls to Avoid”, *J. Chem. Educ.* **72**, 1070–1075 (1995).
22. B. Honig and A. Nicholls, “Classical Electrostatics in Biology and Chemistry”, *Science* **268**, 1144–1149 (1995).

**1996–1998**

23. B. Cipra, “Computer Science Discovers DNA”, in *What’s Happening in the Mathematical Sciences*, pp. 26–37 (P. Zorn, Ed.), American Mathematical Society, Colonial Printing, Cranston, RI (1996).
24. A. Neumaier, “Molecular Modeling of Proteins and Mathematical Prediction of Protein Structure”, *SIAM Review* **39**, 407–460 (1997).
25. K. A. Dill and H. S. Chan, “From Levinthal to Pathways to Funnels”, *Nature Struct. Biol.* **4**, 10–19 (1997).

26. T. Lazaridis and M. Karplus, “‘New View’ of Protein Folding Reconciled with the Old Through Multiple Unfolding Simulations”, *Science* **278**, 1928–1931 (1997).
27. T. Schlick, E. Barth, and M. Mandziuk, “Biomolecular Dynamics at Long Time-steps: Bridging the Timescale Gap Between Simulation and Experimentation”, *Ann. Rev. Biophys. Biomol. Struc.* **26**, 179–220 (1997).
28. E. Barth and T. Schlick, “Overcoming Stability Limitations in Biomolecular Dynamics: I. Combining Force Splitting via Extrapolation with Langevin Dynamics in LN”, *J. Chem. Phys.* **109**, 1617–1632 (1998).
29. M. Gerstein and M. Levitt, “Simulating Water and the Molecules of Life”, *Sci. Amer.* **279**, 101–105 (1998).
30. Y. Duan and P. A. Kollman, “Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution”, *Science* **282**, 740–744 (1998).
31. H. J. C. Berendsen, “A Glimpse of the Holy Grail”, *Science* **282**, 642–643 (1998).
32. L. S. D. Caves, J. D. Evanseck, and M. Karplus, “Locally Accessible Conformations of Proteins: Multiple Molecular Dynamics Simulations of Crambin”, *Prot. Sci.* **7**, 649–666 (1998).
33. W. F. van Gunsteren and A. E. Mark, “Validation of Molecular Dynamics Simulation”, *J. Chem. Phys.* **108**, 6109–6116 (1998).
34. X. Daura, B. Juan, D. Seebach, W. F. Van Gunsteren, and A. Mark, “Reversible Peptide Folding in Solution by Molecular Dynamics Simulation”, *J. Mol. Biol.* **280**, 925–932 (1998).

**1999–2003**

35. A. L. Delcher, S. Kasif, R. D. Fleischmann, J. Peterson, O. White, and S. L. Salzberg, “Alignment of Whole Genomes”, *Nuc. Acids Res.* **27**, 2369–2376 (1999).
36. D. Baker and A. Sali, “Protein Structure Prediction and Structural Genomics”, *Science* **294**, 93–96 (2001).
37. R. Bonneau and D. Baker, “Ab Initio Protein Structure Prediction: Progress and Prospects”, *Annu. Rev. Biophys. Struc.* **30**, 173–189 (2001).
38. J. C. Whisstock and A. M. Lesk, “Prediction of Protein Function from Protein Sequence and Structure”, *Quart. Rev. Biophys.* **36**, 173–189 (2001).
39. H. Kitano, “Systems Biology: A Brief Overview”, *Science* **295**, 1662–1664 (2002).
40. R. M. Karp, “Mathematical Challenges from Genomics and Molecular Biology”, *Notices Amer. Math. Soc.* **49**, 544–553 (2002).
41. M. Karplus and J. A. McCammon, “Molecular Dynamics simulations of Biomolecules”, *Nat. Struc. Biol.* **9**, 307–340 (2003).
42. J. Norberg and L. Nilsson, “Advances in Biomolecular Simulations: Methodology and Applications”, *Quart. Rev. Biophys.* **36**, 257–306 (2003).
43. J. D. Storey and R. Tibshirani, “Statistical Significance for Genomewide Studies”, *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).

44. F. S. Collins, E. D. Green, A. E. Guttmacher, and M. S. Guyer, “A Vision for the Future of Genomics Research”, *Nature* **422**, 835–847 (2003).
45. T. Ideker and D. Lauffenburger, “Building with a Scaffold: Emerging Strategies for High- to Low-Level Cellular Modeling”, *Trends Biotech.* **21**: 255–262 (2003).
46. R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein, “A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data”, *Science* **17**, 449–453 (2003).

2004–

47. J. E. Cohen, “Mathematics is Biology’s Next Microscope, Only Better; Biology is Mathematics’ Next Physics, Only Better”, *PLoS Biology* **2** (e439), 2017–2023 (2004).
48. M. Kellis, N. Patterson, B. Birren, B. Berger, and E. S. Lander, “Methods in Comparative Genomics: Genome Correspondence, Gene Identification and Regulatory Motif Discovery”, *J. Comp. Biol.* **11**, 319–355 (2004).
49. W. C. Winkler, A. Nahvi, A. Roth, J. A. Collins and R. R. Breaker, “Control of Gene Expression by a Natural Metabolite-Responsive Ribozyme”, *Nature* **428**, 281–286 (2004).
50. A. Hastings et al., “Quantitative Bioscience for the 21st Century”, *Bioscience* **55**, 511–517 (2005).
51. J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé and K. Schulten, “Scalable Molecular Dynamics with NAMD”, *J. Comp. Chem.* **26**, 1781–1802 (2005).
52. T. Schlick, “The Critical Collaboration Between Art and Science: Applying *An Experiment on a Bird in an Air Pump* to the Ramifications of Genomics on Society”, *Leonardo* **38** (4), 323–329 (2005).
53. M. Karplus and J. Kuriyan, “Molecular Dynamics and Protein Function”, *Proc. Natl. Acad. Sci. USA* **102**, 6679–6685 (2005).
54. W. F. van Gunsteren et al., “Biomolecular Modeling: Goals, Problems, Perspectives”, *Angew. Chem. Int. Ed.* **45**, 4064–4092 (2006).
55. E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I. K. Moore, J.-P. Z. Wang, and J. Widom, “A Genomic Code for Nucleosome Positioning”, *Nature* **442**, 772–778 (2006).
56. J.-M. Chandonia and S. E. Brenner, “The Impact of Structural Genomics: Expectations and Outcomes”, *Science* **311**, 347–351 (2006).
57. S. A. Adcock and J. A. McCammon, “Molecular dynamics: survey of methods for simulating the activity of proteins”, *Chem. Rev.* **106**: 1589–1615 (2006).
58. M. A. Gerstein et al., “What is a Gene, post ENCODE? History and Updated Definition”, *Genome Research* **17**, 669–681 (2007).
59. E. H. Lee, J. Hsin, M. Sotomayor, G. Comellas, and K. Schulten, “Discovery Through the Computational Microscope”, *Structure* **17**: 1295–1306 (2009).

# Appendix C

## Supplementary Course Texts

1. M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Oxford University Press, New York, NY, 1990.  
[Good advanced reference book for molecular simulations.]
2. A. D. Bates and A. Maxwell. *DNA Topology*, second edition. Oxford University Press, New York, NY, 2005.  
[Beautiful paperback on the higher organizational forms of DNA.]
3. O. M. Becker, A. D. Mackerell Jr., B. Roux, and M. Watanabe, Editors. *Computational Biochemistry and Biophysics*. Marcel Dekker Inc, New York, NY, 2001.  
[Useful resource to computational chemists on various modeling techniques such as molecular mechanics and dynamics, conformational analysis, long-range interactions, implicit solvation, and free energy methods.]
4. J. M. Berg, J. L. Tymoczko, and L. Stryer, *Biochemistry*. W. H. Freeman, New York, NY, latest edition (sixth in May 2006).  
[Wonderful biochemistry textbook, up to date.]
5. H. J. C. Berendsen. *Simulating the Physical World: Hierarchical Modeling from Quantum Mechanics to Fluid Dynamics*. Cambridge University Press, Cambridge, UK, 2007.  
[Introduction to fundamental theory and applications of computer modeling and simulation, from quantum mechanics to fluid dynamics, with practical recommendations and sample Python programs.]
6. V. A. Bloomfield, D. M. Crothers, and I. Tinoco, Jr. *Nucleic Acids: Structures, Properties, and Functions*. University Science Books, Sausalito, CA, 2000.

[Comprehensive account of nucleic acids at an advanced level, with emphasis on biological function and experimental techniques. The first part describes nucleic acid properties on the atomic and molecular levels as deduced by various experimental techniques. The second part presents macromolecular features of nucleic acids in solution (e.g., dynamics behavior, DNA supercoiling). The third part covers noncovalent interactions of nucleic acids and other molecules (ions, drugs, proteins) and higher-order structures due to cellular packing.]

7. C. Branden and J. Tooze. *Introduction to Protein Structure*, second edition. Garland Publishing Inc., New York, NY, 1999. ([www.proteinstructure.com/](http://www.proteinstructure.com/)).
- [A modern and nicely illustrated protein structure textbook dealing with basic structural principles (part 1) and other topics (part 2, broadly grouped under the heading of relationships among structure, function, and engineering). Part 2 includes chapters on transcription regulation, signal transduction, immune regulation, membrane and fibrous proteins, and virus structures.]
8. P. Bratley, B. L. Fox, and L. E. Schrage. *A Guide to Simulation*, second edition. Springer, New York, NY, 1987.  
[Good introduction to Monte Carlo simulations.]
9. C. L. Brooks, III, M. Karplus, and B. M. Pettitt. *Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics*. John Wiley & Sons Inc., New York, NY, 1988.  
[Nice collection on protein simulations.]
10. U. Burkert and N. L. Allinger. *Molecular Mechanics*. American Chemical Society, Washington D.C., 1982.  
[Excellent introduction to molecular mechanics.]
11. C. R. Cantor and P. R. Schimmel. *Biophysical Chemistry*. Parts 1,2,3. W. H. Freeman, San Francisco, CA, 1980.  
[Classic text and reference to many aspects of biophysical chemistry.]
12. D. Chandler. *Introduction to Modern Statistical Mechanics*. Oxford University Press, New York, NY, 1987.  
[Good introduction to Monte Carlo methods.]
13. N. C. Cohen, Editor. *Guidebook on Molecular Modeling in Drug Design*. Academic Press Inc., San Diego, CA, 1996.  
[Reference for molecular modeling as applied to drug design problems, containing contributed chapters by industrial and academic scientists, on problem formulation (database analysis, docking), modeling tools, and medicinal chemistry applications.]
14. C. J. Cramer. *Essentials of Computational Chemistry: Theories and Models*, second edition. John Wiley & Sons Inc., Hoboken, NJ, 2004.  
[Good introduction to quantum mechanical and classical/quantum mechanical modeling and theory.]

15. P. Deuflhard, J. Hermans, B. Leimkuhler, A. E. Mark, S. Reich., R. D. Skeel, Editors. *Computational Molecular Dynamics: Challenges, Methods, Ideas – Proceedings of the 2nd International Symposium on Algorithms for Macromolecular Modelling, Berlin, May 21–24, 1997*, Lecture Notes in Computational Science and Engineering, Vol. 4 (Series Editors M. Griebel, D. E. Keyes, R. M. Nieminen, D. Roose, and T. Schlick), Springer-Verlag, Berlin, 1999.  
[Collection of articles from invited presentations to the May 1997 Berlin workshop on methods for macromolecular modeling. The book contains sections on the following topics: conformational dynamics, thermodynamic modeling, enhanced time-stepping algorithms, quantum-classical simulations, and parallel force field evaluation.]
16. T. E. Creighton, Editor. *Protein Folding*. W. H. Freeman, New York, NY, 1992.  
[Nice collection on general topics regarding protein structure and folding.]
17. D. Eisenberg and D. Crothers. *Physical Chemistry with Applications to the Life Sciences*. The Benjamin/Cummings Publishing Company Inc., Menlo Park, CA, 1979.  
[Wonderful physical chemistry textbook, with useful biomolecular information.]
18. R. Elber, Editor. *Recent Developments in Theoretical Studies of Proteins*. World Scientific Publishing Company Inc., Singapore, 1996.  
[Six chapters on important topics in protein structure and modeling (e.g., reaction path studies, analytical theories of protein folding, ion channels, and protein structure prediction) by the groups of K. Kuczera, R. Elber, J. Straub, D. Thirumalai, R.S. Eisenberg, and P.G. Wolynes.]
19. A. Fersht. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*. W. H. Freeman, New York, NY, 1999.  
[Comprehensive perspective on both enzyme catalysis and protein folding by a pioneer researcher, an updated version of the author's 1995 text on Enzyme Structure and Mechanism; the text reviews protein structure, emphasizing general principles, as well as mentioning recent advances and insights from theoretical approaches.]
20. M. J. Field. *A Practical Introduction to the Simulation of Molecular Systems*, second edition. Cambridge University Press, Cambridge, UK, 2007.  
[Introduction to simulation techniques, including potential energy calculations, conformational/dynamic/thermodynamic sampling, with program modules in Python.]
21. D. Frenkel and B. Smit. *Understanding Molecular Simulations: From Algorithms to Applications*, second edition. Academic Press, San Diego, CA, 2002.  
[Excellent introduction to computer simulation of molecular systems, containing a nice mix of mathematical details and more informal, descriptive text. The focus is on Monte Carlo and molecular dynamics methodologies, including simple algorithms and numerical illustrations. Some important recent methodological advances are also included.]

22. L. M. Giersch and J. King, Editors. *Protein Folding: Deciphering the Second Half of the Genetic Code*. AAAS, Washington D.C., 1990.  
[Interesting and beautifully illustrated collection of articles; best bet: Jane Richardson's origami analogues of protein folding motifs!]
23. H. Gould, J. Tobochnik, and W. Christian. *An Introduction to Computer Simulation Methods: Applications to Physical Systems*, third edition. Addison Wesley, San Francisco, CA, 2007.  
[Good introduction to computer simulations, with a focus on classical mechanics in Part 1 and statistical physics in Part 2. The material is made highly accessible to undergraduates by the inclusion of many simple numerical examples, useful illustrations, and programming segments.]
24. A. Y. Grosberg and A. R. Khokhlov. *Giant Molecules: Here, There, and Everywhere . . .*, second edition. World Scientific Publishing Company, Mountain View, CA, 2009.  
[Lively introduction to polymer physics, with nice illustrations and enticing color plates, aptly fitting a beautiful subject. In the format of a coffee-table book, the authors cover important subjects like the wide range of polymeric subjects, ideal chain models and their properties, Brownian motion, biological polymers, and polymer dynamics. An accompanying CD-ROM animates polymer motion, including reptation and coil collapse.]
25. J. M. Haile. *Molecular Dynamics Simulation: Elementary Methods*. John Wiley & Sons, Inc., New York, NY, 1992.  
[Elementary text on molecular dynamics.]
26. M. H. Kalos and P. A. Whitlock. *Monte Carlo Methods*, second edition. Wiley-VCH, Weinheim, Germany, 2008.  
[Good introduction to Monte Carlo techniques.]
27. A. R. Leach. *Molecular Modelling. Principles and Applications*, second edition. Prentice Hall, Reading, MA, 2001.  
[Broad introduction to many aspects of molecular modeling and computational chemistry techniques, covering basic concepts, quantum and molecular mechanics models, techniques for energy minimization, molecular dynamics, Monte Carlo sampling, protein structure prediction, free energies, solvation, and drug design applications.]
28. B. Leimkuhler, C. Chipot, R. Elber, A. Laaksonen, A. Mark, T. Schlick, C. Schütte, R. Skeel, Editors. *New Algorithms for Macromolecular Simulation – Proceedings of the 4th International Workshop on Algorithms for Macromolecular Modeling, Leicester, UK August 2004*, Lecture Notes in Computational Science and Engineering, Vol. 49 (Series Editors T. J. Barth, M. Griebel, D. E. Keyes, R. M. Nieminen, D. Roose, and T. Schlick), Springer-Verlag, Berlin, 2006.  
[Collection of articles from presentations at the 2004 international workshop on macromolecular modeling, covering biomolecular simulation methods and applications. The contributions are grouped under the following subjects: macromolecular models: from theories to effective algorithms,

- minimization of complex molecular landscapes, dynamical and stochastic-dynamical foundations, free energy computation, fast electrostatics and enhanced solvation models, and quantum-chemical models for macromolecular simulation.]
29. B. Leimkuhler and S. Reich. *Simulating Hamiltonian Dynamics*. Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, UK, 2004.  
[Clear exposition of numerical techniques for simulating the dynamics of conservative systems from a mathematical perspective, including various integration methods, with simple molecular applications and good exercises.]
30. K. B. Lipkowitz and D. B. Boyd, Editors *Reviews in Computational Chemistry*. John Wiley & Sons, Inc., New York, NY, 1990 – present.  
[Nice series of books, with volumes appearing annually with comprehensive reviews and tutorials on many aspects of computational chemistry.]
31. D. Marx and J. Hutter. *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*. Cambridge University Press, Cambridge, UK, 2009.  
[Introduction to various ab initio molecular dynamics techniques, including pseudo-code segments and program design.]
32. J. A. McCammon and S. C. Harvey. *Dynamics of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK, 1987.  
[First book on biomolecular dynamics simulations. Excellent overview of field.]
33. D. A. McQuarrie. *Statistical Mechanics*, second edition. University Science Books, Sausalito, CA, 2000.  
[Good reference text to statistical mechanics.]
34. National Research Council Report. *Mathematical Challenges from Theoretical / Computational Chemistry*. National Academies Press, Washington D.C., 1995. ([www.nap.edu/readingroom/books/mctcc/](http://www.nap.edu/readingroom/books/mctcc/)).  
[Panel report on the opportunities for collaboration, past achievements, and future possibilities between mathematical and chemical scientists.]
35. P. A. Pevzner. *Computational Molecular Biology: An Algorithmic Approach*. MIT Press, Cambridge, MA, 2000.  
[Modern text in computational molecular biology mainly addressed to computer scientists and mathematicians interested in discrete algorithms but accessible to biologists with mathematical grounding. Topics covered include gene hunting, sequencing and mapping, DNA microarray analysis, sequence comparison and alignment, genome evolution, and proteomics.]
36. P. von Ragué Schleyer, Editor-in-Chief, and N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer, III, and P. R. Schreiner, Editors. *Encyclopedia of Computational Chemistry*. John Wiley & Sons Inc., West Sussex, UK, 1998.  
[Comprehensive reference series (five large volumes) written by experts in the field.]

37. D. C. Rapaport. *The Art of Molecular Dynamics Simulation*, second edition. Cambridge University Press, Cambridge, UK, 2004.  
[Elementary text on molecular dynamics focusing on software details.]
38. W. Saenger. *Principles of Nucleic Acid Structure*. Springer Advanced Texts in Chemistry, Springer-Verlag, New York, NY, 1984.  
[Wonderful guide to the richness of DNA structure, with an amazing breadth of topics.]
39. T. Schlick and H. H. Gan, Editors. *Computational Methods for Macromolecules: Challenges and Applications – Proceedings of the 3rd International Workshop on Algorithms for Macromolecular Modeling*, New York, October 12–14, 2000, Lecture Notes in Computational Science and Engineering, Vol. 24 (Series Editors T. J. Barth, M. Griebel, D. E. Keyes, R. M. Nieminen, D. Roose, and T. Schlick), Springer-Verlag, Berlin, 2002.  
[Collection of articles from presentations at the 2000 international workshop on macromolecular modeling, covering biomolecular simulation methods and applications. The contributions are grouped under the following subjects: field perspective (preface), biomolecular dynamics applications, molecular dynamics methods, Monte Carlo methods, other conformational sampling approaches, free energy methods, long-range interactions and fast electrostatics, and statistical approaches to protein structures.]
40. R. R. Sinden. *DNA Structure and Function*. Academic Press, San Diego, CA, 1994.  
[Nice modern textbook on DNA structure.]
41. T. Schwede and M. Peitsch, Editors. *Computational Structural Biology: Methods and Applications*. World Scientific, Hackensack, NJ, 2008.  
[A collection of 27 chapters on structure prediction methods, protein design, drug discovery, new experimental methods, databases, and molecular graphics.]
42. G. E. Schulz and R. H. Schirmer. *Principles of Protein Structure*. Springer Advanced Texts in Chemistry, Springer-Verlag, New York, NY, 1979.  
[Nice advanced text on the rapidly-changing field of protein folding.]
43. J. Šponer and F. Lankáš, Editors. *Computational Studies of RNA and DNA. Challenges and Advances in Computational Chemistry and Physics*. Springer, Dordrecht, The Netherlands, 2006.  
[A collection of 24 chapters on various aspects of modeling DNA, RNA, and complexes by various computational approaches.]
44. W. R. Taylor and A. Aszódi. *Protein Geometry, Classification, Topology and Symmetry: A Computational Analysis of Structure*. Taylor & Francis Group, New York, NY, 2005.  
[Nice introduction to protein structure from a geometric perspective.]
45. W. F. Van Gunsteren and P. K. Weiner, Editors (1989) and W. F. Van Gunsteren, P. K. Weiner, and A. J. Wilkinson, Editors (1993, 1996): *Computer Simulation*

- of Biomolecular Systems: Theoretical and Experimental Applications.* Vol. 1,2,3. Springer-Verlag, New York, NY, 1989, 1993, 1996.  
[Good series on biomolecular simulations, covering both algorithms and applications.]
46. G. A. Voth, Editor. *Coarse-Graining of Condensed Phase and Biomolecular Systems.* Taylor & Francis Group, Boca Raton, FL, 2009.  
[Recent developments in coarse graining modeling of and applications to complex molecular systems, with examples for peptides and proteins at various levels of resolution.]
47. D. Wales. *Energy Landscapes.* Cambridge University Press, Cambridge, UK, 2004.  
[Introduction to energy landscape theory, with applications to clusters, biomolecules and glasses.]
48. A. Warshel. *Computer Simulation of Chemical Reactions in Enzymes and Solutions.* John Wiley & Sons, New York, NY, 1991.  
[Excellent reference text in the field of computational biological chemistry, particularly studies of enzymatic reactions. From basic principles of chemical bonding and enzyme mechanisms, the authors describe the governing force fields for molecular simulations, associated algorithms, various approaches to modeling chemical reactions, and examples of different mechanisms.]
49. Ahmed H. Zewail, Editor. *Physical Biology. From Atoms to Medicine.* Imperial College Press (2008).  
[A collection of 20 chapters that provide a beautiful broad overview into the challenges in our 21st-century biology, including instrumentation and computational advances, protein folding, computer-aided drug discovery, and sampling approaches for biomolecules.]

# Appendix D

## Homework Assignments

*Please Note:* (1) Files mentioned throughout the homeworks can be obtained from the course site. (2) Insight modeling commands may have changed since the time of this writing and may need updating by instructor and/or students.

## Assignment 1: Sequence and Structural Databases, Molecular Modeling Perspective

1. **Molecular Modeling Resources.** Search the web for resources in *molecular modeling*. Look, in particular, for tutorials and instructional material. A good place to start is the NIH site: [cmm.info.nih.gov/modeling/](http://cmm.info.nih.gov/modeling/), which also provides links to many “Related Web Sites”. (Make use of bookmark-type browser utilities to keep useful web sites handy for future use).  
 Submit information from two of the most valuable sites you discover (print-out) along with a description of how you found the material and what you found most useful.
2. **Sequence and Structure Information Databases.** Search the web for protein (amino-acid) sequence and structure databases. Examples of sequence databases are: PIR, Swiss-Prot, GenPept, and NRPR.<sup>1</sup>
  - (a) Plot the amount of available *sequence* database as a function of year, going back as far as possible, to the 1970s. Plot the information on both a regular scale and on a logarithm scale.
  - (b) Similarly, plot the amount of *structural* information available as a function of year, on both a regular and a logarithm scale.
  - (c) Plot the *sequence and structure* information on the *same plot* in both standard and logarithm views. What can you say about the rate of growth

---

<sup>1</sup>PIR was established in 1984 by the National Biomedical Research Foundation (NBRF) and is a good starting point for protein database searching. PIR is somewhat more comprehensive than SwissProt but smaller and better annotated than GenPept (which also includes many hypothetical sequences of unknown function. Since 1999, the NBRF has added a new section to PIR called PATCHX which contains a non-redundant set of all other protein sequences not included in PIR (from other databases), with subsequences removed. Thus, PIR supplemented by PATCHX provides a comprehensive collection of protein sequence data in the public domain. Any search through PIR will automatically include PATCHX. The SwissProt database is useful for searches limited to well annotated sequences, and GenPept is useful for searching all possible sequences, including those that have unknown functions.

The best known structural databases are the Protein Data Bank (PDB) and the Nucleic Acid Database (NDB).

The PDB, managed from 1971 through June 1999 by the Brookhaven National Laboratory, is now operated by the Research Collaboratory for Structural Bioinformatics (RCSB) ([home.rcsb.org/](http://home.rcsb.org/)), a consortium among Rutgers University, the University of California at San Diego, and the National Institute of Standards and Technology. The RCSB has introduced new features, such as a web-based tool for data deposition, fast data processing systems, and new search engines (text-based and data-based), both with extensive reporting capabilities.

The NDB, pioneered in 1992 by the Rutgers RCSB leader Helen Berman, similarly assembles and distributes structural information about nucleic acids ([ndbserver.rutgers.edu/](http://ndbserver.rutgers.edu/)). NDB contains an atlas, an archive, and a sophisticated search engine to access the data.

of sequence and structural information? Discuss these finding in relation to the Human Genome Project.

**3. Early Molecular Modeling Literature and Current Progress.** Read two articles dealing with early molecular modeling work:

- B. J. Alder and T. E. Wainwright, “Studies in Molecular Dynamics. I. General Method”, *J. Chem. Phys.* **31**, 459–466 (1959).
- G. Némethy and H. A. Scheraga, “Theoretical Determination of Sterically Allowed Conformations of a Polypeptide Chain by a Computer Method”, *Biopolymers* **3**, 155–184 (1965).
- A. Rahman and F. H. Stillinger, “Molecular Dynamics Study of Liquid Water”, *J. Chem. Phys.* **55**, 3336–3359 (1971).

Do not worry about not understanding the technical details for now. Then also read later articles describing current progress in the field and another discussing issues in validating simulation results:

- J. A. Board, Jr., L. V. Kalé, K. Schulten, R. D. Skeel, and T. Schlick, “Modeling Biomolecules: Larger Scales, Longer Durations”, *IEEE Comp. Sci. Eng.* **1**, 19–30 (1994).
- W. F. van Gunsteren and A. E. Mark, “Validation of Molecular Dynamics Simulation”, *J. Chem. Phys.* **108**, 6109–6116 (1998).
- F. S. Collins, E. D. Green, A. E. Guttmacher, and M. S. Guyer, “A Vision for the Future of Genomics Research”, *Nature* **422**, 835–847 (2003).
- J. Norberg and L. Nilsson, “Advances in Biomolecular Simulations: Methodology and Applications”, *Quart. Rev. Biophys.* **36**, 257–306 (2003).
- J. E. Cohen, “Mathematics is Biology’s Next Microscope, Only Better; Biology is Mathematics’ Next Physics, Only Better”, *PLoS Biology* **2** (e439), 2017–2023 (2004).
- T. Schlick, “The Critical Collaboration Between Art and Science: Applying *An Experiment on a Bird in an Air Pump* to the Ramifications of Genomics on Society”, *Leonardo* **38** (4), 323–329 (2005).
- W. F. van Gunsteren et al., “Biomolecular Modeling: Goals, Problems, Perspectives”, *Angew. Chem. Int. Ed.* **45**, 4064–4092 (2006).
- M. A. Gerstein et al., “What is a Gene, post ENCODE? History and Updated Definition”, *Genome Research* **17**, 669–681 (2007).

First describe (in about two pages) the difficulties that Alder and Wainwright enumerate in 1959 regarding molecular dynamics simulations. Then discuss the issues that are still serious limiting factors today. Have any of the original limitations been resolved or are likely to be resolved in the near future?

## Assignment 2: Introduction to the Insight II Modeling Package and the PDB File Structure and Retrieval

At this writing, Insight II is available as part of the Discovery Studio from Accelrys ([accelrys.com/products/insight/](http://accelrys.com/products/insight/)). Other molecular modeling packages can be used instead as available.

1. **Introduction to Insight II.** If you are not familiar with Insight II you might start by reading the short manual and the tutorial from the web site.

The manual also contains a short list of basic UNIX commands and a description of simple text editors. Some of the displays on our computers are slightly different from those described in the tutorial but the differences are not critical. For example, in our version the help window automatically follows the tasks invoked from the pulldown menus, and the dial boxes are located on the left, rather than the right, side of the screen. For more information, refer to the *Insight II User Guide*.

*Note: after opening Insight II, ignore the message about detection of the unlicensed mode. Our site does not have license for the Sketch module. Repeated warning messages might occur during the “build in” Insight II Pilot tutorial.*

2. **Running Insight II.** Before starting Insight II, you must define the set of environmental variables by running two commands:

```
source /local/msi/cshrc
```

Alternately, you can insert these lines into your `.cshrc` file and they will be executed by the system automatically every time you log on. In that case, you will only need to specify the command

```
source .cshrc
```

after editing your `.cshrc` file the first time. When this insertion is complete, you can run Insight II by specifying the command

```
insightII
```

at the UNIX prompt. Remember, UNIX commands are case sensitive.

*Note: NYU staff may have already inserted these commands in your `.cshrc` file.*

3. **PDB Structures.** Check the PDB web site for information about the type of stored 3D structures (i.e., proteins, DNA, RNA, DNA/protein complexes, etc.) and the amount in each group. Report your findings.

4. **Retrieval of PDB Files.** Using the web PDB browser, find coordinate files for the crystal forms II and III of Bovine Pancreatic Trypsin Inhibitor (BPTI) among the many BPTI entries. From the PDB Home page, go to **Searching and Browsing PDB** and then choose **PDB’s web Browser**. You can search by specifying the abbreviation “BPTI” in the Compound window. When the search is completed, records containing BPTI (including its mutated forms) will be displayed at the bottom of the page. Note

their ID codes (a number followed by three letters). The two middle letters in this code constitute the name of the subdirectory where the file of interest resides. For example, the subdirectory name for ID code 1abb is ab and the file name pdb1abb.ent. Ftp to `ftp.rcsb.org` and login as anonymous (the password instructions will be on the screen). Change directory to `pub/pdb/data/structures/divided/pdb/ab` and get the desired file with the command

```
get pdb1abb.ent.Z
```

5. **Format of PDB Files.** Read the text in the top of both PDB files and describe the differences in the number of recorded residues, structure resolution, number of solvent molecules, experimental conditions, etc.

Attach to the assignment sheet a printout of a few lines, starting with the word ATOM, from a PDB file; mark with arrows and describe the content of each specific format field. (The PDB browser contains information about the format; see also the original paper on PDB files: *J. Mol. Biol.* **112**, 535–542, 1977).

6. **Displaying a Protein in Insight II.** Retrieve from PDB the file of mutated form of BPTI (ID = 7pti), and start Insight II. From the top menu bar select **Molecule**<sup>2</sup> and then choose **Get**. Press the **PDB** button in **Get File Type** and the **User** button specify the directory. Select the file of the 7pti structure. (Do not press **Heteroatom** button). Execute. The structure of BPTI should now be on the screen. Use **Object** / **DepthCue** and then **Transform** / **Clip** for viewing the protein.

Change the display (**Molecule** / **Display**) to **Backbone** only to speed up the response time. Label the mutated residues (**Molecule** / **Label**).

Repeat the operations described above to display the structure of BPTI's crystal form II (keep both structures on the screen).

Now, overlay both structures by selecting **Overlay** from **Transform**. In few paragraphs describe the structural differences between both forms of BPTI.

7. **Ramachandran Plots.** To create a file listing with all the dihedral angles  $\phi$  and  $\psi$  for a protein, you can use **Protein** / **List** from the **Biopolymer** module. Select **Dihedrals** and the protein (any of the structures). Press **List\_to\_file** button, specify the file name, and execute. From the recorded data create a scatter plot (phase diagram), so that each point corresponds to one  $(\phi, \psi)$  value.

---

<sup>2</sup>The following notation will be used throughout the homework assignments:

- **Pulldown** corresponds to a menu bar pulldown.
- **Command** corresponds to a command from the pulldown menu.
- **Option** corresponds to an option in the dialog box of a given command.

### Summary of Items to Hand in:

- (a) Data with the amount of PDB 3D structures for each category of biomolecules.
- (b) Description of the differences in the informational part of the PDB files for form II and form III of BPTI.
- (c) Explanation of the PDB format for storing atomic coordinates.
- (d) Description of the structural differences between BPTI (form II) and the mutated form (7pti).
- (e) The table with dihedral angles  $\phi$  and  $\psi$  listed for BPTI.
- (f) Scatter plot with points  $(\psi, \phi)$  for each residue of BPTI.

### Background Reading from Coursepack (Appendix B)

- M. Levitt and A. Warshel, “Computer Simulation of Protein Folding”, *Nature* **253**, 694–698 (1975).
- M. Karplus and J. A. McCammon, “The Dynamics of Proteins”, *Sci. Amer.* **254**, 42–51 (1986).
- Y. Duan and P. A. Kollman, “Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution”, *Science* **282**, 740–744 (1998).
- H. J. C. Berendsen, “A Glimpse of the Holy Grail”, *Science* **282**, 642–643 (1998).
- X. Daura, B. Juan, D. Seebach, W. F. Van Gunsteren, and A. Mark, “Reversible Peptide Folding in Solution by Molecular Dynamics Simulation”, *J. Mol. Biol.* **280**, 925–932 (1998).
- R. Bonneau and D. Baker, “Ab Initio Protein Structure Prediction: Progress and Prospects”, *Annu. Rev. Biophys. Struc.* **30**, 173–189 (2001).

## Assignment 3: Construction and Analysis of the Pentapeptide Met-enkephalin with the Insight II Program

1. **Building a Pentapeptide.** To build the molecule met-enkephalin, whose amino acid sequence is **Tyr-Gly-Gly-Phe-Met**, invoke the **Biopolymer** module. From **Residue** select **Append**.

Specify the molecule's name and choose **Extended** to specify the structural motif of the backbone. (At this stage, do not worry whether the structure is correct). Select the first residue, **Tyr**. Then add each of the remaining four residues in turn.

You can *center* the molecule on your screen by clicking on it with the center mouse button and dragging it to the desired position. You can also *rotate* the molecule by pressing the right mouse button and dragging the molecule. By pressing both (center and right) mouse buttons and dragging the molecule you can change its position along the *z*-axis, perpendicular to the screen. You can rotate the molecule around the *z*-axis by dragging the mouse while both left and right mouse buttons are pressed. To change the representation style of the molecule, select **Molecule / Render** and choose any of the options. Note how the speed of executing commands (e.g., translation, rotation) is affected by the representation display.

Now you must amend the ends of the structure you built. Switch to **Protein** and choose **Cap**. Change both the N-terminus and C-terminus moieties to the zwitterionic form ( $\text{NH}_3^+$  and  $\text{COO}^-$ ) to get a proper oligopeptide.

2. **Measurements of Met-enkephalin's Structural Parameters.** Generate a table listing all the dihedral angles in your met-enkephalin molecule by using **Protein / List** from the **Biopolymer** module. Select the appropriate command (**Dihedrals**) and direct the data to a file (**List\_to\_file** button on). This file can be viewed and edited later.

To measure individual bond lengths or distances, bond angles, dihedral angles, etc., use **Measure**. Select the atoms for the measurement by clicking on them with the left mouse button.

3. **Rotameric Structures.** By a *rotamer*, Insight II refers to a different conformational arrangement of a side chain of a given amino acid. The rotameric structures identified in Insight II are correlated with the  $\phi$  and  $\psi$  angles for a given residue (i.e., are sterically compatible).

Select **Manual\_Rotamer** from **Residue** of the **Biopolymer** module. Press the **Evaluate\_Energy** button. Energy for a given rotameric structure will be printed in the information window at the bottom of the screen (you can scroll by pressing on arrows to its right). Keep the **Nonbond Cutoff** value — an option which is displayed on the screen once **Evaluate\_Energy** is chosen — at the default value of 8.0 Å.

Select a residue by clicking on it and sweep through all the rotamers of that residue (while holding other rotamer conformations fixed). **Next** will trigger the execution of the command. In a table, report the rotamer structures for each residue by specifying the dihedral angles  $\chi_1$ ,  $\chi_2$ ,  $\chi_3, \dots$  (**Protein** / **List**), along with associated energies.

Now assemble the pentapeptide structure with the lowest rotamer energy and save it. Thus far we have ignored a global optimization of the structure and have only built it up from low-energy conformations of its components. *We will return to optimization later in the course, after studying minimization techniques.*

4. **Main-chain Structure.** Choose **Protein** / **Secondary** from the **Bio-polymer** module. Change the main-chain configuration by selecting different motifs (**Alpha\_R\_Helix**, **Alpha\_L\_Helix**, **3–10\_Helix**, etc.). For each motif:

- (a) prepare a table with the dihedral angles  $\phi$  and  $\psi$  for each of the residues (**Protein** / **List**),
- (b) list all the hydrogen bonds present in the structure.

**Measure** / **HBonds** and **Molecule** / **Label** will be helpful here.

5. **Torsional Rotation.** Bring back your Met-enkephalin's backbone structure in the extended form. You can use the structure saved in part 3 of this assignment (**File** / **Restore\_Folder**). Change the  $\psi$  dihedral angle on the second residue, **Gly**, to  $60^\circ$  and the  $\phi$  dihedral angle on the forth residue, **Phe**, to  $-60^\circ$ .

Torsional motion around a chosen dihedral angle can be performed with **Transform** / **Torsion** command or by pressing the **Torsion** button on the left of the **Insight II** screen.

First, click with the left mouse button on the bond which constitutes the axis of rotation, then press the **Torsion** button. A little cone, defining the direction of the torsion, will pop up on the screen at one of the bond ends. Now you can change the torsion angle by horizontally dragging the mouse with the center button pressed in. To exit the torsion mode press the **Torsion** button again (the cone will disappear).

Calculate the distance between the N-terminus (N atom) and C-terminus (C atom) atoms. Use the **Measure** / **Distance** command. Keep the **Monitor** button on, and select **Monitor Mode/Add**. **Atom 1** and **Atom 2** can be selected by clicking on them with the left mouse button.

Print a picture of your molecule (keep the distance between the N-terminus and C-terminus atoms on the screen). To save ink, the background color should be changed to white every time you print a color or black/white picture. To do so, go to **Session** / **Environment**, press the **Background** button and change the color to white. Now go to **File** and choose **Export\_Plot**.

Select `postscript`, `Gray_Scale` and optionally `Ball_and_Stick`. Save the file as `postscript` by using `Save_Device_File` (the file will have the “`.ps`” extension). This file can be printed on any `postscript` printer. Hand in your printout as part of the homework.

### Summary of Items to Hand In:

- (a) The table with the dihedral angles for met-enkephalin.
- (b) The table with the dihedral angles and energies for the rotameric structures of met-enkephalin.
- (c) The table with the  $\{\phi, \psi\}$  dihedral angles for each different backbone motif of met-enkephalin.
- (d) The listing of the hydrogen bonds for each of the backbone motif for met-enkephalin.
- (e) Printout of the met-enkephalin structure with the end-to-end link marked.

### Background Reading from Coursepack

- G. Némethy and H. A. Scheraga, “Theoretical Determination of Sterically Allowed Conformations of a Polypeptide Chain by a Computer Method”, *Biopolymers* **3**, 155–184 (1965).
- P. Y. Chou and G. D. Fasman, “Prediction of Protein Conformation”, *Biochemistry* **13**, 222–245 (1974).
- M. Levitt and C. Chothia, “Structural Patterns in Globular Proteins”, *Nature* **261**, 552–558 (1976).

## Assignment 4: Creating Ramachandran Plots from Known Protein Structures and the NDB

1. **Ramachandran Plots.** Our goal is to generate Ramachandran plots for a particular amino acid residue or a group of residues. We have assembled a database of 50 proteins based on the article: M.A. Williams, J.M. Goodfellow, and J.M. Thornton, “Buried Water and Internal Cavities”, *Protein Science* **3**, 1224 (1994). These files can be found in the PDB directory of Insight II prepared for our course.

Each of you must generate two Ramachandran plots. Check the chart below for your particular assignment (on the basis of your last name).

First letter of your last name	Subgroup 1	Subgroup 2
A–N	<b>Ala, Val, Leu, Ile</b>	<b>Gly</b>
O–R	<b>Asp, Asn, Glu, Gln</b>	<b>Pro</b>
S–V	<b>Lys, Arg, His</b>	<b>Ser, Thr</b>
W–Z	<b>Trp, Tyr, Phe</b>	<b>Cys, Met</b>

Each plot should have the data points for the  $(\phi, \psi)$  dihedral angles, corresponding to the assigned group of residues from all the proteins in the database. To find the values of the  $\phi$  and  $\psi$  dihedral angles in a protein you can use **Protein / List** command from the **Biopolymer** module. Record these angles to a file. You can use the Fortran code posted on the website (`aa_select.f`) for searching the **Protein / List** output files (called PDA files) for the dihedral angles of selected residues. Alternatively, you can write a suitable program in a different language.

If you use the code from the website, you will need to edit it to replace all occurrences of the names ALA, VAL, etc. by the abbreviations of the residues you are searching for. These abbreviations must be capitalized. Compile the code and execute it with each PDA file as input. The output file should contain only two numbers per line, the  $\phi$  and  $\psi$  angles for the specified residues. Check the numbers for correctness by comparing a few lines from this output with the numbers in the PDA file.

*Note: For plotting, you must use Insight II so that all plots are uniform in size. We will overlay them in class! Follow the instructions below.*

Prepare a file with your data points collected from all the proteins for each of the assigned groups of residues. These files should have the extension `tbl` (`filename.tbl`). In addition, you must format these files for **Insight II** by inserting the 12 lines as indicated:

---

```

#
TITLE: Phi (deg)
MEASUREMENT TYPE: Ang
UNITS OF MEASUREMENT: Deg
FUNCTION: dihedral
#
TITLE: Psi (deg)
MEASUREMENT TYPE: Ang
UNITS OF MEASUREMENT: Deg
FUNCTION: dihedral
#
#
-34.5      144.9          This is the first line of your numeric data.
:

```

---

Note that at the top of the file there is space for your own comments and you can use as many lines as you need. **Insight II** will start reading the data from the line without the character # on the first column. That first line should be as indicated above.

If you have done everything correctly, you are ready for plotting. Press the **Graph** button on the left of the screen and select **Get**. Specify the data file name, dihedral1 as **X\_Function** and dihedral2 as **Y\_Function**. Keep the **New\_Graph** on and execute. The zigzag appearance of the plot now requires fixing. Move the graph box near the bottom-left corner of the screen. First, connect to the object (your plot) with the command **Transform / Connect**. Second, move the plot by clicking on it and dragging to the desired position. The scatter plot will be produced in the next step. Select **Point (only!)** from the **Graph / Modify\_Display** dialog box. Choose **Star** as the **Point Symbol**, scale it ten times and execute. Use **Graph / Threshold** to change the minimum value to  $-180.0$  and the maximum value to  $180.0$  for both, **X Graph Axis** and **Y Graph Axis**. Scale each axis 4 times with **Graph / Scale\_Axis** command. Change the color of any white elements in your plot with **Graph / Color** command. Change the background color to white. Print your plot (see instructions below) and repeat the procedure for the second data file.

Prepare a transparency for each of the two plots and bring to class. (Hand in the printed version of the plots with your homework.) Also e-mail the  $\{\phi, \psi\}$  files you generated, sending each file separately and specifying your name and the group of residues in the **Subject** line.

2. **The NDB.** The next part of the homework will acquaint you to the Nucleic Acid Database, NDB, on which we will have a guest lecturer.

First, explore the database to discover what is available, and look through the newsletter archives for current update information. Then, describe the structures available in the database and the numbers in each class (e.g., B-DNA, ribozymes). Explore the different features of NDB. There are many exciting structures and utilities.

3. **Sugar Conformations for Canonical A, B, and Z-DNA.** To appreciate the different features of canonical A, B, and Z-DNA forms, choose through the NDB entries one unmodified form in each DNA class with the largest number of residues possible. You can view the structure within NDB, or by porting the PDB file to Insight II and using the **Molecule** / Get from the **Viewer** module (though the latter may be more difficult).

Learn how to use the Report facility in the NDB Query Interface to generate a list of all the sugar pseudorotation angles ( $P$ ) for each deoxyribose in each of the three structures you have chosen for your study. (Remember that there are two sugars per base pair). Print a table for each structure.

Now, on one figure for all the three structures, plot  $P$  versus the residue number. Exclude the two terminal base pairs of each structure for plotting purposes. You may need to connect the points corresponding to each structure for clarity.

Label the pattern clearly to indicate the A, B, and Z-DNA data. Hand in this plot, with a description of the patterns you had identified for the sugar conformation for the three canonical DNA forms, indicating the specific structures you have chosen.

### Summary of Items to Hand In:

- (a) Ramachandran plots for the two subgroups of amino acid residues assigned to you.
- (b) Data on the structure class types and the amount of structures in each class available in NDB.
- (c) The figure with  $P$  versus residue number for the A, B, and Z-DNA forms, with complete discussion.

### Background Reading from Coursepack

- B. Cipra, “Computer Science Discovers DNA”, in *What’s Happening in the Mathematical Sciences*, pp. 26–37 (P. Zorn, Ed.), American Mathematical Society, Colonial Printing, Cranston, RI (1996).
- M. Gerstein and M. Levitt, “Simulating Water and the Molecules of Life”, *Sci. Amer.* **279**, 101–105 (1998).

- J. E. Cohen, “Mathematics is Biology’s Next Microscope, Only Better; Biology is Mathematics’ Next Physics, Only Better”, *PLoS Biology* **2** (e439), 2017–2023 (2004).
- A. Hastings et al., “Quantitative Bioscience for the 21st Century”, *Bioscience* **55**, 511–517 (2005).

---

### Printing Instructions

To save ink, change the background color from black to white at each printing of a color or black/white image. Go to **Session** / **Environment**, press the **Background** button and change the color to white.

Proceed to **File**, choose **Export\_Plot**, **postscript**, **Gray\_Scale** and optionally **Ball\_and\_Stick**. Use **Save\_Device\_File** to save a postscript file (the file will have the “.ps” extension). This file can be printed on any postscript printer you can access. Hand in your printout as part of the homework.

---

**Fortran program for selecting Ala, Ile, Val, and Leu data**  
(see website)

## Assignment 5: Analysis of Protein/DNA Complexes with Insight and NDB: Canonical vs. Protein/Bound DNA, and DNA/Protein Interactions

In this assignment, we will study three nucleic acids structures, two of which have been crystallized with regulatory proteins: a nucleic acid dodecamer, a DNA oligomer bound to a prokaryotic protein in the helix-turn-helix (HTH) motif, and a DNA oligomer (known as the TATA-box sequence) bound to a eukaryotic transcription factor.

NDB ID code	Structure
• <b>BDL078</b>	DNA dodecamer
• <b>PDR010</b>	DNA (20 bp) bound to bacteriophage λ cI repressor
• <b>PDT034</b>	DNA (16 bp) bound to human TATA-Box Binding protein

1. **Structure Downloading.** Download each structure from the NDB, already explored in Assignment 4 ([ndbserver.rutgers.edu:80/](http://ndbserver.rutgers.edu:80/)). Use the **Archives**, **Atlas** or **Search** entry points to the NDB; all are useful. In particular, the **Atlas** allows you to quickly view the structures. The **Search** entry point will be utilized later in this assignment.

Load each structure into **Insight II**. Separate the two complexes into DNA and Protein objects using the **Modify / Unmerge** command from either the **Biopolymer** or **Builder** modules. Both the DNA and Protein objects will be used later in this assignment. Use caution with the wildcard character \* in selecting multiple atoms/residues/nucleotides/ proteins in **Insight II**. A quick test which you can use to test which object you've created with the **Modify / Unmerge** is to blank or blink the object with the **Object / Blank** or **Object / Blink** commands.

In order to create the B-DNA models for specified nucleic acid sequences (see below) and manipulate the downloaded structures (as in the **Unmerge** command), you will need to understand certain general parameters related to the structure as used in both the PDB file and **Insight**. These parameters refer to the DNA sequence, strand and residue labels, and so on. Explore the text in the downloaded files as well as in the structural displays. Information about relevant formatted lines in the coordinate files (such as SEQRES and ATOM) is available from the PDB web site (see Assignment 2).

2. **Canonical vs. Protein/Bound DNA Analysis.** Create an idealized B-DNA structure using the **Nucleotide / Append** command in the **Biopolymer** module corresponding to each of the three nucleic acid sequences in the above structures. The **Nucleotide / Append** command both creates new nucleic acid molecules and appends to existing molecules; when it is

creating a new molecule, the Append Point is None.<sup>3</sup> You will need to use the **Nucleic\_Acid** / **Cap** pulldown menu to replace the phosphate group with a hydroxyl group at the 5' ends of each strand.

**Nucleotide** / **Append** will create your B-DNA model, defining each strand as a separate object. Your downloaded DNAs, however, are each composed of a single object in **Insight II**. To properly superimpose the two structures (the task in the next part), you will need to **Modify** / **Merge** the two strands as one object.

Superimpose each B-DNA model upon its respective crystal structure from the NDB using the **Transform** / **Superimpose** pulldown menu. If your DNAs are each composed of a single object in **Insight II**, you will need to **Modify** / **Merge** the two strands as one object to properly superimpose the two structures. Use the **Heavy** option to avoid superpositioning of hydrogen atoms. After selecting the B-DNA model and the crystal structure in the selection boxes, you will need to click the **End Definition** box to **Execute the Superimpose command**. The root-mean-square deviation (RMSD) value will be printed at the bottom of your screen.

- Record the RMSD values relative to idealized B-DNA for each superimposed model/structure by repeating this procedure for each crystal structure.
- Use the **Search** entry point to the NDB to extract the following parameters for each base pair of the three structures:  $P$  (pseudorotation sugar pucker); the dihedral angles  $\chi$ ,  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ,  $\epsilon$ ,  $\zeta$ ; and the helical parameters twist, tilt, roll, and propeller twist:  $\Omega$ ,  $\tau$ ,  $\rho$ , and  $\omega$ , respectively.

For each conformational variable (excluding those from the end residues), calculate the average ( $\mu$ ) and standard deviation ( $\sigma$ ) from the data per structure. Prepare a table in the following form:

Now discuss your results in terms of the differences noted between the protein-bound DNA and canonical B-DNA (which the BDL078 structure represents):

- Which structure is most deformed from B-DNA? Which parameters display the largest changes from B-DNA (consider both  $\mu$  and  $\sigma$ )? Based on these parameters, what is similar in the way the two complexes deform their recognition sites away from B-DNA (look for similarities in columns 3 and 4 above which are different from column 2)?

---

<sup>3</sup>The **PDR010** structure has an overhanging base at each end, that is a base without a Watson-Crick partner on the other strand. (This procedure promotes crystal formation.) The recommended procedure for creating the overhanging base is to use **Nucleotide** / **Append** to create a 21-base-pair duplex of the correct sequence and then use **Nucleotide** / **Delete** command to delete one base from each strand prior to using **Nucleic\_Acid** / **Cap**.

	<b>BDL078</b> $\mu$ $\sigma$	<b>PDR010</b> $\mu$ $\sigma$	<b>PDT034</b> $\mu$ $\sigma$
$P$			
$\chi$			
$\alpha$			
$\beta$			
$\gamma$			
$\delta$			
$\epsilon$			
$\zeta$			
$\Omega$			
$\tau$			
$\rho$			
$\omega$			
RMSD			

- (d) Are any of the changes observed localized to particular regions in the DNA? (Consider properties with  $\mu$  values similar to B-DNA but large associated  $\sigma$  values). Plot one of these parameters as a function of position (base pair) along the DNA.
- (e) Generate a *side-by-side* picture of the three DNA structures. A recommended utility for this is the **File** / **Export\_Plot** facility.
3. **Analysis of Interface Between Proteins and DNA.** Next, we will examine some of the interactions formed at the interface between the regulatory proteins and their DNA binding sites. Load the PDB files of each DNA/protein complex in turn and unmerge the DNA part (but leave the DNA and protein together in space).

The main tool used here is the **Subset** / **Interface** pulldown in the central **Viewer** module. This menu allows subsets to be defined in one molecule that satisfy a certain spatial relationship with respect to the other molecule. For example, we would like to use this menu to define subsets of atoms in the protein that are near functional groups of the DNA. A contour level of 3.5 Å is useful in this menu for defining interactions between non-hydrogen atoms, since it roughly corresponds to distances for strong interactions.

Open the **Subset** / **Interface** pulldown menu. You can define the **Subset Name** as you please. You can define the **Center of Subset** to be a specific functional group in DNA. For example, DNA:T:C5M refers to atom C5M of thymine's methyl group in the DNA. Define the **Search\_Domain** to be the protein. The **Radius of Subset** should be set to the value 3.5.

For your reference, these are names of some DNA atoms:

- Phosphate groups: Atoms P, O1P, O2P
- Thymine methyl groups: Atom C5M

- Adenine amino groups: Atom N6
- Pyrimidine carbonyls: Atom O2
- Purine amines: Atom N3

- (a) Save listings of these subsets into output files. **Subset** / **List** is recommended for this task.

[Do not be alarmed if you get the error message “Invalid Comparison Object”; it simply means that the comparison could not be performed since no member of the set fulfilled the criteria. If an attempt is made to analyze all atoms **B** that are 3.5 Å from protein **A**, but all atoms of protein **A** are more than 3.51 Å from **B**, this error will occur.]

Combine the analyses of the two complexes (if you can) and construct histograms of the residue types in each subset. That is, from the listings of all residues within 3.5 Å of the atom groups above, count the number of times each residue appears (e.g., Methionine appears 60 times, Glutamate 3) and generate histogram plots as illustrated below; use the one-letter mnemonic for the amino acids. (You may also want to count the frequencies of residues grouped by type, like polar, hydrophobic, charged, etc.).

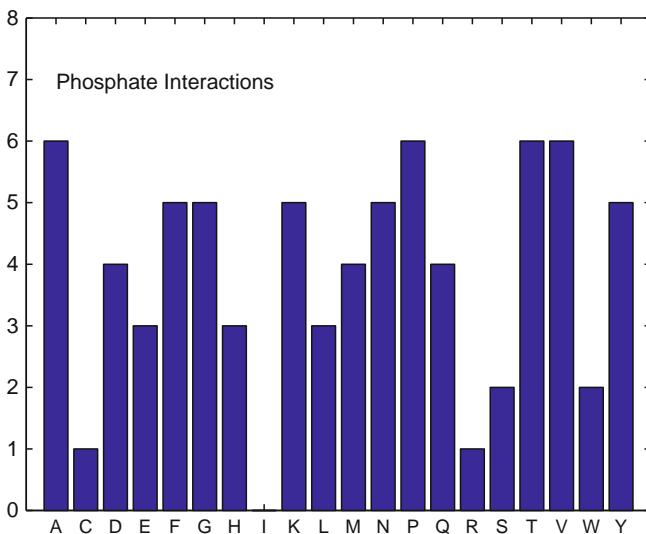


Figure D.1. Sample histogram for protein/DNA interaction analysis.

- (b) Do you observe common patterns in the two complexes? Are certain amino acids likely to be found interacting with a particular functional group? What types of interactions are being formed between these nucleic acid functional groups and the regulatory protein (e.g. attribute to each of the nucleic acid groups above a type of interaction such as

hydrophobic, hydrogen bonding, electrostatic, intercalation/insertion motif, etc.)?

- (c) Is there anything unusual about the subsets formed between the proteins and the O2 carbonyl/N3 amine atoms?
- (d) Is there any relation between the interactions observed in these subsets and the deviations from canonical B-DNA structure observed above (i.e. how do the interactions you observe explain any of the parameter variances you diagnosed)?

*Note: A trick to identify atoms/residues is via [Molecule] / [Color] for assigning a color to an atom/residue. Other labeling tools such as [Molecule] / [Render] and [Molecule] / [Label] can similarly be used.*

- (e) **Bonus Question:**<sup>4</sup> **BDL078 Homologues.** We have used the BDL078 structure as an example of B-DNA in the analysis above. However, sequence-dependent variations in local structure are also important. Therefore, a more sensitive analysis of free versus protein/bound DNA employs the same nucleotide sequence, both with and without bound proteins. There are few cases, however, in which high-resolution DNA structures are available in both the free and protein-bound states. Such analyses are illuminating and show how intrinsic DNA preferences are amplified in the DNA/protein complexes. See recent reports regarding the complex between DNA and the bovine papillomavirus E2 protein in D.M. Crothers (*Proc. Natl. Acad. Sci.* **95**:15163–15165, 1998) and H. Rozenberg *et al.* (*Proc. Natl. Acad. Sci.* **95**:15194–15199, 1998).

Such analyses have not been done with our BDL078 sequence, but there are protein-DNA complexes in the NDB which are closely related to BDL078. This close relationship means that: (i) the related sequence has many similar or closely related residues to BDL078 (e.g., GGGAAAATTT is closely related to GGCATAACTT), and (ii) the protein would bind to BDL078 and this related sequence.

**Determine which protein/DNA complexes these are, and briefly describe what these complexed proteins do.**

### Background Reading from Coursepack

- K. B. Lipkowitz, “Abuses of Molecular Mechanics. Pitfalls to Avoid”, *J. Chem. Educ.* **72**, 1070–1075 (1995).
- S. Lifson, “Potential Energy Functions for Structural Molecular Biology”, in *Methods in Structural Molecular Biology*, pp. 359–385, D. B. Davies, W. Saenger, and S. S. Danyluk, Eds., Plenum Press, London (1981).

---

<sup>4</sup>The correct solution will allow you to drop lowest homework grade in any assignment.

## Assignment 6: MIDTERM: Homology Contest!

### Exploring Sequence/Structure/Function Relationships (& Related Tools/Databases like SCOP, IMAGE, BLAST, NDB, PDB)

With the rapidly growing information on genomic sequences, *comparative modeling* — structure prediction based on sequence similarity — is becoming increasingly valuable. Indeed, structural and functional genomics, the three-dimensional (3D) structure and functional analysis of genomic products, are rising disciplines in bioinformatics. It has been reported, for example, that a sequence homology of larger than 40% usually implies more than 90% 3D-structure overlap (see below for precise definitions of similarity). Thus, with the growing amount of genomic information, we may eventually be able to predict reliably 3D structures of proteins. Since structural similarity is often preserved more strongly than sequence through evolution, reliable homology-based predictions might provide crucial functional properties of new gene products in the near future.

Through this assignment, you will gain some experience in quantifying and analyzing sequence and 3D structure similarity for proteins. You will also explore sequence and structure databases in search of interesting examples, and learn how to use important computational and database resources. You will have to be resourceful in looking for suitable programs for alignment and structure analysis besides those below; no simple recipes will be given here.

*This assignment can be done by teams of two students; choose a partner with complementary skills. You will have to present your results to the class.*

### The 5 Tasks

Find and demonstrate the following four relationships for proteins:

1. [EASY] Two proteins with very *high sequence similarity* (but less than 95%) and very *high structural similarity*. Excluded from consideration are trivial examples, such as involving multiple PDB entries for the same protein.
2. [EASY] Two proteins with very *high sequence similarity* (but less than 95%) and *very high structural similarity* but markedly *different biological/functional properties*.
3. [MODERATE] Two proteins with *low sequence similarity* but *high structural similarity*. Also comment on the *functional properties* of the pair.
4. [HARD] Two proteins with *very high sequence similarity* but *very low structural similarity*. Also comment on the *functional properties* in your example.

For problems 3 and 4 above, the class contest will be won by the students that find the most extreme

examples (i.e., the maximal sequence similarity / minimal structural similarity, minimal sequence similarity / maximal structural similarity).

5. [EASY WARMUP] Search and identify all the determined structures in the PDB/NDB that contain the nucleic acid sequence TATAAAAG. Discuss these structures and their significance.

*For each task, generate color molecular views, report the analyses in detail, and include a description of how you found the example. Also discuss your similarity/dissimilarity criteria (see below), and prepare a class presentation on your results.*

## Ground Rules

1. Homology, or sequence similarity, will be defined by the percentage of sequence identity.
2. 3D-structure similarity will be defined in two ways:
  - (a) the percentage of C<sup>α</sup> atoms of the proteins that “overlap”, i.e., are within 3.5 Å of each other in a rigid-body alignment of the protein;
  - (b) the root-mean-square-deviation (RMSD) between C<sup>α</sup> atoms of the proteins in a rigid-body alignment of the protein. (Recall your experience with RMSD measurements in the previous assignment).

You should first experiment with overlapping several protein structures to determine what RMSD values and/or percentages of C<sup>α</sup> overlap indicate random similarity. *Discuss this in your submission.*

## Tools of the Trade

1. **Sequence and Structure Databases.** You have already navigated through the structural PDB and NDB databases and various sequence databases. Continue to work with these and the RCSB facilities.
2. **SCOP.** This site for the *Structural Classification of Proteins* ([scop.mrc-lmb.cam.ac.uk/scop/](http://scop.mrc-lmb.cam.ac.uk/scop/)) categorizes proteins according to the levels (top-to-bottom) of: class, fold, superfamily, family, domain, and reference PDB structure.
3. **Insight II.** Continue to use Insight II for structure display and analysis.
4. **NCBI Tools like BLAST and Its Cousins.** BLAST is a library of heuristic similarity search programs (Basic Local Alignment Search Tools) that explore relationships involving protein and nucleic-acid sequences and 3D structures. This library contains blastp, blastn, blastx, tblastn, tblastx,

and others, developed at the National Center for Biotechnology Information at the National Library of Medicine of the National Institutes of Health. Get started at their web site [www.ncbi.nlm.nih.gov/BLAST/](http://www.ncbi.nlm.nih.gov/BLAST/). This page leads to the BLAST suites as well as contains usage information. See, for example, Overview, Manual, BLAST FAQs, References.

BLAST, one of the most popular tools among molecular biology researchers, has evolved rapidly since its inauguration in 1990. BLAST searches a database in two stages, finding small sequence lengths that match the target exactly and then attempting to extend the length of the match from this subset of sequences in the database. Not only are the alignment algorithms improving continuously (e.g., allowing alignments of DNA or protein sequences with insertions or deletions in Gapped BLAST; forming families of aligned sequences and quick profiles of them in Position-Specific Iterated (PSI)-BLAST; or incorporating biological-function hypotheses into sequence queries to restrict the analysis to subset of protein sequences as in Pattern-Hit Initiated (PHI)-BLAST), but performance has been greatly accelerated. Algorithmic features include dynamic programming tools, hidden Markov models, and various optimization strategies.

To align two protein or nucleotide sequences, go to the link of **BLAST 2 sequences** ([www.ncbi.nlm.nih.gov/gorf/bl2.html](http://www.ncbi.nlm.nih.gov/gorf/bl2.html)) and set up the computation according to the instructions. Take care to choose the options of the computation with care, and explore different options. The server will send the results to the web browser being used.

Some available programs are:

- blastp: compares an amino acid query sequence against a protein sequence database.
- blastn: compares a nucleotide query sequence against a nucleotide sequence database.
- blastx: compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.
- tblastn: compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames (both strands).
- tblastx: compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

See [www.ncbi.nlm.nih.gov/BLAST/newblast.html#introduction](http://www.ncbi.nlm.nih.gov/BLAST/newblast.html#introduction) for further information.

*Other similarity programs are available (such as MEME and MAST from SDSC); use anything appropriate for the task.*

5. **An Image Library.** The *Image Library of Biological Macromolecules* organized by the Institute for Molecular Biotechnology in Jena, Germany ([www.imb-jena.de/IMAGE.html](http://www.imb-jena.de/IMAGE.html)) offers a colorful library of biomolecular images corresponding to structures available in databases like the NDB

and PDB. Besides detailed colorful illustrations of the structure in a variety of styles, relevant structural information and publication links are available. Basic tutorials on structural biology are under preparation at this site.

## **HINTS for the Assignment**

1. Scan the literature for related papers on comparative or homology modeling but do not repeat known examples. You **CAN** be original.
2. Large changes in 3D structure despite high sequence similarity can result from the following situations:
  - mutations in critical regions of the proteins such as active sites
  - mutations in ligand binding sites (as in immunoglobulins)
  - mutations in regions that connect two secondary-structural elements (as in helix-loop-helix motifs)
  - structure determination of the same system at different environmental conditions (e.g., different solvent, different crystal packing forms for mutant proteins)
  - proteins containing the same subunits but a different number of sub-units, with a structure/fold/topology that depends critically on that number.

Search PDB and SCOP for examples in this spirit.

3. Look for groups of proteins in the same family, or for proteins sharing the same fold in the SCOP site. The structural classification information should generate ideas.
4. General structure alignment via **Insight** is not very sophisticated and may be entirely unsuitable for sequences of disparate lengths and for structures with two similar subdomains adopting a different relative orientation. Search for suitable programs for these cases (e.g., from the RCSB, [home.rcsb.org](http://home.rcsb.org) and from SDSC) and also write/use your own programs to perform certain analyses, such as structure similarity measurements upon alignment (e.g., criterion 2a under **Ground Rules**).

## **Background Reading**

- D. Baker and A. Sali, “Protein Structure Prediction and Structural Genomics” *Science* **294**, 93–96 (2001). [From Coursepack].
- J. C. Whisstock and A. M. Lesk, “Prediction of Protein Function from Protein Sequence and Structure”, *Quart. Rev. Biophys.* **36**, 173–189 (2001). [From Coursepack].
- J.-M. Chandonia and S. E. Brenner, “The Impact of Structural Genomics: Expectations and Outcomes”, *Science* **311**, 347–351 (2006). [From Coursepack].

- B. Honig and A. Nicholls, “Classical Electrostatics in Biology and Chemistry”, *Science* **268**, 1144–1149 (1995) [From Coursepack].
- D. Case, “NMR Refinement”, in P. von Ragué Schleyer (Editor-in Chief), N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, and H. F. Schaefer, III, editors, *Encyclopedia of Computational Chemistry*, volume 3, pages 1866–1876. John Wiley & Sons, West Sussex, England, 1998.

## Assignment 7: Molecular Mechanics Force Fields: Approximations, Variations, and the Assessment of Results with respect to Experiment and other Simulations

1. **Reading.** This assignment deals with the series of four articles below, which raise both general and specific problems in biomolecular simulations. At issue is the validation of conformational predictions by various molecular mechanics force fields. You may also wish to refer to the Lipkowitz article from Assignment 5 (on the pitfalls of molecular mechanics) and the van Gunsteren and Mark article from Assignment 1 (on validating molecular dynamics simulations). Begin by reading these papers (included in the Coursepack, see Appendix B) and thinking about the modeling issues as you read them.

- I. K. Roterman, M. H. Lambert, K. D. Gibson, and H. A. Scheraga, “Comparison of the CHARMM, AMBER and ECEPP Potentials for Peptides. I. Conformational Predictions for the Tandemly Repeated Peptide (Asn-Ala-Asn-Pro)<sub>9</sub>”, *J. Biomol. Struct. Dyn.* **7**, 391–419 (1989a).
- I. K. Roterman, M. H. Lambert, K. D. Gibson, and H. A. Scheraga, “Comparison of the CHARMM, AMBER and ECEPP Potentials for Peptides. II.  $\phi$ - $\psi$  Maps for N-Methyl Amide: Comparisons, Contrasts and Simple Experimental Tests”, *J. Biomol. Struct. Dyn.* **7**, 421–453 (1989b).
- P. A. Kollman and K. A. Dill, “Decisions in Force Field Development: An Alternative to Those Described by Roterman *et al.*”, *J. Biomol. Struct. Dyn.* **8**, 1103–1107 (1991).
- K. B. Gibson and H. A. Scheraga”, “Decisions in Force Field Development: Reply to Kollman and Dill”, *J. Biomol. Struct. Dyn.* **8**, 1109–1111 (1991).

2. **Preparation for Class Discussion.** You will be divided into three groups (assignments will be given in class): (1) the moderators, (2) the ECEPP group, and (3) the AMBER and CHARMM group. Each group will have to prepare material, as described below, for class presentation and discussion. *All materials should be prepared on overhead projector slides.* You should meet with your group members in advance to plan your presentation and debate strategies.

The *moderators* will be in charge of presenting in detail the *facts*: what studies were performed, what questions were asked, and what analyses were made. You should be prepared to answer any background questions (e.g., definitions of polymer quantities analyzed).

The *ECEPP* group will endorse the point of view taken by Roterman, Gibson, Scheraga, and co-workers. Besides understanding your position well, you will need to bring to the debate *concrete examples from the literature* to support your position. Be creative and try to find interesting examples.

The *AMBER* folks and *CHARMMers* will endorse the approach taken in these two molecular packages and, in particular, the point of view taken by Kollman and Dill in their reply to Roterman *et al.* As above, besides understanding well your molecular mechanics packages and position taken in the reply, you will need to bring to the debate *concrete examples from the literature* to support your position. Be creative in your supporting materials and strategies.

3. **Useful Recommendations.** Summarize in brief the useful recommendations and comments that emerged from all the above articles, as well as additional ones, for practitioners of molecular modeling. That is, propose *concrete procedures* that biomolecular simulators can use to gain as much confidence as possible in their conclusions and predictions.

Remember, uncertainties and approximations in numerical modeling and simulations will always exist! The field of modeling biomolecules on modern computers involves as much art as science. But despite their obvious limitations, modeling methodologies are improving continuously. The goal of every practitioner should be to realize the highest possible accuracy as is compatible with the model and methods utilized. Like any calculation, ‘error bars’ in the broad sense should be attributed to the results and conclusions claimed.

4. **Points to Keep in Mind.** Throughout this assignment, think about the following important issues in molecular modeling:

- Accuracy versus approximation
- Theory versus experiment
- Dependence of simulation results on the protocols used
  - starting configuration
  - model assumptions
  - force field
  - algorithms (minimization, adiabatic mapping, etc.)
- Assessment of Results:
  - How can you distinguish between bona fide physical *trends* and numerical *artifacts*?
  - How can you decide whether the model is wrong (energy, assumptions, etc.) or the method is inappropriate?
  - What are appropriate comparisons with experimental results?

### Summary of Items to Hand in:

- (a) Brief description of the issues raised in the four articles regarding molecular mechanics predictions.
- (b) Your work in preparation of the class debate.
- (c) Proposals of procedures to be used to attain the maximum possible confidence from biomolecular simulations.

## Have Fun!

### Background Reading from Coursepack

- J. Skolnick and A. Kolinski, “Simulations of the Folding of a Globular Protein”, *Science* **250**, 1121–1125 (1990).
- F. M. Richards, “The Protein Folding Problem”, *Sci. Amer.* **264**, 54–63 (1991).
- H. A. Scheraga, “Predicting Three-Dimensional Structures of Oligopeptides”, in *Reviews in Computational Chemistry*, K. B. Lipkowitz and D. B. Boyd, Editors, Vol. 3, pp. 73–142, VCH Publishers, New York (1992).
- A. Neumaier, “Molecular Modeling of Proteins and Mathematical Prediction of Protein Structure”, *SIAM Review* **39**, 407–460 (1997).

### Background Reading for Scheraga’s Lecture

- J. Pillardy, Y. A. Arnautova, C. Czaplewski, K. D. Gibson, and H. A. Scheraga, “Conformation-Family Monte Carlo: A New Method for Crystal Structure Prediction”, *Proc. Natl. Acad. Sci., USA* **98**, 12351–12356 (2001).
- J. Pillardy, C. Czaplewski, A. Liwo, W. J. Wedemeyer, J. Lee, D. R. Ripoll, P. Arlukowicz, S. Oldziej, Y. A. Arnautova and H. A. Scheraga, “Development of Physics-Based Energy Functions that Predict Medium-Resolution Structures for Protein of the  $\alpha$ ,  $\beta$ , and  $\alpha/\beta$  Structural Classes”, *J. Phys. Chem. B* **105**, 7299–7311 (2001).
- J. Lee, D. R. Ripoll, C. Czaplewski, J. Pillardy, W. J. Wedemeyer and H. A. Scheraga, “Optimization of Parameters in Macromolecular Potential Energy Functions by Conformational Space Annealing”, *J. Phys. Chem. B* **105**, 7291–7298 (2001).

- A. Liwo, C. Czaplewski, J. Pillardy and H. A. Scheraga, “Cumulant-Based Expressions for the Multibody Terms for the Correlation Between Local and Electrostatic Interactions in the United-Residue Force Field”, *J. Chem. Phys.* **115**, 2323–2347 (2001).
- J. Pillardy, C. Czaplewski, A. Liwo, J. Lee, D. R. Ripoll, R. Kazmierkiewicz, S. Oldziej, W. J. Wedemeyer, K. D. Gibson, Y. A. Arnautova, J. Saunders, Y.-J. Ye and H. A. Scheraga, “Recent Improvements in Prediction of Protein Structure by Global Optimization of a Potential Energy Function”, *Proc. Natl. Acad. Sci., USA* **98**, 2329–2333 (2001).
- J. Pillardy, C. Czaplewski, W. J. Wedemeyer and H. A. Scheraga, “Conformation-Family Monte Carlo (CFMC): An Efficient Computational Method for Identifying the Low-Energy States of a Macromolecule”, *Helv. Chim. Acta* **83**, 2214–2230 (2000).
- J. Lee, J. Pillardy, C. Czaplewski, Y. Arnautova, D. R. Ripoll, A. Liwo, K. D. Gibson, R. J. Wawak, and H. A. Scheraga, “Efficient Parallel Algorithms in Global Optimization of Potential Energy Functions”, *Comput. Physics Commun.* **128**, 399–411 (2000).
- J. Lee, A. Liwo, D. R. Ripoll, J. Pillardy, J. A. Saunders, K. D. Gibson and H. A. Scheraga, “Hierarchical Energy-Based Approach to Protein-Structure Prediction: Blind-Test Evaluation with CASP3 Targets”, *Intl. J. Quantum Chem.* **71**, 90–117 (2000).
- J. Pillardy, R. J. Wawak, Y. A. Arnautova, C. Czaplewski, and H. A. Scheraga, “Crystal Structure Prediction by Global Optimization as a Tool for Evaluating Potentials: Role of the Dipole Moment Correction Term in Successful Predictions”, *J. Am. Chem. Soc.* **122**, 907–921 (2000).
- H. A. Scheraga, J. Lee, J. Pillardy, Y.-J. Ye, A. Liwo, and D. R. Ripoll, “Surmounting the Multiple-Minima Problem in Protein Folding”, *J. Global Optimization* **15**, 235–260 (1999).
- J. Lee, A. Liwo and H. A. Scheraga, “Energy-Based *de novo* Protein Folding by Conformational Space Annealing and an Off-lattice United-Residue Force Field: Application to the 10–55 Fragment of Staphylococcal Protein A and to apo calbindin D9K”, *Proc. Natl. Acad. Sci., USA* **96**, 2025–2030 (1999).
- J. Lee, H. A. Scheraga and S. Rackovsky, “Conformational Analysis of The 20-Residue Membrane-Bound Portion of Melittin by Conformational Space Annealing”, *Biopolymers* **46**, 103–115 (1998).
- R. J. Wawak, J. Pillardy, A. Liwo, K.D. Gibson and H. A. Scheraga, “Diffusion Equation and Distance Scaling Methods of Global Optimization: Applications to Crystal Structure Prediction”, *J. Phys. Chem.* **102**, 2904–2918 (1998).

- A. Liwo, R. Kazmierkiewicz, C. Czaplewski, M. Groth, S. Oldziej, R. J. Wawak, S. Rackovsky, M. R. Pincus, and H. A. Scheraga, “United-Residue Force Field for Off-Lattice Protein-Structure Simulations; III. Origin of Backbone Hydrogen-Bonding Cooperativity in United-Residue Potentials”, *J. Comput. Chem.* **19**, 259–276 (1998).

## Assignment 8: A Bit of Programming: Nonbonded Versus Bonded Energy Computations

This is a small programming assignment. At least one assignment in this modeling course should give you such first-hand experience! If you are a novice in programming, NYU staff, the TA, and the course assistant can help you, so set up an appointment with them early. You will have *two weeks* for this assignment.

### 1. Programming Nonbonded Energy Computations.

We will begin with the nonbonded energy computations since they are most straightforward (but most expensive!)

Write a simple program to compute the nonbonded energy of a system of 1000 atoms. The nonbonded energy, Lennard Jones and Coulomb terms, should have the form:

$$E_{NONB} = \sum_{i < j} \left[ \frac{-A_{ij}}{R_{ij}^3} + \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\sqrt{R_{ij}}} \right]. \quad (\text{D.1})$$

Here  $R_{ij}$  is an interatomic distance *squared*, and  $A_{ij}$ ,  $B_{ij}$ ,  $q_i$ , and  $q_j$  are the familiar energy parameters.

For an atom  $\mathbf{x}_k$  of Cartesian components  $\{x_{k1}, x_{k2}, x_{k3}\}$ ,

$$R_{ij} = (x_{j1} - x_{i1})^2 + (x_{j2} - x_{i2})^2 + (x_{j3} - x_{i3})^2, \quad (\text{D.2})$$

and the interatomic distance  $r_{ij}$  is

$$r_{ij} = \sqrt{R_{ij}}.$$

Set up your program to read in arbitrary atomic coordinate data — a file will be sent to you electronically in case you want to use it<sup>5</sup> — and *repeat the nonbonded energy calculation* 10,000 times for all  $\{i, j\}$  pairs with  $i < j$  for  $j = 1, \dots, 1000$ . The  $R_{ij}$  calculations can be placed in some inline function.

Perform the calculations for the nonbonded energy evaluations for both single and double precision, and record the total CPU time in each case.

Also report how much CPU time and CPU percentage the square-root ( $\sqrt{\phantom{x}}$ ) operation consumes. There are special timing functions that describe the distribution of CPU time among the various program parts.

Describe the machine you are using, the precision, and attach the subroutine and program output, along with the above results.

---

<sup>5</sup>The file can be obtained through the link to the course web site or directly from the author.

## 2. Programming the Bonded Energy Computations.

Next we will write three additional subprograms to compute the bond energy, bond-angle energy, and dihedral-angle energy of a molecular system. For now, we will assume there are 1000 bonds, 1000 bond angles, and 1000 dihedral angles. We will again perform 10,000 energy evaluations of each energy term, with each sweep here involving 1000 internal variables. This large number is necessary to get reliable timing values for the bonded interactions.

You can choose in each case any representative potential form. For example, you may use:

$$E_{BOND} = \sum_{i,j \in S_B} S_{ij} (r_{ij} - \bar{r}_{ij})^2, \quad (\text{D.3})$$

$$E_{BANG} = \sum_{i,j,k \in S_{BA}} K_{ijk} (\cos \theta_{ijk} - \cos \bar{\theta}_{ijk})^2, \quad (\text{D.4})$$

$$E_{TOR} = \sum_{ijkl \in S_{DA}} \left( \frac{V^3_{ijkl}}{2} [1 + \cos(3\tau_{ijkl})] \right), \quad (\text{D.5})$$

where the sets  $S_B$ ,  $S_{BA}$ , and  $S_{DA}$  contain all bonds, bond angles, and dihedral angles, respectively. Here  $\theta$  and  $\tau$  denote a bond angle and dihedral angle, respectively, of a given triplet or quadruplet of atoms. The values with overhead bar symbols indicate reference values.

Since we are interested only in timing for now, you can use any pairs, triplets, or quadruplets in your sample energy routines — even the same sequence — repeatedly, as long as the total number of interactions used to obtain each energy term is 1000.

Some program segments which you may find helpful are posted on the website. The derivative components are present in the angle routines, *but you do not need them* for this assignment. You can find details of the **cosba** and **cosda** subroutines in an article.<sup>6</sup>

For your convenience, an addendum to this assignment also summarizes the basic geometric relations involved in defining internal variables.

Report the CPU time required for each routine in a table, including absolute time as well as percentage of the total time of *bonded* energy components. Again, attach your programs and output to the report of the results.

---

<sup>6</sup>“A Recipe for Evaluating and Differentiating  $\cos\phi$  Expressions”, *J. Comp. Chem.* **10**, 951–956, 1989.

3. **Setting up A Polymer Model.** For obtaining realistic CPU estimates, we will now consider a simple *n*-alkane chain with the chemical formula  $\text{CH}_3-(\text{CH}_2)_m-\text{CH}_3$ , where *m* is an integer. For large *m*, this is polyethylene. For *m* = 2, for example, we have butane, chemical formula  $\text{C}_4\text{H}_{10}$ . To have about 1000 atoms, we will use *m* = 330 for our model calculations.

Determine the number of bonds, bond angles, dihedral angles, and unique interatomic distances (atom pairs) that polyethylene has as a function of *m*. Consider all the distinct possibilities for the bonds and angles. Report these expressions.

Then report how many bonds, bond angles, dihedral angles, and unique atom pairs the polymer has for the case *m* = 330.

4. **Bonded Versus Nonbonded Energy Computations.**

Now we will combine the timing above to estimate the CPU time spent in bonded versus nonbonded energy computations for 10,000 iterations (of energy evaluations) for our polymer of 998 atoms.

Scale the timing you obtained above (10,000 iterations for 1000 atoms for the nonbonded terms, and 10,000 iterations for 1000 bonds, bond angles, and dihedral angles) so that they correspond to the numbers relevant for our polymer with *m* = 330, as determined in item 3 above.

Collect the data in one table which reports the CPU time and percentage required for each of the four subroutines.

What can you conclude? What can you suggest to speed up the nonbonded computations, especially if derivatives are also required?

5. **Extra Credit!**

For extra credit (the grade on this will replace your lowest homework grade), write the four subroutines above specifically for polyethylene. This means that you should use realistic coordinates, as well as correct data structures so that you consider all relevant bonds and angles for this polymer. Similarly, for energy parameters, associate values according to atom, bond, and angle types (e.g., C–C and C–H bonds, C–C–C, H–C–H, and H–C–C bond angles, and rotations about C–C bonds). You can use any resources on **Insight** to help you.

Hand in all programs and results as requested above.

---

**Addendum to Assignment 8:**  
**Definitions of Internal Variables in Molecules**

A bond angle  $\theta_{ijk}$  formed by a bonded triplet of atoms  $i-j-k$  is expressed as an inner product:

$$\cos \theta_{ijk} = \frac{(\mathbf{x}_k - \mathbf{x}_j) \bullet (\mathbf{x}_i - \mathbf{x}_j)}{r_{jk} r_{ji}}, \quad (\text{D.6})$$

or

$$\cos \theta_{ijk} = (\mathbf{r}_{jk} \bullet \mathbf{r}_{ji}) / r_{jk} r_{ji},$$

where the distance vector from atom  $j$  to  $i$  is given by

$$\mathbf{r}_{ji} = \mathbf{x}_i - \mathbf{x}_j = [x_{i1} - x_{j1}, \ x_{i2} - x_{j2}, \ x_{i3} - x_{j3}]^T. \quad (\text{D.7})$$

A dihedral angle  $\tau_{ijkl}$ , defining the rotation of bond  $i-j$  about bond  $j-k$  with respect to  $k-l$ , is expressed as

$$\cos \tau_{ijkl} = \mathbf{n}_{ab} \bullet \mathbf{n}_{bc}. \quad (\text{D.8})$$

The vectors  $\mathbf{n}_{ab}$  and  $\mathbf{n}_{bc}$  denote unit normals to planes spanned by vectors  $\{\mathbf{a}, \mathbf{b}\}$  and  $\{\mathbf{b}, \mathbf{c}\}$ , respectively, where  $\mathbf{a} = \mathbf{r}_{ij}$ ,  $\mathbf{b} = \mathbf{r}_{jk}$ , and  $\mathbf{c} = \mathbf{r}_{kl}$ . Denoting  $\theta_{ab}$  and  $\theta_{bc}$  as angles  $\theta_{ijk}$  and  $\theta_{jkl}$ , respectively, we write:

$$\cos \tau_{ijkl} = \frac{\mathbf{a} \times \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\| \sin \theta_{ab}} \bullet \frac{\mathbf{b} \times \mathbf{c}}{\|\mathbf{b}\| \|\mathbf{c}\| \sin \theta_{bc}}. \quad (\text{D.9})$$

The sign of  $\tau_{ijkl}$  is determined by the sign of the triple scalar product  $\mathbf{a} \bullet (\mathbf{b} \times \mathbf{c})$ .

To simplify potential energy equations (and differentiation when needed) it is convenient to work with inner product expressions and use Lagrange's identity  $(\mathbf{a} \times \mathbf{b}) \bullet (\mathbf{c} \times \mathbf{d}) = (\mathbf{a} \bullet \mathbf{c})(\mathbf{b} \bullet \mathbf{d}) - (\mathbf{b} \bullet \mathbf{c})(\mathbf{a} \bullet \mathbf{d})$ . This produces the alternative expression:

$$\begin{aligned} \cos \tau_{ijkl} &= \frac{(\mathbf{a} \times \mathbf{b}) \bullet (\mathbf{b} \times \mathbf{c})}{[(\mathbf{a} \times \mathbf{b}) \bullet (\mathbf{a} \times \mathbf{b}) (\mathbf{b} \times \mathbf{c}) \bullet (\mathbf{b} \times \mathbf{c})]^{1/2}} \\ &= \frac{(\mathbf{a} \bullet \mathbf{b})(\mathbf{b} \bullet \mathbf{c}) - (\mathbf{a} \bullet \mathbf{c})(\mathbf{b} \bullet \mathbf{b})}{\left\{ [(\mathbf{a} \bullet \mathbf{a})(\mathbf{b} \bullet \mathbf{b}) - (\mathbf{a} \bullet \mathbf{b})^2] [(\mathbf{b} \bullet \mathbf{b})(\mathbf{c} \bullet \mathbf{c}) - (\mathbf{b} \bullet \mathbf{c})^2] \right\}^{1/2}}. \end{aligned} \quad (\text{D.10})$$

According to this convention,  $\tau = 0^\circ$  defines a *cis* coplanar orientation for atoms  $i-j-k-l$ ,  $\tau = 180^\circ$  defines a *trans* coplanar orientation, and a positive sign corresponds to a clockwise rotation of the far bond with respect to the near bond (when viewed along the  $j-k$  bond).

---

(See code segments on website)

**Coordinate file (available electronically) for the 1000-atom molecule**  
**CH<sub>2</sub>OH–(CH<sub>2</sub>)<sub>330</sub>–CH<sub>2</sub>OH:**

Atom	X	Y	Z	ID
1	4.988000	2.012136	-7.818089	O
2	5.915912	2.025234	-7.572307	H
3	4.596542	0.674197	-8.131964	C
4	5.198226	0.305558	-8.962735	H
5	4.751310	0.037135	-7.261160	H
6	3.108016	0.653187	-8.526237	C
7	2.517394	1.015322	-7.710808	H
8	2.956134	1.278341	-9.381230	H
9	2.686321	-0.787809	-8.863891	C
10	2.838205	-1.412964	-8.008898	H
..	.....	.....	.....	.

Etc.

## Assignment 9: TERM PROJECT

### The Successes (Failures?) of Molecular Modeling

The year is 2006. You have graduated and moved on with your life. Due to your outstanding academic record at NYU, you have landed a high-profile job as a staff research scientist for PBS (Public Broadcasting Service) in the nation's capital.

You are now assigned to prepare for an internationally televised scientific program entitled *Biocomputing in the Third Millennium*. In this program, a team of scientific experts will respond to live questions transmitted by comphones from the general public. Since these scientists are busy traveling, consulting, reviewing papers, writing grants, researching, and teaching, your group is in charge of preparing all background information for the panelists.

Specifically, you are told to prepare for the following questions:

*Can the panel describe some concrete examples where computational tools have significantly enhanced our understanding of molecular systems — from small organic systems to macromolecules — by offering new insights, interpretations, and predictions, of practical and scientific importance, that were impossible to obtain by experimental techniques?*

*What modeling/simulation tools were used in each case, and what can be credited to each success (computing power, algorithms, intuition, right time, sheer luck, persistence, etc.)?*

You are promised by your boss a hefty bonus for each complete and satisfactory item provided. However, the minimal requirement (for obtaining a B-level mark on your monthly evaluation form, given that you produce truly outstanding examples) is detailing FOUR “SENSATIONAL” EXAMPLES.

Each example must be clearly described and entered under the following sub-headings: Problem, Methodology, Success, Significance, References. The second item, **Methodology**, requires the most comprehensive coverage, followed by **Significance**. You are asked to attach to your meticulous writeup any visual aids (charts, figures, sketches) that will enhance the presentation, both to a general (nonspecialist) audience and to a highly informed scientist. Creativity is highly desired. Try also to analyze the findings in a larger context.

Back at your ergonomic desk, with your feet up and glancing at regal Washington monuments against a glorious background of blossoming cherry trees with occasional views of ambitious runners and politicians, you recall a molecular modeling course you took in the good ol' days at NYU. Memories come back of many homework assignments inflicted upon you weekly by your

professor — dealing with web resources, sequence and structural databases, Insight, sequence/structure contests, force fields, tedious programming, difficult minimization, and Monte Carlo simulations. You find fragments of lecture notes and transparency copies inside an old purple gym bag and begin to follow up on, and explore, some of those key words, resources, authors, and topics. You also begin to wonder if there are any interesting and instructive examples of *failures in molecular modeling* and decide to pursue those for an extra bonus. (Maybe the boss will let *you* design the next scientific program?)

Your deadline in early May is rapidly approaching and you begin to work early and diligently. The promise that the best examples provided by the crew will be published, if appropriate, in an article provides further motivation for the assignment. You also decide to contact your professor if she is still at NYU when you get stuck or have questions.

You find the project more interesting now, and vow to become *famous* (and maybe even *rich*)!

### Background Reading from Coursepack

- M. S. Friedrichs and P. G. Wolynes, “Toward Protein Tertiary Structure Recognition by Means of Associative Memory Hamiltonians”, *Science* **246**, 371–373 (1989).
- T. Schlick, “Optimization Methods in Computational Chemistry”, in *Reviews in Computational Chemistry*, K. B. Lipkowitz and D. B. Boyd, Editors, Vol. 3, pp. 1–71, VCH Publishers, New York (1992).
- See also an updated version titled “Geometry Optimization” in the *Encyclopedia of Computational Chemistry*, P. von Ragué Schleyer (Editor-in-Chief) and N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, and H. F. Schaefer, III, Editors, Vol. 3, pp. 1136–1157, John Wiley & Sons, West Sussex, England (1998).
- R. A. Abagyan and M. M. Totrov, “Biased Probability Monte Carlo Conformational Searches and Electrostatic Calculations for Peptides and Proteins”, *J. Mol. Biol.* **235**, 983–1002 (1994).

## Assignment 10: Experiments in Molecular Geometry Optimization: Biphenyl Minimization

See **Insight II** and **Discover** manuals for reference.

### 1. Brief introduction to the Discover module of Insight II.

The **Discover**<sup>7</sup> software performs energy minimization and molecular dynamics simulations. This program constitutes a powerful modeling tool since it offers many features such as constrained and restrained minimization, calculation of vibrational frequencies, and analysis tools. Many variations of simulation conditions (e.g., constant temperature, constant pressure) are available.

We will access the **Discover** software from the **Insight II** environment. The **Discover** module of **Insight II** is a convenient interface to the **Discover** program. This module builds **Discover** input files from information provided through graphical interfaces, and it allows users to run **Discover** jobs interactively. Though more advanced users may prefer to use the independent version of **Discover**, the **Insight II** environment is more appropriate for a novice.

Before using **Discover**, make sure that **Insight II** contains all of the necessary information to define the topology, coordinates, and force field parameters. These include, for example, atom types and partial charges (see lecture notes for structure definitions).

If you succeed in displaying the molecule correctly on the screen, the topological and coordinate information is most likely in order. However, selecting the appropriate force field and assigning atom types and parameters is a separate task.

- (a) To select the force field, use **Forcefield** / **Select**.
- (b) To assign atom types, use the **Fix** option for **Potential Action** in **Forcefield** / **Potentials**. Alternatively, first assign atom types with **Atom** / **Potential** in the **Biopolymer** module, and then use the **Accept** option for **Potential Action** in **Forcefield** / **Potentials**.
- (c) To assign charges, use the **Fix** option for both **Partial Chg Action** and **Formal Chg Action** under **Forcefield** / **Potentials**.

Note that after each change in the force field you must assign atom types and charges anew.

---

<sup>7</sup>Note that the name *Discover* has two separate meanings. The first, **Discover**, stands for the software package with minimization and molecular dynamics routines. The second, typed in bold (**Discover**), refers to the module available in **Insight II**.

To check if the assigned atom types and partial charges are correct, you can select Potential or Partial\_charge in **Molecule** / **Label** to label each atom. Once you specify the information about the structure and parameters, you are ready to move to the **Discover** module. (We will not use **Discover\_3** in this course).

The **Constraint** pulldown menu contains various atom-constraining and restraining procedures that you can select. In **Parameters**, you select the simulation type for Discover (**Minimize**, **Dynamics**, etc.), as well as the choice for cutoff parameters for nonbonded interactions, periodic boundary conditions (**Variables**), and dielectric constant (**Set**). Take time to familiarize yourself with the first three pulldown menus **Constraint**, **Parameters**, and **Run**, with **Insight\_help** active, to learn about the various commands they contain.

To start a simulation, go to **Run** / **Run**, select desired options, choose the object for calculations, and execute.

Each **Discover** run is assigned a number in the order of the execution start time. The files created during the execution are identified by the calculation object (molecular system name) and the job integer (appended to the name). The file extension specifies the file type. Examples are listed below.

#### **Discover Input Files:**

- Commands (.inp)
- Cartesian Coordinates (.car)
- Molecular Data (.mdf)
- Force field Parameters (.frc)
- Restraints (.rstrnt)

#### **Discover Output Files:**

- Standard Output (.out)
- Cartesian Coordinates (final structure) (.cor)
- Cartesian Coordinate Archive (multiple frames) (.arc)
- Automatic Potential Parameter Assignment (.prm)
- Discover Dynamics Restart Information (.rst)

You can specify the files to save with **Run** / **Files**. By default, all are saved.

## 2. Setting up biphenyl minimization.

We will begin to learn about potential energy minimization for a simple yet interesting system, biphenyl (see Fig. D.2).

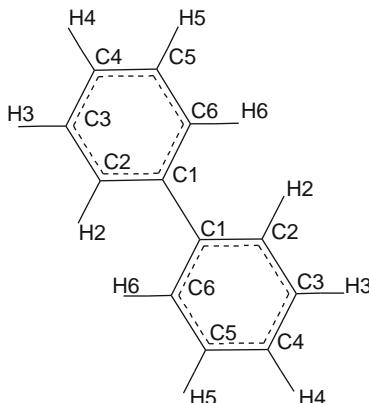


Figure D.2. Biphenyl.

You will receive electronically two files containing the coordinates of biphenyl<sup>8</sup> The first, `biphenyl.car`, includes the structure with coplanar phenyl rings. This configuration was created with the **Builder** module by connecting two benzene rings.

The second file, `biphenyl_distorted.car`, contains the structure with each of the phenyl rings distorted from planarity.

Before displaying these structures, check that the AMBER force field is chosen. This will save work in assigning AMBER force field parameters. It will also permit you to proceed to Discover directly. (Note: For other force fields, you would have to assign parameters through **Forcefield / Potentials**). To open a coordinate file and display a structure, use **Molecule / Get**, specify **Archive** as the **File Type**, and select the desired file.

### 3. Generation of energy profiles by restrained minimization.

A potential energy profile along some molecular coordinate,  $X$  (such as the rotamer dihedral angle  $\chi_1$ ), describes the dependence of the energy, minimized with respect to the remaining coordinates, on  $X$ . The simplest way to generate such a profile is to use minimization with *restraints*. Restraining a coordinate  $X$  to a specified value  $X^0$  can be accomplished by adding harmonic penalty term,

$$E_{\text{rstr}} = K(X - X^0)^2,$$

to the potential energy. After minimization,  $X$  should not deviate significantly from  $X^0$  when the force constant  $K$  is large.<sup>9</sup> For a complete profile,

<sup>8</sup>Files can be obtained through the link to the course web site or directly from the author.

<sup>9</sup>*Constrained*, as opposed to restrained, minimization entails a more complex procedure to guarantee that  $X = X^0$ .

minimum energy values must be calculated for a series of values  $\{X_1^0, X_2^0, X_3^0, \dots\}$  in the range of  $X$ .

For biphenyl, we will analyze the dependence of energy on the torsion angle between the planes of phenyl rings. Four dihedral angles are defined about the C1–C1 bond connecting the two rings. They are specified by the following atom quadruplets {1B:C2, 1B:C1, 1:C1, 1:C6}, {1B:C6, 1B:C1, 1:C1, 1:C2}, {1B:C2, 1B:C1, 1:C1, 1:C2}, and {1B:C6, 1B:C1, 1:C1, 1:C6}. Restraining only one of them will result in a nonplanar geometry of phenyl rings (since the remaining dihedral angles will tend to assume values associated with a lower energy). To ensure that the phenyl-ring planes are not distorted, it is necessary to restrain a *pair* of dihedral angles to the same value. You can choose the first two or the last two atom quadruplets from the list above.

The plot for the full range of the angle,  $[-180.0^\circ, 180.0^\circ]$ , can be created by first computing energy minima for a sequence of values in the range  $[0.0^\circ, 90.0^\circ]$  (e.g.,  $0.0^\circ, 10.0^\circ, 20.0^\circ, \dots, 90.0^\circ$ ), and then using symmetry operations.

Start with the coplanar structure (`biphenyl.car`). Make sure that potential parameters are properly assigned.

Then select **Constraint** / **TorsionForce**. You can now proceed in different ways to calculate the energy values for the profile. For instance, you can make 10 separate minimization runs, each time specifying both restraints (**Intervals** set to 1). Alternatively, you can execute one run specifying the range of values for both restraints (**Intervals** set to 9, **Starting\_Angle** set to 0.0, and **Angle\_Size** set to 90.0). In the latter case, you must extract the appropriate energy values from the output file. For two restraints, defined at ten points each, 100 energy values (corresponding to all restraints) will be listed as output. Extract only those values for which the restraint targets on both angles are identical.

Use **Force Constant** set to the range of 2000–5000.

Switch to **Parameters** / **Minimize** and select Conjugate gradient algorithm with **Gradient tolerance** set to 0.001.

Note that **Parameters** / **Set** and **Parameters** / **Variables** are left at their default values. Now proceed to **Run**.

Another possibility is to use **Run** / **Files** to limit the number of output files. Before executing the **Run** / **Run** command check the restraints and selected minimization options using the **List** option.

After minimization, the dihedral angle might deviate somewhat from the value specified in the restraint. Save the final structure (both dihedral angles are around 90°) to biphenyl.psv using **File / Save\_Folder**.

You can view these structures (frames) with **Trajectory / Get** and **Trajectory / Conformation** from the **Analysis** module and determine the torsion angle value.

Plotting the profile should be done only after completing the next section of the assignment.

#### 4. Unrestrained minimization for biphenyl.

In addition to the restrained minimization calculations, perform unrestrained minimization to find the “global” energy minimum,  $E_{\min}$ , for biphenyl.

Now express the profile energy  $E$  from the previous section relative to the  $E_{\min}$  (i.e.,  $E - E_{\min}$ ), and plot against the dihedral angle for the full range  $[-180.0^\circ, 180.0^\circ]$ .

Note that  $E_{\min}$  may be larger than some  $E$  values. Why is that?

#### 5. Comparison of different force fields.

Repeat the energy profile calculations with the cff91 force field. Plot the results obtained with the AMBER and cff91 force fields on one plot and discuss your findings.

#### 6. Dependence on initial conditions.

Perform unrestrained minimization of biphenyl starting with the structure specified in the biphenyl.psv file from Part 3. Use the cff91 or AMBER force field and any minimization algorithm you wish, but use the Derivative tolerance of 0.001.

Describe the minimization algorithm briefly and discuss your results.

#### 7. Assessment of the performance of various minimization algorithms in Insight II.

For each minimization algorithm offered in Discover record the CPU time required for convergence of the energy gradient to the target values of 10.0, 0.1, 0.001, and 0.00001 kcal/Å. Use the AMBER force field.

For each algorithm, begin with the structure contained in the file biphenyl.distorted.car. Select the desired Algorithm from

**Parameters** / Minimize; set **Iterations** to 5000; specify the first target value of the derivative; execute; and proceed to execute the Run/Run command. After this job is completed, change the derivative tolerance to the next target value and repeat minimization. Extract the computational times and values of minima from each output file. Do not increase the number of **Iterations** above 5000. If the specified convergence is not reached with this threshold, note that in your report.

Repeat this procedure for each of the remaining algorithms. (Remember to start with the structure from `biphenyl_distorted.car` file.) Construct a table comparing the performance of minimization algorithms in the different regions of derivative tolerance (report the timing and energy minimum values).

On the basis of these results, and the information you have learned in class, suggest a simulation schedule to achieve an optimal minimization of a large molecule. Note that for our small system the gradient norm associated with the initial configuration of biphenyl is not extremely large.

### Background Reading from Coursepack

- M. Karplus and G. A. Petsko, “Molecular Dynamics Simulations in Biology”, *Nature* **347**, 631–639 (1990).

## Assignment 11: A Global Optimization Contest!

Our goal is to compute the lowest energy structure for the pentapeptide met-enkephalin, whose sequence is **Tyr–Gly–Gly–Phe–Met**. Many local minima exist for this molecule, so it is a challenge to reach the global minimum. *The student who finds the structure of the lowest energy will receive a prize from the instructor.*

The rules of this contest are:

1. use a molecule with *charged*  $\text{COO}^-$  and  $\text{NH}_3^+$  ends
2. use the **AMBER** force field
3. use the distance dependent dielectric constant (**Discover** module, **Parameters** / **Set** command, **Dist\_Dependent** button on)
4. use 1/2 as the scale factor for 1–4 nonbonded interactions  
(i.e., **Parameters** / **Scale\_Terms** command, **p1\_4** button on, and specify 0.5)

You can use *any* technique mentioned in this course (energy minimization, molecular dynamics, Monte Carlo sampling), as well as any other resources (e.g., web and literature), to find the global minimum of the pentapeptide.

### Be Creative.

Hand in a detailed report describing how you reached the minimum for met-enkephalin and any particular difficulties, or interesting observations, you encountered along the way. Attach the Cartesian coordinate file and the energy value reached.

Also submit a three-dimensional picture of the configuration of lowest energy along with a table specifying all associated bond lengths and bond angle values, and the  $\{\phi, \psi\}$  and  $\chi$  dihedral-angle values per residue.

To qualify for consideration of the prize, send electronically the coordinate file with the minimized structure to the instructor and TA.

Good Luck!

### Background Reading from Coursepack

- K. A. Dill and H. S. Chan, “From Levinthal to Pathways to Funnels”, *Nature Struc. Biol.* **4**, 10–19 (1997).
- T. Lazaridis and M. Karplus, “‘New View’ of Protein Folding Reconciled with the Old Through Multiple Unfolding Simulations”, *Science* **278**, 1928–1931 (1997).

## Assignment 12: Monte Carlo Simulations

1. **Random Number Generators.** Investigate the types of random number generators available on: (a) your local computing environment and (b) a mathematical package that you frequently use. How good are they? Is either one adequate for long molecular dynamics runs? Suggest how to improve them and test your ideas.

To understand some of the defects in linear congruential random number generators, consider the sequence defined by the formula  $y_{i+1} = (ay_i + c) \bmod M$ , with  $a = 65539$ ,  $M = 2^{31}$ , and  $c = 0$ . (This defines the infamous random number generator known as **RANDU** developed by IBM in the 1960s, which subsequent research showed to be seriously flawed). A relatively small number of numbers in the sequence (e.g., 2500) can already reveal a structure in three dimensions when triplets of consecutive random numbers are plotted on the unit cube. Specifically, plot consecutive pairs and triplets of numbers in two and three-dimensional plots, respectively, for an increasing number of generated random numbers in the sequence, e.g., 2500, 50,000, and 1 million. (Hint: Figure D.3 shows results from 2500 numbers in the sequence).

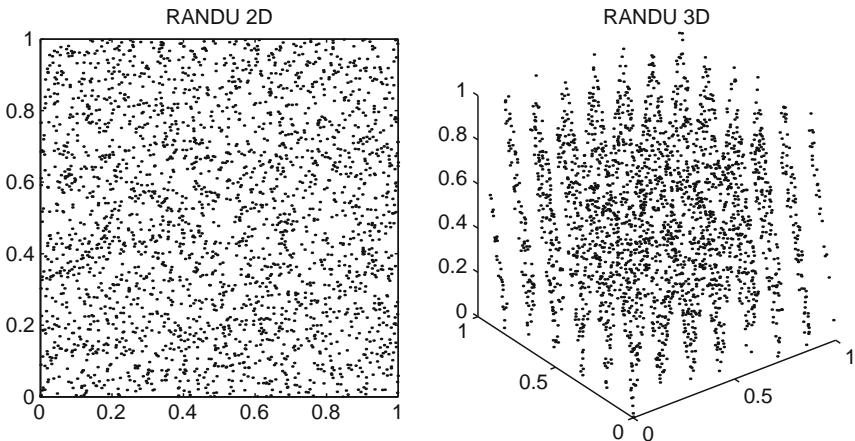


Figure D.3. Plots generated from pairs and triplets of consecutive points in the linear congruential generator known as **RANDU** defined by  $a = 65539$ ,  $M = 2^{31}$ , and  $c = 0$  when 2500 total points in the sequence are generated.

2. **MC Means.** Propose and implement a Monte Carlo procedure to calculate  $\pi$  based on integration. How many MC data points are needed to yield an answer correct up to 5 decimal places? What is the computational time involved? Show a table of your results displaying the number of MC steps, the associated  $\pi$  estimate, and the calculated error.

3. **Gaussian Variates.** You are stranded in an airport with your faithful laptop with one hour to spare until the deadline for emailing your homework assignment to your instructor. The assignment (next item) relies on a *Gaussian random number generator*, but you have forgotten the appropriate formulas involved in the commonly used Box/Muller/Marsaglia transformation approach. Fortunately, however, you remember the powerful Central Limit Theorem in basic probability and decide to form a random Gaussian variate by sampling  $N$  uniform random variates  $\{x_i\}$  on the unit interval as

$$\bar{y} = \sum_{i=1}^N x_i .$$

You quickly program the expression:

$$y = \sqrt{\frac{1}{\sigma^2(\bar{y})}} \sum_{i=1}^N [x_i - \mu(\bar{y})]$$

where above  $\sigma^2$  is the standard deviation of  $\bar{y} = N\sigma^2(x)$  and the mean  $\mu(\bar{y}) = N\mu(x)$ . [Recall that the uniform distribution has a mean of 1/2 and variance of 1/12].

How large should  $N$  be, you wonder. You must finish the assignment in a hurry. To have confidence in your choice, you set up some tests to determine when  $N$  is sufficiently large, and send your resulting routine, along with your testing reports, and results for several choices of  $N$ .

4. **Brownian Motion.** Now you can use the Gaussian variate generator above for propagating *Brownian motion* for a single particle governed by the bi-harmonic potential  $U(x) = kx^4/4$ . Recall that Brownian motion can be mimicked by simulating the following iterative process for the particle's position:

$$x^{n+1} = x^n + \frac{\Delta t}{m\gamma} F^n + R^n$$

where

$$\langle R^i R^j \rangle = \frac{2k_B T \Delta t}{m\gamma} \delta_{ij}, \quad \langle R^i \rangle = 0 .$$

Here  $m$  is the particle's mass;  $\gamma$  is the collision frequency, also equal to  $\xi/m$  where  $\xi$  is the frictional constant; and  $F$  is the systematic force. You are required to test the obtained mean square atomic fluctuations against the known result due to Einstein:

$$\langle x^2 \rangle = 2 \left( \frac{k_B T}{m\gamma} \right) t = 2D t ,$$

where  $D$  is the diffusion constant.

The following parameters may be useful to simulate a single particle of mass  $m = 4 \times 10^{-18}$  kg and radius  $a = 100$  nm in water: by Stokes' law, this particle's friction coefficient is  $\xi = 6\pi\eta a = 1.9 \times 10^{-9}$  kg/s, and  $D = k_B T / \xi = 2.2 \times 10^{-12} \text{m}^2/\text{s}$ . You may, however, need to scale the units appropriately to make the computations reasonable.

Plot the mean square fluctuations of the particle as a function of time, compare to the expected results, and show that for  $t \gg 1/\gamma = 2 \times 10^{-9}$  s the particle's motion is well described by random-walk or diffusion process.

### Background Reading from Coursepack

- T. Schlick, E. Barth, and M. Mandziuk, “Biomolecular Dynamics at Long Timesteps: Bridging the Timescale Gap Between Simulation and Experimentation”, *Ann. Rev. Biophys. Biomol. Struc.* **26**, 179–220 (1997).
- E. Barth and T. Schlick, “Overcoming Stability Limitations in Biomolecular Dynamics: I. Combining Force Splitting via Extrapolation with Langevin Dynamics in LN”, *J. Chem. Phys.* **109**, 1617–1632 (1998).
- L. S. D. Caves, J. D. Evanseck, and M. Karplus, “Locally Accessible Conformations of Proteins: Multiple Molecular Dynamics Simulations of Crambin”, *Prot. Sci.* **7**, 649–666 (1998).
- M. Karplus and J. A. McCammon, “Molecular Dynamics simulations of Biomolecules”, *Nat. Struc. Biol.* **9**, 307–340 (2003).
- M. Karplus and J. Kuriyan, “Molecular Dynamics and Protein Function”, *Proc. Natl. Acad. Sci. USA* **102**, 6679–6685 (2005).
- S. A. Adcock and J. A. McCammon, “Molecular dynamics: survey of methods for simulating the activity of proteins”, *Chem. Rev.* **106**: 1589–1615 (2006).
- E. H. Lee, J. Hsin, M. Sotomayor, G. Comellas, and K. Schulten, “Discovery Through the Computational Microscope”, *Structure* **17**: 1295–1306 (2009).

## Assignment 13: Advanced Exercises in Monte Carlo and Minimization Techniques

1. Study the function:

$$E(x, y) = ax^2 + by^2 + c(1 - \cos \gamma x) + d(1 - \cos \delta y). \quad (\text{D.11})$$

Note that it has many local minima and a global minimum at  $(x, y) = (0, 0)$ . Minimize  $E(x, y)$  with  $a = 1, b = 2, c = 0.3, \gamma = 3\pi, d = 0.4$ , and  $\delta = 4\pi$  by the standard simulated annealing method. Use the starting point  $(1, 1)$  and step perturbations  $\Delta x = 0.15$ , and set  $\beta$  in the range of 3.5 to 4.5. Limit the number of steps to  $\sim 150$ . Now implement the *variant* of the simulated annealing method where acceptance probabilities for steps with  $\Delta E < 0$  are proportional to  $\exp(-\beta E^g \Delta E)$ , with the exponent  $g = -1$ . Analyze and compare the efficiency of the searches in both cases. It will be useful to plot all pairs of points  $(x, y)$  that are generated by the method and distinguish ‘accepted’ from ‘rejected’ points.

2. Devise a different variant of the basic simulated annealing minimization method that would incorporate *gradient* information to make the searches more efficient.
3. Consider the following global optimization deterministic approach based on the *diffusion equation* as first suggested by Scheraga and colleagues (L. Piela, J. Kostrowicki, and H. A. Scheraga, ‘The Multiple-Minima Problem in Conformational Analysis of Molecules. Deformation of the Potential Energy Hypersurface by the Diffusion Equation Method’, *J. Chem. Phys.* **93**, 3339–3346 (1989)).

The basic idea is to deform the energy surface smoothly. That is, we seek to make “shallow” wells in the potential energy landscape disappear iteratively until we reach a global minimum of the deformed function. Then we “backtrack” by successive minimization from the global minimum of the transformed surface in the hope of reaching the global minimum of the real potential energy surface. This idea can be implemented by using the heat equation where  $T$  represents the temperature distribution in space  $x$ , and  $t$  represents time:

$$\frac{\partial^2 T}{\partial x^2} = \frac{\partial T}{\partial t} \quad (\text{D.12})$$

$$T(x, 0) = E(x). \quad (\text{D.13})$$

Here, the boundary condition at time  $t = 0$  equates the initial temperature distribution with the potential energy function  $E(x)$ . Under certain conditions (e.g.,  $E$  is bounded), a solution exists. Physically, the application of this equation exploits the fact that the heat flow (temperature distribution) should eventually settle down.

To formulate this idea, let us for simplicity consider first a one-dimensional problem where the energy function  $E$  depends on a scalar  $x$ . Let  $E^{(n)}(x)$  represent the  $n$ th derivative of  $E$  with respect to  $x$  and define the transformation operator  $\mathcal{S}$  on the energy function  $E$  for  $\beta > 0$  as follows:

$$\mathcal{S}[E(x)] = E(x) + \beta E^{(2)}(x). \quad (\text{D.14})$$

That is, we have:

$$\begin{aligned} \mathcal{S}^0 E &= E \\ \mathcal{S}^1 E &= E + \beta E^{(2)} \\ \mathcal{S}^2 E &= E + 2\beta E^{(2)} + \beta^2 E^{(4)} \\ \mathcal{S}^3 E &= E + 3\beta E^{(2)} + 3\beta^2 E^{(4)} + \beta^3 E^{(8)} \\ &\vdots && \vdots \\ \mathcal{S}^N E &= (1 + \beta d^2/dx^2)^N E. \end{aligned}$$

Now writing  $\beta = t/N$  where  $t$  is the time variable, and letting  $N \rightarrow \infty$ , we write:

$$\exp(td^2/dx^2) E \equiv \exp(A(t)) E = \left[ 1 + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \dots \right]. \quad (\text{D.15})$$

Thus we can define  $T(t)$  as

$$T(t) = \exp(A(t)) = \exp(td^2/dx^2). \quad (\text{D.16})$$

In higher dimensions, let  $x$  represent the collective vector of  $n$  independent variables; we replace the differential operator above  $d^2/dx^2$  by the *Laplacian operator*, that is

$$\Delta = \sum_{i=1}^n \partial^2/\partial x_i.$$

Using this definition, we can also write

$$T(t) = T_1(t) T_2(t) \cdots T_n(t)$$

where

$$T_i = \exp(t\partial^2/\partial x_i).$$

This definition produces the heat equation (D.12, D.13) since

$$\begin{aligned} \frac{\partial T(t)[E(x)]}{\partial t} &= \left[ \frac{dA}{dt} + \frac{2A}{2} \frac{dA}{dt} + \frac{3A^2}{3!} \frac{dA}{dt} + \dots \right] [E] \\ &= \left[ 1 + A + \frac{A^2}{2} + \dots \right] \frac{d^2}{dx^2}[E] \\ &= \frac{\partial^2 T(t)}{\partial x^2}[E(x)]. \end{aligned}$$

In practice, the diffusion equation method for global optimization is implemented by solving the heat equation by Fourier techniques (easy, for example, if we have dihedral potentials only) or by solving for  $T$  up to a sufficiently large time  $t$ . This solution, or approximate solution (representing  $E(x, t)$  for some large  $t$ ), is expected to yield a deformed surface with one (global) minimum. With a local minimization algorithm, we compute the global minimum  $x^*$  of the deformed surface, and then begin an iterative deformation/minimization procedure from  $x^*$  and  $E(x, t)$  so that at each step we deform backwards the potential energy surface and obtain its associated global minimum ( $E(x, t) \rightarrow E(x, t - \Delta t)$  and  $x^* \rightarrow x^1$ ,  $E(x, t - \Delta t) \rightarrow E(x, t - 2\Delta t)$  and  $x^1 \rightarrow x^2$ ,  $\dots$   $E(x, 0) \rightarrow x^k$ ). Of course, depending on how the backtracking is performed, different final solutions can be obtained.

- (a) To experiment with this interesting diffusion-equation approach for global minimization, derive a general form for the deformation operator  $T(t) = \exp(td^2/dx^2)$  on the following special functions  $E(x)$ : (i) polynomial functions of degree  $n$ , and (ii) trigonometric functions  $\sin\omega x$  and  $\cos\omega x$ , where  $\omega$  is a real-valued number (frequency). What is the significance of your result for (ii)?
- (b) Apply the deformation operator  $T(t) = \exp(td^2/dx^2)$  to the quadratic function

$$E(x) = x^4 + ax^3 + bx^2, \quad (\text{D.17})$$

with  $a = 3$  and  $b = 1$ . Evaluate and plot your resulting  $T(t)E(x)$  function at  $t = 0, \Delta t, 2\Delta t, \dots$ , for small time increments  $\Delta t$  until the global minimum is obtained.

- (c) Apply the deformation operator  $T(t)$  for the two-variable function in eq. (D.11). Examine behavior of the deformation as  $t \rightarrow \infty$  as a function of the constants  $a$  and  $b$ . Under what conditions will a unique minimum be obtained as  $t \rightarrow \infty$ ?
4. Use Newton minimization to find the minimum of the two-variable function in equation (D.11) and the one-variable function in equation (D.17). It is sufficient for the line search to use simple bisection:  $\lambda = 1, 0.5$ , etc., or some other simple backtracking strategy. For the quartic function, experiment with various starting points.

**Read remaining paper from Coursepack**

## Assignment 14: Advanced Exercises in Molecular Dynamics

1. Calculate the ‘natural’ time unit for molecular dynamics simulations of biomolecules from the relation: energy = mass \* (length/time)<sup>2</sup>, to obtain the time unit  $\tau$  corresponding to the following units:

$$\begin{aligned} \text{length} & \quad (\text{l}): \quad 1 \text{ \AA} = 10^{-10} \text{ m} \\ \text{mass} & \quad (\text{m}): \quad 1 \text{ amu} = 1 \text{ g/mol} \\ \text{energy} & \quad (\text{v}): \quad 1 \text{ kcal/mol} = 4.184 \text{ kJ/mol}. \end{aligned}$$

Estimate the “quantum mechanical cutoff frequency”,  $\omega_c = kT/\hbar$  at room temperature ( $\sim 300^\circ\text{K}$ ).

2. Derive the amplitude decay rate of  $\gamma/2$  for an underdamped harmonic oscillator due to *friction* by solving the equations of motion:

$$m \frac{d^2x}{dt^2} = -kx - m\gamma \frac{dx}{dt} \quad (\text{D.18})$$

and examining time behavior of the solution.

3. Derive the amplitude decay rate of  $\omega^2(\Delta t)/2$  *intrinsic* to the *implicit-Euler* scheme by solving the discretized form of eq. (D.18).
4. Compare your answer in problem 2 above with behavior of the *explicit-Euler* solution of eq. (D.18).
5. Compare molecular and Langevin dynamics simulations of two water molecules by the Verlet discretization of the equation of motion and its Langevin analog. Use the “SPC” *intermolecular* potential, given by:

$$E = \sum_{\substack{(i,j) \equiv (\text{O},\text{O)} \\ i < j}} \left[ \frac{-A}{r_{ij}^6} + \frac{B}{r_{ij}^{12}} \right] + \sum_{\substack{(k,\ell) \equiv \text{intermolecular} \\ (\text{O},\text{O}), (\text{O},\text{H}), (\text{H},\text{H}) \text{ pairs} \\ k < \ell}} \left[ \frac{Q_k Q_\ell}{r_{k\ell}} \right]$$

where

$$\begin{aligned} A &= 626 \text{ (kcal \AA}^6)/\text{mol} \\ B &= 629 \times 10^3 \text{ (kcal \AA}^{12})/\text{mol} \\ Q_{\text{H}} &= 0.41 e \\ Q_{\text{O}} &= -0.82 e. \end{aligned}$$

A numerical factor of 332 is needed in the electrostatic potential to obtain energies in kcal/mol with the coefficients above. For simplicity, assume that *intramolecular* geometries are rigid:  $r_{\text{OH}} = 1 \text{ \AA}$ ,  $\cos \theta_{\text{HOH}} = -1/3$ . (You can use harmonic soft constraints). Begin by first minimizing the energy of the water dimer and examining the hydrogen bond geometry

(hydrogen-bond distance and angle  $\theta$  between the hydrogen-bond vector and bisector of the acceptor molecule). Then study numerical behavior of the two models/schemes as a function of  $\Delta t$ , and examine the hydrogen bond geometry. Experiment with  $\Delta t = 1, 2, 5$ , and  $10$  fs and use  $\gamma$  values in the range of 1 to  $50$  ps $^{-1}$ . If you are more ambitious, continue to study energy-minimized structures of water clusters of larger sizes and their dynamics. Analyze the hydrogen bonding networks of these clusters.

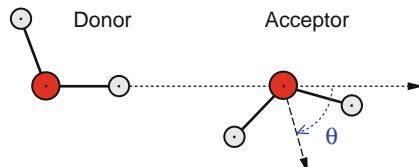


Figure D.4. Hydrogen bond geometry: the angle  $\theta$  is defined between the hydrogen-bond vector and the bisector of the acceptor molecule.

### Some Useful Constants and Conversion Factors

---

Avogadro's Number	$N_A = 6.0221 \times 10^{23} \text{ mol}^{-1}$
Planck's Constant	$h = 6.6261 \times 10^{-34} \text{ Jsec}$
	$\hbar = h/2\pi = 1.055 \times 10^{-34} \text{ Jsec}$
Boltzmann's Constant	$k_B = 1.38066 \times 10^{-23} \text{ JK}^{-1}$
Gas Constant	$R = k_B N_A = 8.3145 \text{ JK}^{-1} \text{ mol}^{-1}$
Atomic Mass Unit, amu	$(1/N_A) = 1 \text{ g/mol} = 1.6605 \times 10^{-27} \text{ kg}$
	$\pi = 3.14159$
	1 kcal = 4.184 kJ

---

## Assignment 15: BONUS PROJECT

### The Scaling of Protein-Conformer Number with Size and Solvability of the Protein Folding Problem

The phone rings one morning as you sip through your Sunluck vanilla latté grande and check your emails in your office at the University of Seabeetle. The editor of the journal *Proteomics Today* is on the line.

Given your expertise in biomolecular modeling, she asks your help in writing a brief *Folding In Silico Perspectives* article for the next issue discussing the following series of interesting papers debating the nature of scaling of the number of protein conformers as function of chain length (exponential or nonexponential) and the solvability of the protein folding problem by computer simulation:

- W. F. van Gunsteren, R. Bürgi, C. Peter, and X. Daura, “The Key to Solving the Protein-Folding Problem Lies in an Accurate Description of the Denatured State”, *Angew. Chem. Int. Ed.* **40**, 352–355 (2001).
- A. R. Dinner and M. Karplus, “Comment on the Communication ‘The Key to Solving the Protein-Folding Problem Lies in an Accurate Description of the Denatured State’ by van Gunsteren et al.”, *Angew. Chem. Int. Ed.* **40**, 4615–4616 (2001).
- W. F. van Gunsteren, R. Bürgi, C. Peter, and X. Daura, “Reply”, *Angew. Chem. Int. Ed.* **40**, 4616–4618 (2001).

Since many of *Proteomics Today* readers and authors perform computer simulations of proteins, both macroscopic and all-atom based, the editor wants you also to mention the strengths and weaknesses of these different approaches.

Though already overloaded with preparing final examinations for your classes, writing grant proposals, supervising your students and postdocs, and reviewing several articles for journals (long-overdue), you agree to take this challenging and potentially rewarding assignment. You immerse yourself in the papers, send your graduate students to collect background articles and information, order through your cellphone Sunluck’s caramel-macchiato enormoso, and decide to make your article not only objective and interesting but also fun to read.

#### Some Suggestions:

- Discuss the Levinthal paradox.
- What is the relation among the number of conformers, timescales, and folding pathways?
- Analyze lattice or other protein simulations reported in the literature to estimate the number of possible conformers and the timescale of protein folding.

# References

- [1] L. Adams and J. L. Nazareth, editors. *Linear and Nonlinear Conjugate Gradient-Related Methods*. SIAM, Philadelphia, PA, 1996.
- [2] M. D. Adams, G. G. Sutton, H. O. Smith, E. W. Myers, and J. C. Venter. The independence of our genome assembly. *Proc. Natl. Acad. Sci. USA*, 100: 3025–3026, 2003.
- [3] P. L. Adams, M. R. Stahley, A. B. Kosek, J. Wang, and S. A. Strobel. Crystal structure of a self-splicing group I intron with both exons. *Nature*, 430:45–50, 2005.
- [4] S. A. Adcock and J. A. McCammon. Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem. Rev.*, 106:1589–1615, 2006.
- [5] D. K. Agrafiotis. Stochastic algorithms for maximizing molecular diversity. *J. Chem. Inf. Comput. Sci.*, 37:841–851, 1997.
- [6] D. K. Agrafiotis. Diversity of chemical libraries. In P. von Ragué Schleyer (Editor-in-Chief), N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, and H. F. Schaefer, III, editors, *Encyclopedia of Computational Chemistry*, volume 1, pages 742–761. John Wiley & Sons, West Sussex, England, 1998.
- [7] D. K. Agrafiotis, V. S. Lobanov, and F. R. Salemme. Combinatorial informatics in the post-genomics era. *Nat. Rev. Drug. Disc.*, 1:337–346, 2002.
- [8] D. K. Agrafiotis, J. C. Myslik, and F. R. Salemme. Advances in diversity profiling and combinatorial series design. *Mol. Div.*, 4:1–22, 1999.
- [9] A. Aguzzi. Prions and antiprions. *Biol. Chem.*, 378:1393–1395, 1997.
- [10] A. Aguzzi, F. Montrasio, and P. S. Kaeser. Prions: Health scare and biological challenge. *Nature Rev. Mol. Cell Biol.*, 2:118–126, 2001.
- [11] P. Ahlquist. RNA-dependent RNA polymerases, viruses, and RNA silencing. *Science*, 296:1270–1273, 2002.

- [12] J. Aishima, R. K. Gitti, J. E. Noah, H. H. Gan, T. Schlick, and C. Wolberger. A Hoogsteen base pair embedded in undistorted B-DNA. *Nuc. Acids Res.*, 30:5244–5252, 2002.
- [13] J. Aishima and C. Wolberger. Crystal structure of the MAT $\alpha$ 2 homeodomain-DNA complex with nonspecifically bound homeodomains. *Nuc. Acids Res.*, 2002.
- [14] A. Aiuti, F. Cattaneo, S. Galimberti, U. Benninghoff, B. Cassani, L. Callegaro, S. Scaramuzza, G. Andolfi, M. Mirolo, I. Brigida, A. Tabucchi, F. Carlucci, M. Eibl, M. Aker, S. Slavin, H. Al-Mousa, A. Al Ghonaium, A. Ferster, A. Duppenthaler, L. Notarangelo, U. Wintergerst, R. Buckley, M. Bregnii, S. Marktel, M. Valsecchi, P. Rossi, F. Ciceri, R. Miniero, C. Bordignon, and M. Roncarolo. Gene therapy for immunodeficiency due to adenosine deaminase deficiency. *New Engl. J. Med.*, 360:447–458, 2009.
- [15] E. Akhmedskaya, N. Bou-Rabee, and S. Reich. A comparison of generalized hybrid Monte Carlo methods with and without momentum flip. *J. Comput. Phys.*, 228:2256–2265, 2009.
- [16] B. Al-Lazikani, J. Jung, Z. Xiang, and B. Honig. Protein structure prediction. *Curr. Opin. Struct. Biol.*, 5:51–56, 2001.
- [17] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science Publishing, New York, NY, fifth edition, 2008.
- [18] I. L. Alberts, Y. Wang, and T. Schlick. DNA polymerase  $\beta$  catalysis: Are different mechanisms possible? *J. Amer. Chem. Soc.*, 129:11100–11110, 2007.
- [19] B. J. Alder and T. E. Wainwright. Studies in molecular dynamics. I. General method. *J. Chem. Phys.*, 31:459–466, 1959.
- [20] C. Alhambra, F. J. Luque, F. Gago, and M. Orozco. *Ab Initio* study of stacking interactions in A- and B-DNA. *J. Phys. Chem. B*, 101:3846–3853, 1997.
- [21] J. F. Allemand, D. Bensimon, R. Lavery, and V. Croquette. Stretched and overwound DNA forms a Pauling-like structure with exposed bases. *Proc. Natl. Acad. Sci. USA*, 95:14152–14157, 1998.
- [22] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Oxford University Press, New York, NY, 1990.
- [23] R. J. Allen, D. Frenkel, and P. R. ten Wolde. Simulating rare events in equilibrium or nonequilibrium stochastic systems. *J. Chem. Phys.*, 124:024102, 2006.
- [24] N. L. Allinger. Calculation of molecular structure and energy by force-field methods. *Adv. Phys. Org. Chem.*, 13:1–85, 1976.
- [25] N. L. Allinger. Conformational analysis. 130. MM2 A hydrocarbon force field utilizing V1 and V2 torsional terms. *J. Amer. Chem. Soc.*, 99:8127–8134, 1977.
- [26] N. L. Allinger and K. Chen. Hyperconjugative effects on carbon-carbon bond lengths in molecular mechanics (MM4). *J. Comput. Chem.*, 17:747–755, 1996.
- [27] N. L. Allinger, J. T. Fermann, W. D. Allen, and H. F. Schaefer, III. The torsional conformations of butane: Definitive energetics from *ab initio* methods. *J. Chem. Phys.*, 106:5143–5150, 1997.
- [28] N. L. Allinger, M. A. Miller, F. A. VanCatledge, and J. A. Hirsch. Conformational analysis. LVII. The calculation of the conformational structures of hydrocarbons by the Westheimer-Hendrickson-Wiberg method. *J. Amer. Chem. Soc.*, 89:4345–4357, 1967.

- [29] N. L. Allinger and J. T. Sprague. Calculation of the structures of hydrocarbons containing delocalized electronic systems by the molecular mechanics method. *J. Amer. Chem. Soc.*, 95:3893–3907, 1973.
- [30] N. L. Allinger, Y. H. Yuh, and J.-H. Lii. Molecular mechanics. The MM3 force field for hydrocarbons. 1. *J. Amer. Chem. Soc.*, 111:8551–8566, 1989.
- [31] C. D. Allis, T. Jenuwein, and D. Reinberg, editors. *Epigenetics*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2009.
- [32] S. A. Allison and J. A. McCammon. Multistep Brownian dynamics: Application to short wormlike chains. *Biopolymers*, 23:363–375, 1984.
- [33] U. Alon. Network motifs: theory and experimental approaches. *Nat. Rev. Genet.*, 8:450–461, 2007.
- [34] A. Altis, M. Otten, P. H. Nguyen, R. Hegger, and G. Stock. Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis. *J. Chem. Phys.*, 128:245102, 2008.
- [35] C. Altona and M. Sundaralingam. Conformational analysis of the sugar ring in nucleosides and nucleotides. a new description using the concept of pseudorotation. *J. Amer. Chem. Soc.*, 94:8205–8212, 1972.
- [36] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen. Essential dynamics of proteins. *Proteins: Struc. Func. Gen.*, 17:412–425, 1993.
- [37] A. Amadei, A. B. M. Linssen, B. L. deGroot, D. M. F. van Aalten, and H. J. C. Berendsen. An efficient method for sampling the essential subspace of proteins. *J. Biomol. Struct. Dynam.*, 13:615–625, 1996.
- [38] J. G. Amar. The Monte Carlo method in science and engineering. *Comput. Sci. Engin.*, 8:9–19, 2006. doi:10.1109/MCSE.2006.34.
- [39] A. Amir-Aslani. Toxicogenomic predictive modeling: Emerging opportunities for more efficient drug discovery and development. *Tech. Forecast. Soc. Change*, 75:905–932, 2008.
- [40] M. Amos. *Theoretical and Experimental DNA Computation*. Natural Computing Series. Springer, New York, NY, 2005.
- [41] L. M. Amzel. Structure-based drug design. *Curr. Opin. Biotech.*, 9:366–369, 1998.
- [42] K. Anand, J. Ziebuhr, P. Wadhwani, J. R. Mesters, and R. Hilgenfeld. Coronavirus main proteinase ( $3CL^{PRO}$ ) structure: Basis for design of anti-SARS drugs. *Science*, 300:1763–1767, 2003.
- [43] H. C. Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.*, 72:2384–2393, 1980.
- [44] H. C. Andersen. Rattle: a ‘velocity’ version of the SHAKE algorithm for molecular dynamics calculations. *J. Comput. Phys.*, 52:24–34, 1983.
- [45] C. R. Anderson. An implementation of the fast multipole algorithm without multipoles. *SIAM J. Sci. Stat. Comput.*, 13:923–947, 1992.
- [46] S. L. Anderson. Random number generators on vector supercomputers and other advanced architectures. *SIAM Rev.*, 32:221–251, 1990.
- [47] A. Andreeva, D. Howorth, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin. SCOP database in 2004: Refinements integrate structure and sequence family data. *Nucl. Acids Res.*, 32:D226–D229, 2004.

- [48] A. Andreeva, D. Howorth, J.-M. Chandonia, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin. Data growth and its impact on the SCOP database: New developments. *Nucl. Acids Res.*, 36:D419–D425, 2008.
- [49] G. Andres Cisneros, L. Perera, M. Garcia-Diaz, K. Bebenek, T. A Kunkel, and L. G. Pedersen. Catalytic mechanism of human DNA polymerase  $\lambda$  with Mg<sup>2+</sup> and Mn<sup>2+</sup> from *ab initio* quantum mechanical/molecular mechanical studies. *DNA Repair*, 7:1824–1834, 2008.
- [50] M. Andronescu, V. Bereg, H.H. Hoos, and A. Condon. RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinformatics*, 9:340, 2008.
- [51] C. B. Anfinsen, E. Haber, M. Sela, and F. H. White, Jr. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. USA*, 47:1309–1314, 1961.
- [52] C. B. Anfinsen and H. A. Scheraga. Experimental and theoretical aspects of protein folding. *Adv. Protein Chem.*, 29:205–301, 1975.
- [53] A. W. Appel. An efficient program for many-body simulation. *SIAM J. Sci. Stat. Comput.*, 6:85–103, 1985.
- [54] J. Åqvist and A. Warshel. Simulation of enzyme reactions using valence bond force fields and other hybrid quantum/classical approaches. *Chem. Rev.*, 93:2523–2544, 1993.
- [55] G. Arents, R. W. Burlingame, B. C. Wang, W. E. Love, and E. N. Moudrianakis. The nucleosomal core histone octamer at 3.1 Å resolution: A tripartite protein assembly and a left-handed superhelix. *Proc. Natl. Acad. Sci. USA*, 88:10148–10152, 1991.
- [56] A. Arkhipov, P. L. Freddolino, and K. Schulten. Stability and dynamics of virus capsids described by coarse-grained modeling. *Structure*, 14:1767–1777, 2006.
- [57] A. Arkhipov, Y. Yin, and K. Schulten. Four-scale description of membrane sculpting by BAR domains. *Biophys. J.*, 95:2806–2821, 2008.
- [58] Y. A. Arnaudova, A. Jagielska, and H. A. Scheraga. A new force field (ECEPP-05) for peptides, proteins, and organic molecules. *J. Phys. Chem. B*, 110:5025 W5044, 2006.
- [59] K. Arora and T. Schlick. Deoxyadenosine sugar puckering pathway simulated by the stochastic difference equation algorithm. *Chem. Phys. Lett.*, 378:1–8, 2003.
- [60] K. Arora and T. Schlick. *In Silico* evidence for DNA polymerase  $\beta$ 's substrate-induced conformational change. *Biophys. J.*, 87:3088–3099, 2004.
- [61] P. G. Arscott, G. Lee, V. A. Bloomfield, and D. F. Evans. Scanning tunneling microscopy of Z-DNA. *Nature*, 339:484–486, 1989.
- [62] G. Arya and T. Schlick. Role of histone tails in chromatin folding revealed by a mesoscopic oligonucleosome model. *Proc. Natl. Acad. Sci. USA*, 103:16236–16241, 2006.
- [63] G. Arya and T. Schlick. Efficient global biopolymer sampling with end-transfer configurational bias Monte Carlo. *J. Chem. Phys.*, 106:044107, 2007.
- [64] G. Arya and T. Schlick. A tale of tails: How histone tails mediate chromatin compaction in different salt and linker histone environments. *J. Phys. Chem. A*, 113:4045–4059, 2009.

- [65] G. Arya, Q. Zhang, and T. Schlick. Flexible histone tails in a new mesoscopic oligonucleosome model. *Biophys. J.*, 91:133–150, 2006.
- [66] K. Ashrafi, F. Y. Chang, J. L. Watts, A. G. Fraser, R. S. Kamath, J. Ahringer, and G. Ruvkun. Genome-wide RNAi analysis of *Caenorhabditis elegans* fat regulatory genes. *Nature*, 421:268–272, 2003.
- [67] A. Askar, B. Space, and H. Rabitz. Subspace method for long time scale molecular dynamics. *J. Phys. Chem.*, 99:7330–7338, 1995.
- [68] J. F. Atkins and R. Gesteland. The 22nd amino acid. *Science*, 296:1409–1410, 2002.
- [69] P. Auffinger, S. Louise-May, and E. Westhof. Multiple molecular dynamics simulations of the anticodon loop of rRNA<sup>Asp</sup> in aqueous solution with counterions. *J. Amer. Chem. Soc.*, 117:6720–6726, 1995.
- [70] P. Auffinger, S. Louise-May, and E. Westhof. Molecular dynamics simulations of the anticodon hairpin of tRNA<sup>Asp</sup>. Structuring effects of C–H· · · O hydrogen bonds and long-range hydration forces. *J. Amer. Chem. Soc.*, 118:1181–1189, 1996.
- [71] P. Auffinger and E. Westhof. Molecular dynamics simulations of nucleic acids. In P. von Ragué Schleyer (Editor-in Chief), N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, and H. F. Schaefer, III, editors, *Encyclopedia of Computational Chemistry*, volume 3, pages 1628–1639. John Wiley & Sons, West Sussex, England, 1998.
- [72] P. Auffinger and E. Westhof. Simulations of the molecular dynamics of nucleic acids. *Curr. Opin. Struct. Biol.*, 8:227–236, 1998.
- [73] P. Auffinger and E. Westhof. RNA solvation: A molecular dynamics simulation perspective. *Biopolymers*, 56:266–274, 2001.
- [74] M. S. Babcock and W. K. Olson. The effect of mathematics and coordinate system on comparability and “dependencies” of nucleic acid structure parameters. *J. Mol. Biol.*, 237:98–124, 1994.
- [75] D. Backowies and W. Thiel. Hybrid models for combined quantum mechanical and molecular mechanical approaches. *J. Phys. Chem.*, 100:10580–10594, 1996.
- [76] S. Bagheri and M. Kashani-Sabet. Ribozymes in the age of molecular therapeutics. *Curr. Mol. Med.*, 4:489–506, 2004.
- [77] I. Bahar, A. R. Atilgan, and B. Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.*, 2:173–181, 1997.
- [78] S.D. Baird, M. Turcotte, R.G. Korneluk, and M. Holcik. Searching for IRES. *RNA*, 12:1755–1785, 2006.
- [79] D. Baker and A. Sali. Protein structure prediction and structural genomics. *Science*, 294:93–96, 2001.
- [80] N. Baker, M. Holst, and E. Wang. Adaptive multilevel finite element solution of the Poisson-Boltzmann equation II. Refinement at solvent-accessible surfaces in biomolecular systems. *J. Comput. Chem.*, 21:1343–1352, 2000.
- [81] N. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. USA*, 98:10037–10041, 2001.
- [82] Y. Bakhtin and C. E. Heitche. Large deviations of random trees. *J. Stat. Phys.*, 132:551–560, 2008.

- [83] D. Baltimore. The preoccupations of twenty-first-century biology. In A. H. Zewail, editor, *Physical Biology: From Atoms to Medicine*, pages 1–6. Imperial College Press, London, UK, 2008.
- [84] J. Balzarini and L. V. Damme. Microbicide drug candidates to prevent HIV infection. *The Lancet*, 369:787–797, 2007.
- [85] N. Ban, P. Nissen, J. Hansen, P. B. Moore, and T. A. Steitz. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, 289:905–920, 2000.
- [86] P. Banáš, P. Jurečka, N. G. Walter, J. Šponer, and M. Otyepka. Theoretical studies of RNA catalysis: Hybrid QM/MM methods and their comparison with MD and QM. *Methods*, 49:202–216, 2009.
- [87] L. Banci, W. Baumeister, U. Heinemann, G. Schneider, I. Silman, D. I Stuart, and J. L. Sussman. An idea whose time has come. [A response to **an idea whose time has gone** by G. A. Petsko]. *Genome Biol.*, 8:107, 2007.
- [88] Y. Bao, C. L. White, and K. Luger. Nucleosome core particles containing a poly(dA•dT) sequence element exhibit a locally distorted DNA structure. *J. Mol. Biol.*, 361:617–624, 2006.
- [89] D. Barash, X. Qian, L. Yang, and T. Schlick. Inherent speedup limitations in multiple-timestep/particle-mesh-Ewald algorithms. *J. Comput. Chem.*, 24:77–88, 2003.
- [90] J. Barnes and P. Hut. A hierarchical  $O(N \log N)$  force-calculation algorithm. *Nature*, 324:446–449, 1986.
- [91] D. Barouch. Challenges in the development of an HIV-1 vaccine. *Nature*, 455:613–619, 2008.
- [92] E. Barth, K. Kuczera, B. Leimkuhler, and R. D. Skeel. Algorithms for constrained molecular dynamics. *J. Comput. Chem.*, 16:1192–1209, 1995.
- [93] E. Barth, M. Mandziuk, and T. Schlick. A separating framework for increasing the timestep in molecular dynamics. In W. F. van Gunsteren, P. K. Weiner, and A. J. Wilkinson, editors, *Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications*, volume III, chapter 4, pages 97–121. ESCOM, Leiden, The Netherlands, 1997.
- [94] E. Barth and T. Schlick. Extrapolation versus impulse in multiple-timestepping schemes: II. Linear analysis and applications to Newtonian and Langevin dynamics. *J. Chem. Phys.*, 109:1632–1642, 1998.
- [95] E. Barth and T. Schlick. Overcoming stability limitations in biomolecular dynamics: I. Combining force splitting via extrapolation with Langevin dynamics in LN. *J. Chem. Phys.*, 109:1617–1632, 1998.
- [96] D. Bashford and D. A. Case. Generalized Born models of macromolecular solvation effects. *Ann. Rev. Phys. Chem.*, 51:129–152, 2000.
- [97] P. Batcho, D. A. Case, and T. Schlick. Optimized particle-mesh Ewald / multiple-timestep integration for molecular dynamics simulations. *J. Chem. Phys.*, 115:4003–4018, 2001.
- [98] P. Batcho and T. Schlick. New splitting formulations for lattice summations. *J. Chem. Phys.*, 115:8312–8326, 2001.

- [99] P. Batcho and T. Schlick. Special stability advantages of position Verlet over velocity Verlet in multiple-timestep integration. *J. Chem. Phys.*, 115:4019–4029, 2001.
- [100] A. D. Bates and A. Maxwell. *DNA Topology*. In Focus. Oxford University Press, New York, NY, 1993.
- [101] R. T. Batey and J. A. Doudna. The parallel universe of RNA folding. *Nature Struct. Biol.*, 5:337–340, 1998.
- [102] R. T. Batey, R. P. Rambo, and J. A. Doudna. Tertiary motifs in RNA structure and folding. *Angew. Chem. Int. Ed.*, 38:2326–2343, 1999.
- [103] W. R. Bauer and C. J. Benham. The free energy, enthalpy, and entropy of native and of partially denatured closed circular DNA. *J. Mol. Biol.*, 234:1184–1196, 1993.
- [104] C. G. Baumann, S. B. Smith, V. A. Bloomfield, and C. Bustamante. Ionic effects on the elasticity of single DNA molecules. *Proc. Natl. Acad. Sci. USA*, 94:6185–6190, 1997.
- [105] M. D. Beachy, D. Chasman, R. B. Murphy, T. A. Halgren, and R. A. Friesner. Accurate Ab initio quantum chemical determination of the relative energetics of peptide conformations and assessment of empirical force fields. *J. Amer. Chem. Soc.*, 119:5908–5920, 1997.
- [106] D. Beard and T. Schlick. Inertial stochastic dynamics: I. long-timestep methods for langevin dynamics. *J. Chem. Phys.*, 112:7313–7322, 2000.
- [107] D. Beard and T. Schlick. Inertial stochastic dynamics: II. influence of inertia on slow kinetic properties of supercoiled DNA. *J. Chem. Phys.*, 112:7323–7338, 2000.
- [108] D. Beard and T. Schlick. Computational modeling predicts the structure and dynamics of the chromatin fiber. *Structure*, 9:105–114, 2001.
- [109] D. Beard and T. Schlick. Modeling salt-mediated electrostatics of macromolecules: The algorithm DiSCO (Discrete Charge Surface Charge Optimization) and its application to the nucleosome. *Biopolymers*, 58:106–115, 2001.
- [110] J. C. Beauchamp and N. W. Isaacs. Methods for X-ray diffraction analysis of macromolecular structures. *Curr. Opin. Chem. Biol.*, 3:525–529, 1999.
- [111] O. M. Becker, D. S. Dhanoa, Y. Marantz, D. Chen, S. Shacham, S. Cheruku, A. Heifetz, P. Mohanty, M. Fichman, and A. Sharadendu. An integrated in silico 3D model-driven discovery of a novel, potent, and selective amidosulfonamide 5-HT1A agonist (PRX-00023) for the treatment of anxiety and depression. *J. Med. Chem.*, 49:3116–3135, 2006.
- [112] D. Beglov and B. Roux. Dominant solvations effects from the primary shell of hydration: Approximation for molecular dynamics simulations. *Biopolymers*, 35:171–178, 1994.
- [113] D. Beglov and B. Roux. Finite representation of an infinite bulk system: Solvent boundary potential for computer simulations. *J. Chem. Phys.*, 100:9050–9063, 1994.
- [114] D. Beglov and B. Roux. Numerical solutions of the hypernetted chain equation for a solute of arbitrary geometry in three dimensions. *J. Chem. Phys.*, 103:360–364, 1995.

- [115] J. Bella, B. Brodsky, and H. M. Berman. Hydration structure of a collagen peptide. *Structure*, 3:893–906, 1995.
- [116] G. Benedetti and S. Moroletti. A graph-topological approach to recognition of pattern and similarity in RNA secondary structures. *Biophys. Chem.*, 59:179–184, 1996.
- [117] Y. Benenson, B. Gil, U. B.-D., R. Adar, and E. Shapiro. An autonomous molecular computer for logical control of gene expression. *Nature*, 429:423–429, 2004.
- [118] H. A. Benjamin and N. R. Cozzarelli. DNA-directed synapsis in recombination: Slithering and random collision of sites. *Proc. R. A. Welch Found. Conf. Chem. Res.*, 29:107–126, 1986.
- [119] K. R. Benjamin, A. P. Abola, R. Kanaar, and N. R. Cozzarelli. Contributions of supercoiling to Tn3 resolvase and phage Mu Gin site-specific recombination. *J. Mol. Biol.*, 256:50–65, 1996.
- [120] D. Bensimon, A. Simon and. A. Chiffaudel, V. Croquette, and A. Bensimon. Alignment and sensitive detection of DNA by a moving interface. *Science*, 265:2096–2098, 1994.
- [121] H. J. C. Berendsen. A glimpse of the holy grail. *Science*, 282:642–643, 1998.
- [122] H. J. C. Berendsen. *Simulating the Physical World: Hierarchical Modeling from Quantum Mechanics to Fluid Dynamics*. Cambridge University Press, Cambridge, UK, 2007.
- [123] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684–3690, 1984.
- [124] H. J. C. Berendsen, D. van der Spoel, and R. van Drunen. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Comm.*, 91:43–56, 1995.
- [125] H. M. Berman. Hydration of DNA: Take 2. *Curr. Opin. Struct. Biol.*, 4:345–350, 1994.
- [126] H. M. Berman. Crystal studies of B-DNA: The answers and the questions. *Biopolymers*, 44:23–44, 1997.
- [127] H. M. Berman, W. K. Olson, D. L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, S.-H. Hsieh, A. R. Srinivasan, and B. Schneider. The nucleic acid database: A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.*, 63:751–759, 1992.
- [128] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res.*, 28:235–242, 2000.
- [129] B. J. Berne and J. E. Straub. Novel methods of sampling phase space in the simulation of biological systems. *Curr. Opin. Struct. Biol.*, 7:181–189, 1997.
- [130] S. Bernèche and B. Roux. Energetics of ion conduction through the K<sup>+</sup> channel. *Nature*, 414:73–77, 2001.
- [131] R. S. Berry, S. A. Rice, and J. Ross. *Physical Chemistry*. Wiley, New York, NY, 1980.
- [132] R. B. Best, N.-V. Buchete, and G. Hummer. Are current molecular dynamics force fields too helical? *Biophys. J.*, 95:L07–L09, 2008.

- [133] D. L. Beveridge, G. Barreiro, K. S. Byun, D. A. Case, T. E. Cheatham, III, S. B. Dixit, E. Giudice, F. Lankas, R. Lavery, J. H. Maddocks, R. Osman, E. Seibert, H. Sklenar, G. Stoll, K. M. Thayer, P. Varnai, and M. A. Young. Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I: Research design and results on d(C<sub>p</sub>G) steps. *Biophys. J.*, 87:3799–3813, 2004.
- [134] J. J. Biesiadecki and R. D. Skeel. Dangers of multiple-time-step methods. *J. Comput. Phys.*, 109:318–328, 1993.
- [135] M. Bixon, H. Dekker, J. D. Dunitz, E. Eser, S. Lifson, C. Mosselman, J. Sicher, and M. Svoboda. Structural and strain-energy consequences of “intra-annular” substitution in the cyclodecane ring. *Chem. Commun.*, pages 360–362, 1967.
- [136] P. Bjelkmar, P. S. Niemelä, I. Vattulainen, and E. Lindahl. Conformational changes and slow dynamics through microsecond polarized atomistic molecular simulation of an integral Kv1.2 ion channel. *PLoS Comp. Biol.*, 5:e1000289, 2009.
- [137] G. M. Blackburn and M. J. Gait, editors. *Nucleic Acids in Chemistry and Biology*. Oxford University Press, New York, NY, 1990.
- [138] V. A. Bloomfield. DNA condensation by multivalent cations. *Biopolymers*, 44:269–282, 1997.
- [139] V. A. Bloomfield, D. M. Crothers, and I. Tinoco, Jr. *Nucleic Acids: Structures, Properties, and Functions*. University Science Press, New York, NY, 2000.
- [140] J. Board, A. John, Z. S. Hakura, W. D. Elliott, and W. T. Rankin. Scalable variants of multipole-accelerated algorithms for molecular dynamics applications. In *Proceedings, Seventh SIAM Conference on Parallel Processing for Scientific Computing*, pages 295–300, Philadelphia, PA, 1995. SIAM.
- [141] J. Board, A. John, C. W. Humphres, C. G. Lambert, W. T. Rankin, and A. Y. Toukmaji. Ewald and multipole methods for periodic *N*-body problems. In *Proceedings, Eighth SIAM Conference on Parallel Processing for Scientific Computing*, Philadelphia, PA, 1997. SIAM. CD-ROM.
- [142] J. A. Board, Jr., J. W. Causey, T. F. Leathrum, Jr., A. Windemuth, and K. Schulten. Accelerated molecular dynamics simulations with the parallel fast multiple algorithm. *Chem. Phys. Lett.*, 198:89–94, 1992.
- [143] J. A. Board, Jr., L. V. Kalé, K. Schulten, R. D. Skeel, and T. Schlick. Modeling biomolecules: Larger scales, longer durations. *IEEE Comput. Sci. Eng.*, 1:19–30, Winter 1994.
- [144] M. D. Bojin and T. Schlick. A quantum mechanical investigation of possible mechanisms for the nucleotidyl transfer reaction catalyzed by DNA polymerase  $\beta$ . *J. Phys. Chem. B*, 111:11244–11252, 2007.
- [145] T. C. Boles, J. H. White, and N. R. Cozzarelli. Structure of plectonemically supercoiled DNA. *J. Mol. Biol.*, 213:931–951, 1990.
- [146] P. G. Bolhuis. Rare events via multiple reaction channels sampled by path replica exchange. *J. Chem. Phys.*, 129:114108, 2008.
- [147] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler. Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.*, 53:291–318, 2002.
- [148] M. Bon, G. Vernizzi, H. Orland, and A. Zee. Topological classification of RNA structures. *J. Mol. Biol.*, 379:900–911, 2008.

- [149] R. Bonneau, J. Tsai, I. Ruczinski, and D. Baker. Functional inferences from blind *ab initio* protein structure predictions. *J. Struc. Biol.*, 134:186–190, 2001.
- [150] S. Boresch, S. Ringhofer, P. Höchtl, and O. Steinhauser. Towards a better description and understanding of biomolecular solvation. *Biophys. Chem.*, 78:43–68, 1999.
- [151] S. Borman. Reducing time to drug discovery. *Chem. Eng. News*, 77:33–48, 1998.
- [152] E. E. Borrero and F. A. Escobedo. Optimizing the sampling and staging for simulations of rare events via forward flux sampling schemes. *J. Chem. Phys.*, 129:024115, 2008.
- [153] J. M. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization. Theory and Examples*, volume 3 of *Canadian Mathematical Society (CMS) Books in Mathematics*. Springer-Verlag, New York, NY, 2000.
- [154] A. H. Boschitsch, M. O. Fenley, and W. K. Olson. A fast adaptive multipole algorithm for calculating screened Coulomb (Yukawa) interactions. *J. Comput. Phys.*, 151:212–241, 1999.
- [155] G. C. Boulougouris and D. Frenkel. Novel Monte Carlo scheme for systems with short-ranged interactions. *J. Chem. Phys.*, 122:244105, 2005.
- [156] K. J. Bowers, E. Chow, H. Xu, R. O. Dror, M. P. Eastwood, B. A. Gregersen, J. L. Klepeis, I. Kolossvary, M. A. Moraes, F. D. Sacerdoti, J. K. Salmon, Y. Shan, and D. E. Shaw. Scalable algorithms for molecular dynamics simulations on commodity clusters. In *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing (SC06), Tampa, Florida*, New York, 2006. ACM Press.
- [157] J. U. Bowie, R. Lüthy, and D. Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253:164–170, 1991.
- [158] G. R. Bowman and V. S. Pande. The roles of entropy and kinetics in structure prediction. *PLoS ONE*, 4:e5840, 2009. doi:10.1371/journal.pone.0005840.
- [159] D. B. Boyd. Computer-aided molecular design. In A. Kent (Executive) and C. M. Hall (Administrative), editors, *Encyclopedia of Library and Information Science*, volume 59, pages 54–84. Marcel Dekker, New York, NY, 1997. Supplement 22.
- [160] D. B. Boyd. Rational drug design: Controlling the size of the haystack. *Mod. Drug Dis.*, 1:41–47, 1998.
- [161] S. F. Brady, K. J. Stauffer, W. C. Lumma, G. M. Smith, H. G. Ramjit, S. D. Lewis, B. J. Lucas, S. J. Gardell, E. A. Lyle, S. D. Appleby, J. J. Cook, M. A. Holahan, M. T. Stranieri, J. J. Lynch, Jr., J. H. Lin, I.-W. Chen, K. Vastag, A. M. Naylor-Olsen, and J. P. Vacca. Discovery and development of the novel potent orally active thrombin inhibitor N-(9-Hydroxy-9-fluorencarboxy)prolyl *trans*-4-Aminocyclohexylmethyl amide (L-372,460): Coapplication of structure-based design and rapid multiple analogue synthesis on solid support. *J. Med. Chem.*, 41(3):401–406, 1998.
- [162] R. S. Braich, N. Chelyapov, C. Johnson, P. W. K. Rothemund, and L. Adleman. Solution of a 20-variable 3-SAT problem on a DNA computer. *Science*, 296:499–502, 2002.
- [163] C. Branden and J. Tooze. *Introduction to Protein Structure*. Garland Publishing Inc., New York, NY, second edition, 1999. ([www.proteinstructure.com/](http://www.proteinstructure.com/)).

- [164] A. Brandt and A. A. Lubrecht. Multilevel matrix multiplication and fast solution of integral equations. *J. Comput. Phys.*, 90:348–370, 1990.
- [165] P. Bratley, B. L. Fox, and L. E. Schrage. *A Guide to Simulation*. Springer-Verlag, New York, NY, second edition, 1987.
- [166] R. R. Breaker. Natural and engineered nucleic acids as tools to explore biology. *Nature*, 432:838–845, 2004.
- [167] S. E. Brenner. A tour of structural genomics. *Nat. Genet.*, 2:801–809, 2001.
- [168] S. E. Brenner, C. Chothia, and T. J. P. Hubbard. Population statistics of protein structures: Lessons from structural classifications. *Curr. Opin. Struct. Biol.*, 7:369–376, 1997.
- [169] K. J. Breslauer, R. Frank, H. Blöcker, and L. A. Marky. Predicting DNA duplex stability from the base sequence. *Proc. Natl. Acad. Sci. USA*, 83:3746–3750, 1986.
- [170] I. Brierley, S. Pennell, and R.J. Gilbert. Viral RNA pseudoknots: versatile motifs in gene expression and replication. *Nat. Rev. Microbiol.*, 5:598–610, 2007.
- [171] R. Brimacombe. The bacterial ribosome at atomic resolution. *Structure*, 8:R195–R200, 2000.
- [172] P. Brion and E. Westhof. Hierarchy and dynamics of RNA folding. *Ann. Rev. Biophys. Biomol. Struc.*, 26:113–137, 1997.
- [173] J. E. Brody. What to serve for dinner, when dinner is on Mars. *The New York Times*, pages F1–2, 19 May 1998.
- [174] B. R. Brooks, C. L. Brooks, III, Jr. A. D. MacKerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Calflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. CHARMM: the biomolecular simulation program. *J. Comput. Chem.*, 30:1545–1614, 2009.
- [175] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 4:187–217, 1983.
- [176] C. L. Brooks, III. Viewing protein folding from many perspectives. *Proc. Natl. Acad. Sci. USA*, 99:1099–1100, 2002.
- [177] C. L. Brooks, III. With a little help ... *Nature*, 420:33–34, 2002.
- [178] C. L. Brooks, III, M. Karplus, and B. M. Pettitt. *Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics*, volume 71 of *Advances in Chemical Physics*. John Wiley & Sons, New York, NY, paperback edition, 1990.
- [179] C. L. Brooks, III, J. N. Onuchic, and D. J. Wales. Statistical thermodynamics: Taking a walk on a landscape. *Science*, 293:612–613, 2001.
- [180] A. Brünger, C. L. Brooks, III, and M. Karplus. Stochastic boundary conditions for molecular dynamics simulations of ST2 water. *Chem. Phys. Lett.*, 105:495–500, 1982.
- [181] A. T. Brünger, P. D. Adams, and L. M. Rice. New applications of simulated annealing in X-ray crystallography and solution NMR. *Structure*, 5:325–336, 1997.

- [182] Z. Bryant, M. D. Stone, J. Gore, S. B. Smith, N. R. Cozzarelli, and C. Bustamante. Structural transitions and elasticity from torque measurements on DNA. *Nature*, 424:338–341, 2003.
- [183] M. Bucciantini, E. Giannoni, F. Chiti, F. Baroni, L. Formigli, J. Zurdo, N. Taddei, G. Ramponi, C. M. Dobson, and M. Stefani. Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature*, 416:507–511, 2002.
- [184] F. Bueche. The viscoelastic properties of plastics. *J. Chem. Phys.*, 22:603–609, 1954.
- [185] U. Burkert and N. L. Allinger. *Molecular Mechanics*, volume 177 of *American Chemical Society Monograph*. ACS, Washington D. C., 1982.
- [186] A. M. Burkhoff and T. D. Tullius. The unusual conformation adopted by the adenine tracts in kinetoplast DNA. *Cell*, 48:935–943, 1987.
- [187] C. Bustamante. *In singulo* biochemistry: When less is more. *Ann. Rev. Biochem.*, 77:45–50, 2008.
- [188] C. Bustamante, Z. Bryant, and S. B. Smith. Ten years of tension: Single-molecule DNA mechanics. *Nature*, 421:423–427, 2003.
- [189] C. Bustamante and D. Keller. Scanning force microscopy in biology. *Physics Today*, 48:32–38, 1995.
- [190] C. Bustamante, J. F. Marko, E. D. Siggia, and S. Smith. Entropic elasticity of  $\lambda$ -phage DNA. *Science*, 265:1599–1600, 1994.
- [191] E. C. Butcher, E. L. Berg, and E. J. Kunkel. Systems biology in drug discovery. *Nat. Biotech.*, 22:1253–1259, 2004.
- [192] R. H. Byrd, P. Lu, and J. Nocedal. A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Stat. Comput.*, 16:1190–1208, 1995.
- [193] R. H. Byrd, J. Nocedal, and R. B. Schnabel. Representations of quasi-Newton matrices and their use in limited memory methods. *Math. Prog.*, 63:129–156, 1994.
- [194] R. H. Byrd, J. Nocedal, and C. Zhu. Towards a discrete Newton method with memory for large-scale optimization. In G. Di Pillo and F. Giannessi, editors, *Nonlinear Optimization and Applications*. Plenum, 1996.
- [195] C. R. Calladine and H. R. Drew. *Understanding DNA. The Molecule and How It Works*. Academic Press, San Diego, CA, second edition, 1997.
- [196] F. Calvo. Non-genetic global optimization methods in molecular science: An overview. *Comp. Mat. Sci.*, 45:8–15, 2009.
- [197] C. R. Cantor and P. R. Schimmel. *Biophysical Chemistry*, volume 1–3. W. H. Freeman and Company, San Francisco, 1980.
- [198] P. R. Caron, M. D. Mullican, R. D. Mashal, K. P. Wilson, M. S. Su, and M. A. Murcko. Chemogenomic approaches to drug discovery. *Curr. Opin. Chem. Biol.*, 5:464–470, 2001.
- [199] J. M. Carothers, S. C. Oestreich, J. H. Davis, and J. W. Szostak. Informational complexity and functional activity of RNA structures. *J. Amer. Chem. Soc.*, 126:5130–5137, 2004.
- [200] J. C. Carrington and V. Ambros. Role of microRNAs in plant and animal development. *Science*, 301:336–338, 2003.

- [201] A. P. Carter, W. M. Clemons, D. E. Brodersen, R. J. Morgan-Warren, B. T. Wimberly, and V. Ramakrishnan. Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature*, 407:340–348, 2000.
- [202] D. A. Case. Normal mode analysis of protein dynamics. *Curr. Opin. Struct. Biol.*, 4:385–290, 1994.
- [203] D. A. Case. NMR refinement. In P. von Ragué Schleyer (Editor-in Chief), N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, and H. F. Schaefer, III, editors, *Encyclopedia of Computational Chemistry*, volume 3, pages 1866–1876. John Wiley & Sons, West Sussex, England, 1998.
- [204] D. A. Case, T. E. Cheatham, III, T. Darden, H. Gohlke, R. Luo, K.M. Jr. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods. The Amber biomolecular simulation programs. *J. Comput. Chem.*, 26:1668 W1688, 2005.
- [205] L. Castagnoli, M. Scarpa, M. Kokkinidis, D.W. Banner, D. Tsernoglou, and G. Cesareni. Genetic and structural analysis of the CoIE1 Rop (Rom) protein. *Embo. J.*, 8:621–629, 1989.
- [206] J. H. Cate, M. M. Yusupov, C. Zh. Yusupova, T. N. Earnest, and H. F. Noller. X-ray crystal structure of 70S ribosome functional complexes. *Science*, 285:2095–2104, 1999.
- [207] T. Caulfield and K. Burgess. Combinatorial chemistry. Focused diversity and diversity of focus. *Curr. Opin. Chem. Biol.*, 5:241–242, 2001.
- [208] M. Cavazzana-Calvo and A. Fischer. Gene therapy for severe combined immunodeficiency: Are we there yet? *J. Clin. Inves.*, 117:1456–1465, 2007.
- [209] L. S. D. Caves, J. D. Evanseck, and M. Karplus. Locally accessible conformations of proteins: Multiple molecular dynamics simulations of crambin. *Prot. Sci.*, 7:649–666, 1998.
- [210] T. R. Cech. The ribosome is a ribozyme. *Science*, 289:878–879, 2000.
- [211] L. Cerchia and V. De Franciscis. Noncoding RNAs in cancer medicine. *J. Biomed. Biotechnol.*, 2006:73104, 2006.
- [212] M. R. Chance, A. R. Bresnick, S. K. Burley, J.-S. Jiang, C. D. Lima, A. Sali, S. C. Almo, J. B. Bonanno, J. A. Buglino, S. Boulton, H. Chen, N. Eswar, G. He, R. Huang, V. Ilyin, L. McMahan, U. Pieper, S. Ray, M. Vidal, and L. K. Wang. Structural genomics: A pipeline for providing structures for the biologist. *Prot. Sci.*, 11:723–738, 2002.
- [213] D. Chandler. *Introduction to Modern Statistical Mechanics*. Oxford University Press, New York, NY, 1987.
- [214] J.-M. Chandonia and S. E. Brenner. The impact of structural genomics: Expectations and outcomes. *Science*, 311:347–351, 2006.
- [215] R. Chandrasekaran and S. Arnott. The structure of B-DNA in oriented fibers. *J. Biomol. Struct. Dynam.*, 13:1015–1027, 1996.
- [216] T. E. Cheatham, III. Simulation and modeling of nucleic acid structure, dynamics and interactions. *Curr. Opin. Struct. Biol.*, 14:360–367, 2004.
- [217] T. E. Cheatham, III. Molecular modeling and atomistic simulation of nucleic acids. *Ann. Rev. Comp. Chem.*, 1:75–89, 2005.

- [218] T. E. Cheatham, III, P. Cieplak, and P. A. Kollman. A modified version of the cornel *et al.* force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dynam.*, 16:845–862, 1999.
- [219] T. E. Cheatham, III and P. A. Kollman. Observation of the A-DNA to B-DNA transition during unrestrained molecular dynamics in aqueous solution. *J. Mol. Biol.*, 259:434–444, 1996.
- [220] T. E. Cheatham, III, J. L. Miller, T. Fox, T. A. Darden, and P. A. Kollman. Molecular dynamics simulations of solvated biomolecular systems: The particle mesh Ewald method leads to stable trajectories of DNA, RNA, and proteins. *J. Amer. Chem. Soc.*, 117:4193–4194, 1995.
- [221] T. E. Cheatham, III and M. A. Young. Molecular dynamics simulation of nucleic acids: Successes, limitations, and promise. *Biopolymers*, 56:232–256, 2001.
- [222] Y. Chebaro, X. Dong, R. Laghaei, P. Derreumaux, and N. Mousseau. Replica exchange molecular dynamics simulations of coarse-grained proteins in implicit solvent. *J. Phys. Chem. B.*, 113:267–274, 2009.
- [223] J. Chen and C. L. Brooks, III. Implicit modeling of nonpolar solvation for simulating protein folding and conformational transitions. *Phys. Chem. Chem. Phys.*, 10:471–481, 2008.
- [224] L. Chen and M. W. Deem. Monte Carlo methods for small molecule high-throughput experimentation. *J. Chem. Inf. Comput. Sci.*, 41:950–957, 2001.
- [225] S.-J. Chen. RNA folding: Conformational statistics, folding kinetics, and ion electrostatics. *Ann. Rev. Biophys.*, 37:197–214, 2008.
- [226] W. W. Chen, J. S. Yand, and E. Shakhnovich. A knowledge-based move set for protein folding. *Prot. Struc. Funct. Bioinf.*, 66:682–688, 2007.
- [227] Z. Chen, Y. Li, E. Chen, D. L. Hall, P. L. Darke, C. Culberson, J. A. Shafer, and L. C. Kuo. Crystal structure at 1.9-Å resolution of human immunodeficiency (HIV) II protease complexed with L-735,524, an orally bioavailable inhibitor of the HIV proteases. *J. Biol. Chem.*, 269:26344–26348, 1994.
- [228] D. Y. Cherny, B. P. Belotserkovskii, M. D. Frank-Kamenetskii, M. Egholm, O. Buchardt, R. H. Berg, and P. E. Nielsen. DNA unwinding upon strand-displacement binding of a thymine-substituted polyamide to double-stranded DNA. *Proc. Natl. Acad. Sci. USA*, 90:1667–1670, 1993.
- [229] S. Chew, P. Chen, N. Link, K. Galindo, K. Pogue, and J. Abrams. Genome-wide silencing in *Drosophila* captures conserved apoptotic effectors. *Nature*, 460:123–127, 2009.
- [230] D. N. Chin, F. Sussman, H. M. Chun, and R. Czerninski. A simple solvation model along with a multibody dynamics strategy MBO(N)D produces stable DNA simulations that are faster than traditional atomistic methods. *Mol. Sim.*, 24:449–463, 2000.
- [231] G. Chirico and J. Langowski. Calculating hydrodynamic properties of DNA through a second-order Brownian dynamics algorithm. *Macromolecules*, 25:769–775, 1992.
- [232] G. Chirico and J. Langowski. Kinetics of DNA supercoiling studied by Brownian dynamics simulation. *Biopolymers*, 34:415–433, 1994.
- [233] G. Chirico and J. Langowski. Brownian dynamics simulations of supercoiled DNA with bent sequences. *Biophys. J.*, 71:955–971, 1996.

- [234] T. K. Chiu and R. E. Dickerson. 1 Å crystal structures of *B*-DNA reveal sequence-specific binding and groove-specific bending of DNA by magnesium and calcium. *J. Mol. Biol.*, 301:915–945, 2000.
- [235] W. W. Chiu, R. M. Kinney, and T. W. Dreher. Control of translation by the 5'- and 3'-terminal regions of the dengue virus genome. *J. Virol.*, 79:8303–8315, 2005.
- [236] C. H. Cho and M. E. Nuttall. Emerging techniques for the discovery and validation of therapeutic targets for skeletal diseases. *Expert Opin. Ther. Targets*, 6:679–689, 2002.
- [237] C. Chothia. One thousand families for the molecular biologist. *Nature*, 357:543–544, 1992.
- [238] C. Chothia, T. Hubbard, S. Brenner, H. Barns, and A. Murzin. Protein folds in the all- $\beta$  and all- $\alpha$  classes. *Ann. Rev. Biophys. Biomol. Struc.*, 26:597–627, 1997.
- [239] V. B. Chu, Y. Bai, J. Lipfert, D. Herschlag, and S. Doniach. Evaluation of ion binding to DNA duplexes using a size-modified Poisson-Boltzmann theory. *Biophys. J.*, 93:3202–3209, 2007.
- [240] V. B. Chu, Y. Bai, J. Lipfert, D. Herschlag, and S. Doniach. A repulsive field: Advances in the electrostatics of the ion atmosphere. *Curr. Opin. Chem. Biol.*, 12:619–625, 2008.
- [241] V. B. Chu and D. Herschlag. Understanding RNA's secrets: Advances in the biology, physics, and modeling of complex RNAs. *Curr. Opin. Struct. Biol.*, 18:305–314, 2008.
- [242] A. Chworos, I. Sevcenac, A. Koifman, P. Weinkam, E. Oroudjev, H. Hansma, and L. Jaeger. Building programmable jigsaw puzzles with RNA. *Science*, 306:2068–2072, 2004.
- [243] P. Cieplak, F.-Y. Dupradeau, Y. Duan, and J. Wang. Polarization effects in molecular mechanical force fields. *J. Phys.: Condens. Matter*, 21:333102, 2009.
- [244] P. Cieplak, P. Kollman, and T. Lybrand. A new water potential including polarization: Application to gas-phase, liquid, and crystal properties of water. *J. Chem. Phys.*, 92:6755–6760, 1990.
- [245] B. Cipra. *What's Happening in the Mathematical Sciences*. American Mathematical Society, Cranston, RI, 1996. (Series Editor: P. Zorn).
- [246] B. A. Cipra. Molecular biologists team up with mathematicians to investigate DNA. *SIAM News*, 23:16, 1990.
- [247] J. B. Clarage, T. Romo, B. K. Andrews, B. M. Pettitt, and G. N. Philipps, Jr. A sampling problem in molecular dynamics simulations of macromolecules. *Proc. Natl. Acad. Sci. USA*, 92:3288–3292, 1995.
- [248] G. M. Clore and A. M. Gronenborn. New methods of structure refinement for macromolecular structure determination by NMR. *Proc. Natl. Acad. Sci. USA*, 95:5891–5898, 1998.
- [249] G. M. Clore and C. D. Schwieters. Theoretical and computational advances in biomolecular NMR spectroscopy. *Curr. Opin. Struct. Biol.*, 12:146–153, 2002.
- [250] P. Cluzel, A. Lebrun, C. Heller, R. Lavery, J.-L. Viovy, D. Chatenay, and F. Caron. DNA: An extensible molecule. *Science*, 271:792–794, 1996.
- [251] F. E. Cohen. Protein misfolding and prion diseases. *J. Mol. Biol.*, 293:313–320, 1999.

- [252] J. E. Cohen. Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better. *PLoS Biol.*, 2:2017–2023, 2004.
- [253] M. S. Cohen, N. Hellmann, J. A. Levy, K. DeCook, and J. Lange. The spread, treatment, and prevention of HIV-1: Evolution of a global pandemic. *J. Clin. Inves.*, 118:1244–1254, 2008.
- [254] N. C. Cohen, editor. *Guidebook on Molecular Modeling in Drug Design*. Academic Press, San Diego, CA, 1996.
- [255] B. D. Coleman and D. Swigon. Theory of supercoiled elastic rings with self-contact and its application to DNA plasmids. *J. Elasticity*, 60:173–221, 2000.
- [256] B. D. Coleman, D. Swigon, and I. Tobias. Elastic stability of DNA configurations: II. Supercoiled plasmids with self-contact. *Phys. Rev. E*, 61:759–770, 2000.
- [257] F. S. Collins. Opportunities for research and NIH. *Science*, 327:36–37, 2010.
- [258] F. S. Collins, E. D. Green, A. E. Guttmacher, and M. S. Guyer. A vision for the future of genomics research. *Nature*, 422:835–847, 2003.
- [259] F. S. Collins and K. G. Jegalian. Deciphering the code of life. *Sci. Amer.*, 281:86–91, 1999.
- [260] I. Coluzza, A. De Simon, F. Fraternali, and D. Frenkel. Multi-scale simulations provide supporting evidence for the hypothesis of intramolecular protein translocation in GroEL/GroES complexes. *PLoS Comp. Biol.*, 4:e1000006, 2008.
- [261] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. *LANCELOT: A FORTRAN Package for Large-Scale Nonlinear Optimization (Release A)*, volume 17 of *Springer Series in Computational Mathematics*. Springer-Verlag, New York, NY, 1992.
- [262] L. Lo Conte, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin. SCOP database in 2002: Refinements accommodate structural genomics. *Nucl. Acids Res.*, 30:264–267, 2002.
- [263] W. J. Cook, W. H. Cunningham, W. R. Pulleyblank, and A. Schrijver. *Combinatorial Optimization*. John Wiley & Sons, New York, NY, 1998.
- [264] B. Cooke and S. C. Schmidler. Preserving the Boltzmann ensemble in replica-exchange molecular dynamics. *J. Chem. Phys.*, 129:164112, 2008.
- [265] 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. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Amer. Chem. Soc.*, 117:5179–5197, 1995.
- [266] A. F. W. Coulson and J. Moult. A unifold, mesofold, and superfold model of protein fold use. *Proteins: Struc. Func. Gen.*, 46:61–71, 2002.
- [267] J. Couzin. Small RNAs make big splash. *Science*, 298:2296–2297, 2002.
- [268] N. R. Cozzarelli. Revisiting the independence of the publicly and privately funded drafts of the human genome. *Proc. Natl. Acad. Sci. USA*, 100:3021, 2003.
- [269] N. R. Cozzarelli and J. C. Wang, editors. *DNA Topology and Its Biological Effects*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1990.
- [270] C. J. Cramer. *Essentials of Computational Chemistry: Theories and Models*. John Wiley & Sons Inc., Hoboken, NJ, second edition, 2004.
- [271] D. Cremer and J. A. Pople. A general definition of ring puckering coordinates. *J. Amer. Chem. Soc.*, 97:1354–1358, 1975.

- [272] F. H. C. Crick. *What Mad Pursuit: A Personal View of Scientific Discovery*. Alfred P. Sloan Foundation Series. Basic Books, New York, NY, 1988.
- [273] S. Cronin, N. Nehme, S. Limmer, S. Liegeois, J. Pospisilik, D. Schramek, A. Leibbrandt, R. Simoes, S. Gruber, U. Puc, I. Ebersberger, T. Zoranovic, G. Neely, A. von Haeseler, D. Ferrandon, and J. Penninger. Genome-wide RNAi screen identifies genes involved in intestinal pathogenic bacterial infection. *Science*, 325:340–343, 2009.
- [274] D. Crothers. DNA curvature and deformation in protein-DNA complexes: A step in the right direction. *Proc. Natl. Acad. Sci. USA*, 95:15163–15165, 1998.
- [275] D. Crothers and D. Eisenberg. *Physical Chemistry with Applications to the Life Sciences*. Benjamin/Cummings, Menlo Park, CA, 1979.
- [276] D. Crothers, T. E. Haran, and J. G. Nadeau. Intrinsically bent DNA. *J. Biol. Chem.*, 265:7093–7096, 1990.
- [277] J. A. Cruz and E. Westhof. The dynamic landscape of RNA architecture. *Cell*, 136:604–609, 2009.
- [278] P. Csermely, V. Agoston, and S. Pongor. The efficiency of multi-target drugs: The network approach might help drug design. *Trends in Pharm. Sci.*, 26:178–182, 2005.
- [279] G. Călugăreanu. Sur les classes d'isotopie des noeuds tridimensionnels et leurs invariants. *Czechoslovak Math. J.*, 11:588–624, 1961.
- [280] G. Dahlquist and Å. Björck. *Numerical Methods*. Prentice Hall, Englewood Cliffs, New Jersey, 1974.
- [281] S. Dalal, S. Balasubramanian, and L. Regan. Protein alchemy: Changing  $\beta$ -sheet into  $\alpha$ -helix. *Nature Struc. Biol.*, 4:548–552, 1997.
- [282] A. D. Daniels and G. E. Scuseria. What is the best alternative to diagonalization of the Hamiltonian in large scale semiempirical calculations? *J. Chem. Phys.*, 110:1321–1328, 1999.
- [283] T. Darden, L. Perera, L. Li, and L. Pedersen. New tricks for modelers from the crystallography toolkit: The particle mesh Ewald algorithm and its use in nucleic acid simulations. *Structure*, 7:R55–R60, 1999.
- [284] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald: An  $N \cdot \log(N)$  method for Ewald sums in large systems. *J. Chem. Phys.*, 98:10089–10092, 1993.
- [285] X. Daura, B. Jaun, D. Seebach, W. F. Van Gunsteren, and A. Mark. Reversible peptide folding in solution by molecular dynamics simulation. *J. Mol. Biol.*, 280:925–932, 1998.
- [286] X. Daura, B. Oliva, E. Querol, F. X. Avilés, and O. Tapia. On the sensitivity of MD trajectories to changes in water-protein interaction parameters: The potato carboxypeptidase inhibitor in water as a test case for the GROMOS force field. *Proteins: Struc. Func. Gen.*, 25:89–103, 1996.
- [287] C. A. Davey, D. F. Sargent, K. Luger, A. W. Mäder, and T. J. Richmond. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *J. Mol. Biol.*, 319:1097–1113, 2002.
- [288] K. Davies. *Cracking the Genome: Inside the Race to Unlock Human DNA*. The Free Press (A Simon & Schuster Division), New York, NY, 2001.

- [289] L. A. Day, R. L. Wiseman, and C. J. Marzec. Structure models for DNA in filamentous viruses with phosphates near the center. *Nucl. Acids Res.*, 7:1393–1403, 1979.
- [290] P. S. de Laplace. *Oeuvres Complètes de Laplace. Théorie Analytique des Probabilités*, volume VII. Gauthier-Villars, Paris, France, third edition, 1820.
- [291] O. Norberto de Souza and R. L. Ornstein. Effect of periodic box size on aqueous molecular dynamics simulation of a DNA dodecamer with particle-mesh Ewald method. *Biophys. J.*, 72:2395–2397, 1997.
- [292] O. Norberto de Souza and R. L. Ornstein. Inherent DNA curvature and flexibility correlate with TATA box functionality. *Biopolymers*, 46:403–415, 1998.
- [293] S. W. DeLeeuw, J. W. Perram, and E. R. Smith. Simulation of electrostatic systems in periodic boundary conditions. I. Lattice sums and dielectric constant. *Proc. Roy. Soc. Lond. A*, 373:27–56, 1980.
- [294] C. Dellago and P. G. Bolhuis. Transition path sampling simulations of biological systems. *Top. Curr. Chem.*, 268:291–317, 2007.
- [295] R. S. Dembo and T. Steihaug. Truncated-Newton algorithms for large-scale unconstrained optimization. *Math. Prog.*, 26:190–212, 1983.
- [296] J. W. Demmel. *Applied Numerical Linear Algebra*. SIAM, Philadelphia, PA, 1997.
- [297] J. E. Dennis, Jr. and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1983. (Reprinted by SIAM, 1996).
- [298] P. Derreumaux and T. Schlick. Long-time integration for peptides by the dynamics driver approach. *Proteins: Struc. Func. Gen.*, 21:282–302, 1995.
- [299] P. Derreumaux and T. Schlick. The loop opening/closing motion of the enzyme triosephosphate isomerase. *Biophys. J.*, 74:72–81, January 1998.
- [300] P. Derreumaux and G. Vergoten. Influence of the spectroscopic potential energy function SPASIBA on molecular dynamics of proteins: Comparison with the AMBER potential. *J. Mol. Struct.*, 286:55–64, 1993.
- [301] P. Derreumaux and G. Vergoten. A new spectroscopic molecular mechanics force field. Parameters for proteins. *J. Chem. Phys.*, 102:8586–8605, 1995.
- [302] P. Derreumaux, G. Zhang, B. Brooks, and T. Schlick. A truncated-Newton method adapted for CHARMM and biomolecular applications. *J. Comput. Chem.*, 15:532–552, 1994.
- [303] M. Deserno, C. Holm, and S. May. Fraction of condensed counterions around a charged rod: Comparison of Poisson-Boltzmann theory and computer simulations. *Macromolecules*, 33:199–206, 2000.
- [304] P. Deuflhard, M. Dellnitz, O. Junge, and Ch. Schütte. Computation of essential molecular dynamics by subdivision techniques: I. Basic concepts. Technical Report SC 96–45, Konrad-Zuse-Zentrum für Informationstechnik Berlin, Takustraße 7, D-14195, Berlin-Dahlem, December 1996.
- [305] D. J. Dichmann, Y. Li, and J. H. Maddocks. Hamiltonian formulations and symmetries in rod mechanics. In J. P. Mesirow, K. Schulter, and D. W. Sumners, editors, *Mathematical Applications to Biomolecular Structure and Dynamics*, volume 82 of *IMA Volumes in Mathematics and Its Applications*, pages 71–113, New York, NY, 1996. Springer-Verlag.

- [306] R. E. Dickerson. Definitions and nomenclature of nucleic acid structure parameters. *EMBO J.*, 8:1–4, 1989.
- [307] R. E. Dickerson, M. Bansal, C. R. Calladine, S. Diekmann, W. N. Hunter, O. Kennard, E. von Kitzing, R. Lavery, H. C. M. Nelson, W. K. Olson, W. Saenger, Z. Shakked, H. Sklenar, D. M. Soumpasis, C.-S. Tung, A. H.-J. Wang, and V. B. Zhurkin. Definitions and nomenclature of nucleic acid structure parameters. *J. Mol. Biol.*, 208:787–791, 1989.
- [308] R. E. Dickerson and T. K. Chiu. Helix bending as a factor in protein/DNA recognition. *Biopolymers*, 44:361–403, 1997.
- [309] R. E. Dickerson and H. R. Drew. Structure of a B-DNA dodecamer. II. influence of base sequence on helix structure. *J. Mol. Biol.*, 149:761–786, 1981.
- [310] R. E. Dickerson, D. S. Goodsell, and M. L. Kopka. MPD and DNA bending in crystals and in solution. *J. Mol. Biol.*, 256:108–125, 1996.
- [311] R. E. Dickerson, D. S. Goodsell, and S. Neidle. “... the tyranny of the lattice ...”. *Proc. Natl. Acad. Sci. USA*, 91:3579–3583, 1994.
- [312] A. D. DiGabriele, T. A., and Steitz. A DNA dodecamer containing an adenine tract crystallizes in a unique lattice and exhibits a new bend. *J. Mol. Biol.*, 231:1024–1039, 1993.
- [313] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan. Principles of protein folding — A perspective from simple exact models. *Protein Science*, 4:561–602, 1995.
- [314] K. A. Dill and H. S. Chan. From Levinthal to pathways to funnels. *Nature Struct. Biol.*, 4:10–19, 1997.
- [315] F. Ding, S. Sharma, P. Chalasani, V. V. Demidov, N. E. Broude, and N. V. Dokholyan. Ab initio RNA folding by discrete molecular dynamics: From structure prediction to folding mechanisms. *RNA*, 14:1164–1173, 2008.
- [316] H. Q. Ding, N. Karasawa, and W. A. Goddard, III. Atomic level simulations on a million particles: The cell multipole method for Coulomb and London nonbond interactions. *J. Chem. Phys.*, 97:4309–4315, 1992.
- [317] H. Q. Ding, N. Karasawa, and W. A. Goddard, III. The reduced cell multipole method for Coulomb interactions in periodic systems with million-atom unit cells. *Chem. Phys. Lett.*, 196:6–10, 1992.
- [318] A. R. Dinner and M. Karplus. Comment on the communication “The key to solving the protein-folding problem lies in an accurate description of the denatured state” by van Gunsteren et al. *Angew. Chem. Int. Ed.*, 40:4615–4616, 2001.
- [319] A. R. Dinner, R. Lazaridis, and M. Karplus. Understanding  $\beta$ -hairpin formation. *Proc. Natl. Acad. Sci. USA*, 96:9068–9073, 1999.
- [320] P. A. M. Dirac. Quantum mechanics of many-electron systems. *Proc. Royal Soc. London*, A123:714–733, 1929.
- [321] S. B. Dixit, D. L. Beveridge, D. A. Case, T. E. Cheatham, III, E. Giudice, F. Lankas, R. Lavery, J. H. Maddocks, R. Osman, H. Sklenar, K. M. Thayer, and P. Varnai. Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. II: Sequence context effects on the dynamical structures of the 10 unique dinucleotide steps. *Biophys. J.*, 89:3721–3740, 2005.

- [322] S. L. Dixon and H. O. Villar. Investigation of classification methods for the prediction of activity in diverse chemical libraries. *J. Comput.-Aided Mol. Design*, 13:533–545, 1999.
- [323] C. Djerassi. *The Pill, Pygmy Chimps, and Degas' Horse. The Remarkable Autobiography of the Award-Winning Scientist Who Synthesized the Birth Control Pill*. Basic Books, New York, NY, 1992.
- [324] M. Dlakic and R. E. Harrington. The effects of sequence context on DNA curvature. *Proc. Natl. Acad. Sci. USA*, 93:3847–3852, 1996. [Erratum appeared in *Proc. Natl. Acad. Sci. USA*, 93:8796, 1996.]
- [325] C. M. Dobson. Getting out of shape. *Nature*, 418:729–730, 2002.
- [326] C. M. Dobson. Protein folding and misfolding: From atoms to organisms. In A. H. Zewail, editor, *Physical Biology: From Atoms to Medicine*, pages 289–335. Imperial College Press, London, UK, 2008.
- [327] E. A. Doherty, R. T. Batey, B. Masquida, and J. A. Doudna. A universal mode of helix packing in RNA. *Nat. Struc. Biol.*, 8:339–343, 2001.
- [328] S. Doniach and P. Eastman. Protein dynamics simulations from nanoseconds to microseconds. *Curr. Opin. Struct. Biol.*, 9:157–163, 1999.
- [329] A. J. Dooley, N. Shindo, B. Taggart, J. G. Park, and Y. P. Pang. From genome to drug lead: Identification of a small-molecule inhibitor of the SARS virus. *Bioorg. Med. Chem. Lett.*, 16:830–833, 2006.
- [330] B. Dorigo, T. Schalch, A. Kulangara, S. Duda, R. R. Schroeder, and T. J. Richmond. Nucleosome arrays reveal the two-start organization of the chromatin fiber. *Science*, 306:1571–1573, 2004.
- [331] J. A. Doudna. Ribozymes: The hammerhead swings into action. *Curr. Biol.*, 8:R495–R497, 1998.
- [332] J. A. Doudna. Structural genomics of RNA. *Nature Struc. Biol.*, 7:954–956, 2000. (Structural Genomics Supplement).
- [333] R. B. Dover, L. F. Schneemeyer, and R. M. Fleming. Discovery of a useful thin-film dielectric using a composition-spread approach. *Nature*, 392:162–164, 1998.
- [334] H. R. Drew, R. M. Wing, T. Takano, C. Broka, S. Tanaka, K. Itakura, and R. E. Dickerson. Structure of a B-DNA dodecamer: Conformation and dynamics. *Proc. Natl. Acad. Sci. USA*, 78:2179–2183, 1981.
- [335] J. Drews. Drug discovery: A historical perspective. *Science*, 287:1960–1964, 2000.
- [336] R. J. Driscoll, M. G. Younquist, and J. D. Baldeschwieler. Atomic-scale imaging of DNA using scanning tunneling microscopy. *Nature*, 346:294–296, 1990.
- [337] R. O. Dror, D. H. Arlow, D. W. Borhani, M. . Jensen, S. Piana, and D. E. Shaw. Identification of two distinct inactive conformations of the 2-adrenergic receptor reconciles structural and biochemical observations. *Proc. Natl. Acad. Sci. USA*, 106:4689–4694, 2009.
- [338] Y. Duan and P. A. Kollman. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, 282:740–744, 23 October 1998.
- [339] Y. Duan, P. A. Kollman, and S. C. Harvey. Protein folding and beyond. In E. Keinam and I. Schechter, editors, *Chemistry for the 21st Century*. Wiley-VCH, Weinheim, Germany, 2000.

- [340] Y. Duan, L. Wang, and P. A. Kollman. The early stage of folding of villin headpiece subdomain observed in a 200-nanosecond fully solvate molecular dynamics simulation. *Proc. Natl. Acad. Sci. USA*, 95:9897–9902, 1998.
- [341] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, and P. Kollman. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.*, 24:1999–2012, 2003.
- [342] Z.-H. Duan and R. Krasny. An Ewald summation based multipole method. *J. Chem. Phys.*, 113:3492–3495, 2000.
- [343] Z.-H. Duan and R. Krasny. An adaptive tree code for computing nonbonded potential energy in classical molecular systems. *J. Comput. Chem.*, 22:184–195, 2001.
- [344] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Phys. Lett. B*, 195:216–222, 1987.
- [345] H. J. Dyson and P. E. Wright. Insights into protein folding from NMR. *Annu. Rev. Phys. Chem.*, 47:369–395, 1996.
- [346] H. J. Dyson and P. E. Wright. Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.*, 12:54–60, 2002.
- [347] D. J. Earl and M. W. Deem. Monte Carlo simulations. *Methods Mol. Biol.*, 443:25–36, 2008.
- [348] A. S. Edison. Linus Pauling and the planar peptide bond. *Nat. Struc. Biol.*, 8:201–202, 2001.
- [349] E. H. Egelman. Does a stretched DNA structure dictate the helical geometry of Rec-A-like filaments? *J. Mol. Biol.*, 309:539–542, 2001.
- [350] M. Egholm, O. Buchardt, P. E. Nielsen, and R. H. Berg. Peptide nucleic-acids (PNA)—Oligonucleotide analogs with an achiral peptide backbone. *J. Amer. Chem. Soc.*, 114:1895–1897, 1992.
- [351] M. Egholm, P. E. Nielsen, O. Buchardt, and R. H. Berg. Recognition of guanine and adenine in DNA by cytosine and thymine containing peptide nucleic-acids (PNA). *J. Amer. Chem. Soc.*, 114:9677–9678, 1992.
- [352] M. Egli. DNA-cation interactions: Quo vadis? *Chem. Biol.*, 9:277–286, 2002.
- [353] L. Ehrlich, C. Münkel, G. Chirico, and J. Langowski. A Brownian dynamics model for the chromatin fiber. *CABIOS*, 13(3):271–279, 1997.
- [354] R. Elber. A milestones study of the kinetics of an allosteric transition: atomically detailed simulations of deoxy *Scapharca* hemoglobin. *Biophys. J.*, 92:L85–L87, 2007.
- [355] R. Elber, J. Meller, and R. Olender. Stochastic path approach to compute atomically detailed trajectories: Application to the folding of C peptide. *J. Phys. Chem. B*, 103:899–911, 1999.
- [356] A. H. Elcock, R. R. Gabdoulline, R. C. Wade, and J. A. McCammon. Computer simulation of protein-protein association kinetics: Acetylcholinesterase-Fasciculin. *J. Mol. Biol.*, 291:149–162, 1999.
- [357] L. Eldén and L. Wittmeyer-Koch. *Numerical Analysis*. Academic Press, Inc., San Diego, CA, 1990.

- [358] L. O. Elkin. Rosalind Franklin and the double helix. *Physics Today*, 56:42–48, 2003.
- [359] A. D. Ellington and J. W. Szostak. *In Vitro* selection of RNA molecules that bind specific ligands. *Nature*, 346:818–822, 1990.
- [360] A. Elofsson and L. Nilsson. How consistent are molecular dynamics simulations? Comparing structure and dynamics in reduced and oxidized *Escherichia coli* Thioredoxin. *J. Mol. Biol.*, 233:766–780, 1991.
- [361] A. Engel. New frontiers in high-resolution electron microscopy. In T. Schwede and M. Peitsch, editors, *Computational Structural Biology. Methods and Applications*, pages 623–654. World Scientific, Singapore, 2008.
- [362] S. W. Englander, L. Mayne, and M. M. G. Krishna. Protein folding and misfolding: Mechanism and principles. *Quar. Rev. Biophys.*, 40:287–326, 2007.
- [363] M. Enserink. Full-genome sequencing paved the way from spores to a suspect. *Science*, 321:898–899, 2008.
- [364] D. L. Ensign, P. M. Kasson, and V. S. Pande. Heterogeneity even at the speed limit of folding: Large-scale molecular dynamics study of a fast-folding variant of the villin headpiece. *J. Mol. Biol.*, 374:806–816, 2007.
- [365] D. L. Ensign and V. S. Pande. The Fip35 WW domain folds with structural and mechanistic heterogeneity in molecular dynamics simulations. *Biophys. J.*, 96:L53–L55, 2009.
- [366] Y. Erlich, K. Chang, A. Gordon, R. Ronen, O. Navon, M. Rooks, and G. J. Hannon. DNA Sudoku — harnessing high-throughput sequencing for multiplexed specimen analysis. *Gen. Res.*, 19:1243–1253, 2009.
- [367] D. L. Ermak and J. A. McCammon. Brownian dynamics with hydrodynamic interactions. *J. Chem. Phys.*, 69:1352–1360, 1978.
- [368] R. R. Ernst, G. Bodenhausen, and A. Wokaum. *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*, volume 14 of *International Series of Monographs on Chemistry*. Clarendon Press, Oxford, New York, NY, 1987.
- [369] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen. A smooth particle mesh Ewald method. *J. Chem. Phys.*, 103:8577–8593, 1995.
- [370] W. E. Evans and M. V. Relling. Pharmacogenomics: Translating functional genomics into rational therapeutics. *Science*, 286:487–491, 1999.
- [371] C. S. Ewig, T. S. Thacher, and A. T. Hagler. Derivation of class II force fields. 7. nonbonded force field parameters for organic compounds. *J. Phys. Chem. B*, 103:6998–7014, 1999.
- [372] V. A. Eyrich, D. M. Standley, and R. A. Friesner. Prediction of protein tertiary structure to low resolution: Performance for a large and structurally diverse test set. *J. Mol. Biol.*, 14:725–742, 1999.
- [373] C. Ezzell. Proteins rule. *Sci. Amer.*, 286:40–47, 2002.
- [374] H. R. Faber and B. W. Matthews. A mutant lysozyme displays five different crystal conformations. *Nature*, 348:263–265, 1990.
- [375] A. S. Fauci, M. I. Johnston, C. W. Dieffenbach, D. R. Burton, S. M. Hammer, J. A. Hoxie, M. Martin, J. Overbaugh, D. I. Watkins, A. Mahmoud, and W. C. Greene. HIV vaccine research: The way forward. *Science*, 321:530–532, 2008.

- [376] M. Feig and C. L. Brooks, III. Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Curr. Opin. Struct. Biol.*, 14:217–224, 2004.
- [377] M. Feig, W. Im, and C. L. Brooks, III. Implicit solvation based on generalized Born theory in different dielectric environments. *J. Chem. Phys.*, 120:903–911, 2004.
- [378] M. Feig, Jr. A. D. MacKerell, and C. L. Brooks, III. Force field influence on the observation of  $\pi$ -helical protein structures in molecular dynamics simulations. *J. Phys. Chem. B*, 107:2831–2836, 2003.
- [379] M. Feig and B. M. Pettitt. Experiment vs. force fields: DNA conformation from molecular dynamics simulations. *Phys. Chem. B*, 101(38):7361–7363, 1997.
- [380] M. Feig and B. M. Pettitt. Crystallographic water sites from a theoretical perspective. *Curr. Biol.*, 6:1351–1354, 1998.
- [381] M. Feig and B. M. Pettitt. Structural equilibrium of DNA represented with different force fields. *Biophys. J.*, 75:134–149, 1998.
- [382] S. E. Feller, Y. Zhang, R. W. Pastor, and B. R. Brooks. Constant pressure molecular dynamics simulation: The Langevin piston method. *J. Chem. Phys.*, 103: 4613–4621, 1995.
- [383] G. Felsenfeld and M. Grasdine. Controlling the double helix. *Nature*, 421: 448–453, 2003.
- [384] P. W. Fenimore, H. Frauenfelder, B. H. McMahon, and F. G. Parak. Slaving: Solvent fluctuations dominate protein dynamics and functions. *Proc. Natl. Acad. Sci. USA*, 99:16047–16051, 2002.
- [385] M. O. Fenley, K. Chua, A. H. Boschitsch, and W. K. Olson. A fast adaptive multipole method for computation of electrostatic energy in simulations of polyelectrolyte DNA. *J. Comput. Chem.*, 17:976–991, 1996.
- [386] M. O. Fenley, W. K. Olson, I. Tobias, and G. S. Manning. Electrostatic effects in short superhelical DNA. *Biophys. Chem.*, 50:255–271, 1994.
- [387] W. A. Fenton and A. L. Horwich. GroEL-mediated protein folding. *Protein Sci.*, 6:743–760, 1997.
- [388] D. Fera, N. Kim, N. Shiffeldrim, J. Zorn, U. Laserson, N. Kim, and T. Schlick. RAG: RNA-As-Graphs web resource. *BMC Bioinformatics*, 5:88, 2004.
- [389] P. Ferrara, J. Apostolakis, and A. Cafisch. Targeted molecular dynamics simulations of protein unfolding. *J. Phys. Chem. B*, 104:4511–4518, 2000.
- [390] A. R. Ferré-D’Amaré and J. A. Doudna. RNA folds: Insights from recent crystal structures. *Ann. Rev. Biophys. Biomol. Struc.*, 28:57–73, 1999.
- [391] A. R. Ferré-D’Amaré, K. Zhou, and J. A. Doudna. Crystal structure of a hepatitis delta virus ribozyme. *Nature*, 395:567–574, 1998.
- [392] A. M. Ferrenberg, D. P. Landau, and Y. J. Wong. Monte Carlo simulations: Hidden errors from “good” random number generators. *Phys. Rev. Lett.*, 69:3382–3384, 1992.
- [393] M. Ferrer, T. A. Kapoor, T. Strassmaier, W. Weissenhorn, J. J. Skehel, D. Oprian, S. L. Schreiber, D. C. Wiley, and S. C. Harrison. Selection of gp41-mediated HIV-1 cell entry inhibitors from biased combinatorial libraries of non-natural binding elements. *Nature Struc. Biol.*, 6:953–960, 1999.

- [394] A. Fersht. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*. W. H. Freeman and Company, New York, NY, 1999.
- [395] A. R. Fersht. Perspectives. *Nature*, 9:650–654, 2008.
- [396] A. R. Fersht and V. Daggett. Protein folding and unfolding at atomic resolution. *Cell*, 108:573–582, 2002.
- [397] J. Fiaux, E. B. Bertelsen, A. L Horwich, and K. Wüthrich. NMR analysis of a 900K GroEL–GroES complex. *Nature*, 418:207–211, 2002.
- [398] M. J. Field. *A Practical Introduction to the Simulation of Molecular Systems*. Cambridge University Press, Cambridge, UK, second edition, 2007.
- [399] F. Figueirido, R. M. Levy, R. Zhou, and B. J. Berne. Large scale simulation of macromolecules in solution: Combining the periodic fast multipole method with multiple time step integrators. *J. Chem. Phys.*, 106:9835–9849, 1997. (Erratum published in *J. Chem. Phys.* 107:7002, 1997).
- [400] D. Filmore. Taming the beast. *Mod. Drug Dis.*, 4:40–46, 2001.
- [401] J. T. Finch, L. C. Lutter, D. Rhodes, A. S. Brown, B. Rushton, M. Levitt, and A. Klug. Structure of nucleosome core particles of chromatin. *Nature*, 269:29–36, 1977.
- [402] D. Fincham. Optimisation of the Ewald sum for large systems. *Mol. Sim.*, 13:1–9, 1994.
- [403] B. G. Fitch, A. Rayshubskiy, M. Eleftheriou, T. J. C. Ward, M. Giampapa, M. C. Pitman, and R. S. Germain. Blue matter: approaching the limits of concurrency for classical molecular dynamics. In *Supercomputing, 2006. SC'06. Proceedings of the ACM/IEEE SC 2006 Conference*, pages 44–44, 2006.
- [404] M. Fixman. Construction of Langevin forces in the simulation of hydrodynamic interaction. *Macromolecules*, 19:1204–1207, 1986.
- [405] D. Flatters and R. Lavery. Sequence-dependent dynamics of TATA-box binding sites. *Biophys. J.*, 75:372–381, 1998.
- [406] D. Flatters, M. Young, D. L. Beveridge, and R. Lavery. Conformational properties of the TATA-box binding sequence of DNA. *J. Biomol. Struct. Dynam.*, 14:757–765, 1997.
- [407] R. Fletcher. *Practical Methods of Optimization*. John Wiley & Sons, Tiptree, Essex, Great Britain, second edition, 1987.
- [408] J. Florián, M. F. Goodman, and A. Warshel. Computer simulation of the chemical catalysis of DNA polymerases: Discriminating between alternative nucleotide insertion mechanisms for T7 DNA polymerase. *J. Amer. Chem. Soc.*, 125:8163–8177, 2003.
- [409] J. Florián, M. F. Goodman, and A. Warshel. Computer simulations of protein functions: searching for the molecular origin of the replication fidelity of DNA polymerases. *Proc. Natl. Acad. Sci. USA*, 102:6819–6824, 2005.
- [410] P. J. Flory. *Statistical Mechanics of Chain Molecules*. Oxford University Press, New York, NY, 1988. (Reprinted version of the 1969 text with an added excerpt from Flory's Nobel address).
- [411] C. A. Floudas and P. Pardalos, editors. *Optimization in Computational Chemistry and Molecular Biology: Local and Global Approaches*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000.

- [412] F. Fogolari, A. Brigo, and H. Molinari. The Poisson-Boltzmann equation for biomolecular electrostatics: a tool for structural biology. *J. Mol. Recognit.*, 15:377–392, 2002.
- [413] M. C. Foley and T. Schlick. Simulations of dna pol  $\lambda$  R517 mutants indicate 517's crucial role in ternary complex stability and suggest DNA slippage origin. *J. Amer. Chem. Soc.*, 130:3967–3977, 2008.
- [414] N. Foloppe, B. Hartmann, L. Nilsson, and A. D. MacKerell, Jr. Intrinsic conformational energetics associated with the glycosyl torsion in DNA: A quantum mechanical study. *Biophys. J.*, 82:1554–1569, 2002.
- [415] N. Foloppe and A. D. MacKerell, Jr. All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phased macromolecular target data. *J. Comput. Chem.*, 21:86–104, 2000.
- [416] J.J. Forman, P.A. Clemons, S.L. Schreiber, and S.J. Haggarty. SpectralNET—an application for spectral graph analysis and visualization. *BMC Bioinformatics*, 6:260, 2005.
- [417] A. Forsgren, P. E. Gill, and M. H. Wright. Interior methods for nonlinear optimization. *SIAM Rev.*, 44:525–597, 2002.
- [418] J. Frank. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. Academic Press, San Diego, CA, 1996.
- [419] J. Frank. How the ribosome works. *Sci. Amer.*, 86:428–439, 1998.
- [420] M. D. Frank-Kamenetskii. *Unravelling DNA*. VCH Publishers, New York, NY, 1993. (Translated from Russian by L. Liapin).
- [421] M. D. Frank-Kamenetskii. Triplex DNA structures. *Ann. Rev. Biochem.*, 64:65–95, 1995.
- [422] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes. The energy landscapes and motions of proteins. *Science*, 254:1598–1603, 1991.
- [423] H. Frauenfelder and P. G. Wolynes. Biomolecules: Where the physics of complexity and simplicity meet. *Phys. Today*, 47:58–64, 1994.
- [424] P. L. Freddolino, F. Liu, M. Gruebele, and K. Schulten. Ten-microsecond molecular dynamics simulation of a fast-folding WW domain. *Biophys. J.*, 94:L75–L77, 2008.
- [425] P. L. Freddolino, A. S. Arkhipov, S. B. Larson, A. McPherson, and K. Schulten. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure*, 14:437–449, 2006.
- [426] P. L. Freddolino, S. Park, B. Roux, and K. Schulten. Force field bias in protein folding simulations. *Biophys. J.*, 96:3772–3780, 2009.
- [427] P. L. Freddolino and K. Schulten. Common structural Transitions in explicit-solvent simulations of villin headpiece folding. *Biophys. J.*, 97:2338–2347, 2009.
- [428] D. Frenkel and B. Smit. *Understanding Molecular Simulations. From Algorithms to Applications*. Academic Press, San Diego, CA, second edition, 2002.
- [429] M. Friedrichs, R. Zhou, S. R. Edinger, and R. A. Friesner. Poisson-Boltzmann analytical gradients for molecular modeling calculations. *J. Phys. Chem. B*, 103:3057–3061, 1999.

- [430] R. A. Friesner and J. R. Gunn. Computational studies of protein folding. *Annu. Rev. Biophys. Biomol. Struc.*, 25:315–342, 1996.
- [431] F. B. Fuller. The writhing number of a space curve. *Proc. Natl. Acad. Sci. USA*, 68:815–819, 1971.
- [432] F. B. Fuller. Decomposition of the linking number of a closed ribbon: A problem from molecular biology. *Proc. Natl. Acad. Sci. USA*, 75:3557–3561, 1978.
- [433] E. A. Galburd and B. L. Stoddard. Time-resolved macromolecular crystallography. *Phys. Today*, 54:33–39, 1989.
- [434] R. C. Gallo. A reflection on HIV/AIDS research after 25 years. *Retrovirology*, 3:72, 2006.
- [435] H. H. Gan, S. Pasquali, and T. Schlick. Exploring the repertoire of RNA secondary motifs using u graph theory: Implications for RNA design. *Nuc. Acids Res.*, 31:2926–2943, 2003.
- [436] H. H. Gan, R. A. Perlow, S. Roy, J. Ko, M. Wu, J. Huang, S. Yan, A. Nicoletta, J. Vafai, D. Sun, L. Wang, J. E. Noah, S. Pasquali, and T. Schlick. Analysis of protein sequence/structure similarity relationships. *Biophys. J.*, 83:2781–2791, 2002.
- [437] H. H. Gan, A. Tropsha, and T. Schlick. Lattice protein folding with two and four-body statistical potentials. *Proteins: Struc. Func. Gen.*, 43:161–174, 2001.
- [438] H.H. Gan, D. Fera, J. Zorn, M. Tang, N. Shiffeldrim, U. Laserson, N. Kim, and T. Schlick. RAG: RNA-As-Graphics database – concepts, analysis, and features. *Bioinformatics*, 20:1285–1291, 2004.
- [439] R. M. Ganunis, H. Guo, and T. D. Tullius. Effect of the crystallizing agent 2-methyl-2,4-pentanediol on the structure of adenine tract DNA in solution. *Biochemistry*, 35:13729–13732, 1996.
- [440] F. Gao, E. Bailes, D. L. Robertson, Y. Chen, C. M. Rodenburg, S. F. Michael, L. B. Cummins, L. O. Arthur, M. Peeters, G. M. Shaw, P. M. Sharp, and B. H. Hahn. Origin of HIV-1 in the chimpanzee *pan troglodytes troglodytes*. *Nature*, 397:436–441, 1999.
- [441] J. Gao, S. Ma, D. T. Major, K. Nam, J. Pu, and D. G. Truhlar. Mechanisms and free energies of enzymatic reactions. *Chem. Rev.*, 106:3188–3209, 2006.
- [442] J. Gao and B. Xu. Applications of nanomaterials inside cells. *Nano Today*, 4:37–51, 2009.
- [443] A. E. Garcia and J. N. Onuchic. Folding a protein in a computer: an atomic description of the folding/unfolding of protein A. *Proc. Natl. Acad. Sci. USA*, 100:13898–13903, 2003.
- [444] A. E. Garcia and D. Pascheck. Simulation of the pressure and temperature folding/unfolding equilibrium of a small RNA hairpin. *J. Amer. Chem. Soc.*, 130:815–817, 2008.
- [445] A. E. García and K. Y. Sanbonmatsu.  $\alpha$ -Helical stabilization by side chain shielding of backbone hydrogen bonds. *Proc. Natl. Acad. Sci. USA*, 99:2782–2787, 2002.
- [446] B. García-Archilla, J.M. Sanz-Serna, and R.D. Skeel. Long-time-step methods for oscillatory differential equations. *SIAM J. Sci. Comput.*, 20:930–963, 1998.

- [447] C. W. Gear. *Numerical Initial Value Problems in Ordinary Differential Equations*. Prentice Hall, Englewood Cliffs, New Jersey, 1971.
- [448] P. Gedeck and P. Willet. Visual and computational analysis of structure-activity relationships in high-throughput screening data. *Curr. Opin. Chem. Biol.*, 5:389–395, 2001.
- [449] C. A. Gelfand, G. E. Plum, S. Mielewczyk, D. P. Remeta, and K. J. Breslauer. A quantitative method for evaluating the stabilities of nucleic acid complexes. *Proc. Natl. Acad. Sci. USA*, 96:6113–6118, 1999.
- [450] J. Gevertz, H. H. Gan, and T. Schlick. *In Vitro* RNA random pools are not structurally diverse: A computational analysis. *RNA*, 11:853–863, 2005.
- [451] A. K. Ghose, V. N. Viswanadhan, and J. J. Wendoloski. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *J. Comb. Chem.*, 1:55–68, 1999.
- [452] N. Ghosh and Q. Cui. pKa of residue 66 in *Staphylococal nuclease*. I. insights from QM/MM simulations with conventional sampling. *J. Phys. Chem.*, 112: 8387–8397, 2008.
- [453] S. Ghosh, A. Nie, J. An, and Z. Huang. Structure-based virtual screening of chemical libraries for drug discovery. *Curr. Opin. Chem. Biol.*, 10:194–202, 2006.
- [454] W. W. Gibbs. Evolution in a bottle: Synthetic life oozes closer to reality. self-replicating RNAs advance science another step toward artificial life. *Sci. Amer.*, 300:18–21, 2009.
- [455] K. B. Gibson and H. A. Scheraga. Decisions in force field development: Reply to Kollman and Dill. *J. Biomol. Struct. Dyn.*, 8:1109–1111, 1991.
- [456] J. C. Gilbert and C. Lemarechal. Some numerical experiments with variable-storage quasi-Newton algorithms. *Math. Prog. B*, 45:407–435, 1989.
- [457] J. C. Gilbert and J. Nocedal. Global convergence properties of conjugate gradient methods for optimization. Technical Report 1268, Institut National de Recherche en Informatique et en Automatique, January 1991.
- [458] P. E. Gill and M. W. Leonard. Reduced-Hessian quasi-Newton methods for unconstrained optimization. *SIAM J. Optim.*, 12:209–237, 2001.
- [459] P. E. Gill, W. Murray, and M. H. Wright. *Practical Optimization*. Academic Press, London, England, 1983.
- [460] P. M. W. Gill. A new expansion of the Coulomb interaction. *Chem. Phys. Lett.*, 270:193–195, 1997.
- [461] P. Gkekka and L. Sarkisov. Spontaneous formation of a barrel-stave pore in a coarse-grained model of the synthetic LS3 peptide and a DPPC lipid bilayer. *J. Phys. Chem. B*, 113:6–8, 2009.
- [462] N. M. Glykos, G. Cesareni, and M. Kokkinidis. Protein plasticity to the extreme: Changing the topology of a 4-helical bundle with a single amino acid substitution. *Struc. Fold. Design*, 7:597–603, 1999.
- [463] N. Gō and H. Taketomi. Respective roles of short- and long-range interactions in protein folding. *Proc. Natl. Acad. Sci. USA*, 75:559–563, 1978.
- [464] S. Goedecker. Linear scaling electronic structure methods. *Rev. Mod. Phys.*, 71:1085–1123, 1999.

- [465] V. Gogonea, D. Suárez, A. van der Vaart, and K. M. Merz, Jr. New developments in applying quantum mechanics to proteins. *Curr. Opin. Struct. Biol.*, 11:217–223, 2001.
- [466] B. L. Golden, H. Kim, and E. Chase. Crystal structure of a phage Twort Group I ribozyme-product complex. *Nat. Struct. Mol. Biol.*, 12:82–89, 2005.
- [467] G. H. Golub and C. F. van Loan. *Matrix Computations*. John Hopkins University Press, Baltimore, MD, second edition, 1986.
- [468] O. Gonzales and J. C. Simo. On the stability of symplectic and energy-momentum conserving algorithms for nonlinear Hamiltonian systems with symmetry. *Comput. Meth. App. Mech. Engin.*, 134:197, 1994.
- [469] J. A. González and R. Pino. A random number generator based on unpredictable chaotic functions. *Comput. Phys. Comm.*, 120:109–114, 1999.
- [470] H. Gonzlez-Daz, Y. Gonzlez-Daz, L. Santana, F.M. Ubeira, and E. Uriarte. Proteomics, networks and connectivity indices. *Proteomics*, 8:750–778, 2008.
- [471] H. Gonzlez-Daz, S. Vilar, L. Santana, and E. Uriarte. Medicinal chemistry and bioinformatics—current trends in drugs discovery with networks topological indices. *Curr. Top. Med. Chem.*, 7:1015–1029, 2007.
- [472] L. Goodman, V. Pophristic, and F. Weinhold. Origin of methyl internal rotation barriers. *Acc. Chem. Res.*, 32:983–993, 1999.
- [473] A. A. Gorin, V. B. Zhurkin, and W. K. Olson. B-DNA twisting correlates with base pair morphology. *J. Mol. Biol.*, 247:34–48, 1995.
- [474] H. Gould, J. Tobochnik, and W. Christian. *An Introduction to Computer Simulation Methods: Applications to Physical Systems*. Addison-Wesley, San Francisco, CA, third edition, 2007.
- [475] P. Grayson, E. Tajkhorshid, and K. Schulten. Mechanisms of selectivity in channels and enzymes studied with interactive molecular dynamics. *Biophys. J.*, 85:36, 2003.
- [476] P. Green. Whole-genome disassembly. *Proc. Natl. Acad. Sci. USA*, 99:4143–4144, 2002.
- [477] L. Greengard. *The Rapid Evaluation of Potential Fields in Particle Systems*. MIT Press, Cambridge, Massachusetts, 1988.
- [478] L. Greengard. Fast algorithms for classical physics. *Science*, 265:909–914, 1994.
- [479] L. Greengard and V. Rokhlin. A fast algorithm for particle simulation. *J. Comput. Phys.*, 73:325–348, 1987.
- [480] L. Greengard and V. Rokhlin. On the evaluation of electrostatic interactions in molecular modeling. *Chemica Scripta*, 29A:139–144, 1989.
- [481] L. Greengard and V. Rokhlin. A new version of the fast multipole method for the Laplace equation in three dimensions. *Acta Numerica*, 6:229–269, 1997.
- [482] A. Griewank and G. F. Corliss, editors. *Automatic Differentiation of Algorithms: Theory, Implementation, and Applications*. SIAM, Philadelphia, PA, 1991.
- [483] S. A. Grigoryev, G. Arya, S. Correll, C. L. Woodcock, and T. Schlick. Evidence for heteromeric chromatin fibers from analysis of nucleosome interactions. *Proc. Natl. Acad. Sci. USA*, 106:13317–13322, 2009.

- [484] J. M. Grimes, J. N. Burroughs, P. Gouet, J. M. Diprose, R. Malby, S. Ziéntara, P. P. C. Mertens, and D. I. Stuart. The atomic structure of the bluetongue virus core. *Nature*, 395:470–478, 1998.
- [485] H. Grubmüller. Predicting slow structural transitions in macromolecular systems: Conformational flooding. *Phys. Rev. E*, 52:2893–2906, 1995.
- [486] H. Grubmüller, H. Heller, A. Windemuth, and K. Schulten. Generalized Verlet algorithm for efficient molecular dynamics simulations with long-range interactions. *Mol. Sim.*, 6:121–142, 1991.
- [487] P. Güntert. Structure calculation of biological macromolecules from NMR data. *Quart. Rev. Biophys.*, 31:145–237, 1998.
- [488] F. Guo, A. R. Gooding, and T. R. Cech. Structure of the tetrahymena ribozyme: base triple sandwich and metal ion at the active site. *Mol Cell.*, 16:351–362, 2004.
- [489] O. Guvench and Jr. A. D. MacKerell. Comparison of protein force fields for molecular dynamics simulations. In A. Kukol, editor, *Methods in Molecular Biology*, volume 443, pages 63–88. Humana Press, Totowa, NJ, 2008.
- [490] W.W. Hager and H. Zhang. A new conjugate gradient method with guaranteed descent and an efficient line search. *SIAM J. Opt.*, 16:170, 2005.
- [491] P. J. Hagerman. Flexibility of DNA. *Ann. Rev. Biophys. Biophys. Chem.*, 17: 265–286, 1988.
- [492] P. J. Hagerman. Straightening out the bends in curved DNA. *Biochim. Biophys. Acta*, 1131:125–132, 1992.
- [493] P. J. Hagerman. Flexibility of RNA. *Ann. Rev. Biophys. Biomol. Struc.*, 26: 139–156, 1997.
- [494] J. M. Haile. *Molecular Dynamics Simulations: Elementary Methods*. John Wiley & Sons, New York, NY, 1992.
- [495] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems*, volume 14 of *Springer Series in Computational Mathematics*. Springer-Verlag, New York, NY, second edition, 1996.
- [496] P. J. Hajduk and J. Greer. A decade of fragment-based drug design: Strategic advances and lessons learned. *Nat. Rev. Drug Disc.*, 6:211–219, 2007.
- [497] T. A. Halgren. Merck molecular force field: I. Basis, form, scope, parameterization and performance of MMFF94. *J. Comput. Chem.*, 17:490–519, 1996.
- [498] T. A. Halgren. Merck molecular force field: II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. *J. Comput. Chem.*, 17:520–552, 1996.
- [499] T. A. Halgren. MMFF VI. MMFF94s option for energy minimization studies. *J. Comput. Chem.*, 20:720–729, 1999.
- [500] T. A. Halgren. MMFF VII. Characterization of MMFF94, MMFF94s, and other widely available force fields for conformational energies and for interaction energies and geometries. *J. Comput. Chem.*, 20:730–748, 1999.
- [501] T. A. Halgren and W. Damm. Polarizable force fields. *Curr. Opin. Struct. Biol.*, 11:236–242, 2001.
- [502] K. B. Hall. RNA in motion. *Curr. Opin. Chem. Biol.*, 12:612–618, 2008.
- [503] M. Hamada, K. Tsuda, T. Kudo, T. Kin, and K. Asai. Mining frequent stem patterns from unaligned RNA sequences. *Bioinformatics*, 22:2480–2487, 2006.

- [504] D. Hamelberg, C. A. F. de Oliveira, and J. A. McCammon. Sampling of slow diffusive conformational transitions with accelerated molecular dynamics. *J. Chem. Phys.*, 127:155102, 2007.
- [505] P. Hammarström, F. Schneider, and J. W. Kelly. *Trans*-suppression of misfolding in an amyloid disease. *Science*, 293:2459–2462, 2001.
- [506] W. Han, C.-K. Wan, and Y.-D. Wu. Toward a coarse-grained protein model coupled with a coarse-grained solvent model: solvation free energies of amino acid side chains. *J. Chem. Theo. Comp.*, 4:1891–1901, 2008.
- [507] M. Hann and R. Green. Cheminformatics – A new name for an old problem? *Curr. Opin. Chem. Biol.*, 3:379–383, 1999.
- [508] U. H. E. Hansmann. Parallel-tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.*, 281:140–150, 1997.
- [509] U. H. E. Hansmann and Y. Okamoto. New Monte Carlo algorithms for protein folding. *Curr. Opin. Struct. Biol.*, 9:177–183, 1999.
- [510] T. Hansson, C. Oostenbrink, and W. F. van Gunsteren. Molecular dynamics simulations. *Curr. Opin. Struct. Biol.*, 12:190–196, 2002.
- [511] B. Hao, W. Gong, T. K. Ferguson, C. M. James, J. A. Krzycki, and M. K. Chan. A new UAG-encoded residue in the structure of a methanogen methyltransferase. *Science*, 296:1462–1466, 2002.
- [512] M.-H. Hao and W. K. Olson. Modeling DNA supercoils and knots with B-spline functions. *Biopolymers*, 28:873–900, 1989.
- [513] M. H. Hao, M. R. Pincus, S. Rackovsky, and H. A. Scheraga. Unfolding and refolding of the native structure of bovine pancreatic trypsin inhibitor studied by computer simulations. *Biochemistry*, 32:9614–9631, 1993.
- [514] T. E. Haran, J. D. Kahn, and D. M. Crothers. Sequence elements responsible for DNA curvature. *J. Mol. Biol.*, 244:135–143, 1994.
- [515] T. E. Haran and U. Mohanty. The unique structure of A-tracts and intrinsic DNA bending. *Quart. Rev. Biophys.*, 42:41–81, 2009.
- [516] J. Harms, F. Schluenzen, R. Zarivach, A. Bashan, S. Gat, I. Agmon, H. Bartels, F. Franceschi, and A. Yonath. High resolution structure of the large ribosomal subunit from a mesophilic eubacterium. *Cell*, 107:679–688, 2001.
- [517] H. S. Harned and B. B. Owen. *The Physical Chemistry of Electrolytic Solutions*. American Chemical Society Monograph Series. Reinhold Publishing Corporation, New York, NY, second edition, 1950.
- [518] W. E. Harte, Jr., S. Swaminathan, and D. L. Beveridge. Molecular dynamics of HIV-1 protease. *Proteins: Struc. Func. Gen.*, 13:175–194, 1992.
- [519] F. U. Hartl and M. H.-Hartl. Molecular chaperones in the cytosol: Nascent chain to folded protein. *Science*, 295:1852–1858, 2002.
- [520] S. C. Harvey, M. Dlakic, J. Griffith, R. Harrington, K. Park, D. Sprous, and W. Zacharias. What is the basis of sequence-directed curvature in DNAs containing A-tracts? *J. Biomol. Struct. Dynam.*, 13:301–307, 1995.
- [521] S. C. Harvey and H. A. Gabb. Conformational transitions using molecular dynamics with minimum biasing. *Biopolymers*, 33:1167–1172, 1993.

- [522] S. C. Harvey and M. Prabhakaran. Ribose puckering: Structure, dynamics, energetics and the pseudorotation cycle. *J. Amer. Chem. Soc.*, 108:6128–6136, 1986.
- [523] S. C. Harvey, M. Prabhakaran, B. Mao, and J. A. McCammon. Phenylalanine transfer RNA: Molecular dynamics simulation. *Science*, 223:1189–1191, 1984.
- [524] S. C. Harvey, R. K.-Z. Tan, and T. E. Cheatham, III. The flying ice cube: Velocity rescaling in molecular dynamics leads to violation of energy equipartition. *J. Comput. Chem.*, 19:726–740, 1998.
- [525] Y. Hashem and P. Auffinger. A short guide for molecular dynamics simulations of RNA systems. *Methods*, 47:187–197, 2009.
- [526] Y. Hashem, E. Westhof, and P. Auffinger. Milestones in molecular dynamics simulations of RNA systems. In T. Schwede and M. Peitsch, editors, *Computational Structural Biology*, pages 363–399. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2008.
- [527] M. A. El Hassan and C. R. Calladine. Conformational characteristics of DNA: Empirical classifications and a hypothesis for the conformational behaviour of dinucleotide steps. *Phil. Trans. Math. Phys. Engin. Sci.*, 355:43–100, 1997.
- [528] S. A. Hassan, F. Guarnieri, and E. L. Mehler. A general treatment of solvent effects based on screened Coulomb potentials. *J. Phys. Chem. B*, 104:6478–6489, 2000.
- [529] W. A. Hasteline. Beyond chicken soup. *Sci. Amer.*, 285:56–63, 2001.
- [530] A. Hastings, P. Arzberger, B. Bolker, S. Collins, A. Irves, N. Johnson, and M. Palmer. Quantitative bioscience for the 21st century. *Bioscience*, 55:511–517, 2005.
- [531] H. A. Hauptman. The phase problem of X-ray crystallography. *Phys. Today*, 42:24–29, 1989.
- [532] D. M. Hayes, P.A. Kollman, and S. Rothenberg. A conformational analysis of  $\text{H}_3\text{PO}_4$ ,  $\text{H}_3\text{PO}_4^-$ ,  $\text{HPO}_4^{2-}$  and related model compounds. *J. Amer. Chem. Soc.*, 99:2150–2154, 1977.
- [533] T. Haynes, D. Knisley, and J. Knisley. Using a neural network to identify secondary RNA structures quantified by graph invariants. *Comm. in Math. Comp. Chem.*, 60:277, 2008.
- [534] T. Haynes, D. Knisley, E. Seier, and Y. Zou. A quantitative analysis of secondary RNA structure using domination based parameters on trees. *BMC Bioinformatics*, 7:108, 2006.
- [535] S. Hayward and B. L. deGroot. Normal modes and essential dynamics. *Methods Mol. Biol.*, 443:89–106, 2008.
- [536] D. J. Hazuda, P. Felock, M. Witmer, A. Wolfe, K. Stillmock, J. A. Grobler, A. Espeseth, L. Gabryelski, W. Schleif, C. Blau, and Michael D. Miller. Inhibitors of strand transfer that prevent integration and inhibit HIV-1 replication in cells. *Science*, 287:646–650, 2000.
- [537] C. E. Hecht. *Statistical Thermodynamics and Kinetic Theory*. W. H. Freeman, New York, NY, 1990.
- [538] C. E. Heitsch. Analyzing the branching degree of RNA viral genomes: a hepatitis C case study. *The Ninth Annual International Conference on Research in Computational Molecular Biology (RECOMB 2005)*, 2005.

- [539] D. M. Held, J. D. Kissel, J. T. Patterson, D. G. Nickens, and D. H. Burke. HIV-1 inactivation by nucleic acid aptamers. *Front Biosci.*, 11:89–112, 2006.
- [540] W. A. Hendrickson. Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science*, 254:51–58, 1991.
- [541] W. A. Hendrickson and C. Ogata. Phase determination from multiwavelength anomalous diffraction measurements. *Meth. Enzymol.*, 276:494–523, 1997.
- [542] D. K. Hendrix, S. E. Brenner, and S. R. Holbrook. RNA structural motifs: building blocks of a modular biomolecule. *Q. Rev. Biophys.*, 38:221–243, 2005.
- [543] G. Henkelman and H. Jónsson. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.*, 113:9978–9985, 2000.
- [544] C. M. Henry. Pharmacogenomics. *Chem. Engin. News*, 79:37–42, 2001.
- [545] T. Hermann and D. J. Patel. Stitching together RNA tertiary architectures. *J. Mol. Biol.*, 294:828–849, 1999.
- [546] T. Hermann and D. J. Patel. Adaptive recognition by nucleic acid aptamers. *Science*, 287:820–825, 2000.
- [547] T. L. Hill. *An Introduction to Statistical Thermodynamics*. Dover, New York, NY, 1986.
- [548] B. E. Hingerty, R. H. Ritchie, T. L. Ferrell, and J. E. Turner. Dielectric effects in biopolymers: The theory of ionic saturation revisited. *Biopolymers*, 24:427–439, 1985.
- [549] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization. Algorithms I*, volume 305 of *Grundlehren der mathematischen Wissenschaften. A Series of Comprehensive Studies in Mathematics*. Springer-Verlag, Berlin and Heidelberg, 1993.
- [550] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization. Algorithms II*, volume 306 of *Grundlehren der mathematischen Wissenschaften. A Series of Comprehensive Studies in Mathematics*. Springer-Verlag, Berlin and Heidelberg, 1993.
- [551] J. Hizver, H. Rozenberg, F. Frolov, D. Rabinovich, and Z. Shakked. DNA bending by an adenine-thymine tract and its role in gene regulation. *Proc. Natl. Acad. Sci. USA*, 98:8490–8495, 2001.
- [552] P. Hobza and J. Šponer. Structure, energetics, and dynamics of the nucleic acid base pairs: Nonempirical Ab Initio calculations. *Chem. Rev.*, 99:3247–3276, 1999.
- [553] R. W. Hockney and J. W. Eastwood. *Computer Simulation Using Particles*. McGraw-Hill, New York, NY, 1981.
- [554] R. W. Hockney and J. W. Eastwood. *Computer Simulation Using Particles*. Institute of Physics, London, England, 1988.
- [555] E. Hodgkin and K. Andrew-Cramer. Compound collections get focused. *Modern Drug Discovery*, 3:55–60, 2000.
- [556] C.-J. Höglberg, A. M. Nikitin, and A. P. Lyubartsev. Modification of the CHARMM force field for DMPC lipid bilayer. *J. Comput. Chem.*, 29:2359–2369, 2008.
- [557] S. R. Holbrook. RNA structure: The long and the short of it. *Curr. Opin. Struct. Biol.*, 15:302–308, 2005.

- [558] S. R. Holbrook and S.-H. Kim. RNA crystallography. *Biopolymers*, 44:3–21, 1997.
- [559] B. L. Holian, O. E. Percus, T. T. Warnock, and P. A. Whitlock. Pseudorandom number generator for massively parallel molecular-dynamics applications. *Phys. Rev. E*, 50:1607–1615, 1994.
- [560] L. Holm and C. Sander. Mapping the protein universe. *Science*, 273:595–602, 1996.
- [561] A. Holmgren and C.-I. Bränden. Crystal structure of chaperone protein PapD reveals an immunoglobulin fold. *Nature*, 342:248–251, 1989.
- [562] S. K. Holmgren, K. M. Taylor, L. E. Bretscher, and R. T. Raines. Code for collagen’s stability deciphered. *Nature*, 392, 1998.
- [563] M. Holst, N. Baker, and E. Wang. Adaptive multilevel finite element solution of the Poisson-Boltzmann equation I. Algorithms and examples. *J. Comput. Chem.*, 21:1319–1342, 2000.
- [564] B. Honig. Protein folding: From the Levinthal paradox to structure prediction. *J. Mol. Biol.*, 293:283–293, 1999.
- [565] B. Honig and A. Nicholls. Classical electrostatics in biology and chemistry. *Science*, 268:1144–1149, 1995.
- [566] W. Hoover. Classical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A*, 31:1695–1697, 1985.
- [567] P. J. Horn and C. L. Peterson. Chromatin higher order folding: Wrapping up transcription. *Science*, 297:1824–1827, 2002.
- [568] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Struct., Funct., Bioinf.*, 65:712–725, 2006.
- [569] M. P. Horvath and S. C. Schultz. DNA G-quartets in a 1.86 Å resolution structure of an *Oxytricha Nova* telomeric protein-DNA complex. *J. Mol. Biol.*, 310:367–377, 2001.
- [570] W. A. Houry, D. Frishman, C. Eckerskorn, F. Lottspeich, and F. U. Hartl. Identification of *in vivo* substrates of the chaperonin GroEL. *Nature*, 402:147–154, 1999.
- [571] K. Howard. The bioinformatics gold rush. *Sci. Amer.*, 283:58–63, 2000.
- [572] H. Hu, Z. Y. Lu, and W. T. Yang. QM/MM minimum free-energy path: Methodology and application to triosephosphate isomerase. *J. Chem. Theory Comput.*, 3:390–406, 2007.
- [573] H. Hu and W. Yang. Free energies of chemical reactions in solution and in enzymes with Ab Initio quantum mechanics/molecular mechanics methods. *Annu. Rev. Phys. Chem.*, 59:573–601, 2008.
- [574] H. Hu and W. Yang. Development and application of *ab initio* QM/MM methods for mechanistic simulation of reactions in solution and in enzymes. *J. Mol. Struct.: THEOCHEM*, 898:17–30, 2009.
- [575] J. Hu, A. Ma, and R. Dinner. Monte Carlo simulations of biomolecules: the MC module in CHARMM. *J. Comput. Chem.*, 27:203–216, 2006.
- [576] H. Huang, R. Chopra, G. L. Verdine, and S. C. Harrison. Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: Implications for drug resistance. *Science*, 282:1669–1675, 1998.

- [577] J. Huang, T. Schlick, and A. Vologodskii. Dynamics of site juxtaposition in supercoiled DNA. *Proc. Natl. Acad. Sci. USA*, 98:968–973, 2001.
- [578] N. V. Hud and J. Feigon. Localization of divalent metal ions in the minor groove of DNA A-tracts. *J. Amer. Chem. Soc.*, 119:5756–5757, 1997.
- [579] D. E. Humphreys, R. A. Friesner, and B. J. Berne. A multiple-time-step molecular dynamics algorithm for macromolecules. *J. Phys. Chem.*, 98(27):6885–6892, 1994.
- [580] P. H. Hünenberger and J. A. McCammon. Effect of artificial periodicity in simulations of biomolecules under Ewald boundary conditions: A continuum electrostatics study. *Biophys. Chem.*, 78:69–88, 1999.
- [581] P. H. Hünenberger and J. A. McCammon. Ewald artifacts in computer simulations of ionic solvation and ion-ion interaction: A continuum electrostatics study. *J. Chem. Phys.*, 110:1856–1872, 1999.
- [582] P. A. Hunt. QSAR using 2D descriptors and TRIPoS' SIMCA. *J. Comput.-Aided Mol. Design*, 13:453–467, 1999.
- [583] S. Huo and J. E. Straub. The MaxFlux algorithm for calculating variationally optimized reaction paths for conformational transitions in many body systems at finite temperature. *J. Chem. Phys.*, 107:5000–5006, 1997.
- [584] T. Ideker, T. Galitski, and L. Hood. A new approach to decoding life: Systems biology. *Ann. Rev. Genom. Hum. Genet.*, 2:343–372, 2001.
- [585] J. Chen W. Im and C. L. Brooks, III. Application of torsion angle molecular dynamics for efficient sampling of protein conformations. *J. Comput. Chem.*, 26:1565–1578, 2005.
- [586] W. Im, D. Beglov, and B. Roux. Continuum solvation model: Computation of electrostatic forces from numerical solutions to the Poisson-Boltzmann equation. *Comput. Phys. Comm.*, 111:59–75, 1998.
- [587] W. Im, J. Chen, and C. L. Brooks, III. Peptide and protein folding and conformational equilibria: Theoretical treatment of electrostatics and hydrogen bonding with implicit solvent models. *Adv. Protein Chem.*, 72:173–197, 2006.
- [588] M. Ingelman-Sundberg. Pharmacogenomic biomarkers for prediction of severe adverse drug reactions. *N. Eng. J. Med.*, 358:637–639, 2008.
- [589] J. Inglese, D. S. Auld, A. Jadhav, R. L. Johnson, A. Simeonov, A. Yasgar, W. Zheng, and C. P. Austin. Quantitative high-throughput screening qHTS: A titration-based approach that efficiently identifies biological activities in large chemical libraries. *Proc. Natl. Acad. Sci. USA*, 103:11473–11478, 2006.
- [590] V. M. Ingram. Hemoglobin: The chemical difference between normal and sickle cell hemoglobin. *Nature*, 180:326–328, 1957.
- [591] The International Warfarin Pharmacogenetics Consortium. Estimation of the warfarin dose with clinical and pharmacogenetic data. *N. Engl. J. Med.*, 360:753–764, 2009.
- [592] B. Isralewitz, J. Baudry, J. Gullingsrud, D. Kosztin, and K. Schulten. Steered molecular dynamics investigations of protein function. *J. Mol. Graph. Model.*, 19:13–25, 2001.
- [593] B. Isralewitz, M. Gao, and K. Schulten. Steered molecular dynamics and mechanical functions of proteins. *Curr. Opin. Struct. Biol.*, 11:224–230, 2001.

- [594] J. A. Izaguirre. *Longer Time Steps for Molecular Dynamics*. PhD thesis, University of Illinois at Urbana-Champaign, 1999. Also UIUC Technical Report UIUCDCS-R-99-2107. Available via [www.cs.uiuc.edu/research/tech-reports.html](http://www.cs.uiuc.edu/research/tech-reports.html).
- [595] J. A. Izaguirre, D. P. Catarello, J. M. Wozniak, and R. D. Skeel. Langevin stabilization of molecular dynamics. *J. Comput. Phys.*, 114:2090–2098, 2001.
- [596] J. A. Izaguirre and S. S. Hampton. Shadow hybrid Monte Carlo: An efficient propagator in phase space of macromolecules. *J. Chem. Phys.*, 200:581–604, 2004.
- [597] S. Izrailev, A. R. Crofts, E. A. Berry, and K. Schulten. Steered molecular dynamics simulation of the Rieske subunit motion in the cytochrome *bc*<sub>1</sub> complex. *Biophys. J.*, 77:1753–1768, 1999.
- [598] A. Jack and M. Levitt. Refinement of large structures by simultaneous minimization of energy and R factor. *Acta Crystallogr.*, A34:931–935, 1978.
- [599] L. Jaeger and A. Chworos. The architectonics of programmable RNA and dna nanostructures. *Curr. Opin. Struct. Biol.*, 16:531–543, 2006.
- [600] L. Jaeger, E. Westhof, and N. B. Leontis. TectoRNA: modular assembly units for the construction of RNA nano-objects. *Nucl. Acids Res.*, 29:455–463, 2001.
- [601] M. Jain, C. Arvanitis, K Chu, W. Dewey, E. Leonhardt, M. Trinh, C. D. Sundberg, J. M. Bishop, and D. W. Felsher. Sustained loss of a neoplastic phenotype by brief inactivation of *MYC*. *Science*, 297:102–104, 2002.
- [602] T. L. James, H. Liu, N. B. Ulyanov, S. Farr-Jones, H. Zhang, D. G. Donne, K. Kaneko, D. Groth, I. Mehlhorn, S. B. Prusiner, and F. E. Cohen. Solution structure of a 142-residue recombinant prion protein corresponding to the infectious fragment of the scrapie isoform. *Proc. Natl. Acad. Sci. USA*, 94:10086–10091, 1997.
- [603] D. Janežič and F. Merzel. An efficient symplectic integration algorithm for molecular dynamics simulations. *J. Chem. Info. Comput. Sci.*, 35:321–326, 1995.
- [604] D. Janežič and B. Orel. Implicit Runge-Kutta method for molecular dynamics integration. *J. Chem. Info. Comput. Sci.*, 33:252–257, 1993.
- [605] R. M. Jendrejack, M. D. Graham, and J. J. de Pablo. Hydrodynamic interactions in long chain polymers: Application of the Chebyshev polynomial approximation in stochastic simulations. *J. Chem. Phys.*, 113:2894–2900, 2000.
- [606] M. Ø Jensen, E. Tajkhorshid, and K. Schulten. The mechanism of glycerol conduction in aquaglyceroporins. *Structure*, 9:1083–1093, 2001.
- [607] H. Jian. *A Combined Wormlike-Chain and Bead Model for Dynamic Simulations of Long DNA*. PhD thesis, New York University, Department of Physics, New York, NY, October 1997.
- [608] H. Jian, T. Schlick, and A. Vologodskii. Internal motion of supercoiled DNA: Brownian dynamics simulations of site juxtaposition. *J. Mol. Biol.*, 284:287–296, 1998.
- [609] L. Jiang, E. A. Althoff, F. R. Clemente, L. Doyle, D. Röthlisberger, A. Zanghellini, J. L. Gallaher, J. L. Betker, F. Tanaka, C. F. Barbas III, D. Hilvert, K. N. Houk, B. L. Stoddard, and D. Baker. De novo computational design of retro-aldol enzymes. *Science*, 319:1387–1391, 2008.

- [610] S. Jo, M. Vargyas, J. Vasko-Szedlar, B. Roux, and W. Im. PBEQ-Solver for online visualization of electrostatic potential of biomolecules. *Nucl. Acids Res.*, 36:W270–W275, 2008.
- [611] M. Johnston and A. Fauci. An HIV vaccine - Challenges and prospects. *N. Engl. J. Med.*, 359:888–890, 2008.
- [612] M. I. Johnston and A. S. Fauci. An HIV vaccine — Evolving concepts. *The New England J. Med.*, 356:2073–2081, 2007.
- [613] V. F. R. Jones. Knot theory and statistical mechanics. *Sci. Amer.*, 263:98–103, 1990.
- [614] S. Ó. Jónsdóttir and K. Rasmussen. The consistent force field. part 6: an optimized set of potential energy functions for primary amines. *New J. Chem.*, 24:243–247, 2000.
- [615] I. K. Jordan, F. A. Kondrashov, I. A. Adzhubei, Y. I. Wolf, E. V. Koonin, A. S. Kondrashov, and S. Sunyaev. A universal trend of amino acid gain and loss in protein evolution. *Nature*, 433:633–638, 2005.
- [616] P. Jordan, P. Fromme, H. T. Witt, O. Klukas, W. Saenger, and N. Krauß. Three-dimensional structure of cyanobacterial photosystem I at 2.5 Å resolution. *Nature*, 411:909–917, 2001.
- [617] W. L. Jorgensen, J. Chandrasekar, J. Madura, R. Impey, and M. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79:926–935, 1983.
- [618] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Amer. Chem. Soc.*, 118:11225–11236, 1996.
- [619] W. L. Jorgensen and J. Tirado-Rives. Monte Carlo vs. molecular dynamics for conformational sampling. *J. Phys. Chem.*, 100:14508–14513, 1996.
- [620] W. L. Jorgensen and J. Tirado-Rives. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc. Natl. Acad. Sci. USA*, 102:6665–6670, 2005.
- [621] G. F. Joyce, W. C. Still, and K. T. Chapman. Combinatorial chemistry. Searching for a winning combination (Editorial overview). *Curr. Opin. Chem. Biol.*, 1:3–4, 1997.
- [622] H. F. Judson. *The Eighth Day of Creation. Makers of the Revolution in Biology*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1996. (Expanded edition).
- [623] D. Junghans and U. H. E. Hansmann. Numerical comparison of Wang-Landau sampling and parallel tempering for met-enkephelin. *Intl. J. Mod. Phys. C*, 17: 817–824, 2006.
- [624] J. Kaiser. Death prompts a review of gene therapy vector. *Science*, 317:580, 2007.
- [625] M. H. Kalos and P. A. Whitlock. *Monte Carlo Methods*. Wiley-VCH, Weinheim, Germany, second edition, 2008.
- [626] R. S. Kamath, A. G. Fraser, Y. Dong, G. Poulin, R. Durbin, M. Gotta, A. Kanapin, N. Le Bot, S. Moreno, M. Sohrmann, D. P. Welchman, P. Zipperlen, and J. Ahringer. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature*, 421:231–237, 2003.

- [627] H. Kamberaj and A. van der Vaart. Multiple scaling replica exchange for the conformational sampling of biomolecules in explicit water. *J. Chem. Phys.*, 127:234102, 2007.
- [628] S. C. L. Kamerlin, M. Haranczyk, and A. Warshel. Progress in *ab initio* QM/MM free-energy simulations of electrostatic energies in proteins: Accelerated QM/MM studies of pK, redox reactions and solvation free energies. *J. Phys. Chem. B*, 113:1253–1272, 2009.
- [629] G. A. Kaminski, R. A. Friesner, J. Tirado-Rives, and W. L. Jorgensen. Evaluation and reparameterization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B*, 105:6474–6487, 2001.
- [630] G. A. Kaminski, H. A. Stern, B. J. Berne, R. A. Friesner, Y. X. Cao, R. B. Murphy, R. Zhou, and T. A. Halgren. Development of a polarizable force field for proteins via *Ab Initio* quantum chemistry: First generation model and gas phase tests. *J. Comput. Chem.*, 23:1515–1531, 2002.
- [631] R. Kanaar and N. R. Cozzarelli. Roles of supercoiled DNA structure in DNA transactions. *Curr. Opin. Struc. Bio.*, 2:369–379, 1992.
- [632] F. Kang. The Hamiltonian way for computing Hamiltonian dynamics. In R. Spigler, editor, *Applied and Industrial Mathematics*, pages 17–35. Kluwer Academic, Dordrecht, The Netherlands, 1990.
- [633] Y. Karklin, R.F. Meraz, and S.R. Holbrook. Classification of non-coding RNA using graph representations of secondary structure. *Pac. Symp. Biocomput.*, pages 4–15, 2005.
- [634] J. Karle. Macromolecular structure from anomalous dispersion. *Phys. Today*, 42:20–22, 1989.
- [635] R. M. Karp. Mathematical challenges from genomics and molecular biology. *Notices Amer. Math. Soc.*, 49:544–553, 2002.
- [636] M. Karplus and J. Kuriyan. Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. USA*, 102:6679–6685, 2005.
- [637] N. W. Kelley, X. Huang, S. Tam, C. Spiess, J. Frydman, and V. S. Pande. The predicted structure of the headpiece of the Huntingtin protein and its implications on Huntingtin aggregation. *J. Mol. Biol.*, 388:919–927, 2009.
- [638] J. Khandogin, A. Hu, and D. M. York. Electronic structure properties of solvated biomolecules: A quantum approach for macromolecular characterization. *J. Comput. Chem.*, 21:1562–1571, 2000.
- [639] J. Khandogin and D. M. York. Quantum mechanical characterization of nucleic acids in solution: A linear-scaling study of charge fluctuations in DNA and RNA. *J. Phys. Chem. B*, 106:7693–7703, 2002.
- [640] E. K. Kick, D. C. Roe, A. G. Skillman, G. Liu, T. J. A. Ewing, Y. Sun, I. D. Kuntz, and J. A. Ellman. Structure-based design of combinatorial chemistry yield low nanomolar inhibitors of cathepsin D. *Chem. Biol.*, 4:297–307, 1997.
- [641] C. L. Kielkopf, S. Ding, P. Kuhn, and D. C. Rees. Conformational flexibility of B-DNA at 0.74 Å resolution: d(CCAGTACTGG)<sub>2</sub>. *J. Mol. Biol.*, 296:787–801, 2000.

- [642] N. Kim. *Exploring RNA Structure Space Using Multidisciplinary Approaches with Applications for Novel RNA Design*. PhD thesis, New York University, Department of Chemistry (Program in Computational Biology), New York, NY, May 2009.
- [643] N. Kim, H. H. Gan, and T. Schlick. Designing structured RNA pools for *in vitro* selection of RNAs. *RNA*, 13:478–492, 2007.
- [644] N. Kim, J. A. Izzo, S. Elmetwaly, H. H. Gan, and T. Schlick. Computational generation and screening of RNA motifs in large sequence pools. *Nucl. Acids Res.*, 2010. doi: 10.1093/nar/gkq282.
- [645] N. Kim, N. Shiffeldrim, H.H. Gan, and T. Schlick. Candidates for novel RNA topologies. *J. Mol. Biol.*, 341:1129–1144, 2004.
- [646] N. Kim, J. Sup Shin, S. Elmetwaly, H. H. Gan, and T. Schlick. RAGPOOLS: RNA-As-Graph-Pools — A web server for assisting the design of structured RNA pools for *in vitro* selection. *Bioinfor.*, 23:2959–2960, 2007.
- [647] Y. C. Kim and G. Hummer. Coarse-grained models for simulations of multiprotein complexes: application to ubiquitin binding. *J. Mol. Biol.*, 375:1416–1433, 2008.
- [648] J. G. Kirkwood. Statistical mechanics of fluid mixtures. *J. Chem. Phys.*, 3:300–313, 1935.
- [649] H. Kitano. A robustness-based approach to systems-oriented drug design. *Nat. Rev. Drug Disc.*, 6:202–210, 2007.
- [650] A. Kitao and N. Gō. Investigating protein dynamics in collective coordinate space. *Curr. Opin. Struct. Biol.*, 9:164–169, 1999.
- [651] I. Klapper. Biological applications of the dynamics of twisted elastic rods. *J. Comput. Phys.*, 125:325–337, 1996.
- [652] J. B. Klauda, B. R. Brooks, Jr. A. D. MacKerell, R. M. Venable, and R. W. Pastor. An *ab initio* study on the torsional surface of alkanes and its effect on molecular simulations of alkanes and a DPPC bilayer. *J. Phys. Chem. B*, 109:5300–5311, 2005.
- [653] B. J. Klein and G. R. Pack. Calculations of the spatial distribution of charge density in the environment of DNA. *Biopolymers*, 22:2331–2352, 1983.
- [654] D. J. Klein, T. M. Schmeing, P. B. Moore, and T. A. Steitz. The kink-turn: A new RNA secondary structure motif. *EMBO J.*, 20:4214–4221, 2001.
- [655] M. L. Klein and W. Shinoda. Large-scale molecular dynamics simulations of self-assembling systems. *Science*, 321:798–800, 2008.
- [656] K. V. Klenin, M. D. Frank-Kamenetskii, and J. Langowski. Modulation of intramolecular interactions in superhelical DNA by curved sequences: A Monte-Carlo simulation study. *Biophys. J.*, 68:81–88, 1995.
- [657] K. V. Klenin, A. V. Vologodskii, V. V. Anshelevich, A. M. Dykhne, and M. D. Frank-Kamenetskii. Computer simulation of DNA supercoiling. *J. Mol. Biol.*, 217:413–419, 1991.
- [658] J. L. Klepeis, K. Lindorff-Larsen, R. O. Dror, and D. E. Shaw. Long-timescale molecular dynamics simulations of protein structure and function. *Curr. Opin. Struct. Biol.*, 19:120–127, 2009.
- [659] D. K. Klimov and D. Thirumalai. Viscosity dependence of the folding rates of proteins. *Phys. Rev. Lett.*, 79:317–320, 1997.

- [660] D. K. Klimov and D. Thirumalai. Stretching single-domain proteins: Phase diagram and kinetics of force-induced unfolding. *Proc. Natl. Acad. Sci. USA*, 96:6166–6170, 1999.
- [661] J. Kling. Out of Malaysia: Finding natural products to fight AIDS. *Mod. Drug Dis.*, 2:31–36, 1999.
- [662] R. D. Knight and L. F. Landweber. The early evolution of the genetic code. *Cell*, 101:569–572, 2000.
- [663] D. E. Knuth. *The Art of Computer Programming. Volume 2: Seminumerical Methods*. Addison-Wesley, Reading, Massachusetts, second edition, 1981.
- [664] P. Koehl and M. Levitt. A brighter future for protein structure prediction. *Nature Struct. Biol.*, 6:108–111, 1999.
- [665] P. Koehl and M. Levitt. Theory and simulation: Can theory challenge experiment? *Curr. Opin. Struct. Biol.*, 9:155–156, 1999.
- [666] F. E. Koehn and G. T. Carter. The evolving role of natural products in drug discovery. *Nat. Rev. Drug Disc.*, 4:206–220, 2005.
- [667] N. Koga and S. Takada. Folding-based molecular simulations reveal mechanisms of the rotary motor F<sub>1</sub>-ATPase. *Proc. Natl. Acad. Sci. USA*, 103:5367–5372, 2008.
- [668] P. A. Kollman and K. A. Dill. Decisions in force field development: An alternative to those described by Roterman *et al.* *J. Biomol. Struct. Dyn.*, 8:1103–1107, 1991.
- [669] Y. Kong, Y. Shen, T. E. Warth, and J. Ma. Conformational pathways in the gating of *Escherichia coli* mechanosensitive channel. *Proc. Natl. Acad. Sci. USA*, 99:5999–6004, 2002.
- [670] J. H. Konnert and W. A. Hendrickson. A restrained-parameter thermal-factor refinement procedure. *Acta Crystallogr.*, A36:344–350, 1980.
- [671] M. W. Konrad and J. I. Bolonick. Molecular dynamics simulation of DNA stretching is consistent with the tension observed for extension and strand separation and predicts a novel ladder structure. *J. Amer. Chem. Soc.*, 118:10989–10994, 1996.
- [672] E. V. Koonin, L. Aravind, and A. S. Kondrashov. The impact of comparative genomics on our understanding of evolution. *Cell*, 101:573–576, 2000.
- [673] G. Koren, J. Cairns, D. Chitayat, A. Gaedigk, and S. J. Leeder. Pharmacogenetics of morphine poisoning in a breastfed neonate of a codeine-prescribed mother. *Lancet*, 368:704, 2006.
- [674] R. Kornberg and J. O. Thomas. Chromatin structure: Oligomers of histones. *Science*, 184:865–868, 1974.
- [675] N. Korolev, A. P. Lyubartsev, A. Laaksonen, and L. Nordenskiöld. On the competition between water, sodium ions, and spermine in binding to DNA: A molecular dynamics simulation study. *Biophys. J.*, 82:2860–2875, 2002.
- [676] A. Korostelev, R. Bertram, and M. S. Chapman. Simulated-annealing real-space refinement as a tool in model building. *Acta Cryst.*, D58:761–767, 2002.
- [677] C. Korth, B. C. H. May, F. E. Cohen, and S. B. Prusiner. Acridine and phenothiazine derivatives as pharmacotherapeutics for prion disease. *Proc. Natl. Acad. Sci. USA*, 98:9836–9841, 2001.
- [678] K. M. Kosikov, A. A. Gorin, V. B. Zhurkin, and W. K. Olson. DNA stretching and compression: Large-scale simulations of double helical structures. *J. Mol. Biol.*, 289:1301–1326, 1999.

- [679] D. Kosztin, T. C. Bishop, and K. Schulten. Binding of the estrogen receptor to DNA: The role of waters. *Biophys. J.*, 73:557–570, 1997.
- [680] R. Z. Kramer, J. Bella, B. Brodsky, and H. M. Berman. The crystal and molecular structure of a collagen-like peptide with a biologically relevant sequence. *J. Mol. Biol.*, 311:131–147, 2001.
- [681] M. Kröger, A. Alba-Perez, M. Laso, and H. C. Öttinger. Variance reduced Brownian simulation of a bead-spring chain under steady shear flow considering hydrodynamic interaction effects. *J. Chem. Phys.*, 113:4767–4773, 2000.
- [682] M. Kruithof, F.-T. Chen, A. Routh, C. Logie, D. Rhodes, and J. van Noort. Single-molecule force microscopy reveals a highly compliant helical folding for the 30-nm chromatin fiber. *Nat. Struct. Mol. Biol.*, 16:534–540, 2009.
- [683] A. Kryshtafovych, K. Fidelis, and J. Moult. Progress from CASP6 to CASP7. *Proteins: Struc. Func. Gen.*, 69 (Suppl. 8):194–207, 2007.
- [684] R. Kubo. The fluctuation-dissipation theorem. *Rep. Prog. Phys.*, 29:255–284, 1966.
- [685] W. Kühlbrandt and K. A. Williams. Analysis of macromolecular structure and dynamics by electron cryo-microscopy. *Curr. Opin. Chem. Biol.*, 3:537–543, 1999.
- [686] P. Kumar, H.-S. Ban, S.-S. Kim, H. Wu, T. Pearson, D. L. Greiner, A. Laouar, J. Yao, V. Haridas, K. Habiro, Y.-G. Yang, J.-H. Jeong, K.-Y. Lee, Y.-H. Kim, S. W. Kim, M. Peipp, G. H. Fey, N. Manjunath, L. D. Shultz, and S.-K. Lee. T cell-specific siRNA delivery suppresses HIV-1 infection in humanized mice. *Cell*, 134:577–586, 2008.
- [687] P. D. Kwong, R. Wyatt, J. Robinson, R. W. Sweet, J. Sodroski, and W. A. Hendrickson. Structure of an HIV gp 120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature*, 393:648–659, 1998.
- [688] L. J. LaBerge and J. C. Tully. A rigorous procedure for combining molecular dynamics and Monte Carlo simulation algorithms. *Chem. Phys.*, 260:183–191, 2000.
- [689] A. Laederach. Informatics challenges in structured RNAs. *Brief. Bioinf.*, 8:294–303, 2007.
- [690] C. Laing, S. Jung, A. Iqbal, and T. Schlick. Tertiary motifs revealed in analyses of higher-order RNA junctions. *J. Mol. Biol.*, 393:67–82, 2009.
- [691] C. Laing and T. Schlick. Analysis of four-way junctions in RNA structures. *J. Mol. Biol.*, 390:547–559, 2009.
- [692] M. L. Lamb, K. W. Burdick, S. Toba, M. M. Young, A. G. Skillman, X. Zou, J. R. Arnold, and I. D. Kuntz. Design, docking, and evaluation of multiple libraries against multiple targets. *Proteins: Struc. Func. Gen.*, 42:296–318, 2001.
- [693] C. G. Lambert, T. A. Darden, and J. A. Board, Jr. A multipole-based algorithm for efficient calculation of forces and potentials in macroscopic periodic assemblies of particles. *J. Comput. Phys.*, 126:274–285, 1996.
- [694] L. D. Landau and E. M. Lifshitz. *Course of Theoretical Physics*, volume 5. Pergamon Press, Oxford, third edition, 1980.
- [695] E. Lander. The new genomics: Global views of biology. *Science*, 274:536–539, 1996.

- [696] C. H. Langley and N. L. Allinger. Molecular mechanics (MM4) and *ab initio* study of amide-amide and amide-water dimers. *J. Phys. Chem. A*, 107:5208–5216, 2003.
- [697] J. Langowski. Salt effects on internal motions of superhelical and linear pUC8 DNA. Dynamic light scattering studies. *Biophys. Chem.*, 27:263–271, 1987.
- [698] J. Langowski, W. K. Olson, S. C. Pedersen, I. Tobias, and T. Westcott. DNA supercoiling, localized bending, and thermal fluctuations. *Trends Bio. Sci.*, 21:50, 1996.
- [699] F. Lankáš, J. Šponer, P. Hobza, and J. Langowski. Sequence-dependent elastic properties of DNA. *J. Mol. Biol.*, 299:695–709, 2000.
- [700] B. A. Larder and D. K. Stammers. Closing in on HIV drug resistance. *Nature Struct. Biol.*, 6:103–106, 1999.
- [701] A. Larshminarayanan and V. Sasisekharan. Stereochemistry of nucleic acids and polynucleotides. IV. Conformational energy of base-sugar units. *Biopolymers*, 8:475–488, 1969.
- [702] U. Laserson, H. H. Gan, and T. Schlick. Searching for 2D RNA geometries in bacterial genomes. In *Proceedings of the Twentieth Annual ACM Symposium on Computational Geometry*, pages 373–377, New York, 2004. ACM Press.
- [703] U. Laserson, H. H. Gan, and T. Schlick. Exploring the connection between synthetic and natural RNAs in genomes via a novel computational approach. In B. Leimkuhler, C. Chipot, R. Elber, A. Laaksonen, A. Mark, T. Schlick, C. Schüette, and R. D. Skeel, editors, *New Algorithms for Macromolecular Simulation, Proceedings of the Fourth International Workshop on Algorithms for Macromolecular Modelling, Leicester, UK, August 2004*, volume 49 of *Lecture Notes in Computational Science and Engineering*. Springer-Verlag, Berlin, Germany, 2005.
- [704] U. Laserson, H. H. Gan, and T. Schlick. Predicting candidate genomic sequences that correspond to synthetic functional RNA motifs. *Nucl. Acids Res.*, 33: 6057–6069, 2005.
- [705] R. Lavery, K. Zakrzewska, D. Beveridge, T.C. Bishop, D.A. Case, T. Cheatham, S. Dixit, B. Jayaram, F. Lankas, C. Laughton, J. Maddocks, A. Michon, R. Osman, M. Orozco, A. Perez, T. Singh, N. Spackova, and J. Sponer. A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nuc. Acids Res.*, 38:299–313, 2010.
- [706] A. M. Law and W. D. Kelton. *Simulation Modeling and Analysis*. McGraw-Hill Series in Industrial Engineering and Management Science. McGraw-Hill, Boston, MA, third edition, 2000.
- [707] T. Lazaridis and M. Karplus. “New view” of protein folding reconciled with the old through multiple unfolding simulations. *Science*, 278:1928–1931, 1997.
- [708] S.Y. Le, R. Nussinov, and J. Maizel. Tree graphs of RNA secondary structures and their comparisons. *Comput. Biomed. Res.*, 22:461–473, 1989.
- [709] A. M. Leach. *Molecular Modelling. Principles and Applications*. Pearson Education Limited, Harlow, England, second edition, 2001.
- [710] J. LeBarron, R. A. Grassucci, T. R. Shaikh, W. T. Baxter, J. Sengupta, and J. Frank. Exploration of parameters in cryo-RM leading to an improved density map of the *e. coli* ribosome. *J. Struc. Biol.*, 0:0–0, 2008.

- [711] A. Lebrun and R. Lavery. Modelling extreme stretching of DNA. *Nucl. Acids Res.*, 24:2260–2267, 1996.
- [712] A. Lebrun and R. Lavery. Unusual DNA conformations. *Curr. Opin. Struct. Biol.*, 7:348–356, 1998.
- [713] P. L'Ecuyer. Maximally-equidistributed combined Tausworthe generators. *Math. Comput.*, 65:203–213, 1996.
- [714] P. L'Ecuyer. Bad lattice structures for vectors of non-successive values produced by some linear recurrences. *Informs J. Comput.*, 9:57–60, 1997.
- [715] P. L'Ecuyer. Random number generation. In J. Banks, editor, *Handbook on Simulation*, chapter 4, pages 93–137. John Wiley & Sons, New York, NY, 1998.
- [716] P. L'Ecuyer. Good parameter sets for combined multiple recursive random number generators. *Oper. Res.*, 47:159–164, 1999.
- [717] P. L'Ecuyer. Tables of maximally-equidistributed combined LFSR generators. *Math. Comput.*, 68:261–269, 1999.
- [718] P. L'Ecuyer. Pseudorandom number generation. In S. G. Henderson and B. L. Nelson, editors, *Handbooks in Operations Research and Management Science: Simulation*, chapter 3, pages 55–81. Elsevier Science, Amsterdam, The Netherlands, 2006.
- [719] P. L'Ecuyer. Pseudorandom number generators. In E. Platen and P. Jaeckel, editors, *Encyclopedia of Quantitative Finance*, Simulation Methods in Financial Engineering. Wiley, New York, NY, 2009.
- [720] P. L'Ecuyer and T. H. Andres. A random number generator based on the combination of four LCGs. *Mathematics and Computers in Simulation*, 44:99–107, 1997.
- [721] P. L'Ecuyer and F. Panneton.  $F_2$ -linear random number generators. In C. Alexopoulos, D. Goldsman, and J. R. Wilson, editors, *Advancing the Frontiers of Simulation: A Festschrift in Honor of George Samuel Fishman*, pages 169–193. Springer-Verlag, New York, NY, 2009.
- [722] P. L'Ecuyer and R. Simard. TestU01: A C library for empirical testing of random number generators. *ACM Trans. Math. Softw.*, 33, 2007.
- [723] B. Lee and F. M. Richards. The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.*, 55:379–400, 1971.
- [724] J. F. Lee, J. R. Hesselberth, L. A. Meyers, and A. D. Ellington. Aptamer database. *Nucl. Acids Res.*, 32:D95–100, 2004. (Database issue).
- [725] J. H. Lee, M. D. Canny, A. De Erkenez, D. Krilleke, Y. S. Ng, D. T. Shima, A. Pardi, and F. Jucker. A therapeutic aptamer inhibits angiogenesis by specifically targeting the heparin binding domain of VEGF165. *Proc. Natl. Acad. Sci. USA*, 102:18902–18907, 2005.
- [726] T.-S. Lee, D. M. York, and W. Yang. Linear-scaling semiempirical quantum calculations for macromolecules. *J. Chem. Phys.*, 105:2744–2750, 1996.
- [727] P. D. Leeson and B. Springthorpe. The influence of drug-like concepts on decision-making in medical chemistry. *Nat. Rev. Drug Disc.*, 6:881–890, 2007.
- [728] R. B. Lehoucq, D. C. Sorensen, and C. Yang. *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*. SIAM, Philadelphia, PA, 1998. [www.caam.rice.edu/software/ARPACK/indexold.html](http://www.caam.rice.edu/software/ARPACK/indexold.html).

- [729] H. Lei and Y. Duan. Improved sampling methods for molecular simulation. *Curr. Opin. Struct. Biol.*, 17:187–191, 2007.
- [730] H. Lei, C. Wu, H. M. Liu, and Y. Duan. Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA*, 104:4925–4930, 2007.
- [731] B. Leimkuhler and S. Reich. *Simulating Hamiltonian Dynamics*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge, UK, 2004.
- [732] B. Leimkuhler and S. Reich. A Metropolis adjusted Nosé-Hoover thermostat. *Math. Mod. Num. Anal.*, 2009. In Press.
- [733] B. Leimkuhler and R. D. Skeel. Symplectic numerical integrators in constrained Hamiltonian systems. *J. Comput. Phys.*, 112:117–125, 1994.
- [734] D. A. Leonard, N. Rajaram, and T. K. Kerppola. Structural basis of DNA bending and oriented heterodimer binding by the basic leucine zipper domains of Fos and Jun. *Proc. Natl. Acad. Sci. USA*, 94:4913–4918, 1997.
- [735] N. B. Leontis, R. B. Altman, H. M. Berman, S. E. Brenner, J. W. Brown, D. R. Engelke, S. C. Harvey, S. R. Holbrook, F. Jossinet, S. E. Lewis, F. Major, D. H. Mathews, J. Richardson, J. R. Williamson, and E. Westhof. The RNA Ontology Consortium: An open invitation to the RNA community. *RNA*, 12:533–541, 2006.
- [736] N. B. Leontis, A. Lescoute, and E. Westhof. The building blocks and motifs of RNA architecture. *Curr. Opin. Struct. Biol.*, 16:279–287, 2006.
- [737] N. B. Leontis and E. Westhof. Conserved geometrical base-pairing patterns in RNA. *Quart. Rev. Biophys.*, 31:399–455, 1998.
- [738] N. B. Leontis and E. Westhof. Geometric nomenclature and classification of RNA base pairs. *RNA*, 7:499–512, 2001.
- [739] N. B. Leontis and E. Westhof. Analysis of RNA motifs. *Curr. Opin. Struct. Biol.*, 13:300–308, 2003.
- [740] A. Lescoute and E. Westhof. The interaction networks of structured RNAs. *Nucl. Acids Res.*, 22:6587–6604, 2006.
- [741] S. D. Levene, H.-M. Wu, and D. M. Crothers. Bending and flexibility of kinetoplast DNA. *Biochemistry*, 25:3988–3995, 1986.
- [742] I. N. Levine. *Quantum Chemistry*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, fourth edition, 1991.
- [743] C. Levinthal. Are there pathways for protein folding? *J. Chim. Physique*, 65:44–45, 1969.
- [744] C. Levinthal. How to fold graciously. In P. Debrunner, J. C. M. Tsibris, and E. Münch, editors, *Mossbauer Spectroscopy in Biological Systems, Proceedings of a Meeting held at Allerton House, Monticello, Illinois*, page 22, Urbana, Illinois, 1969. University of Illinois Press.
- [745] M. Levitt. How many base-pairs per turn does DNA have in solution and in chromatin? Some theoretical calculations. *Proc. Natl. Acad. Sci. USA*, 75:640–644, 1978.
- [746] M. Levitt. Computer simulation of DNA double-helix dynamics. *Cold Spring Harbor Symp. Quant. Biol.*, 47:251–275, 1983.

- [747] M. Levitt. The birth of computational structural biology. *Nat. Struct. Biol.*, 8:392–393, 2001.
- [748] M. Levitt and S. Lifson. Refinement of protein conformations using a macromolecular energy minimization procedure. *J. Mol. Biol.*, 46:269–279, 1969.
- [749] M. Levitt and A. Warshel. Computer simulation of protein folding. *Nature*, 253:694–698, 1975.
- [750] M. Levitt and A. Warshel. Extreme conformational flexibility of the furanose ring in DNA and RNA. *J. Amer. Chem. Soc.*, 100:2607–2613, 1978.
- [751] R. M. Levy and E. Gallicchio. Computer simulations with explicit solvent: Recent progress in the thermodynamic decomposition of free energies, and in modeling electrostatic effects. *Annu. Rev. Phys. Chem.*, 49:531–567, 1998.
- [752] R. M. Levy, R. P. Sheridan, J. W. Keepers, G. S. Dubey, S. Swaminathan, and M. Karplus. Molecular dynamics of myoglobin at 298°K. Results from a 300-ps computer simulation. *Biophys. J.*, 48:509–518, 1985.
- [753] S. Levy, G. Sutton, P. C. Ng, L. Feuk, A. L. Halpern, B. P. Walenz1, N. Axelrod, J. Huang, E. F. Kirkness, G. Denisov, Y. Lin, J. R. MacDonald, A. W. C. Pang, M. Shago, T. B. Stockwell, A. Tsiamouri, V. Bafna, V. Bansal, S. A. Kravitz, D. A. Busam, K. Y. Beeson, T. C. McIntosh, K. A. Remington, J. F. Abril, J. Gill, J. Borman, Y. H. Rogers, M. E. Frazier, S. W. Scherer, R. L. Strausberg, and J. C. Venter. The diploid genome sequence of an individual human. *PLoS Biol.*, 5:2113–2144, 2007.
- [754] J. P. Lewis, C. W. Carter, Jr., J. Hermans, W. Pan, T.-S. Lee, and W. Yang. Active species for the ground-state complex of cytidine deaminase: A linear-scaling quantum mechanical investigation. *J. Amer. Chem. Soc.*, 120:5407–5410, 1998.
- [755] J. P. Lewis, P. Ordejón, and O. F. Sankey. An electronic structure based molecular dynamics for large biomolecular systems: Applications to the 10 basepair Poly(dG)\*Poly(dC) DNA double helix. *Phys. Rev. B*, 55:6880–6887, 1997.
- [756] J. P. Lewis, N. H. Pawley, and O. F. Sankey. Theoretical investigation of the cyclic peptide system cyclo[(D-Ala-Glu-D-Ala-Gln)<sub>m=1–4</sub>]. *J. Phys. Chem. B*, 101:10576–10583, 1997.
- [757] H. Li, W. X. Li, and S. W. Ding. Induction and suppression of RNA silencing by an animal virus. *Science*, 296:1319–1321, 2002.
- [758] J. Li, C. J. Cramer, and D. G. Truhlar. Application of a universal solvation model to nucleic acid bases: Comparison of semiempirical molecular orbital theory, ab initio Hartree–Fock theory, and density functional theory. *Biophys. Chem.*, 78:147–155, 1999.
- [759] P. T. X. Li, J. Vieregg, and I. Tinoco, Jr. How RNA unfolds and refolds. *Ann. Rev. Biochem.*, 77:77–100, 2008.
- [760] Z. Li and H. A. Scheraga. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc. Natl. Acad. Sci. USA*, 84:6611–6615, 1987.
- [761] X. Liang, H. Kuhn, and M.D. Frank-Kamenetskii. Monitoring single-stranded DNA secondary structure formation by determining the topological state of DNA catenanes. *Biophys. J.*, 90:2877–2889, 2006.
- [762] K. Liberek, A. Lewandowska, and S. Zietkiewicz. Chaperones in control of protein disaggregation. *EMBO J.*, 27:328–335, 2008.

- [763] E. Liepinsh, G. Otting, and K. Wuthrich. NMR observation of individual molecules of hydration water bound to DNA duplexes: Direct evidence for a spine of hydration water present in aqueous solution. *Nucl. Acids Res.*, 20:6549–6553, 1992.
- [764] S. Lifson. Potential energy functions for structural molecular biology. In D. B. Davies, W. Saenger, and S. S. Danyluk, editors, *Methods in Structural Molecular Biology*, pages 359–385. Plenum Press, London, England, 1981.
- [765] S. Lifson. Theoretical foundation for the empirical force field method. *Gazzetta Chimica Italiana*, 116:687–692, 1986.
- [766] S. Lifson and A. Warshel. Consistent force field for calculations of conformations, vibrational spectra, and enthalpies of cycloalkane and *n*-alkane molecules. *J. Chem. Phys.*, 49:5116–5129, 1968.
- [767] J.-H. Lii and N. L. Allinger. Intensities of infrared bands in molecular mechanics (MM3). *J. Comput. Chem.*, 13:1138–1141, 1992.
- [768] J.-H. Lii and N. L. Allinger. Directional hydrogen bonding in the MM3 force field: II. *J. Comput. Chem.*, 19:1001–1016, 1998.
- [769] P. Lin, L. C. Pedersen, V. K. Batra, W. A. Beard, S. H. Wilson, and L. G. Pedersen. Energy analysis of chemistry for correct insertion by DNA polymerase  $\beta$ . *Proc. Natl. Acad. Sci. USA*, 103:13294–13299, 2006.
- [770] E. Lindahl, B. Hess, and D. van der Spoel. GROMACS 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Model.*, 7:306–317, 2001.
- [771] J. Liphardt, B. Onoa, S. B. Smith, I. Tinoco, Jr., and C. Bustamante. Reversible unfolding of single RNA molecules by mechanical force. *Science*, 292:733–737, 2001.
- [772] C. Lipinski and A. Hopkins. Navigating chemical space for biology and medicine. *Nature*, 432:855–861, 2004.
- [773] K. B. Lipkowitz. Abuses of molecular mechanics. Pitfalls to avoid. *J. Chem. Educ.*, 72:1070–1075, 1995.
- [774] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large-scale optimization. *Math. Prog. B*, 45:503–528, 1989.
- [775] D. J. Liu and L. A. Day. Pf1 virus structure: Helical coat protein and DNA with paraxial phosphates. *Science*, 265:671–674, 1994.
- [776] H. Liu, M. Elstner, E. Kaxiras, T. Frauenheim, J. Hermans, and W. Yang. Quantum mechanics simulation of protein dynamics on long timescale. *Proteins: Struct. Func. Gen.*, 44:484–489, 2001.
- [777] H. Liu, S. Farr-Jones, N. B. Ulyanov, M. Llinas, S. Marqusee, D. Groth, F. E. Cohen, S. B. Prusiner, and T. L. James. Solution structure of a syrian hamster prion protein rPrP(90–231). *Biochemistry*, 38:5362–5377, 1999.
- [778] H. Liu, Y. Zhang, and W. Yang. How is the active-site of enolase organized to achieve overall efficiency in catalyzing a two step reaction. *J. Amer. Chem. Soc.*, 122:6560–6570, 2000.
- [779] Q. Liu, L. Wang, A. G. Frutos, A. E. Condon, R. M. Corn, and L. M. Smith. DNA computing on surfaces. *Science*, 403:175–179, 2000.
- [780] X. Liu, K. Fan, and W. Wang. The number of protein folds and their distribution over families in nature. *Proteins: Struc. Func. Bioinf.*, 54:491–499, 2004.

- [781] A. Liwo, C. Czaplewski, S. Oldziej, and H. A. Scheraga. Computational techniques for efficient conformational sampling of proteins. *Curr. Opin. Struct. Biol.*, 18:134–139, 2008.
- [782] O. Llorca, E. A. McCormack, G. Hynes, J. Grantham, J. Cordell, J. L. Carrascosa, K. R. Willison, J. J. Fernandez, and J. M. Valpuesta. Eukaryotic type II chaperonin CCT interacts with actin through specific subunits. *Nature*, 402:693–696, 1999.
- [783] R. J. Loncharich, B. R. Brooks, and R. W. Pastor. Langevin dynamics of peptides: The frictional dependence of isomerization rates of N-acetylalanyl-N'-methylamide. *Biopolymers*, 32:523–535, 1992.
- [784] S. Louise-May, P. Auffinger, and E. Westhof. Calculations of nucleic acid conformations. *Curr. Opin. Struct. Biol.*, 6:289–298, 1996.
- [785] V. Lounnas, S. K. Lüdemann, and R. C. Wade. Towards molecular dynamics simulation of large proteins with a hydration shell at constant pressure. *Biophys. Chem.*, 78:157–182, 1999.
- [786] X. Lu, M. D. Simon, J. V. Chodaparambil, J. C. Hansen, K. M. Shokat, and K. Luger. The effect of H3K79 dimethylation and H4K20 trimethylation on nucleosome and chromatin structure. *Nat. Struct. Mol. Biol.*, 15:1122–1124, 2008.
- [787] Y. Lu and J. Liu. Functional DNA nanotechnology: Emerging applications of DNAzymes and aptamers. *Curr. Opin. Biotech.*, 17:580–588, 2006.
- [788] R. Ludwig. Water: From clusters to the bulk. *Angew. Chem. Int. Ed.*, 40:1808–1827, 2001.
- [789] D. G. Luenberger. *Linear and Nonlinear Programming*. Addison Wesley, Reading, Massachusetts, 1984.
- [790] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389:251–260, 1997.
- [791] L. M. Luheshi, D. C. Crowther, and C. M. Dobson. Protein misfolding and disease: From the test tube to the organism. *Curr. Opin. Chem. Biol.*, 12:25–31, 2008.
- [792] B. A. Luty, M. E. David, I. G. Tironi, and W. F. Van Gunsteren. A comparison of particle-particle particle-mesh and Ewald methods for calculating electrostatic interactions in periodic molecular systems. *Mol. Sim.*, 14:11–20, 1994.
- [793] B. A. Luty, I. G. Tironi, and W. F. Van Gunsteren. Lattice-sum methods for calculating electrostatic interactions in molecular simulations. *J. Chem. Phys.*, 103:3014–3021, 1995.
- [794] E. Lyman, J. Pfaendtner, and G. A. Voth. Systematic multiscale parameterization of heterogeneous elastic network models of proteins. *Biophys. J.*, 95:4183–4192, 2008.
- [795] B. Ma and N. L. Allinger. Calculation of  $r_z$  structures from  $r_s$  structures. *J. Mol. Struct.*, 413–414:395–404, 1997.
- [796] B. Ma, J.-H. Lii, and N. L. Allinger. Molecular polarizabilities and induced dipole moments in molecular mechanics. *J. Comput. Chem.*, 21:813–825, 2000.
- [797] B. Ma, J.-H. Lii, K. Chen, and N. L. Allinger. A molecular mechanics study of the cholesteryl acetate crystal: Evaluation of interconversion among  $r_g$ ,  $r_z$ , and  $r_\alpha$  bond lengths. *J. Amer. Chem. Soc.*, 119:2570–2573, 1997.

- [798] B. Ma, J.-H. Lii, H. F. Schaefer, III, and N. L. Allinger. Systematic comparison of experimental, quantum mechanical, and molecular mechanical bond lengths for organic molecules. *J. Phys. Chem.*, 100:8763–8769, 1996.
- [799] M. W. MacArthur and J. M. Thornton. Deviations from planarity of the peptide bond in peptides and proteins. *J. Mol. Biol.*, 264:1180–1195, 1996.
- [800] D. MacDonald, K. Herbert, X. Zhang, T. Polgruto, and P. Lu. Solution structure of an A-tract DNA bend. *J. Mol. Biol.*, 306:1081–1098, 2001.
- [801] A. Machado-Lima, H.A. del Portillo, and A.M. Durham. Computational methods in noncoding RNA research. *J. Math. Biol.*, 56:15–49, 2008.
- [802] A. D. MacKerell, Jr. Empirical force fields for biological macromolecules: overview and issues. *J. Comput. Chem.*, 25:1584–1604, 2004.
- [803] A. D. MacKerell, Jr., N. Banavali, and N. Foloppe. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers*, 56:257–265, 2001.
- [804] A. D. MacKerell, Jr. and N. K. Banavali. All-atom empirical force field for nucleic acids: II. Application to molecular dynamics simulations of DNA and RNA in solution. *J. Comput. Chem.*, 21:105–120, 2000.
- [805] A. D. MacKerell, Jr., D. Bashford, M. Bellott, R. L. Dunbrack, Jr., J. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, III., B. Roux, M. Schlenkrich, J. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. An all-atom empirical potential for molecular modeling and dynamics of proteins. *J. Phys. Chem. B*, 102:3586–3616, 1998.
- [806] A. D. MacKerell, Jr., M. Feig, and C. L. Brooks, III. Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.*, 25:1400–1415, 2004.
- [807] A. D. MacKerell, Jr., M. Feig, and C. L. Brooks, III. Improved treatment of the protein backbone in empirical force fields. *J. Amer. Chem. Soc.*, 126:698–699, 2004.
- [808] A. D. MacKerell, Jr. and L. Nilsson. Molecular dynamics simulations of nucleic acid-protein complexes. *Curr. Opin. Struct. Biol.*, 18:194–199, 2008.
- [809] A. D. MacKerell, Jr., J. Wiorkiewicz-Kuczera, and M. Karplus. An all-atom empirical energy function for the simulation of nucleic acids. *J. Amer. Chem. Soc.*, 117:11946–11975, 1995.
- [810] B. Maddox. *Rosalind Franklin. The Dark Lady of DNA*. HarperCollins, New York, NY, 2002.
- [811] J. Maddox. Statistical mechanics by numbers. *Nature*, 334:561, 1989.
- [812] J. Maddox. Towards the calculation of DNA. *Nature*, 339:557, 1989.
- [813] J. D. Madura, M. E. Davis, M. K. Gilson, R. C. Wade, B. A. Luty, and J. A. McCammon. Biological applications of electrostatic calculations and Brownian dynamics simulations. In K. B. Lipkowitz and D. B. Boyd, editors, *Reviews in Computational Chemistry*, volume V, pages 229–267. VCH Publishers, New York, NY, 1994.

- [814] A. Maguire, K. High, A. Auricchio, J. Wright, E. Pierce, F. Testa, F. Mingozzi, J. Bennicelli, G. Ying, S. Rossi, A. Fulton, K. Marshall, S. Banfi, D. Chung, J. Morgan, B. Hauck, O. Zelenaya, X. Zhu, L. Raffini, F. Coppievers, E. De Baere, K. Shindler, N. Volpe, E. Surace, C. Acerra, A. Lyubarsky, T. Redmond, E. Stone, J. Sun, J. W. McDonnell, B. Leroy, F. Simonelli, and J. Bennett. Age-dependent effects of RPE65 gene therapy for Leber's congenital amaurosis: A phase 1 dose-escalation trial. *The Lancet*, 374:1597–1605, 2009.
- [815] L. J. Maher, III. Mechanisms of DNA bending. *Curr. Opin. Struct. Biol.*, 2:688–694, 1998.
- [816] G. Maisuradze, A. Liwo, and H. Scheraga. Principal component analysis for protein folding dynamics. *J. Mol. Biol.*, 385:312–329, 2009.
- [817] W. Makalowski. Not junk after all. *Science*, 300:1246–1247, 2003.
- [818] S. Makino, T. J. A. Ewing, and I. D. Kuntz. DREAM++: Flexible docking program for virtual combinatorial libraries. *J. Comput.-Aided Mol. Design*, 13:513–532, 1999.
- [819] M. Mandal and R.R. Breaker. Adenine riboswitches and gene activation by disruption of a transcription terminator. *Nat. Struct. Mol. Biol.*, 11:29–35, 2004.
- [820] Y. Mandel-Gutfreund, H. Margalit, R. L. Jernigan, and V. B. Zhurkin. A role for CH · · · O interactions in protein-DNA recognition. *J. Mol. Biol.*, 277:1129–1140, 1998.
- [821] M. Mandziuk and T. Schlick. Resonance in the dynamics of chemical systems simulated by the implicit-midpoint scheme. *Chem. Phys. Lett.*, 237:525–535, 1995.
- [822] G. S. Manning. The molecular theory of polyelectrolyte solutions with applications to the electrostatic properties of polynucleotides. *Quart. Rev. Biophys.*, 179:181–246, 1978.
- [823] G. S. Manning, K. K. Ebralidse, A. D. Mirzabekov, and A. Rich. An estimate of the extent of folding of nucleosomal DNA by laterally asymmetric neutralization of phosphate groups. *J. Biomol. Struct. Dynam.*, 6:877–889, 1989.
- [824] G. S. Manning and J. Ray. Counterion condensation revisited. *J. Biomol. Struct. Dynam.*, 16:461–476, 1998.
- [825] C. Mao, T. LaBean, J. H. Reif, and N. C. Seeman. Logical computation using algorithmic self-assembly of DNA triple crossover molecules. *Nature*, 407:493–496, 2000.
- [826] J. R. Maple, M.-J. Hwang, T. P. Stockfisch, U. Dinur, M. Waldman, C. S. Ewing, and A. T. Hagler. Derivation of class II force fields. I. Methodology and quantum force field for the alkyl functional group and alkane molecules. *J. Comput. Chem.*, 15:162–182, 1994.
- [827] P. Maragakis, K. Lindorff-Larsen, M. P. Eastwood, R. O. Dror, J. L. Klepeis, I. T. Arkin, M. Ø Jensen, H. Xu, N. Trbovic, R. A. Friesner, A. G. Palmer, III, and D. E. Shaw. Microsecond molecular dynamics simulation shows effect of slow loop dynamics on backbone amide order parameter of proteins. *J. Phys. Chem. B*, 112:6155–6158, 2008.
- [828] L. Maragliano and E. Vanden-Eijnden. Single-sweep methods for free energy calculations. *J. Chem. Phys.*, 128:184110, 2008.
- [829] J. C. Marini, S. D. Levine, D. M. Crothers, and P. T. Englund. Bent helical structure in kinetoplast DNA. *Proc. Natl. Acad. Sci. USA*, 79:7664–7668, 1982.

- [830] J. F. Marko and E. D. Siggia. Fluctuations and supercoiling of DNA. *Science*, 265:506–508, 1994.
- [831] P. R. L. Markwick, G. Bouvignies, and M. Blackledge. Exploring multiple time-scale motions in protein GB3 using accelerated molecular dynamics and NMR spectroscopy. *J. Amer. Chem. Soc.*, 129:4724–4730, 2007.
- [832] G. Marsaglia and L.-H Tsay. Matrices and the structure of random number sequences. *Lin. Alg. Appl.*, 67:147–156, 1985.
- [833] J. Martin, T. Langer, R. Boteva, A. Schramel, A. L. Horwich, and F.-U. Hartl. Chaperonin-mediated protein folding at the surface of groEL through a ‘molten globule’-like intermediate. *Nature*, 352:36–42, 1991.
- [834] J. A. Martino, V. Katritch, and W. K. Olson. Influence of nucleosome structure on the three-dimensional folding of idealized minichromosomes. *Struc. Fold. Design*, 7:1009–1022, 1999.
- [835] G. J. Martyna, A. Hughes, and M. E. Tuckerman. Molecular dynamics algorithms for path integrals at constant pressure. *J. Chem. Phys.*, 110:3275–3290, 1999.
- [836] G. J. Martyna, M. E. Tuckerman, D. J. Tobias, and M. L. Klein. Explicit reversible integrators for extended systems dynamics. *Mol. Phys.*, 87:1117–1157, 1996.
- [837] M. Mascagni. Some methods of parallel pseudorandom number generation. In M. T. Heath, A. Ranade, and R S. Schreiber, editors, *Algorithms for Parallel Processing*, volume 105 of *IMA Volumes in Mathematics and Its Applications*, pages 277–288. Springer-Verlag, New York, NY, 1999.
- [838] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.
- [839] D. H. Mathews and D. H. Turner. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol.*, 317:191–203, 2002.
- [840] D. Matsuda and T. W. Dreher. The tRNA-like structure of Turnip yellow mosaic virus RNA is a 3'-translational enhancer. *Virology*, 321:36–46, 2004.
- [841] J. Maupetit, P. Derreumaux, and P. Tufféry. A fast method for large-scale *de novo* peptide and miniprotein structure prediction. *J. Comput. Chem.*, 31:726–738, 2009.
- [842] D. S. Maxwell, J. Tirado-Rives, and W. L. Jorgensen. A comprehensive study of the rotational energy profiles of organic systems by *ab initio* MO theory, forming a basis for peptide torsional potentials. *J. Comput. Chem.*, 16:984–1010, 1995.
- [843] K. M. Mazur. Accurate DNA dynamics without accurate long-range electrostatics. *J. Amer. Chem. Soc.*, 120:10928–10937, 1998.
- [844] S. J. McBryant, J. Klonoski, T. C. Sorensen, S. S. Norskog, S. Williams, M. G. Resch, J. A. Toombs, III, S. E. Hobday, and J. C. Hansen. Determinants of histone H4 N-terminal domain function during nucleosomal array oligomerization. *J. Biol. Chem.*, 284:16716–16722, 2009.
- [845] J. A. McCammon, B. R. Gelin, and M Karplus. Dynamics of folded proteins. *Nature*, 267:585–590, 1977.
- [846] J. A. McCammon and S. C. Harvey. *Dynamics of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, MA, 1987.
- [847] J. A. McCammon, B. M. Pettitt, and L. R. Scott. Ordinary differential equations of molecular dynamics. *Computers Math. Applic.*, 28:319–326, 1994.

- [848] J. J. McCarthy and R. Hilfiker. The use of single-nucleotide polymorphism maps in pharmacogenomics. *Nature Biotech.*, 18:505–508, 2000.
- [849] J. L. McCauley. *Chaos, Dynamics, and Fractals: an Algorithmic Approach to Deterministic Chaos*. Cambridge University Press, Cambridge, 1994.
- [850] L. McFail-Isom, C. C. Sines, and L. D. Williams. DNA structure: Cations in charge? *Curr. Opin. Struct. Biol.*, 9:298–304, 1999.
- [851] J. D. McGhee, J. M. Nickol, G. Felsenfeld, and D. C. Rau. Higher order structure of chromatin: Orientation of nucleosomes within the 30 nm chromatin solenoid is independent of species and spacer length. *Cell*, 33:831–841, 1983.
- [852] A. McPherson. Macromolecular crystals. *Sci. Amer.*, 260:62–69, 1989.
- [853] D. A. McQuarrie. *Statistical Mechanics*. University Science Books, Sausalito, CA, second edition, 2000.
- [854] E. L. Mehler. The Lorentz-Debye-Sack theory and dielectric screening of electrostatic effects in proteins and nucleic acids. In J. S. Murray and K. Sen, editors, *Molecular Electrostatic Potential: Concepts and Applications*, volume 3 of *Theoretical and Computational Chemistry*, chapter 9, pages 371–405. Elsevier Science, Amsterdam, 1996.
- [855] E. L. Mehler and F. Guarnieri. A self-consistent, microenvironment modulated screened Coulomb potential approximation to calculate pH-dependent electrostatic effects in proteins. *Biophys. J.*, 77:3–22, 1999.
- [856] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953.
- [857] M. Mezei. Optimal position of the solute for simulations. *J. Comput. Chem.*, 18:812–815, 1997.
- [858] T. F. Miller, E. Vanden-Eijnden, and D. Chandler. Solvent coarse-graining and the string method applied to the hydrophobic collapse of a hydrated chain. *Proc. Natl. Acad. Sci. USA*, 104:14559–14564, 2007.
- [859] M. Mills and I. Andrcioaei. An experimentally guided umbrella sampling protocol for biomolecules. *J. Chem. Phys.*, 129:114101, 2008.
- [860] P. Minary, M. E. Tuckerman, and G. J. Martyna. Dynamical spatial warping: A novel method for the conformational sampling of biophysical structure. *SIAM J. Sci. Comp.*, 30:2055–2083, 2008.
- [861] L. Mirny and E. Shakhnovich. Protein folding theory: From lattice to all-atom models. *Ann. Rev. Biophys. Biomol. Struc.*, 30:361–396, 2001.
- [862] R.A. Miron and K. A. Fichthorn. Accelerated molecular dynamics with the bond-boost method. *J. Chem. Phys.*, 119:6210–6216, 2003.
- [863] A. Mironov, V. Epshteyn, and E. Nudler. Transcriptional approaches to riboswitch studies. *Methods Mol. Biol.*, 540, 2009.
- [864] A. S. Mironov, I. Gusarov, R. Rafikov, L. E. Lopez, K. S., R. A. Krneva, D. A. Perumov, and E. Nudler. Sensing small molecules by nascent RNA: A mechanism to control transcription in bacteria. *Cell*, 111:747–756, 2002.
- [865] A. D. Mirzabekov and A. Rich. Asymmetric lateral distribution of unshielded phosphate groups in nucleosomal DNA and its role in DNA bending. *Proc. Natl. Acad. Sci. USA*, 76:1118–1121, 1979.

- [866] B. Mishra and T. Schlick. The notion of error in Langevin dynamics: 1. Linear analysis. *J. Chem. Phys.*, 105:299–318, 1996.
- [867] P. R. E. Mittl and M. G. Grütter. Structural genomics: Opportunities and challenges. *Curr. Opin. Chem. Biol.*, 5:402–408, 2001.
- [868] S. Miyamoto and P. A. Kollman. SETTLE: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.*, 13:952–962, 1992.
- [869] S. Miyazawa and R. L. Jernigan. A new substitution matrix for protein sequence searches based on contact frequencies in protein structures. *Prot. Engin.*, 6:267–278, 1993.
- [870] K. Moffat. Time-resolved biochemical crystallography: A mechanistic perspective. *Chem. Rev.*, 101:1569–1581, 2001.
- [871] F. A. Momany, R. F. McGuire, A. W. Burgess, and H. A. Scheraga. Energy parameters in polypeptides. VII. Geometric parameters partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids. *J. Phys. Chem.*, 79:2361–2381, 1975.
- [872] R. Montangue and R. Batey. Structure of the S-adenosylmethionine riboswitch regulatory mRNA element. *Nature*, 441:1772–1775, 2006.
- [873] R. Montangue and R. Batey. Riboswitches: Emerging themes in RNA structure and function. *Annu. Rev. Biophys.*, 37:117–133, 2008.
- [874] G. T. Montelione and S. Anderson. Structural genomics: Keystone for a human proteome project. *Nature Struc. Biol.*, 6:11–12, 1999.
- [875] P. B. Moore. Structural motifs in RNA. *Ann. Rev. Biochem.*, 68:287–300, 1999.
- [876] J. L. Morales and J. Nocedal. Automatic preconditioning by limited memory quasi-Newton updating. *SIAM J. Opt.*, 10:1079–1096, 2000. (also Technical Report 97/07, Optimization Technology Center, Northwestern University, 1997).
- [877] J. J. Moré and S. J. Wright. *Optimization Software Guide*, volume 14 of *Frontiers in Applied Mathematics*. SIAM, Philadelphia, PA, 1993. See [www.mcs.anl.gov/otc/Guide/](http://www.mcs.anl.gov/otc/Guide/) and [www.mcs.anl.gov/otc/Guide/SoftwareGuide/](http://www.mcs.anl.gov/otc/Guide/SoftwareGuide/) for updated information on the software guide.
- [878] R. T. Morrison and R. N. Boyd. *Organic Chemistry*. Allyn and Bacon, Inc., Newton, MA, fourth edition, 1983.
- [879] P. M. Morse. Diatomic molecules according to the wave mechanics. II. Vibrational levels. *Phys. Rev.*, 34:57–64, 1929.
- [880] J. Moult, K. Fidelis, A. Kryshtafovych, and B. Rost. Critical assessment of methods of protein structure prediction – Round VII. *Proteins: Struc. Func. Gen.*, 69 (Suppl. 8):3–9, 2007.
- [881] D. W. Mount. *Bioinformatics. Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2001.
- [882] Y. Mu, D. S. Kosov, and G. Stock. Conformational dynamics of trialanine in water. 2. comparison of AMBER, CHARMM, GROMOS, and OPLS force fields to NMR and infrared experiments. *J. Phys. Chem. B*, 107:5064–5073, 2003.
- [883] F. Mühlbacher, H. Schiessel, and C. Holm. Tail-induced attraction between nucleosome core particles. *Phys. Rev. E*, 74:031919, 2006.

- [884] K. B. Mullis. *Dancing Naked in the Mind Field*. Pantheon Books, New York, NY, 1998.
- [885] B. Munos. Lessons from 60 years of pharmaceutical innovation. *Nature Rev.*, 8:959–968, 2009.
- [886] J. B. Murray, D. P. Terwey, L. Maloney, A. Karpeisky, N. Usman, L. Beigelman, and W. G. Scott. The structural basis for hammerhead ribozyme self-cleavage. *Cell*, 92:665–673, 1998.
- [887] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. Scop: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247:536–540, 1995.
- [888] E. W. Myers, G. G. Sutton, H. O. Smith, M. D. Adams, and J. C. Venter. On the sequencing and assembly of the human genome. *Proc. Natl. Acad. Sci. USA*, 99:4145–4146, 2002.
- [889] A. Nahvi, N. Sudarsan, M. S. Ebert, X. Zou, K. L. Brown, and R. R. Breaker. Genetic control by a metabolite binding mRNA. *Chem. Biol.*, 9:1043–1049, 2002.
- [890] L. Naldini. A comeback for gene therapy. *Science*, 326:805–806, 2009.
- [891] S. G. Nash and J. Nocedal. A numerical study of the limited memory BFGS method and the truncated-Newton method for large-scale optimization. *SIAM J. Opt.*, 1:358–372, 1991.
- [892] X. Nassif. A furtive pathogen revealed. *Science*, 287:1767–1768, 2000.
- [893] S. Neidle. *DNA Structure and Recognition*. Oxford University Press, Oxford, England, 1994.
- [894] S. Neidle. New insights into sequence-dependent DNA structure. *Nature Struct. Biol.*, 5:754–756, 1998.
- [895] S. Neidle, editor. *Oxford Handbook of Nucleic Acid Structure*. Oxford University Press, Oxford, England, 1999.
- [896] H. C. Nelson, J. T. Finch, B. F. Luisi, and A. Klug. The structure of an oligo(dA) · oligo(dT) tract and its biological implications. *Nature*, 330:221–226, 1987.
- [897] G. Némethy, K. D. Gibson, K. A. Palmer, C. N. Yoon, G. Paterlini, A. Zagari, S. Rumsey, and H. A. Scheraga. Energy parameters in polypeptides. 10. improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. *J. Phys. Chem.*, 96:6472–6484, 1992.
- [898] G. Némethy, M. S. Pottle, and H. A. Scheraga. Energy parameters in polypeptides. 9. Updating of geometrical parameters, nonbonded interactions, and hydrogen bond interactions for the naturally occurring amino acids. *J. Phys. Chem.*, 87:1883–1887, 1983.
- [899] N. Nevins, J.-H. Lii, and N. L. Allinger. Molecular mechanics (MM4) calculations on conjugated hydrocarbons. *J. Comput. Chem.*, 17:695–729, 1996.
- [900] M. Newman. The physics of networks. *Phys. Today*, 61:33–38, 2008.
- [901] K.L. Ng and S.K. Mishra. De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics*, 23:1321–1330, 2007.

- [902] S. Ng Kwang Loong and S.K. Mishra. Unique folding of precursor microRNAs: quantitative evidence and implications for de novo identification. *RNA*, 13:170–187, 2007.
- [903] A. Nicholls and B. Honig. A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson-Boltzmann equation. *J. Comput. Chem.*, 12:435–445, 1991.
- [904] S. Nie, Y. Xing, G. J. Kim, and J. W. Simons. Nanotechnology applications in cancer. *Annu. Rev. Biomed. Eng.*, 9:257–288, 2007.
- [905] P. E. Nielsen. DNA analogues with nonphosphodiester backbones. *Ann. Rev. Biophys. Biomol. Struct.*, 24:167–183, 1995.
- [906] P. E. Nielsen. A new molecule of life. *Sci. Amer.*, 299:64–71, 2008.
- [907] P. E. Nielsen, M. Egholm, R. H. Berg, and O. Buchardt. Sequence-selective recognition of DNA by strand displacement with a thymine-substituted polyamide. *Science*, 254:1497–1500, 1991.
- [908] M. Nilges, P. Markwick, T. Malliavin, W. Rieping, and M. Habeck. New frontiers in characterizing structure and dynamics by NMR. In T. Schwede and M. Peitsch, editors, *Computational Structural Biology. Methods and Applications*, pages 655–679. World Scientific, Singapore, 2008.
- [909] L. Nilsson and M. Karplus. Empirical energy functions for energy minimization and dynamics of nucleic acids. *J. Comput. Chem.*, 7:591–616, 1986.
- [910] P. Nissen, J. Hansen, N. Ban, P. B. Moore, and T. A. Steitz. The structural basis of ribosome activity in peptide bond synthesis. *Science*, 289:920–930, 2000.
- [911] P. Nissen, J. A. Ippolito, N. Ban, P. B. Moore, and T. A. Steitz. RNA tertiary interactions in the large ribosomal subunit: The A-minor motif. *Proc. Natl. Acad. Sci. USA*, 98:4899–4903, 2001.
- [912] K. T. No, S. G. Kim, K.-H. Cho, and H. A. Scheraga. Description of hydration free energy density as a function of molecular physical properties. *Biophys. Chem.*, 78:127–145, 1999.
- [913] M. E. M. Noble, J. A. Endicott, and L. N. Johnson. Protein kinase inhibitors: Insights into drug design from structure. *Science*, 303:1800–1805, 2004.
- [914] J. Nocedal. Updating quasi-Newton matrices with limited storage. *Mathematics of Computation*, 35:773–782, 1980.
- [915] J. Nocedal. Theory of algorithms for unconstrained optimization. *Acta Numerica*, 1:199–242, 1992.
- [916] J. Nocedal. Large-scale unconstrained optimization. In A. Watson and I. Duff, editors, *The State of the Art in Numerical Analysis*, pages 311–338. Oxford University Press, 1997.
- [917] J. Nocedal, A. Sartenaer, and C. Zhu. On the behavior of the gradient norm in the steepest descent method. Technical report, CERFACS, Toulouse, France (May 2000), 2000.
- [918] J. Nocedal and S. Wright. *Numerical Optimization*. Springer Verlag, New York, NY, 1999.
- [919] F. Noé and S. Fischer. Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr. Opin. Struct. Biol.*, 8:154–162, 2008.

- [920] F. Noé, I. Horenko, C. Schütte, and J. C. Smith. Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. *J. Chem. Phys.*, 126:155102, 2007.
- [921] W. G. Noid, J.-W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, and H. C. Anderson. The multiscale coarse-grained method: I. a rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.*, 128:244114, 2008.
- [922] H. F. Noller. RNA structure: Reading the ribosome. *Science*, 309:1508–1514, 2005.
- [923] J. Norberg and L. Nilsson. On the truncation of long-range electrostatic interactions in DNA. *Biophys. J.*, 79:1537–1553, 2000.
- [924] S. Nosé. A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.*, 52:255–268, 1984.
- [925] S. Nosé. Constant temperature molecular dynamics methods. *Prog. Theor. Phys. Suppl.*, 103:1–46, 1991.
- [926] E. Nudler. Flipping riboswitches. *Cell*, 126:19–22, 2006.
- [927] A. Nyberg and T. Schlick. Increasing the time step in molecular dynamics. *Chem. Phys. Lett.*, 198:538–546, 1992.
- [928] R. E. Odeh and J. O. Evans. The percentage points of the normal distribution. *App. Stat.*, 23:96–97, 1974.
- [929] M. Ogihara and A. Ray. DNA computing on a chip. *Science*, 403:143–144, 2000.
- [930] K. Okada and S. Okada. X-ray crystallographic analysis and semiempirical computations. In P. von Ragué Schleyer (Editor-in Chief), N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, and H. F. Schaefer, III, editors, *Encyclopedia of Computational Chemistry*, volume 5, pages 3223–3247. John Wiley & Sons, West Sussex, England, 1998.
- [931] Y. Okamoto. Generalized-ensemble algorithms: enhanced sampling techniques for Monte Carlo and molecular dynamics simulations. *J. Mol. Graph. Mod.*, 22:425–439, 2004.
- [932] K. Okazaki, N. Koga, S. Takada, J. N. Onuchic, and P. G. Wolynes. Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA*, 103:11844–11849, 2006.
- [933] S. Okumoto, L. L. Looger, K. D. Micheva, R. J. Reimer, S. J. Smith, and W. B. Frommer. Detection of glutamate release from neurons by genetically encoded surface-displayed FRET nanosensors. *Proc. Natl. Acad. Sci. USA*, 102:8740–8745, 2005.
- [934] A. Okur, B. Stockbine, V. Hornak, and C. Simmerling. Using PC clusters to evaluate the transferability of molecular mechanics force fields for proteins. *J. Comput. Chem.*, 24:21–31, 2003.
- [935] S. Oliver. Proteomics: Guilt-by-association goes global. *Nature*, 403:601–602, 2000.
- [936] M. V. Olson. Dr. watson's base pairs. *Nature*, 452:819–820, 2008.
- [937] W. K. Olson. How flexible is the furanose ring? 2. an updated potential energy estimate. *J. Amer. Chem. Soc.*, 104:278–286, 1982.

- [938] W. K. Olson. Theoretical studies of nucleic acid conformation: Potential energies, chain statistics, and model building. In S. Neidle, editor, *Topics in Nucleic Acid Structures: Part 2*, pages 1–79. Macmillan Press, London, England, 1982.
- [939] W. K. Olson. Simulating DNA at low resolution. *Curr. Opin. Struct. Biol.*, 6:242–256, 1996.
- [940] W. K. Olson, M. S. Babcock, A. Gorin, G.-H. Liu, N. L. Markey, J. A. Martino, S. C. Pedersen, A. R. Srinivasan, I. Tobias, T. P. Westcott, and P. Zhang. Flexing and folding of double helical DNA. *Biol. Chem.*, 55:7–29, 1995.
- [941] W. K. Olson, S. K. Burley, R. E. Dickerson, M. Gerstein, S. C. Harvey, U. Heinemann, X.-J. Lu, S. Neidle, Z. Shakked, M. Suzuki, X.-S. Tung, H. Sklenar, J. Westbrook, E. Westhof, C. Wolberger, and H. Berman. A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.*, 313:229–237, 2001. Available also through posting on NDB Archives, [ndb-server.rutgers.edu/NDB/archives/index.html](http://ndb-server.rutgers.edu/NDB/archives/index.html) and [www.idealibrary.com/links/doi/10.1006/jmbi.2001.4987/](http://www.idealibrary.com/links/doi/10.1006/jmbi.2001.4987/).
- [942] W. K. Olson, A. A. Gorin, X. J. Lu, L. M. Hock, and V. B. Zhurkin. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. USA*, 95:11163–11168, 1998.
- [943] W. K. Olson, N. L. Markey, R. L. Jernigan, and V. B. Zhurkin. Influence of fluctuations on DNA curvature. A comparison of flexible and static wedge models of intrinsically bent DNA. *J. Mol. Biol.*, 232:530–554, 1993.
- [944] W. K. Olson and J. L. Sussman. How flexible is the furanose ring? 1. A comparison of experimental and theoretical studies. *J. Amer. Chem. Soc.*, 104:270–278, 1982.
- [945] W. K. Olson, T. P. Westcott, J. A. Martino, and G.-H. Liu. Computational studies of spatially constrained DNA chains. In J. P. Mesirov, K. Schulten, and D. W. Sumners, editors, *Mathematical Approaches to Biomolecular Structure and Dynamics*, volume 82 of *IMA Volumes in Mathematics and Its Applications*, New York, NY, 1996. Springer-Verlag.
- [946] W. K. Olson and V. B. Zhurkin. Twenty years of DNA bending. In R. H. Sarma and M. H. Sarma, editors, *Biological Structure and Dynamics: Proceedings of the Ninth Conversation in the Discipline Biomolecular Stereodynamics*, pages 341–370, Schenectady, NY, 1996. Adenine Press.
- [947] W. K. Olson and V. B. Zhurkin. Modeling DNA deformations. *Curr. Opin. Struct. Biol.*, 10:286–297, 2000.
- [948] J. N. Onuchic. Contacting the protein folding funnel with NMR. *Proc. Natl. Acad. Sci. USA*, 94:7129–7131, 1997.
- [949] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes. Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.*, 48:545–600, 1997.
- [950] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.*, 25:1656–1676, 2004.
- [951] P. Ordejón, D. A. Drabold, R. M. Martin, and M. P. Grumbach. Linear system-size scaling methods for electronic structure calculations. *Phys. Rev. B*, 51:1456–1476, 1995.

- [952] C. A. Orengo, A. E. Todd, and J. M. Thornton. From protein structure to function. *Curr. Opin. Struct. Biol.*, 9:374–382, 1999.
- [953] M. Orozco, A. Noy, and A. Pérez. Recent advances in the study of nucleic acid flexibility by molecular dynamics. *Curr. Opin. Struct. Biol.*, 18:185–193, 2008.
- [954] M. L. Overton. *Numerical Computing with IEEE Floating Point Arithmetic*. SIAM, Philadelphia, PA, 2001.
- [955] E. Paci and M. Karplus. Unfolding proteins by external forces and temperature: The importance of topology and energetics. *Proc. Natl. Acad. Sci. USA*, 97:6521–6526, 2000.
- [956] E. Paci, K. Lindorff-Larsen, C. M. Dobson, M. Karplus, and M. Vendruscolo. Transition state contact orders correlate with protein folding rates. *J. Mol. Biol.*, 352:495–500, 2005.
- [957] A. C. Pan and B. Roux. Building Markov state models along pathways to determine free energies and rates of transitions. *J. Chem. Phys.*, 129:064107, 2008.
- [958] P. M. Pardalos, D. Shalloway, and G. Xue, editors. *Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding*, volume 23 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*. American Mathematical Society, Providence, Rhode Island, 1996.
- [959] J. Park, S. A. Teichmann, T. Hubbard, and C. Chothia. Intermediate sequences increase the detection of homology between sequences. *J. Mol. Biol.*, 273:349–354, 1997.
- [960] S. C. Park and K. W. Miller. Random number generators: Good ones are hard to find. *Comm. ACM*, 31:1192–1201, 1988.
- [961] G. N. Parkinson, M. P. H. Lee, and S. Neidle. Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature*, 417:876–880, 2002.
- [962] R. G. Parr and W. Yang. Density-functional theory of the electronic structure of molecules. *Ann. Rev. Phys. Chem.*, 46:701–728, 1995.
- [963] M. Parrinello. Eppur si muove. In A. H. Zewail, editor, *Physical biology: 4D visualization of complexity*, chapter 11, pages 247–266. Imperial College Press, London, UK, 2008.
- [964] M. Parrinello and A. Rahman. Crystal structure and pair potentials: A molecular-dynamics study. *Phys. Rev. Lett.*, 45:1196–1199, 1980.
- [965] S. Pasquali, H. H. Gan, and T. Schlick. Modular RNA architecture revealed by computational analysis of existing pseudoknots and ribosomal RNAs. *Nuc. Acids Res.*, 33:1384–1398, 2005.
- [966] R. W. Pastor. Techniques and applications of Langevin dynamics simulations. In G. R. Luckhurst and C. A. Veracini, editors, *The Molecular Dynamics of Liquid Crystals*, pages 85–138. Kluwer Academic, Dordrecht, The Netherlands, 1994.
- [967] R. W. Pastor, B. R. Brooks, and A. Szabo. An analysis of the accuracy of Langevin and molecular dynamics algorithms. *Mol. Phys.*, 65:1409–1419, 1988.
- [968] D. J. Patel. A molecular propeller. *Nature*, 417:807–808, 2002.
- [969] D. J. Patel, B. Mao, Z. Gu, B. E. Hingerty, A. Gorin, A. K. Basu, and S. Broyde. Nuclear magnetic resonance solution structures of covalent aromatic amine-DNA adducts and their magnetic relevance. *Chem. Res. Toxic.*, 11:391–407, 1998.

- [970] S. Patel and C. L. Brooks, III. Fluctuating charge force fields: recent developments and applications from small molecules to macromolecular biological systems. *Mol. Sim.*, 32:231–249, 2006.
- [971] G. A. Patikoglou, J. L. Kim, L. Sun, S.-H. Yang, T. Kodadek, and S. K. Burley. TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. *Genes & Devt.*, 13:3217–3230, 1999.
- [972] L. Pauling. The nature of bond orbitals and the origin of potential barriers to internal rotation in molecules. *Proc. Natl. Acad. Sci.*, 44:211–216, 1958.
- [973] L. Pauling. *The Nature of the Chemical Bond*. third edition, Cornell University Press, New York, NY, 1960.
- [974] L. Pauling and R. B. Corey. Configurations of polypeptide chains with favored orientations around single bonds: Two new pleated sheets. *Proc. Natl. Acad. Sci. USA*, 37:729–740, 1951.
- [975] L. Pauling and R. B. Corey. A proposed structure for nucleic acids. *Proc. Natl. Acad. Sci. USA*, 39:84–97, 1953.
- [976] L. Pauling, R. B. Corey, and H. R. Branson. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. USA*, 37:205–211, 1951.
- [977] L. Pauling, H. A. Itano, S. J. Singer, and I. C. Wells. Sickle cell anemia, a molecular disease. *Science*, 110:543–548, 1949.
- [978] L. Pauling and E. B. Wilson, Jr. *Introduction to Quantum Mechanics with Applications to Chemistry*. Dover, New York, NY, 1985.
- [979] K. Pawłowski, A. Bierzyński, and A. Godzik. Structural diversity in a family of homologous proteins. *J. Mol. Biol.*, 258:349–366, 1996.
- [980] D. A. Pearlman and S. H. Kim. Determinations of atomic partial charges for nucleic acid constituents from x-ray diffraction data. I. 2'-deoxycytidine-5'-monophosphate. *Biopolymers*, 24:327–357, 1985.
- [981] N. D. Pearson and C. D. Prescott. RNA as a drug target. *Chem. & Biol.*, 4:409–414, 1997.
- [982] S. Pennell, E. Manktelow, A. Flatt, G. Kelly, S.J. Smerdon, and I. Brierley. The stimulatory RNA of the Visna-Maedi retrovirus ribosomal frameshifting signal is an unusual pseudoknot with an interstem element. *RNA*, 14:1366–1377, 2008.
- [983] M. B. Pepys, J. Herbert, W. L. Hutchinson, G. A. Tennent, H. J. Lachmann, J. R. Gallimore, L. B. Lovat, T. Bartfai, A. Alanine, C. Hertel, T. Hoffmann, R. Jakob-Roetne, R. D. Norcross, J. A. Kemp, K. Yamamura, M. Suzuki, G. W. Taylor, S. Murray, D. Thompson, A. Purvis, S. Kolstoe, S. P. Wood, and P. N. Hawkins. Targeted pharmacological depletion of serum amyloid P component for treatment of human amyloidosis. *Nature*, 417:254–259, 2002.
- [984] A. Peracchi. Prospects for antiviral ribozymes and deoxyribozymes. *Rev. Med. Virol.*, 14:47–64, 2004.
- [985] O. E. Percus and M. H. Kalos. Random number generators for SIMD parallel processors. *J. Paral. Dist. Comput.*, 6:477–497, 1989.
- [986] O. E. Percus and J. K. Percus. Intrinsic relations in the structure of linear congruential generators modulo  $2^\beta$ . *Stat. Prob. Let.*, 15:381–383, 1992.

- [987] A. Pérez, J. Luque, and M. Orozco. Dynamics of B-DNA on the microsecond time scale. *J. Amer. Chem. Soc.*, 129:14739–14745, 2007.
- [988] A. Pérez, I. Marchán, D. Svozil, J. Sponer, T. E. Cheatham, III, C. A. Laughton, and M. Orozco. Refinement of the AMBER force field for nucleic acids: Improving the description of  $\alpha/\gamma$  conformers. *Biophys. J.*, 92:3817–3829, 2007.
- [989] T. Schlick and O. Perisic. Mesoscale simulations of two nucleosome-repeat length oligonucleosomes. *Phys. Chem. Chem. Phys.*, 11:10729–10737, 2009.
- [990] T. T. Perkins, S. R. Quake, D. E. Smith, and S. Chu. Relaxation of a single DNA molecule observed by optical microscopy. *Science*, 264:822–825, 1994.
- [991] J. W. Perram, H. G. Petersen, and S. W. De Leeuw. An algorithm for the simulation of condensed matter which grows as the  $\frac{3}{2}$  power of the number of particles. *Mol. Phys.*, 65:875–893, 1988.
- [992] C. S. Peskin and T. Schlick. Molecular dynamics by the backward Euler's method. *Comm. Pure App. Math.*, 42:1001–1031, 1989.
- [993] G. A. Petsko. An idea whose time has gone. *Genome Biol.*, 8:107, 2007.
- [994] G. A. Petsko and D. Ringe. Observation of unstable species in enzyme-catalyzed transformations using protein crystallography. *Curr. Opin. Chem. Biol.*, 4:89–94, 2000.
- [995] A. T. Phan, J.-L. Leroy, and M. Guéron. Determination of the residence time of water molecules hydrating  $B'$ -DNA and B-DNA, by one-dimensional zero-enhancement nuclear Overhauser effect spectroscopy. *J. Mol. Biol.*, 286:505–519, 1999.
- [996] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten. Scalable molecular dynamics with NAMD. *J. Comput. Chem.*, 26:1781–1802, 2005.
- [997] L. Piela, J. Kostrowicki, and H. A. Scheraga. The multiple-minima problem in conformational analysis of molecules. deformation of the potential energy hypersurface by the diffusion equation method. *J. Phys. Chem.*, 93:3339–3346, 1989.
- [998] D. S. Pilch, G. E. Plum, and K. J. Breslauer. The thermodynamics of DNA structures which contain lesions or guanine tetrads. *Curr. Opin. Struct. Biol.*, 5:334–342, 1995.
- [999] O. Pillai, A. B. Dhanikula, and R. Panchagnula. Drug delivery: An odyssey of 100 years. *Curr. Opin. Chem. Biol.*, 5:439–446, 2001.
- [1000] J. Pillardy, C. Czaplewski, A. Liwo, J. Lee, D. R. Ripoll, R. Kaźmierkiewicz, S. Oldziej, W. J. Wedemeyer, K. D. Gibson, Y. A. Arnautova, J. Saunders, Y.-J. Ye, and H. A. Scheraga. Recent improvements in prediction of protein structure by global optimization of a potential energy function. *Proc. Natl. Acad. Sci. USA*, 98:2329–2333, 2001.
- [1001] T. Pinou, T. Schlick, B. Li, and H. G. Dowling. Addition of Darwin's third dimension to phyletic trees. *J. Theor. Biol.*, 182:505–512, 1996.
- [1002] R. M. Pitzer. The barrier to internal rotation in ethane. *Acc. Chem. Res.*, 16:207–210, 1983.
- [1003] M. Pizza, V. Scarlato, V. Masignani, M. M. Giuliani, B. Aricó, M. Comanducci, G. T. Jennings, L. Baldi, E. Bartolini, B. Capecci, C. L. Galeotti, E. Luzzi,

- R. Manetti, E. Marchetti, M. Mora, S. Nuti, G. Ratti, L. Santini, S. Savino, M. Scarselli, E. Storni, P. Zuo, M. Broeker, E. Hundt, B. Knapp, E. Blair, T. Mason, H. Tettelin, D. W. Hood, A. C. Jeffries, N. J. Saunders, D. M. Granoff, J. C. Venter, E. R. Moxon, G. Grandi, and R. Rappuoli. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science*, 287:1816–1820, 2000.
- [1004] R. H. A. Plasterk. RNA silencing: The genome’s immune system. *Science*, 296:1263–1265, 2002.
- [1005] H. W. Pley, K. M. Flaherty, and D. B. McKay. Three-dimensional structure of a hammerhead ribozyme. *Nature*, 372:68–74, 1994.
- [1006] M. J. Plotkin. *Medicine Quest: In Search of Nature’s Healing Secrets*. Viking Penguin, New York, NY, 2000.
- [1007] G. E. Plum and K. J. Breslauer. Calorimetry of proteins and nucleic acids. *Curr. Opin. Struct. Biol.*, 5:682–690, 1995.
- [1008] G. E. Plum, D. S. Pilch, S. F. Singleton, and K. J. Breslauer. Nucleic acid hybridization: Triplex stability and energetics. *Ann. Rev. Biophys. Biomol. Struct.*, 24:319–350, 1995.
- [1009] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. H. Kim, L. C. Goumnerova, P. M. Black, C. Lau, J. C. Allen, D. Zagzag, J. M. Olson, T. Curran, C. Wetmore, J. A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D. N. Louis, J. P. Mesirov, E. S. Lander, and T. R. Golub. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415:436–442, 2002.
- [1010] V. Popovits and L. Goodman. Hyperconjugation not steric repulsion leads to the staggered structure of ethane. *Nature*, 411:565–568, 2001.
- [1011] J. A. Pople. Quantum chemical models (Nobel lecture). *Angew. Chem. Int. Ed.*, 38:1894–1902, 1999.
- [1012] M. J. D. Powell. Restart procedures for the conjugate gradient method. *Math. Prog.*, 12:241–254, 1977.
- [1013] M. Prabhakaran, S. C. Harvey, B. Mao, and J. A. McCammon. Molecular dynamics of phenylalanine transfer RNA. *J. Biomol. Struct. Dynam.*, 1:357–369, 1983.
- [1014] N. V. Prabhu, J. S. Perkyns, H. D. Blatt, P. E. Smith, and B. M. Pettitt. Comparison of the potentials of mean force for alanine tetrapeptide between integral equation theory and simulation. *Biophys. Chem.*, 78:113–126, 1999.
- [1015] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in Fortran 77: The Art of Scientific Computing*, volume 1 of *Fortran Numerical Recipes*. Cambridge University Press, New York, NY, second edition, 1992.
- [1016] D. J. Price and C. L. Brooks, III. Modern protein force fields behave comparably in molecular dynamics simulations. *J. Comput. Chem.*, 23:1045–1057, 2002.
- [1017] M. A. Price and T. D. Tullius. How the structure of an adenine tract depends on sequence context: A new model for the structure of  $T_nA_n$  DNA sequences. *Biochemistry*, 32:127–136, 1993.
- [1018] L. J. Prins, D. N. Reinhoudt, and P. Timmerman. Nonvalent synthesis using hydrogen bonding. *Angew. Chem. Int. Ed.*, 40:2382–2426, 2001.

- [1019] P. Procacci, M. Marchi, and G. J. Martyna. Electrostatic calculations and multiple time scales in molecular dynamics simulation of flexible molecular systems. *J. Chem. Phys.*, 108:8799–8803, 1998.
- [1020] D. Pruss, B. Bartholomew, J. Persinger, J. Hayes, G. Arents E. N. Moudrianakis, and A. P. Wolffe. An asymmetric model for the nucleosome: A binding site for linker histones inside the DNA gyres. *Science*, 274:614–617, 1996.
- [1021] M. Ptashne. How gene activators work. *Sci. Amer.*, 260:41–47, 1989.
- [1022] A. M. Pyle and J. B. Green. RNA folding. *Curr. Opin. Struct. Biol.*, 5:303–310, 1995.
- [1023] H. Qian and J. A. Schellman. Transformed Poisson-Boltzmann relations and ionic distributions. *J. Phys. Chem. B*, 104:11528–11540, 2000.
- [1024] X. Qian. *Biomolecular Structure and Dynamics: Algorithm Development and Applications to DNA-Transcription and Promoter Elements*. PhD thesis, New York University, Department of Chemistry, New York, NY, May 2002.
- [1025] X. Qian and T. Schlick. Efficient multiple-timestep integrators with distance-based force splitting for particle-mesh-Ewald molecular dynamics simulations. *J. Chem. Phys.*, 116:5971–5983, 2002.
- [1026] X. Qian, D. Strahs, and T. Schlick. Dynamic simulations of 13 TATA variants refine kinetic hypotheses of sequence/activity relationships. *J. Mol. Biol.*, 308:681–703, 2001.
- [1027] X. Qian, D. Strahs, and T. Schlick. A new program for optimizing periodic boundary models of solvated biomolecules (PBCAID). *J. Comput. Chem.*, 22:1843–1850, 2001.
- [1028] G. Quarta, N. Kim, J. A. Izzo, and T. Schlick. Analysis of riboswitch structure and function by an energy landscape framework. *J. Mol. Biol.*, 393:993–1003, 2009.
- [1029] R. Radhakrishnan, K. Arora, Y. Wang, W. A. Beard, S. H. Wilson, and T. Schlick. Regulation of DNA repair fidelity by molecular checkpoints: “gates” in DNA polymerase  $\beta$ ’s substrate selection. *Biochem.*, 45:15142–15156, 2006.
- [1030] R. Radhakrishnan and T. Schlick. Biomolecular free energy profiles by a shooting/umbrella sampling protocol, “BOLAS”. *J. Chem. Phys.*, 121:2436–2444, 2004.
- [1031] R. Radhakrishnan and T. Schlick. Orchestration of cooperative events in DNA synthesis and repair mechanism unraveled by transition path sampling of DNA polymerase  $\beta$ ’s closing. *Proc. Natl. Acad. Sci. USA*, 101:5970–5975, 2004.
- [1032] R. Radhakrishnan and T. Schlick. Fidelity discrimination in DNA polymerase  $\beta$ : differing closing profiles for a mismatched G:A versus matched G:C base pair. *J. Amer. Chem. Soc.*, 127:13245–13252, 2005.
- [1033] R. Radhakrishnan and T. Schlick. Correct and incorrect nucleotide incorporation pathways in dna polymerase  $\beta$ ’s. *Biochem. Biophys. Res. Comm.*, 350:521–529, 2006.
- [1034] A. Rahman and F. H. Stillinger. Molecular dynamics study of liquid water. *J. Chem. Phys.*, 55:3336–3359, 1971.
- [1035] A. Rahman and F. H. Stillinger. Improved simulation of liquid water by molecular dynamics. *J. Chem. Phys.*, 60:1545–1557, 1974.

- [1036] G. Ramachandran and T. Schlick. Solvent effects on supercoiled DNA dynamics explored by Langevin dynamics simulations. *Phys. Rev. E*, 51:6188–6203, 1995.
- [1037] G. Ramachandran and T. Schlick. Beyond optimization: Simulating the dynamics of supercoiled DNA by a macroscopic model. In P. M. Pardalos, D. Shalloway, and G. Xue, editors, *Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding*, volume 23 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 215–231, Providence, Rhode Island, 1996. American Mathematical Society.
- [1038] D.C. Rapaport. *The Art of Molecular Dynamics Simulation*. Cambridge University Press, Cambridge, England, second edition, 2004.
- [1039] T. D. Rasmussen, P. Ren, J. W. Ponder, and F. Jensen. Force field modeling of conformational energies: importance of multipole moments and intramolecular polarization. *Int. J. Quant. Chem.*, 107:1390–1395, 2007.
- [1040] J. Ray and G. S. Manning. Counterion and coion distribution functions in the counterion condensation theory of polyelectrolytes. *Macromolecules*, 32:4588–4595, 1999.
- [1041] E.A. Rødland. Pseudoknots in RNA secondary structures: representation, enumeration, and prevalence. *J. Comput. Biol.*, 13:1197–1213, 2006.
- [1042] J. S. Read and G. F. Joyce. A ribozyme composed of only two different nucleotides. *Nature*, 420:841–844, 2002.
- [1043] R. J. Read and D. E. Wemmer. Biophysical methods. Bigger, better, faster and automatically too? (Editorial overview). *Curr. Opin. Struct. Biol.*, 9:591–593, 1999.
- [1044] P. R. Reilly. *Abraham Lincoln's DNA and Other Adventures in Genetics*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2000.
- [1045] J. Ren, R. M. Esnouf, A. L. Hopkins, E. Y. Jones, I. Kirby, J. Keeling, C. K. Ross, B. A. Larder, D. I. Stuart, and D. K. Stammers. 3'-Azido-3'-deoxythymidine drug resistance mutations in HIV-1 reverse transcriptase can induce long range conformational changes resistance. *Proc. Natl. Acad. Sci. USA*, 95:9518–9523, 1998.
- [1046] F. J. Resende and B. V. Costa. Using random number generators in Monte Carlo simulations. *Phys. Rev. E*, 58:5183–5184, 1998.
- [1047] G. Rhodes. *Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models*. Academic Press, San Diego, CA, second edition, 2000.
- [1048] S. A. Rice, M. Nagasawa, and H. Morawetz. *Polyelectrolyte Solutions: A Theoretical Introduction*, volume 2 of *Molecular Biology: An International Series of Monographs and Textbooks*. Academic Press, New York, NY, 1961.
- [1049] A. Rich. The rise of single-molecule DNA chemistry. *Proc. Natl. Acad. Sci. USA*, 95:13999–14000, 1998.
- [1050] T. J. Richmond and C. A. Davey. The structure of DNA in the nucleosome core. *Nature*, 423:145–150, 2003.
- [1051] D. R. Ripoll, J. A. Vila, and H. A. Scheraga. Folding of the villin headpiece subdomain from random structures. Analysis of the charge distribution as a function of the pH. *J. Mol. Biol.*, 339:915–925, 2004.

- [1052] K. Rippe, P. H. von Hippel, and J. Langowski. Action at a distance: DNA-looping and initiation of transcription. *Trends Bio. Sci.*, 20(12):500–506, Dec. 1996.
- [1053] J. B. Ristaino, C. T. Groves, and G. R. Parra. PCR amplification of the Irish potato famine pathogen from historic specimens. *Nature*, 411:695–697, 2001.
- [1054] N. A. Roberts, J. A. Martin, D. Kinchington, A. V. Broadhurst, J. C. Craig, I. B. Duncan, S. A. Galpin, B. K. Handa, J. Kay, A. Kröhn, R. W. Lambert, J. H. Merrett, J. S. Mills, K. E. B. Parkes, S. Redshaw, A. J. Ritchie, D. L. Taylor, G. J. Thomas, and P. J. Machin. Rational design of peptide-based HIV proteinase inhibitors. *Science*, 248:358–361, 1990.
- [1055] P. J. J. Robinson, L. Fairall, V. A. T. Huynh, and D. Rhodes. EM measurements define the dimensions of the “30-nm” chromatin fiber: Evidence for a compact, interdigitated structure. *Proc. Natl. Acad. Sci. USA*, 103:6506–6511, 2006.
- [1056] R. A. Robinson and R. H. Stokes. *Electrolyte Solutions: The Measurement and Interpretation of Conductance, Chemical Potential and Diffusion in Solutions of Simple Electrolytes*. Butterworth & Co., London, England, second edition, 1965.
- [1057] W. Rocchia, E. Alexov, and B. Honig. Extending the applicability of the nonlinear Poisson-Boltzmann equation: Multiple dielectric constants and multivalent ions. *J. Phys. Chem. B*, 105:6507–6514, 2001.
- [1058] W. Rocchia, S. Sridharan, A. Nicholls, E. Alexov, A. Chiabrera, and B. Honig. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: Applications to the molecular systems and geometric objects. *J. Comput. Chem.*, 23:128–137, 2002.
- [1059] M. C. Roco, R. S. Williams, and P. Alivisatos, editors. *Nanotechnology Research Directions: IWGN (Interagency Working Group on Nanoscience, Engineering and Technology) Workshop Report. Vision for Nanotechnology Research and Development in the Next Decade*, Loyola College, Maryland, 1999. International Technology Research Institute, World Technology (WTEC) Division, Loyola College. URL: [itri.loyola.edu/nano/IWGN.Research.Directions/](http://itri.loyola.edu/nano/IWGN.Research.Directions/); Also published in hard copy by Kluwer Academic Press, February 2000.
- [1060] T. Rodinger, P. Howell, and R. Poms. Distributed replica sampling. *J. Chem. Ther. Comp.*, 2:725–731, 2006.
- [1061] A. D. Rodrigues and J. H. Lin. Screening of drug candidates for their drug–drug interaction potential. *Curr. Opin. Chem. Biol.*, 5:396–401, 2001.
- [1062] J. Rogal and P. G. Bolhuis. Multiple state transition path sampling. *J. Chem. Phys.*, 129:224107, 2008.
- [1063] A. Roitberg and R. Elber. Modeling of side chains in peptides and proteins: Application of the locally enhanced sampling and the simulated annealing methods to find minimum energy conformations. *J. Chem. Phys.*, 95:9277–9287, 1991.
- [1064] A. E. Roitberg, A. Okur, and C. Simmerling. Coupling of replica exchange simulations to a non-Boltzmann structure reservoir. *J. Phys. Chem. B*, 111:2415–2418, 2007.
- [1065] V. Rokhlin. Rapid solution of integral equations of classical potential theory. *J. Comput. Phys.*, 60:187–207, 1985.
- [1066] G. Rose. Protein folding and the Paracelsus challenge. *Nature Struc. Biol.*, 4:512–514, 1997.

- [1067] S. M. Ross. *A Course in Simulation*. Macmillan Publishing Company, New York, NY, 1990.
- [1068] I. K. Roterman, M. H. Lambert, K. D. Gibson, and H. A. Scheraga. A comparison of the CHARMM, AMBER and ECEPP potentials for peptides. I. Conformational predictions for the tandemly repeated peptide (Asn-Ala-Asn-Pro)<sub>9</sub>. *J. Biomol. Struct. Dyn.*, 7:391–419, 1989.
- [1069] I. K. Roterman, M. H. Lambert, K. D. Gibson, and H. A. Scheraga. A comparison of the CHARMM, AMBER and ECEPP potentials for peptides. II.  $\phi$ - $\psi$  maps for N-methyl amide: Comparisons, contrasts and simple experimental tests. *J. Biomol. Struct. Dyn.*, 7:421–453, 1989.
- [1070] E. Rothstein. DNA teaches history a few lessons of its own. *New York Times*, 1998. Sunday, May 24 (under Ideas & Trends of the Week in Review).
- [1071] J. Rotne and S. Prager. Variational treatment of hydrodynamic interaction in polymers. *J. Chem. Phys.*, 50:4831–4837, 1969.
- [1072] P. E. Rouse, Jr. Dilute solutions of coiling polymers. *J. Chem. Phys.*, 21: 1272–1280, 1953.
- [1073] B. Roux and T. Simonson. Implicit solvent models. *Biophys. Chem.*, 78:1–20, 1999.
- [1074] H. Rozenberg, D. Rabinovich, F. Frolow, R. S. Hegde, and Z. Shakkeb. Structural code for DNA recognition revealed in crystal structures of papillomavirus E2-DNA targets. *Proc. Natl. Acad. Sci. USA*, 95:15194–15199, 1998.
- [1075] M. Rueda, P. Chacon, and M. Orozco. Thorough validation of protein normal mode analysis: a comparative study with essential dynamics. *Structure*, 15:565–575, 2007.
- [1076] M. Rueda, E. Cubero, C. A. Laughton, and M. Orozco. Exploring the counterion atmosphere around DNA: What can be learned from molecular dynamics simulations? *Biophys. J.*, 87:800–811, 2004.
- [1077] D. Rugar and P. Hansma. Atomic force microscope. *Physics Today*, 43:23–30, 1990.
- [1078] R. D. Ruth. A canonical integration technique. *IEEE Trans. Nucl. Sci.*, 30:2669–2671, 1983.
- [1079] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comput. Phys.*, 23:327–341, 1977.
- [1080] W. Saenger. *Principles of Nucleic Acid Structure*. Springer Advanced Texts in Chemistry. Springer-Verlag, New York, NY, 1984.
- [1081] C. Sagui and T. A. Darden. Molecular dynamics simulations of biomolecules: Long-range electrostatic effects. *Ann. Rev. Biophys. Biomol. Struc.*, 28:155–179, 1999.
- [1082] C. Sagui and T. A. Darden. Multigrid methods for classical molecular dynamics simulations of biomolecules. *J. Chem. Phys.*, 114:6578–6591, 2001.
- [1083] K. Sakamoto, H. Gouzu, K. Komiya, D. Kiga, S. Yokoyama, T. Yokomori, and M. Hagiya. Molecular computation by DNA hairpin formation. *Science*, 288:1223–1226, 2000.

- [1084] K. Salehi-Ashtiani, A. Luptk, A. Litovchick, and J.W. Szostak. A genomewide search for ribozymes reveals an HDV-like sequence in the human CPEB3 gene. *Science*, 313:1788–1792, 2006.
- [1085] F. A. Samatey, K. Imada, S. Nagashima, F. Vonderviszt, T. Kumasaka, M. Yamamoto, and K. Namba. Structure of the bacterial flagellar protofilament and implications for a switch for supercoiling. *Nature*, 410:331–337, 2001.
- [1086] R. Sánchez and A. Šali. Advances in comparative protein-structure modelling. *Curr. Opin. Struct. Biol.*, 7:206–214, 1997.
- [1087] R. Sánchez and A. Šali. Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl. Acad. Sci. USA*, 95:13597–13602, 1998.
- [1088] B. Sandak. Multiscale fast summation of long-range charge and dipolar interactions. *J. Comput. Chem.*, 22:717–731, 2001.
- [1089] A. Sandu and T. Schlick. Masking resonance artifacts in force splitting methods for biomolecular simulations by extrapolative Langevin dynamics. *J. Comput. Phys.*, 151:74–113, May 1999. (Special Volume on Computational Biophysics).
- [1090] J. M. Sanz-Serna and M. P. Calvo. *Numerical Hamiltonian Problems*. Chapman & Hall, London, England, 1994.
- [1091] R. Sarma. *Ramachandran: A Biography of Gopalasamudram Narayana Ramachandran, the Famous Indian Biophysicist*. Adenine Press, Schenectady, NY, 1998.
- [1092] H. F. Schaefer. Methylene: A paradigm for computational quantum chemistry. *Science*, 231:1100–1107, 1986.
- [1093] J. M. Schafer, E.-S. Lee, R. C. Dardes, D. Bentrem, R. M. O'Regan, A. De Los Reyes, and V. C. Jordan. Analysis of cross-resistance of the selective estrogen receptor modulators arzoxifene (LY353381) and LY117018 in tamoxifen-stimulated breast cancer xenografts. *Clin. Cancer Res.*, 7:2505–2512, 2001.
- [1094] T. Schalch, S. Duda, D. F. Sargent, and T. J. Richmond. X-ray structure of a tetranucleosome and its implications for the chromatin fibre. *Nature*, 436:138–141, 2005.
- [1095] J. A. Schellman. Flexibility of DNA. *Biopolymers*, 13:217–226, 1974.
- [1096] J. A. Schellman. The flexibility of DNA. I. Thermal fluctuations. *Biophys. Chem.*, 11:321–328, 1980.
- [1097] J. A. Schellman and S. C. Harvey. Static contributions to the persistence length of DNA and dynamic contributions to DNA curvature. *Biophys. Chem.*, 55:95–114, 1995.
- [1098] J. A. Schellman and C. Schellman. The conformation of polypeptide chains in proteins. In H. Neurath, editor, *The Proteins*, volume 2, pages 1–137. Academic Press, New York, NY, second edition, 1964.
- [1099] C. A. Schiffer, J. W. Caldwell, P. A. Kollman, and R. M. Stroud. Protein structure prediction with a combined solvation free energy-molecular mechanics force field. *Mol. Sim.*, 10:121–149, 1993.
- [1100] J. F. Schildbach, A. W. Karzai, B. E. Raumann, and R. T. Sauer. Origins of DNA-binding specificity: Role of protein contacts with the DNA backbone. *Proc. Natl. Acad. Sci. USA*, 96:811–817, 1999.

- [1101] T. Schlick. *Modeling and Minimization Techniques for Predicting Three-Dimensional Structures of Large Biological Molecules*. PhD thesis, New York University, Courant Institute of Mathematical Sciences, New York, NY, October 1987.
- [1102] T. Schlick. A modular strategy for generating starting conformations and data structures of polynucleotide helices for potential energy calculations. *J. Comput. Chem.*, 9(8):861–889, 1988.
- [1103] T. Schlick. A recipe for evaluating and differentiating  $\cos \phi$  expressions. *J. Comput. Chem.*, 10:951–956, 1989.
- [1104] T. Schlick. Optimization methods in computational chemistry. In K. B. Lipkowitz and D. B. Boyd, editors, *Reviews in Computational Chemistry*, volume III, pages 1–71. VCH Publishers, New York, NY, 1992.
- [1105] T. Schlick. Modeling superhelical DNA: Recent analytical and dynamic approaches. *Curr. Opin. Struct. Biol.*, 5:245–262, 1995.
- [1106] T. Schlick. Modeling superhelical DNA: Recent analytical and dynamic approaches. *Curr. Opin. Struct. Biol.*, 5:245–262, 1995.
- [1107] T. Schlick. Pursuing Laplace’s vision on modern computers. In J. P. Mesirov, K. Schulten, and D. W. Sumners, editors, *Mathematical Applications to Biomolecular Structure and Dynamics*, volume 82 of *IMA Volumes in Mathematics and Its Applications*, pages 219–247, New York, NY, 1996. Springer-Verlag.
- [1108] T. Schlick. Geometry optimization. In P. von Ragué Schleyer (Editor-in Chief), N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, and H. F. Schaefer, III, editors, *Encyclopedia of Computational Chemistry*, volume 2, pages 1136–1157. John Wiley & Sons, West Sussex, England, 1998.
- [1109] T. Schlick. Computational molecular biophysics today: A confluence of methodological advances and complex biomolecular applications. *J. Comput. Phys.*, 151:1–8, May 1999. (Special Volume on Computational Biophysics).
- [1110] T. Schlick. Time-trimming tricks for dynamic simulations: Splitting force updates to reduce computational work. *Structure*, 9:R45–R53, 2001.
- [1111] T. Schlick. Engineering teams up with computer-simulation and visualization tools to probe biomolecular mechanisms. *Biophys. J.*, 85:1, 2003.
- [1112] T. Schlick. The critical collaboration between art and science: Applying *an experiment on a bird in an air pump* to the ramifications of genomics on society. *Leonardo*, 38:323–329, 2005.
- [1113] T. Schlick. RNA — the cousin left behind becomes a star. In J. Šponer and F. Lankáš, editors, *Computational Studies of DNA and RNA*, pages 259–281. Springer Verlag, Dordrecht, The Netherlands, 2006.
- [1114] T. Schlick. From macroscopic to mesoscopic models of chromatin folding. In J. Fish, editor, *Bridging The Scales in Science in Engineering*, pages 514–535. Oxford University Press, New York, NY, 2009.
- [1115] T. Schlick. Mathematical and biological scientists assess the state-of-the-art in RNA science at an IMA workshop “RNA in Biology, Bioengineering and Biotechnology”. *Intl. J. Mult. Sci. Eng.*, 8 (4), 2010; Also available as IMA report <http://www.ima.umn.edu/2007-2008/W10.29-11.2.07/activities/Schlick-Tamar/RNAReport.pdf>.

- [1116] T. Schlick. Molecular-dynamics based approaches for enhanced sampling of long-time, large-scale conformational changes in biomolecules. *F1000 Biol. Rep.*, 1:51, 2009.
- [1117] T. Schlick. Monte Carlo, harmonic approximation, and coarse-graining approaches for enhanced sampling of biomolecular structure. *F1000 Biol. Rep.*, 1:48, 2009.
- [1118] T. Schlick, E. Barth, and M. Mandziuk. Biomolecular dynamics at long timesteps: Bridging the timescale gap between simulation and experimentation. *Ann. Rev. Biophys. Biomol. Struct.*, 26:179–220, 1997.
- [1119] T. Schlick, D. Beard, J. Huang, D. Strahs, and X. Qian. Computational challenges in simulating large DNA over long times. *IEEE Comput. Sci. Eng.*, 2(6):38–51, November/December 2000. (Special Issue on Computational Chemistry).
- [1120] T. Schlick, S. Figueira, and M. Mezei. A molecular dynamics simulation of a water droplet by the implicit-Euler/Langevin scheme. *J. Chem. Phys.*, 94:2118–2129, 1991.
- [1121] T. Schlick and A. Fogelson. TNPACK — A truncated Newton minimization package for large-scale problems: I. Algorithm and usage. *ACM Trans. Math. Softw.*, 14:46–70, 1992.
- [1122] T. Schlick and A. Fogelson. TNPACK — A truncated Newton minimization package for large-scale problems: II. Implementation examples. *ACM Trans. Math. Softw.*, 14:71–111, 1992.
- [1123] T. Schlick and H. H. Gan. Methods for macromolecular modeling ( $M^3$ ): Assessment of progress and future perspectives. In T. Schlick and H. H. Gan, editors, *Computational Methods for Macromolecules: Challenges and Applications — Proceedings of the 3rd International Workshop on Algorithms for Macromolecular Modelling, New York, October 12–14, 2000*, volume 24 of *Lecture Notes in Computational Science and Engineering (Series Eds. M. Griebel, D.E. Keyes, R. M. Nieminen, D. Roose, and T. Schlick)*, pages 1–25, Berlin, 2002. Springer-Verlag.
- [1124] T. Schlick, B. Li, and M.-H. Hao. Calibration of the timestep for molecular dynamics of supercoiled DNA modeled by B-splines. In R. H. Sarma and M. H. Sarma, editors, *Structural Biology: The State of the Art, Proceedings of the Eighth Conversation in the Discipline of Biomolecular Stereodynamics, Volume I*, pages 157–174, Schenectady, NY, 1994. Adenine Press.
- [1125] T. Schlick, B. Li, and W. K. Olson. The influence of salt on DNA energetics and dynamics. *Biophys. J.*, 67:2146–2166, 1994.
- [1126] T. Schlick, M. Mandziuk, R.D. Skeel, and K. Srinivas. Nonlinear resonance artifacts in molecular dynamics simulations. *J. Comput. Phys.*, 139:1–29, 1998.
- [1127] T. Schlick and W. K. Olson. Supercoiled DNA energetics and dynamics by computer simulation. *J. Mol. Biol.*, 223:1089–1119, 1992.
- [1128] T. Schlick and W. K. Olson. Trefoil knotting revealed by molecular dynamics simulations of supercoiled DNA. *Science*, 257:1110–1115, 1992.
- [1129] T. Schlick, W. K. Olson, T. Westcott, and J. P. Greenberg. On higher buckling transitions in supercoiled DNA. *Biopolymers*, 34:565–598, 1994.
- [1130] T. Schlick and M. L. Overton. A powerful truncated Newton method for potential energy functions. *J. Comput. Chem.*, 8:1025–1039, 1987.

- [1131] T. Schlick and G. Parks. DOE computational sciences education project, 1994. Chapter on Mathematical Optimization. URL: [csep1.phy.ornl.gov/](http://csep1.phy.ornl.gov/).
- [1132] T. Schlick and C. S. Peskin. Can classical equations simulate quantum-mechanical behavior? A molecular dynamics investigation of a diatomic molecule with a Morse potential. *Comm. Pure App. Math.*, 42:1141–1163, 1989.
- [1133] T. Schlick, R. D. Skeel, A. T. Brünger, L. V. Kalé, J. A. Board, Jr., J. Hermans, and K. Schulten. Algorithmic challenges in computational molecular biophysics. *J. Comput. Phys.*, 151:9–48, May 1999. (Special Volume on Computational Biophysics).
- [1134] T. Schlick and L. Yang. Long-timestep biomolecular dynamics simulations: LN performance on a polymerase  $\beta$  / DNA system. In A. Brandt, J. Bernholc, and K. Binder, editors, *Multiscale Computational Methods in Chemistry and Physics*, volume 177 of *NATO Science Series. Series III: Computer and Systems Sciences*, pages 293–305, Amsterdam, The Netherlands, 2001. IOS Press.
- [1135] F. Schlüzen, A. Tocilj, R. Zarivach, J. Harms, M. Gluehmann, D. Janell, A. Bashan, H. Bartels, I. Agmon, F. Franceschi, and A. Yonath. Structure of functionally activated small ribosomal subunit at 3.3 Å resolution. *Cell*, 102:615–623, 2000.
- [1136] K. E. Schmidt and M. A. Lee. Implementing the fast multipole method in three dimensions. *J. Stat. Phys.*, 63:1223–1235, 1991.
- [1137] R. B. Schnabel and T. Chow. Tensor methods for unconstrained optimization. *SIAM J. Opt.*, 1:293–315, 1991.
- [1138] B. Schneider and H. M. Berman. Hydration of the DNA bases is local. *Biophys. J.*, 69:2661–2669, 1995.
- [1139] B. Schneider, D. Cohen, and H. M. Berman. Hydration of DNA bases: Analysis of crystallographic data. *Biopolymers*, 32:725–750, 1992.
- [1140] B. Schneider, D. M. Cohen, L. Schleifer, A. R. Srinivasan, W. K. Olson, and H. M. Berman. A systematic method for studying the spatial distribution of water molecules around nucleic acid bases. *Biophys. J.*, 65:2291–2303, 1993.
- [1141] B. Schneider, K. Patel, and H. M. Berman. Hydration of the phosphate group in double-helical DNA. *Biophys. J.*, 75:2422–2434, 1998.
- [1142] E. A. Schulze and D. B. Bartel. One sequence, two ribozymes: Implications for the emergence of new ribozyme folds. *Science*, 289:448–452, 2000.
- [1143] S. C. Schultz, G. C. Shields, and T. A. Steitz. Crystal structure of a CAP-DNA complex: The DNA is bent by 90°. *Science*, 253:1001–1007, 1991.
- [1144] J. M. Schurr, H. P. Babcock, and J. A. Gebe. Effect of anisotropy of the bending rigidity on the supercoiling free energy of small circular DNAs. *Biopolymers*, 36:633–641, 1995.
- [1145] P. Schuster, P. F. Stadler, and A. Renner. RNA structures and folding: From conventional to new issues in structure predictions. *Curr. Opin. Struct. Biol.*, 7:229–235, 1997.
- [1146] C. N. Schutz and A. Warshel. What are the dielectric “constants” of proteins and how to validate electrostatic models? *Proteins: Struc. Func. Gen.*, 44:400–417, 2001.

- [1147] T. Schwartz, M. A. Rould, K. Lowenhaupt, A. Herbert, and A. Rich. Crystal structure of the Z domain of the human editing enzyme ADAR1 bound to left-handed Z-DNA. *Science*, 284:1841–1845, 1999.
- [1148] E. Schwegler and M. Challacombe. Linear scaling computation of the Hartree-Fock exchange matrix. *J. Chem. Phys.*, 105:2726–2734, 1996.
- [1149] B. Sclavi, M. Sullivan, M. R. Chance, M. Brenowitz, and S. A. Woodson. RNA folding at millisecond intervals by synchrotron hydroxyl radical footprinting. *Science*, 279:1940–1943, 1998.
- [1150] K. A. Scott, P. J. Bond, A. Ivetac, A. P. Chetwynd, S. Khalid, and M. S. P. Sansom. Coarse-grained MD simulations of membrane protein-bilayer self-assembly. *Structure*, 16:621–630, 2008.
- [1151] M. R. Scott, R. Will, J. Ironside, H.-Oanh B. Nguyen, P. Tremblay, S. J. DeArmond, and S. B. Prusiner. Compelling transgenicetic evidence for transmission of bovine spongiform encephalopathy prions to humans. *Proc. Natl. Acad. Sci. USA*, 96:15137–15142, 1999.
- [1152] R. A. Scott and H. A. Scheraga. Method for calculating internal rotation barriers. *J. Chem. Phys.*, 42:2209–2215, 1965.
- [1153] R. A. Scott and H. A. Scheraga. Conformational analysis of macromolecules. II. the rotational isomeric states of the normal hydrocarbons. *J. Chem. Phys.*, 44:3054–3069, 1966.
- [1154] W. G. Scott, J. T. Finch, and A. Klug. The crystal structure of an all-RNA hammerhead ribozyme: A proposed mechanism for RNA catalytic cleavage. *Cell*, 81:991–1002, 1995.
- [1155] N. Seeman. DNA in a material world. *Nature*, 421:427–431, 2003.
- [1156] N. C. Seeman. DNA nanotechnology: Novel DNA constructions. *Annu. Rev. Biophys. Biomol. Struct.*, 27:225–248, 1998.
- [1157] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I. K. Moore, J. P. Wang, and J. Widom. A genomic code for nucleosome positioning. *Nature*, 442:772–778, 2006.
- [1158] G. L. Seibel, U. C. Singh, and P. A Kollman. A molecular dynamics simulation of double-helical B-DNA including counterions and water. *Proc. Natl. Acad. Sci. USA*, 82:6537–6540, 1985.
- [1159] E. Selsing, R. D. Wells, C. J. Alden, and S. Arnott. Bent DNA: Visualization of a base-paired and stacked A-B conformational junction. *J. Biol. Chem.*, 254:5417–5422, 1979.
- [1160] D. Sen and W. Gilbert. Cationic switches in the formation of DNA structures containing guanine-quartets. In R. H. Sarma and M. H. Sarma, editors, *Structure and Function: Proceedings of the Seventh Conversation in Biomolecular Stereodynamics*. Adenine Press, Schenectady, New York, 1992.
- [1161] S. Sen and L. Nilsson. Molecular dynamics of duplex systems involving PNA: Structural and dynamical consequences of nucleic acid backbone. *J. Amer. Chem. Soc.*, 120:619–631, 1998.
- [1162] H. Senderowitz and W. C. Still. MC(JBW): Simple but smart Monte Carlo algorithm for free energy simulations of multiconformational molecules. *J. Comput. Chem.*, 19:1736–1745, 1998.

- [1163] H. M. Senn and W. Thiel. QM/MM methods for biological systems. *Top. Curr. Chem.*, 268:173–290, 2007.
- [1164] H. M. Senn and W. Thiel. QM/MM studies of enzymes. *Curr. Opin. Chem. Biol.*, 11:182–187, 2007.
- [1165] H. M. Senn and W. Thiel. QM/MM methods for biomolecular systems. *Angew. Chem. Int. Ed.*, 48:1198–1229, 2009.
- [1166] A. Serganov, A. Polonskaia, A. T. Phan, R. R. Breaker, and D. J. Patel. Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. *Nature*, 441:1167–1171, 2006.
- [1167] E.I. Shakhnovich and A.M. Gutin. Implications of thermodynamics on protein folding for evolution of primary sequences. *Nature*, 346:773–775, 1990.
- [1168] D. F. Shanno and K. H. Phua. Remark on Algorithm 500: Minimization of unconstrained multivariate functions. *ACM Trans. Math. Softw.*, 6:618–622, 1980.
- [1169] D. E. Shaw, M. M. Deneroff, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, J. C. Chao, M. P. Eastwood, J. Gagliardo, J. Grossman, C. R. Ho, D. J. Ierardi, I. Kolossvry, J. L. Klepeis, T. Layman, C. McLeavy, M. A. Moraes, R. Mueller, E. C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, and S. C. Wang. Anton: A special-purpose machine for molecular dynamics simulation. In *Proceedings of the 34th annual international symposium on Computer architecture*, pages 1–12, San Diego, CA, 2007. ACM.
- [1170] J.-E. Shea and C. L. Brooks, III. From folding theories to folding proteins: A review and assessment of simulation studies of protein folding and unfolding. *Annu. Rev. Phys. Chem.*, 52:499–535, 2001.
- [1171] E. C. Sherer, S. A. Harris, R. Soliva, M. Orozco, and C. A. Laughton. Molecular dynamics studies of DNA A-tract structure and flexibility. *J. Amer. Chem. Soc.*, 121:5981–5991, 1999.
- [1172] C. D. Sherrill, B. G. Sumpter, M. O. Sinnokrot, M. S. Marshall, E. G. Hohenstein, R. C. Walker, and I. R. Gould. Assessment of standard force field models against high-quality *ab initio* potential curves for prototypes of  $\pi-\pi$ , CH/ $\pi$ , and SH/ $\pi$  interactions. *J. Comput. Chem.*, 30:2187–2193, 2009. doi:10.1002/jcc.21226.
- [1173] H. Shi and P. Moore. The crystal structure of yeast phenylalanine tRNA at 1.93 Å resolution: A classic structure revisited. *RNA*, 6:1091–1105, 2000.
- [1174] M. M. Shi, D. Mehren, and K. Dacus. Pharmacogenomics: Changing the health care paradigm. *Mod. Drug Disc.*, 4:27–32, 2001.
- [1175] Y. Shi, A. E. Borovik, and J. E. Hearst. Elastic rod model incorporating shear and extension, generalized nonlinear schrödinger equations, and novel closed-form solutions for supercoiled DNA. *J. Chem. Phys.*, 103:3166–3183, 1995.
- [1176] J. Shimada, H. Kaneko, and T. Takada. Performance of fast multipole methods for calculating electrostatic interactions in biomolecular simulations. *J. Comput. Chem.*, 15:28–43, 1994.
- [1177] M. Shirts and V. Pande. Screen savers of the world unite! *Science*, 290:1903–1904, 2000.
- [1178] I. A. Shkel, O. V. Tsodikov, and M. T. Record, Jr. Complete asymptotic solution of cylindrical and spherical Poisson-Boltzmann equations at experimental salt concentrations. *J. Phys. Chem. B*, 104:5161–5170, 2000.

- [1179] B. K. Shoichet. Virtual screening of chemical libraries. *Nature*, 432:862–865, 2004.
- [1180] M. Shtilerman, G. H. Lorimer, and S. W. Englander. Chaperonin function: Folding by forced unfolding. *Science*, 284:822–825, 1999.
- [1181] W. Shu, X. Bo, Z. Zheng, and S. Wang. A novel representation of RNA secondary structure based on element-contact graphs. *BMC Bioinformatics*, 9:188, 2008.
- [1182] X. Shui, L. McFail-Isom, G. G. Hu, and L. D. Williams. The B-DNA dodecamer at high resolution reveals a spine of water on sodium. *Biochemistry*, 37:8341–8355, 1998.
- [1183] A. Siddiqui-Jain, C. L. Grand, D. J. Bearss, and L. H. Hurley. Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc. Natl. Acad. Sci. USA*, 99:11593–11598, 2002.
- [1184] G. Siegal, J. van Duynhoven, and M. Baldus. Biomolecular NMR: Recent advances in liquids, solids and screening. *Curr. Opin. Chem. Biol.*, 3:530–536, 1999.
- [1185] P. E. M. Siegbahn, M. R. A. Blomberg, and M. L. Blomberg. Theoretical study of the energetics of proton pumping and oxygen reduction in cytochrome oxidase. *J. Phys. Chem. B*, 107:10946–10955, 2003.
- [1186] J. C. Simo and N. Tarnow. The discrete energy-momentum method. Conserving algorithms for nonlinear elastodynamics. *Z. Angew. Math. Phys.*, 43:757–793, 1992.
- [1187] J. C. Simo, N. Tarnow, and K. K. Wang. Exact energy-momentum conserving algorithms and symplectic schemes for nonlinear dynamics. *Comput. Meth. App. Mech. Engin.*, 100:63–116, 1994.
- [1188] B. Simon and M. Sattler. De novo structure determination from residual dipolar couplings by NMR spectroscopy. *Angew. Chem. Int. Ed.*, 41:437–440, 2002.
- [1189] T. Simonson. Accurate calculation of the dielectric constant of water from simulations of a microscopic droplet in vacuum. *Chem. Phys. Lett.*, 250:450–454, 1996.
- [1190] T. Simonson. Macromolecular electrostatics: Continuum models and their growing pains. *Curr. Opin. Struct. Biol.*, 11:243–252, 2001.
- [1191] R. R. Sinden. *DNA Structure and Function*. Academic Press, San Diego, CA, 1994.
- [1192] D. Sindhikara, Y. Meng, and A. E. Roitberg. Exchange frequency in replica exchange molecular dynamics. *J. Chem. Phys.*, 128:024103, 2008.
- [1193] S. Singh, B. K. Malik, and D. K. Sharma. Molecular drug targets and structure based drug design: A holistic approach. *Bioinformation*, 1:314–320, 2006.
- [1194] U. C. Singh and P. A. Kollman. A combined Ab Initio quantum mechanical and molecular mechanical method for carrying out simulations on complex molecular systems: Applications to the CH<sub>3</sub>Cl + Cl<sup>-</sup> exchange reaction and gas phase protonation of polyethers. *J. Comput. Chem.*, 7:718–730, 1986.
- [1195] U.C. Singh and P.A. Kollman. An approach to computing electrostatic charges for molecules. *J. Comput. Chem.*, 5:129–145, 1984.
- [1196] Frontiers in chemistry: Single molecules, 1999. *Science* **283**: 1667–1695 (1999), special compendium of articles.
- [1197] R. D. Skeel. What makes molecular dynamics work? *SIAM J. Sci. Comput.*, 31:1363–1378, 2009.

- [1198] R. D. Skeel, I. Tezcan, and D. J. Hardy. Multiple grid methods for classical molecular dynamics. *J. Comput. Chem.*, 23:673–684, 2002.
- [1199] R. D. Skeel, G. Zhang, and T. Schlick. A family of symplectic integrators: Stability, accuracy, and molecular dynamics applications. *SIAM J. Sci. Comput.*, 18(1):202–222, January 1997.
- [1200] Robert D. Skeel. Integration schemes for molecular dynamics and related applications. In M. Ainsworth, J. Levesley, and M. Marletta, editors, *The Graduate Student's Guide to Numerical Analysis*, volume 26 of *Springer Series in Computational Mathematics*, pages 119–176. Springer-Verlag, New York, NY, 1999.
- [1201] N. B. Slater. Classical motion under a morse potential. *Nature*, 180:1352–1353, 1957.
- [1202] R. H. Smith. Nanotechnology gains momentum. *Mod. Drug Dis.*, 4:33–38, 2001.
- [1203] S. B. Smith, Y. Cui, and C. Bustamante. Overstretching B-DNA: The elastic response of individual double-stranded and single-stranded DNA molecules. *Science*, 271:795–798, 1996.
- [1204] M. Snir. A note on N-body computations with cutoffs. *Theory Comput. Sys.*, 37:295–318, 2004.
- [1205] C. D. Snow, H. Nguyen, V. S. Pande, and M. Gruebele. Absolute comparison of simulated and experimental protein folding dynamics. *Nature*, 420:102–106, 2002.
- [1206] T. A. Soares, P. H. Hünenberger, M. A. Kastenholz, V. Kräutler, T. Lenz, R. D. Lins, C. Oostenbrink, and W. F. van Gunsteren. An improved nucleic acid parameter set for the GROMOS force field. *J. Comput. Chem.*, 26:725–737, 2005.
- [1207] G. A. Soukup and R. R. Breaker. Engineering precision RNA molecular switches. *Proc. Natl. Acad. Sci. USA*, 96:3584–3589, 1999.
- [1208] G. A. Soukup and R. R. Breaker. Allosteric nucleic acid catalysts. *Curr. Opin. Struct. Biol.*, 10:318–325, 2000.
- [1209] B. Space, H. Rabitz, and A. Askar. Long time scale molecular dynamics subspace integration method applied to anharmonic crystals and glasses. *J. Chem. Phys.*, 99:9070–9079, 1993.
- [1210] J. T. Sprague, J. C. Tai, Y. Yuh, and N. L. Allinger. The MMP2 calculational method. *J. Comput. Chem.*, 8:581–603, 1987.
- [1211] D. Sprous and S. C. Harvey. Action at a distance in supercoiled DNA: Effects of sequences on slither, branching and intermolecular concentration. *Biophys. J.*, 70:1893–1908, 1996.
- [1212] D. Sprous, M. A. Young, and D. L. Beveridge. Molecular Dynamics Studies of Axis Bending in d(G<sub>5</sub>-(GA<sub>4</sub>T<sub>4</sub>C)<sub>2</sub>-C<sub>5</sub>) and d(G<sub>5</sub>-(GT<sub>4</sub>A<sub>4</sub>C)<sub>2</sub>-C<sub>5</sub>): Effects of Sequence Polarity on DNA Curvature. *J. Mol. Biol.*, 285:1623–1632, 1999.
- [1213] D. Sprous, W. Zacharias, Z. A. Wood, and S. C. Harvey. Dehydrating agents sharply reduce curvature in DNAs containing A-tracts. *Nucl. Acids Res.*, 23: 1816–1821, 1995.
- [1214] D. Sprous, II, R. K.-Z. Tan, and S. C. Harvey. Molecular modeling of closed circular DNA thermodynamic ensembles. *Biopolymers*, 39:248–258, 1996.
- [1215] A. Srinivasan and W. K. Olson. Polynucleotide conformation in real solution — a preliminary theoretical estimate. *Fed. Amer. Soc. Exp. Bio.*, 39:2199, 1980.

- [1216] A. R. Srinivasan and W. K. Olson. Molecular models of nucleic acid triple helices. II. PNA and 2'-5' backbone complexes. *J. Amer. Chem. Soc.*, 120:492–499, 1998.
- [1217] G. Srinivasan, C. M. James, and J. A. Krzycki. Pyrrolysine encoded by UAG in Archaea: Charging of a UAG-decoding specialized tRNA. *Science*, 296:1459–1462, 2002.
- [1218] D. K. Stammers, D. O’N. Somers, C. K. Ross, I. Kirby, P. H. Ray, J. E. Wilson, M. Norman, J. S. Ren, R. M. Esnouf, E. F. Garman, E. Y. Jones, and D. I. Stuart. Crystals of HIV-1 reverse transcriptase diffracting to 2.2 Å resolution. *J. Mol. Biol.*, 242:586–588, 1994.
- [1219] R. Stehr, N. Kepper, K. Rippe, and G. Wedemann. The effect of the internucleosomal interaction potential on the folding of the chromatin fiber. *Biophys. J.*, 95:3677–3691, 2008.
- [1220] P. J. Steinbach and B. R. Brooks. New spherical-cutoff methods for long-range forces in macromolecular simulation. *J. Comput. Chem.*, 15:667–683, 1994.
- [1221] R. Steinbrook. The AIDS epidemic - A progress report from Mexico City. *N. Engl. J. Med.*, 359:885–887, 2008.
- [1222] S. D. Stellman, B. Hingerty, S. B. Broyde, E. Subramanian, T. Sato, and R. Langridge. Structure of guanosine-3', 5'-cytidine monophosphate. I. Semi-empirical potential energy calculations and model-building. *Biopolymers*, 12:1731–2750, 1973.
- [1223] D. Stigter. Interactions of highly charged colloidal cylinders with applications to double-stranded DNA. *Biopolymers*, 16:1435–1448, 1977.
- [1224] C. U. Stirnimann and M. G. Güttler. New frontiers in X-ray crystallography. In T. Schwede and M. Peitsch, editors, *Computational Structural Biology. Methods and Applications*, pages 601–622. World Scientific, Singapore, 2008.
- [1225] E. Stofer, C. Chipot, and R. Lavery. Free energy calculations of Watson-Crick base pairing in aqueous solution. *J. Amer. Chem. Soc.*, 121:9503–9508, 1999.
- [1226] A. J. Stone. *The Theory of Intermolecular Forces*. Oxford University Press, Oxford, England, 1996.
- [1227] C. Störmer. Sur les trajectoires des corpuscules électrisés dans l'espace. *Archives des Sciences Physiques et Naturelles*, 24:5–18, 113–158, 221–247, 1907. (This reference is the first in a three-part essay. The second part appeared in the same journal in 1911 [pages 190, 277, 415, and 501], and the third part appeared in the 1912 volume of the journal, pages 51–69).
- [1228] G. Storz. An expanding universe of noncoding RNAs. *Science*, 296:1260–1263, 2002.
- [1229] D. Strahs, X. Qian, D. Barash, and T. Schlick. Sequence-dependent solution structure of 13 TATA/TBP complexes. *Biopolymers*, 69:216–243, 2003.
- [1230] D. Strahs and T. Schlick. A-tract bending: Insights into experimental structures by computational models. *J. Mol. Biol.*, 301:643–666, 2000.
- [1231] W. B. Streett, D. J. Tildesley, and G. Saville. Multiple time step methods and an improved potential function for molecular dynamics simulations of molecular liquids. In Peter Lykos, editor, *Computer Modeling of Matter*, volume 86 of *ACS Symposium Series*, pages 144–158. ACS, Washington, D. C., 1978.
- [1232] W. B. Streett, D. J. Tildesley, and G. Saville. Multiple time step methods in molecular dynamics. *Mol. Phys.*, 35:639–648, 1978.

- [1233] T. Strick, J.-F. Allemand, V. Croquette, and D. Bensimon. The manipulation of single biomolecules. *Physics Today*, 54:46–51, October 2001.
- [1234] R. S. Struthers, J. Rivier, and A. T. Hagler. Theoretical simulation of conformation, energetics, and dynamics in the design of GnRH analogs. *Trans. Amer. Cryst. Assoc.*, 20:83–96, 1984. Proceedings of the Symposium on Molecules in Motion, University of Kentucky, Lexington, Kentucky, May 20–21, 1984.
- [1235] L. Stryer. *Biochemistry*. W. H. Freeman, New York, NY, 5 edition, 2001.
- [1236] S. J. Stuart, R. Zhou, and B. J. Berne. Molecular dynamics with multiple time scales: The selection of efficient reference system propagators. *J. Chem. Phys.*, 105:1426–1436, 1996.
- [1237] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics methods for protein folding. *Chem. Phys. Lett.*, 314:141–151, 1999.
- [1238] J. Sulston and G. Ferry. *The Common Thread: A Story of Science, Politics, Ethics and the Human Genome*. Joseph Henry Press, Washington D. C., 2002.
- [1239] D. W. Sumners. Lifting the curtain: Using topology to probe the hidden action of enzymes. *Notices Amer. Math. Soc.*, 42:528–537, 1995.
- [1240] J. Sun, Q. Zhang, and T. Schlick. Electrostatic mechanism of nucleosomal array folding revealed by computer simulation. *Proc. Natl. Acad. Sci. USA*, 102:8180–8185, 2005.
- [1241] Y. Sun, T. J. A. Ewing, A. G. Skillman, and I. D. Kuntz. CombiDOCK: Structure-based combinatorial docking and library design. *J. Comput.-Aided Mol. Design*, 12:597–604, 1998.
- [1242] C. R. Sweet, P. Petrine, V. S. Pande, and J. A. Izquierre. Normal mode partitioning of Langevin dynamics for biomolecules. *J. Chem. Phys.*, 128:145101, 2008.
- [1243] R. D. Swindoll and J. M. Haile. A multiple time-step method for molecular dynamics simulations of fluids of chain molecules. *J. Chem. Phys.*, 53:289–298, 1984.
- [1244] W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Applications to small water clusters. *J. Chem. Phys.*, 76:637–649, 1982.
- [1245] J. W. Szostak. RNA gets a grip on translation. *Nature*, 419:890–891, 2002.
- [1246] J. C. Tai and N. L. Allinger. Effect of inclusion of electron correlation in MM3 studies of cyclic conjugated compounds. *J. Comput. Chem.*, 19:475–487, 1998.
- [1247] E. Tajkhorshid, P. Nollert, M. Ø Jensen, L. J. W. Miercke, J. O’Connell, R. M. Stroud, and K. Schulten. Control of the selectivity of the aquaporin water channel family by global orientational tuning. *Science*, 296:525–530, 2002.
- [1248] F. Tama, M. Valle, J. Frank, and C. L. Brooks, III. Dynamic reorganization of the functionally active ribosome explored by normal mode analysis and cryo-electron microscopy. *Proc. Natl. Acad. Sci. USA*, 100:9319–9323, 2003.
- [1249] R. K.-Z. Tan and S. C. Harvey. Molecular mechanics model of supercoiled DNA. *J. Mol. Biol.*, 205:573–591, 1989.
- [1250] R. K.-Z. Tan, D. Sprous, and S. C. Harvey. Molecular dynamics simulations of small DNA plasmids: Effects of sequence and supercoiling on intramolecular motions. *Biopolymers*, 39:259, 1996.

- [1251] J. Tang and R. R. Breaker. Structural diversity of self-cleaving ribozymes. *Proc. Natl. Acad. Sci. USA*, 97:5784–5789, 2001.
- [1252] Y. Tao and W. Zhang. Recent developments in cryo-electron microscopy reconstruction of single particles. *Structure*, 10:616–622, 2000.
- [1253] S. Tara, A. H. Elcock, P. D. Kirchhoff, J. M. Briggs, Z. Radic, P. Taylor, and J. A. McCammon. Rapid binding of a cationic active site inhibitor to wild type and mutant mouse acetylcholinesterase: Brownian dynamics simulation including diffusion in the active site gorge. *Biopolymers*, 46:465–474, 1998.
- [1254] W. H. Taylor and K. Lin. A tangled problem. *Nature*, 421:25, 2002.
- [1255] V. Tereshko, G. Minasov, and M. Egli. A “Hydra-Ion” spine in B-DNA minor groove. *J. Amer. Chem. Soc.*, 121:3590–3595, 1999.
- [1256] M. C. Tesi, E. J. Janse van Rensburg, E. Orlandini, D. W. Sumners, and S. G. Whittington. Knotting and supercoiling in circular DNA: A model incorporating the effect of added salt. *Phys. Rev. E*, 49:868–872, 1994.
- [1257] D. Thirumalai and G. H. Lorimer. Chaperonin-mediated protein folding. *Ann. Rev. Biophys. Biomol. Struct.*, 30:245–269, 2001.
- [1258] J. T. Thomas. The scientific and humane legacy of Max Perutz (1914 – 2002). *Angew. Chem. Int. Ed.*, 41:3155–3166, 2002.
- [1259] J. M. Thornton, C. A. Orengo, A. E. Todd, and F. M. G. Pearl. Protein folds, functions and evolution. *J. Mol. Biol.*, 293:333–342, 1999.
- [1260] B. Tidor, K. K. Irikura, B. R. Brooks, and M. Karplus. Dynamics of DNA oligomers. *J. Biomol. Struct. Dynam.*, 1:231–252, 1983.
- [1261] I. Tinoco, Jr. and C. Bustamante. How RNA folds. *J. Mol. Biol.*, 293:271–281, 1999.
- [1262] M. M. Tirion. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phy. Rev. Lett.*, 77:1905–1908, 1996.
- [1263] D. J. Tobias. Electrostatic calculations: Recent methodological advances and applications to membranes. *Curr. Opin. Struct. Biol.*, 11:253–261, 2001.
- [1264] D. J. Tobias and C. L. Brooks, III. Conformational equilibrium in the alanine dipeptide in the gas phase and aqueous solution: A comparison of theoretical results. *J. Chem. Phys.*, 89:5115–5126, 1988.
- [1265] J. R. Tolman. Dipolar couplings as a probe of molecular dynamics and structure in solution. *Curr. Opin. Struct. Biol.*, 11:532–539, 2001.
- [1266] N. Toor, K.S. Keating, S. D. Taylor, and A. M. Pyle. Crystal structure of a self-spliced group II intron. *Science*, 320:77–82, 2008.
- [1267] A. Y. Toukmaji and J. A. Board, Jr. Ewald summation techniques in perspective: A survey. *Comput. Phys. Commun.*, 95:73–92, 1996.
- [1268] J. J. Toulme, C. Di Primo, and D. Boucard. Regulating eukaryotic gene expression with aptamers. *FEBS Lett.*, 567:55–62, 2004.
- [1269] S. Toxvaerd. Comment on constrained molecular dynamics of macromolecules. *J. Chem. Phys.*, 87:6140–6143, 1987.
- [1270] T. L. Trapane and E. E. Lattman. Seventh meeting on the critical assessment of techniques for protein structure prediction. *Proteins: Struc. Func. Gen.*, 69 (Suppl. 8):1–2, 2007.

- [1271] D. J. Tremethick. Higher-Order structures of chromatin: The elusive 30 nm fiber. *Cell*, 128:651–654, 2007.
- [1272] E. N. Trifonov, R. K.-Z. Tan, and S. C. Harvey. Static persistence length of DNA. In W. K. Olson, M. H. Sarma, R. H. Sarma, and M. Sundaralingam, editors, *Structure and Expression: DNA Bending and Curvature*, volume 3. Adenine Press, Schenectady, New York, 1987.
- [1273] V. Tsui and D. A. Case. Theory and applications of the Generalized Born solvation model in macromolecular simulations. *Biopolymers*, 56:275–291, 2001.
- [1274] H. Tsuru and M. Wadati. Elastic model of highly supercoiled DNA. *Biopolymers*, 25:2083–2096, 1986.
- [1275] B. J. Tucker and R. R. Breaker. Riboswitches as versatile gene control elements. *Curr. Opin. Struct. Biol.*, 15:342–348, 2005.
- [1276] M. E. Tuckerman and B. J. Berne. Molecular dynamics in systems with multiple time scales: Systems with stiff and soft degrees of freedom and with short and long range forces. *J. Comput. Chem.*, 95:8362–8364, 1992.
- [1277] M. E. Tuckerman, B. J. Berne, and G. J. Martyna. Reversible multiple time scale molecular dynamics. *J. Chem. Phys.*, 97:1990–2001, 1992.
- [1278] M. E. Tuckerman, B. J. Berne, and A. Rossi. Molecular dynamics algorithm for multiple time scales: Systems with disparate masses. *J. Chem. Phys.*, 94:1465–1469, 1991.
- [1279] M. E. Tuckerman, B.J. Berne, and G. J. Martyna. Molecular dynamics algorithm for multiple time scales: Systems with long range forces. *J. Chem. Phys.*, 94:6811–6815, 1991.
- [1280] M. E. Tuckerman, G. J. Martyna, and B. J. Berne. Molecular dynamics algorithm for condensed systems with multiple time scales. *J. Chem. Phys.*, 93:1287–1291, 1990.
- [1281] C. Tuerk and L. Gold. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249:505–570, 1990.
- [1282] T. Tuschl, C. Gohlke, T. M. Jovin, E. Westhof, and F. Eckstein. A three-dimensional model for the hammerhead ribozyme based on fluorescence measurements. *Science*, 266:785–789, 1994.
- [1283] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403:623–627, 2000.
- [1284] L. Ulanovsky and E. N. Trifonov. Estimation of wedge components in curved DNA. *Nature*, 326:720–722, 1987.
- [1285] N. B. Ulyanov and V. B. Zhurkin. Sequence-dependent anisotropic flexibility of B-DNA: A conformational study. *J. Biomol. Struct. Dynam.*, 2:361–385, 1984.
- [1286] V. M. Unger. Electron cryomicroscopy. *Curr. Opin. Struct. Biol.*, 11:548–554, 2001.
- [1287] I. Usón and G. M. Sheldrick. Advances in direct methods for protein crystallography. *Curr. Opin. Struct. Biol.*, 9:643–648, 1999.

- [1288] A. Valouev, J. Ichikawa, T. Tonthat, J. Stuart, S. Ranade, H. Peckham, K. Zeng, J. A. Malek, G. Costa, K. McKernan, A. Sidow, A. Fire, and S. M. Johnson. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Gen. Res.*, 18, 2008.
- [1289] D. M. F. van Aalten, B. L. deGroot, J. B. C. Findlay, H. J. C. Berendsen, and A. Amadei. A comparison of techniques for calculating protein essential dynamics. *J. Comput. Chem.*, 18:169–181, 1997.
- [1290] M. J. van Dongen, J. F. Doreleijers, G. A. van der Marel, J. H. van Boom, C. W. Hilbers, and S. S. Wijmenga. Structure and mechanism of formation of the H-y5 isomer of an intramolecular DNA triple helix. *Nat. Struc. Biol.*, 6:854–859, 1999.
- [1291] W. F. van Gunsteren. Constrained dynamics of flexible molecules. *Mol. Phys.*, 40:1015–1019, 1980.
- [1292] W. F. van Gunsteren and H. J. C. Berendsen. Algorithms for macromolecular dynamics and constraint dynamics. *Mol. Phys.*, 34:1311–1327, 1977.
- [1293] W. F. van Gunsteren and H. J. C. Berendsen. Algorithms for Brownian dynamics. *Mol. Phys.*, 45:637–647, 1982.
- [1294] W. F. van Gunsteren, R. Bürgi, C. Peter, and X. Daura. The key to solving the protein-folding problem lies in an accurate description of the denatured state. *Angew. Chem. Int. Ed.*, 40:352–355, 2001.
- [1295] W. F. van Gunsteren, R. Bürgi, C. Peter, and X. Daura. Reply. *Angew. Chem. Int. Ed.*, 40:4616–4618, 2001.
- [1296] W. F. van Gunsteren and M. Karplus. Effect of constraints on the dynamics of macromolecules. *Macromolecules*, 15:1528–1543, 1982.
- [1297] K. van Holde and J. Zlatanova. Chromatin fiber structure, where is the problem now? *Sem. Cell Dev. Bio.*, 18:651–658, 2007.
- [1298] E. Vanden-Eijnden, M. Venturoli, G. Ciccotti, and R. Elber. On the assumptions underlying milestoneing. *J. Chem. Phys.*, 129:174102, 2008.
- [1299] J. VandeVondele and U. Rothlisberger. Canonical adiabatic free energy sampling (CAFES): A novel method for the exploration of free energy surface. *J. Phys. Chem. B*, 106:203–208, 2002.
- [1300] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and Jr. A. D. MacKerell. CHARMM general force field: a force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.*, 31:671–690, 2010.
- [1301] D. F. Veber, F. H. Drake, and M. Gowen. The new partnership of genomics and chemistry for accelerated drug development. *Curr. Opin. Chem. Biol.*, 1:151–156, 1997.
- [1302] L. Verlet. Computer ‘experiments’ on classical fluids: I. Thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.*, 159(1):98–103, July 1967.
- [1303] M. A. Viswamitra, B. S. Reddy, G.H.-Y. Lin, and M. Sundaralingam. Stereochemistry of nucleic acids and their constituents. XVII. Crystal and molecular structure of deoxycytidine 5'-phosphate monohydrate. A possible puckering for the furanoside ring in B-deoxyribonucleic acid. *J. Amer. Chem. Soc.*, 93:4565–4573, 1971.

- [1304] C. T. Vogelson. Advances in drug delivery systems. *Mod. Drug Dis.*, 4:49–52, 2001.
- [1305] A. V. Vologodskii. *Topology and physics of circular DNA*. CRC Press, Boca Raton, Florida, 1992.
- [1306] A. V. Vologodskii, V. V. Anshelevich, A. V. Lukashin, and M. D. Frank-Kamenetskii. Statistical mechanics of supercoils and the torsional stiffness of the DNA double helix. *Nature*, 280:294–298, 1979.
- [1307] A. V. Vologodskii and N. R. Cozzarelli. Conformational and thermodynamic properties of supercoiled DNA. *Ann. Rev. Biophys. Biomol. Struc.*, 23:609–643, 1994.
- [1308] A. V. Vologodskii and N. R. Cozzarelli. Supercoiling, knotting, looping, and other large-scale conformational properties of DNA. *Curr. Opin. Struct. Biol.*, 4:372–375, 1994.
- [1309] A. V. Vologodskii and N. R. Cozzarelli. Modeling of long-range electrostatic interactions in DNA. *Biopolymers*, 35:289–296, 1995.
- [1310] A. V. Vologodskii and N. R. Cozzarelli. Effect of supercoiling on the juxtaposition and relative orientation of DNA sites. *Biophys. J.*, 70:2548–2556, 1996.
- [1311] A. V. Vologodskii, S. D. Levene, K. V. Klenin, M. D. Frank-Kamenetskii, and N. R. Cozzarelli. Conformational and thermodynamic properties of supercoiled DNA. *J. Mol. Biol.*, 227:1224–1243, 1992.
- [1312] K. Voltz, J. Trylska, V. Tozzini, V. Kurkal-Siebert, J. Langowski, and J. C. Smith. Coarse-grained force field for the nucleosome from self-consistent multiscaling. *J. Comput. Chem.*, 29:1429–1439, 2008.
- [1313] P. H. von Hippel. From “simple” DNA-protein interactions to the macromolecular machines of gene expression. *Ann. Rev. Biophys. Biomol. Struc.*, 36:79–105, 2007.
- [1314] Y. N. Vorobjev and J. Hermans. ES/IS: Estimation of conformational free energy by combining dynamics simulations with explicit solvent with an implicit solvent continuum model. *Biophys. Chem.*, 78:195–205, 1999.
- [1315] A. F. Voter. A method for accelerating the molecular dynamics simulation of infrequent events. *J. Chem. Phys.*, 106, 1997.
- [1316] J. Vrebalov, D. Ruezinsky, V. Padmanabhan, R. White, D. Medrano, R. Drake, W. Schuch, and J. Giovannoni. A MADS-box gene necessary for fruit ripening at the tomato *ripening-inhibitor (Rin)* locus. *Science*, 296:343–346, 2002.
- [1317] J. Šponer, J. Leszczyński, and P. Hobza. Nature of nucleic acid-base stacking: Nonempirical ab Initio and empirical potential characterization of 10 stacked base dimers. Comparison of stacked and H-bonded base pairs. *J. Phys. Chem.*, 100:5590–5596, 1996.
- [1318] J. Šponer, J. Leszczyński, and P. Hobza. Structures and energies of hydrogen-bonded DNA base pairs: A nonempirical study with inclusion of electron correlation. *J. Phys. Chem.*, 100:1965–1974, 1996.
- [1319] J. Šponer and N. Špačková. Molecular dynamics simulations and their application to four-stranded DNA. *Methods*, 43:278–290, 2007.
- [1320] R. Štefl, T. E. Cheatham, III, N. Špačková, E. Fadrná, I. Berger, J. Koča, and J. Šponer. Formation pathways of guanine-quadruplex DNA revealed by molecular dynamics and thermodynamic analysis of substates. *Biophys. J.*, 85:1787–1804, 2003.

- [1321] R. C. Wade, M. E. Davis, B. A. Luty, J. D. Madura, and J. A. McCammon. Gating of the active site of triose phosphate isomerase: Brownian dynamics simulations of flexible peptide loops in the enzyme. *Biophys. J.*, 64:9–15, 1993.
- [1322] R. C. Wade, B.A. Luty, E. Demchuk, J. D. Madura, M. E. Davis, J. M. Briggs, and J. A. McCammon. Simulation of enzyme-substrate encounter with gated active sites. *Struct. Biol.*, 1:65–69, 1994.
- [1323] M. Wadman. James watson's genome sequenced at high speed. *Nature*, 452:788, 2008.
- [1324] S. Wallin and H. S. Chan. Conformational entropic barriers in topology-dependent protein folding: perspectives from a simple native-centric polymer model. *J. Phys. Condens. Matter*, 18:S307–S328, 2006.
- [1325] D. M. Walsh, I. Klyubin, J. V. Fadeeva, W. K. Cullen, R. Anwyl, M. S. Wolfe, M. J. Rowan, and D. J. Selkoe. Naturally secreted oligomers of amyloid  $\beta$  protein potently inhibit hippocampal long-term potentiation *in vivo*. *Nature*, 416:535–539, 2002.
- [1326] S. Walter and J. Buchner. Molecular chaperones—Cellular machines for protein folding. *Angew. Chem. Int. Ed.*, 41:1098–1113, 2002.
- [1327] A. H. Wang, G. J. Quigley, F. J. Kolpak, J. L. Crawford, J. H. van Boom, G. van der Marel, and A. Rich. Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature*, 282:680–686, 1979.
- [1328] F. Wang and D. P. Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.*, 86:2050–2053, 2001.
- [1329] H. Y. Wang and R. LeSar. An efficient fast-multipole algorithm based on an expansion in the solid harmonics. *J. Chem. Phys.*, 104:4173–4179, 1996.
- [1330] J. Wang, P. Cieplak, and P. A. Kollman. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.*, 21:1049–1074, 2000.
- [1331] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development and testing of a general Amber force field. *J. Comput. Chem.*, 25:1157–1174, 2004.
- [1332] J.-C. Wang, S. Pal, and K. A. Fichthorn. Accelerated molecular dynamics of rare events using the local boost method. *Phys. Rev. B*, 63:085403, 2001.
- [1333] L. Wang, S. Broyde, and Y. Zhang. Polymerase-tailored variations in the water-mediated and substrate-assisted mechanism for nucleotidyl transfer: insights from a study of T7 DNA polymerase. *J. Mol. Biol.*, 389:787–796, 2009.
- [1334] L. Wang, X. Yu, P. Hu, S. Broyde, and Y. Zhang. A water-mediated and substrate-assisted catalytic mechanism for *sulfolobus solfataricus* DNA polymerase IV. *J. Amer. Chem. Soc.*, 129:4731–4747, 2007.
- [1335] W. Wang, O. Donini, C. M. Reyes, and P. A. Kollman. Biomolecular simulations: Recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. *Ann. Rev. Biophys. Biomol. Struc.*, 30:211–243, 2001.
- [1336] Y. Wang, K. Arora, and T. Schlick. Subtle but variable conformational rearrangements in the replication cycle of *sulfolobus solfataricus* P2 DNA polymerase IV may accommodate lesion bypass. *Protein Sci.*, 15:135–151, 2006.

- [1337] Y. Wang and D. J. Patel. Solution structure of the human telomeric repeat d[AG<sub>3</sub>(T<sub>2</sub>AG<sub>3</sub>)<sub>3</sub>] G-tetraplex. *Structure*, 1:263–282, 1993.
- [1338] Y. Wang and T. Schlick. Quantum mechanics/molecular mechanics investigation of the chemical reaction in Dpo4 reveals water-dependent pathways and requirements for active site reorganization. *J. Amer. Chem. Soc.*, 130:13240–13250, 2008.
- [1339] Z. Wang and R. M. Harshey. Crucial role for DNA supercoiling in Mu transposition: A kinetic study. *Proc. Natl. Acad. Sci. USA*, 91:699–703, 1994.
- [1340] Z.-X. Wang. How many fold types of protein are there in Nature? *Proteins: Struct. Func. Gen.*, 26:186–191, 1996.
- [1341] A. Warshel. The consistent force field and its quantum mechanical extension. In G. A. Segal, editor, *Modern Theoretical Chemistry*, volume 7. Plenum Press, New York, NY, 1977.
- [1342] A. Warshel. *Computer Modeling of Chemical Reactions in Enzymes and Solutions*. John Wiley & Sons, New York, NY, 1991.
- [1343] A. Warshel, M. Kato, and A. V. Pisliakov. Polarizable force fields: History, test cases, and prospects. *J. Chem. Theor. Comput.*, 3:2034–2045, 2007.
- [1344] A. Warshel and M. Levitt. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of carbonium ion in the reaction of lysozyme. *J. Mol. Biol.*, 103:227–249, 1976.
- [1345] A. Warshel, M. Levitt, and S. Lifson. Consistent force field for calculations of vibrational spectra, and conformations of some amides and lactam rings. *J. Mol. Spect.*, 33:84–89, 1970.
- [1346] A. Warshel and S. T. Russell. Calculations of electrostatic interactions in biological systems and in solutions. *Q. Rev. Biophys.*, 17:283–422, 1984.
- [1347] A. Warshel, P. K. Sharma, M. Kato, and W. W. Parson. Modeling electrostatic effects in proteins. *Biochim. Biophys. Acta*, 1764:1647–1676, 2006.
- [1348] A. Warshel, P. K. Sharma, M. Kato, Y. Xiang, H. Liu, and M. H. M. Olson. Electrostatic basis for enzyme catalysis. *Chem. Rev.*, 106:3210–3235, 2006.
- [1349] M. Watanabe and M. Karplus. Simulations of macromolecules by multiple time-step methods. *J. Phys. Chem.*, 99:5680–5697, 1995.
- [1350] M.S. Waterman and T.F. Smith. RNA secondary structure: A complete mathematical analysis. *Math. Biosci.*, 42:257–266, 1978.
- [1351] R. H. Waterston, E. S. Lander, and J. E. Sulston. On the sequencing of the human genome. *Proc. Natl. Acad. Sci. USA*, 99:3712–3716, 2002.
- [1352] R. H. Waterston, E. S. Lander, and J. E. Sulston. More on the sequencing of the human genome. *Proc. Natl. Acad. Sci. USA*, 100:3022–3024, 2003.
- [1353] J. D. Watson. *The Double Helix. A Personal Account of the Discovery of the Structure of DNA*. Norton Critical Edition G.S. Stent (Editor), Norton & Company, New York, NY, 1980.
- [1354] J. D. Watson and F. H. C. Crick. Genetical implications of the structure of deoxyribonucleic acid. *Nature*, 171:964–967, 1953.
- [1355] J. D. Watson and F. H. C. Crick. A structure for deoxyribose nucleic acid. *Nature*, 171:737–738, 1953.
- [1356] J. D. Watson and F. H. C. Crick. The structure of DNA. *Cold Spr. Harb. Symp. Quant. Biol.*, XVIII:123–131, 1953.

- [1357] M. Weber, S. Kube, L. Wlater, and P. Deuflhard. Stable computational of probability densities for metastable dynamical systems. *SIAM Mult. Model. Sim.*, 6:396–416, 2007.
- [1358] Z. Wei, G. Li, and L. Qi. New nonlinear conjugate gradient formulas for large-scale unconstrained optimization problems. *App. Math. Comput.*, 179:407–430, 2006.
- [1359] S. J. Weiner, P. A. Kollman, D. T. Nguyen, and D.A. Case. An all atom force field for simulations of proteins and nucleic acids. *J. Comput. Chem.*, 7:230–252, 1986.
- [1360] F. Weinhold. A new twist on molecular shape. *Nature*, 411:539–541, 2001.
- [1361] J. J. Wendoloski, S. J. Kimatian, C. E. Schutt, and F. R. Salemme. Molecular dynamics simulation of a phospholipid micelle. *Science*, 243:636–638, 1989.
- [1362] B. G. Wensley, S. Batey, F. A. Bone, Z. M. Chan, N. R. Tumelty, A. Steward, L. G. Kwa, A. Borgia, and J. Clarke. Experimental evidence for a frustrated energy landscape in a three-helix-bundle protein family. *Nature*, 463:685–688, 2010.
- [1363] B. Werth. *The Billion-Dollar Molecule: One Company's Quest for the Perfect Drug*. Simon & Schuster, New York, NY, 1994.
- [1364] G. Wess, M. Urmann, and B. Sickenberger. Medicinal chemistry: Challenges and opportunities. *Angew. Chem. Int. Ed.*, 40:3341–3350, 2001.
- [1365] J. Westbrook, Z. Feng, L. Chen, H. Yang, and H. M. Berman. The Protein Data Bank and structural genomics. *Nucl. Acids Res.*, 31:489–491, 2003.
- [1366] T. P. Westcott, I. Tobias, and W. K. Olson. Elasticity theory and numerical analysis of DNA supercoiling: An application to DNA looping. *J. Phys. Chem.*, 99:17926–317935, 1995.
- [1367] E. Westhof and L. Jaeger. RNA pseudoknots. *Curr. Opin. Struct. Biol.*, 2:327–333, 1992.
- [1368] D. A. Wheeler, M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y. J. Chen, V. Makhijani, G. T. Roth, X. Gomes, K. Tartaro, F. Niazi, C. L. Turcotte, G. P. Irzyk, J. R. Lupski, C. Chinault, X. Z. Song, Y. Liu, Y. Yuan, L. Nazareth, X. Qin, D. M. Muzny, M. Margulies, G. M. Weinstock, R. A. Gibbs, and J. M. Rothberg. The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452:872–876, 2008.
- [1369] J. H. White. Self-linking and the Gauss integral in higher dimensions. *Amer. J. Math.*, 91:693–728, 1969.
- [1370] J. H. White. An introduction to the geometry and topology of DNA structures. In M. S. Waterman, editor, *Mathematical Methods for DNA Sequences*, chapter 9. CRC Press, Boca Raton, Florida, 1989.
- [1371] H. Wille, M. D. Michelitsch, V. Guénebaut, S. Supattapone, A. Serban, F. E. Cohen, D. A. Agard, and S. B. Prusiner. Structural studies of the scrapie prion protein by electron crystallography. *Proc. Natl. Acad. Sci. USA*, 99:3563–3568, 2002.
- [1372] P. Willett. Structural similarity measures for database searching. In P. von Ragué Schleyer (Editor-in Chief), N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, and H. F. Schaefer, III, editors, *Encyclopedia of Computational Chemistry*, volume 4, pages 2748–2756. John Wiley & Sons, West Sussex, England, 1998.
- [1373] L. D. Williams and L. J. Maher, III. Electrostatic mechanisms of DNA deformation. *Annu. Rev. Biophys. Biomol. Struct.*, 29:497–521, 2000.

- [1374] J. R. Williamson. Small subunit, big science. *Nature*, 407:306–307, 2000.
- [1375] S. Willmann, A. N. Edginton, K. Coboeken, G. Ahr, and J. Lippert. Risk to the breast-fed neonate from codeine treatment to the mother: A quantitative mechanistic modeling study. *Clin. Pharmacol. Ther.*, 86:634–643, 2009.
- [1376] E. K. Wilson. Computers customize combinatorial libraries. *Chem. Eng. News*, 76:31–37, 1998.
- [1377] E. O. Wilson. *Consilience. The Unity of Knowledge*. Alfred A. Knopf, New York, NY, 1998.
- [1378] M. Wilson, J. DeRisi, H. H. Kristensen, P. Imboden, S. Rane, P. O. Brown, and G. K. Schoolnik. Exploring drug-induced alterations in gene expression in *Mycobacterium tuberculosis* by microarray hybridization. *Proc. Natl. Acad. Sci. USA*, 96:12833–12838, 1999.
- [1379] B. T. Wimberly, D. E. Brodersen, W. M. Clemons Jr., R. J. Morgan-Warren, A. P. Carter, C. Vonrhein, T. Hartsch, and V. Ramakrishnan. Structure of the 30S ribosomal subunit. *Nature*, 407:327–339, 2000.
- [1380] R. Wing, H. Drew, T. Takano, C. Broka, S. Tanaka, K. Itakura, and R. E. Dickerson. Crystal structure analysis of a complete turn of B-DNA. *Nature*, 287:755–758, 1980.
- [1381] W. Winkler, A. Nahvi, and R. R. Breaker. Thiamine derivatives bind messenger RNAs directly to regulate bacterial expression. *Nature*, 419:952–956, 2002.
- [1382] J. Woda, B. Schneider, K. Patel, K. Mistry, and H. M. Berman. An analysis of the relationship between hydration and protein-DNA interactions. *Biophys. J.*, 75:2170–2177, 1998.
- [1383] C. R. Woese, S. Winker, and R. R. Gutell. Architecture of ribosomal RNA: Constraints on the sequence of tetra-loops. *Proc. Natl. Acad. Sci. USA*, 87:8467–8471, 1990.
- [1384] A. Wolffe. *Chromatin Structure and Function*. Academic Press Inc., San Diego, CA, 1995.
- [1385] P. G. Wolynes. Folding funnels and energy landscapes of larger proteins within the capillarity approximation. *Proc. Natl. Acad. Sci. USA*, 94:6170–6175, 1997.
- [1386] P. G. Wolynes. Recent successes of the energy landscape theory of protein folding and function. *Quart. Rev. Biophys.*, 38:405–410, 2005.
- [1387] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai. Navigating the folding routes. *Science*, 267:1619–1620, 1995.
- [1388] H. Wong, J.-M. Victor, and J. Mozziconacci. An all-atom model of the chromatin fiber containing linker histones reveals a versatile structure tuned by the nucleosomal repeat length. *PLoS ONE*, 2:e877, 2007.
- [1389] K. Wong. The mammals that conquered the seas. *Sci. Amer.*, 286:70–79, 2002.
- [1390] M. H. Wright. What, if anything, is new in optimization? In J. M. Ball and J. C. R. Hunt, editors, *ICIAM'99: Proceedings of the 4th International Congress on Industrial and Applied Mathematics*, pages 259–270, Oxford, England, 2000. Oxford University Press. Also available in modified form as Technical Report 00-4-08, Bell Laboratories, Computing Sciences Research Center, Murray Hill, New Jersey 07074; [cm.bell-labs.com/cm/cs/doc/00/4-08.ps.gz](http://cm.bell-labs.com/cm/cs/doc/00/4-08.ps.gz).

- [1391] P. E. Wright and H. J. Dyson. Intrinsically unstructured proteins: Reassessing the protein structure-function paradigm. *J. Mol. Biol.*, 293:321–331, 1999.
- [1392] B. Wu, P. Dröge, and C. A. Davey. Site selectivity of platinum anticancer therapeutics. *Nat. Chem. Biol.*, 4:110–112, 2008.
- [1393] X. Wu and S. Wang. Self-guided molecular dynamics simulation for efficient conformational search. *J. Phys. Chem. B*, 102:7238–7250, 1998.
- [1394] X. Wu and S. Wang. Enhancing systematic motion in molecular dynamics simulation. *J. Chem. Phys.*, 110:9401–9410, 1999.
- [1395] X. Wu, S. Wang, and B. R. Brooks. Direct observation of the folding and unfolding of  $\beta$ -hairpin in explicit water through computer simulation. *J. Amer. Chem. Soc.*, 124:5282–5283, 2002.
- [1396] K. Wüthrich. *NMR of Proteins and Nucleic Acids*. (The George Fisher Baker Non-Resident Lectureship in Chemistry at Cornell University series). Wiley Interscience, New York, NY, 1986.
- [1397] D. Xie and T. Schlick. Efficient implementation of the truncated Newton method for large-scale chemistry applications. *SIAM J. Opt.*, 10(1):132–154, 1999.
- [1398] D. Xie and T. Schlick. Remark on the updated truncated Newton minimization package, *Algorithm 702. ACM Trans. Math. Softw.*, 25(1):108–122, 1999.
- [1399] D. Xie and T. Schlick. Visualization of chemical databases using the singular value decomposition and truncated-Newton minimization. In C. A. Floudas and P. Pardalos, editors, *Optimization in Computational Chemistry and Molecular Biology: Local and Global Approaches*, pages 267–286. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000.
- [1400] D. Xie and T. Schlick. A more lenient stopping rule for line search algorithms. *Opt. Math. Softw.*, 17:683–700, 2002.
- [1401] D. Xie, L. R. Scott, and T. Schlick. Analysis of the SHAKE-SOR algorithm for constrained molecular dynamics simulations. *Methods and Applications of Analysis*, 7(3):577–590, 2000. (Special Issue dedicated to Cathleen Morawetz).
- [1402] D. Xie, A. Tropsha, and T. Schlick. A data projection approach using the singular value decomposition and energy refinement. *J. Chem. Inf. Comput. Sci.*, 40(1):167–177, 2000.
- [1403] Y. Xin, C. Laing, N. B. Leontis, and T. Schlick. Annotation of tertiary interactions in RNA structures reveals variations and correlations. *RNA*, 14:2465–2477, 2008.
- [1404] R. Xu, B. Ayers, D. Cowburn, and T. W. Muir. Chemical ligation of folded recombinant proteins: Segmental isotopic labeling of domains for NMR studies. *Proc. Natl. Acad. Sci. USA*, 96:388–393, 1999.
- [1405] Z. Xu, A. L. Horwich, and P. B. Sigler. The crystal structure of the asymmetric GroEL-GroES-(ADP)<sub>7</sub> chaperonin complex. *Nature*, 388:741–750, 1997.
- [1406] B. I. Yakobson and R. E. Smalley. Fullerene nanotubes: C<sub>1,000,000</sub> and beyond. *American Scientist*, 85(4):324–337, 1997.
- [1407] H. Yamakawa. *Modern Theory of Polymer Solutions*. Harper and Row Publishers, New York, NY, 1971.
- [1408] L. Yang, W. A. Beard, S. H. Wilson, S. Broyde, and T. Schlick. Polymerase  $\beta$  simulations suggest that Arg258 rotation is a slow step rather than large subdomain motion *per se*. *J. Mol. Biol.*, 317:651–671, 2002.

- [1409] W. Yang and T. Lee. A density-matrix divide-and-conquer approach for electronic structure calculations of large molecules. *J. Chem. Phys.*, 163:5674–5678, 1995.
- [1410] Y. Yang, I. Tobias, and W. K. Olson. Finite element analysis of DNA supercoiling. *J. Chem. Phys.*, 98:1673–1686, 1993.
- [1411] T. Yoda, Y. Sugita, and Y. Okamoto. Secondary-structure preferences of force fields for proteins evaluated by generalized-ensemble simulations. *Chem. Phys.*, 307:269–283, 2004.
- [1412] Y. Yonetani, Y. Maruyama, F. Hirata, and H. Kono. Comparison of DNA hydration patterns obtained using two distinct computational methods, molecular dynamics simulation and three-dimensional reference interactions site model theory. *J. Chem. Phys.*, 128:185102, 2008.
- [1413] D. York and W. Yang. The Fast Fourier Poisson method for calculating Ewald sums. *J. Chem. Phys.*, 101:3298–3300, 1994.
- [1414] D. M. York, T.-S. Lee, and W. Yang. Parameterization and efficient implementation of a solvent model for linear-scaling semiempirical quantum-mechanical calculations of biological macromolecules. *Chem. Phys. Lett.*, 263:297–304, 1996.
- [1415] D. M. York, T.-S. Lee, and W. Yang. Quantum-mechanical study of aqueous polarization effects on biological macromolecules. *J. Amer. Chem. Soc.*, 118:10940–10941, 1996.
- [1416] D. M. York, W. Yang, H. Lee, T. Darden, and L. G. Pederson. Toward the accurate modeling of DNA: The importance of long-range electrostatics. *J. Amer. Chem. Soc.*, 117:5001–5002, 1995.
- [1417] M. A. Young and D. L. Beveridge. Molecular dynamics simulations of an oligonucleotide duplex with adenine tracts phased by a full helix turn. *J. Mol. Biol.*, 281:675–687, 1998.
- [1418] M. A. Young, S. Gonfloni, G. Superti-Furga, B. Roux, and J. Kuriyan. Dynamic coupling between SH2 and SH3 domains of c-Src and Hck underlies their inactivation by C-terminal tyrosine phosphorylation. *Cell*, 105:115–126, 2001.
- [1419] M. A. Young, B. Jayaram, and D. L. Beveridge. Intrusion of counterions into the spine of hydration in the minor groove of B-DNA: Fractional occupancy of electronegative pockets. *J. Amer. Chem. Soc.*, 119:59–69, 1997.
- [1420] M. A. Young, B. Jayaram, and D. L. Beveridge. Local dielectric environment of B-DNA in solution: Results from a 14 ns molecular dynamics trajectory. *J. Phys. Chem. B*, 102:7666–7669, 1998.
- [1421] M. A. Young, G. Ravishankar, and D. L. Beveridge. A 5-nanosecond molecular dynamics trajectory for B-DNA: Analysis of structure, motions, and solvation. *Biophys. J.*, 73:2313–2336, 1997.
- [1422] M. A. Young, J. Srinivasan, I. Goljer, S. Kumar, D. L. Beveridge, and P. H. Bolton. Structure determination and analysis of local bending in an A-tract DNA duplex: Comparison of results from crystallography, nuclear magnetic resonance, and molecular dynamics simulation on d(CGCAAAATGCG). *Methods in Enzymology*, 261:121–144, 1995.
- [1423] H. Yu. Extending the size limit of protein nuclear magnetic resonance. *Proc. Natl. Acad. Sci. USA*, 96:332–334, 1999.

- [1424] G. Yuan. Modified nonlinear conjugate gradient methods with sufficient descent condition for large-scale optimization problems. *Opt. Lett.*, 3:11–21, 2009. doi: 10.1007/s11590-008-0086-5.
- [1425] G. C. Yuan and J. S. Liu. Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Comp. Biol.*, 4:e13, 2008.
- [1426] M.-R. Yun, R. Lavery, N. Mousseau, K. Zakrzewska, and P. Derreumaux. ARTIST: an activated method in internal coordinate space for sampling protein energy landscapes. *Proteins*, 63:967–975, 2006.
- [1427] M. M. Yusupov, G. Zh. Yusupova, A. Baucom, K. Lieberman, T. N. Earnest, J. H. D. Cate, and H. F. Noller. Crystal structure of the ribosome at 5.5 Å resolution. *Science*, 292:883–896, 2001.
- [1428] B. Zagrovic, E. J. Sorin, and V. Pande.  $\beta$ -hairpin folding simulations in atomistic detail using an implicit solvent model. *J. Mol. Biol.*, 313:151–169, 2001.
- [1429] R. Zahn, A. Liu, T. Lührs, R. Riek, C. von Schroetter, F. López Garcia, M. Billeter, L. Calzolai, G. Wider, and Kurt Wüthrich. NMR solution structure of the human prion protein. *Proc. Natl. Acad. Sci. USA*, 97:145–150, 2000.
- [1430] F. N. Zaidi, U. Nath, and J. B. Udgaonkar. Multiple intermediates and transition states during protein unfolding. *Nature Struc. Biol.*, 4:1016–1024, 1997.
- [1431] M. J. Zaki. Efficient mining frequent trees in a forest: Algorithms and applications. *IEEE Trans. Know. Data Eng.*, 17:1021–1035, 2005.
- [1432] P. D. Zamore. Ancient pathways programmed by small RNAs. *Science*, 296: 1265–1269, 2002.
- [1433] A. H. Zewail. Physical biology: 4D visualization of complexity. In A. H. Zewail, editor, *Physical Biology: From Atoms to Medicine*, pages 23–50. Imperial College Press, London, UK, 2008.
- [1434] C. Zhang and C. DeLisi. Estimating the number of protein folds. *J. Mol. Biol.*, 284:1301–1305, 1998.
- [1435] G. Zhang and T. Schlick. LIN: A new algorithm combining implicit integration and normal mode techniques for molecular dynamics. *J. Comput. Chem.*, 14:1212–1233, 1993.
- [1436] G. Zhang and T. Schlick. The Langevin/implicit-Euler/Normal-Mode scheme (LIN) for molecular dynamics at large time steps. *J. Chem. Phys.*, 101:4995–5012, 1994.
- [1437] G. Zhang and T. Schlick. Implicit discretization schemes for Langevin dynamics. *Mol. Phys.*, 84:1077–1098, 1995.
- [1438] J. Zhang, Y. Xiao, and Z. Wei. Nonlinear conjugate gradient methods with sufficient descent condition for large-scale unconstrained optimization. *Math. Prob. Engin.*, 2009, 2009. 16 pages, doi: 10.1155/2009/243290.
- [1439] L. Y. Zhang, E. Gallicchino, R. A. Friesner, and R. M. Levy. Solvent models for protein-ligand binding: Comparison of implicit solvent Poisson and surface generalized Born models with explicit solvent simulations. *J. Comput. Chem.*, 22:591–607, 2001.
- [1440] M. Zhang and R. D. Skeel. Cheap implicit symplectic integrators. *Appl. Num. Math.*, 25:297–302, 1997.

- [1441] Q. Zhang, D. Beard, , and T. Schlick. Constructing irregular surfaces to enclose macromolecular complexes for mesoscale modeling using the discrete surface charge optimization (DiSCO) algorithm. *J. Comput. Chem.*, 24:2063–2074, 2003.
- [1442] Y. Zhang. Pseudobond Ab Initio QM/MM approach and its applications to enzyme reactions. *Theor. Chem. Acc.*, 116:43–50, 2006.
- [1443] Y. Zhang, T. Lee, and W. Yang. A pseudo-bond approach to combining quantum mechanical and molecular mechanical methods. *J. Chem. Phys.*, 110:46–54, 1999.
- [1444] Y. Zhang, H. Liu, and W. Yang. Free energy calculation on enzyme reactions with an efficient iterative procedure to determine minimum energy paths on a combined Ab Initio QM/MM potential energy surface. *J. Chem. Phys.*, 112:3483–3492, 2000.
- [1445] Q. Zhao and W. Yang. Analytical energy gradients and geometry optimization in the divide-and-conquer method for large molecules. *J. Chem. Phys.*, 102:9598–9603, 1995.
- [1446] Y. Zhao, B. L. Kormos, D. L. Beveridge, and A. M. Baranger. Molecular dynamics simulations studies of a protein-RNA complex with a selectively modified binding interface. *Biopolymers*, 81:256–269, 2006.
- [1447] J. Z. Zhou. Structure-directed combinatorial library design. *Curr. Opin. Chem. Biol.*, 12:379–385, 2008.
- [1448] R. Zhou and B. J. Berne. A new molecular dynamics method combining the reference system propagator algorithm with a fast multipole method for simulating proteins and other complex systems. *J. Chem. Phys.*, 103:9444–9459, 1995.
- [1449] R. Zhou, E. Harder, H. Xu, and B. J. Berne. Efficient multiple time step method for use with Ewald and particle mesh Ewald for large biomolecular systems. *J. Chem. Phys.*, 115:2348–2358, 2001.
- [1450] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-BFGS-B, FORTRAN subroutines for large scale bound constrained optimization. *ACM Trans. Math. Softw.*, 23:550–560, 1997.
- [1451] K. Zhu, M. R. Shirts, R. A. Friesner, and M. P. Jacobson. Multiscale Optimization of a Truncated Newton Minimization Algorithm and Application to Proteins and Protein-Ligand Complexes. *J. Chem. Theory Comput.*, 3:640–648, 2007.
- [1452] X. Zhuang. Single-molecule RNA science. *Ann. Rev. Biophys. Biomol. Struct.*, 34:399–414, 2005.
- [1453] X. Zhuang and M. Rief. Single-molecule folding. *Curr. Opin. Struct. Biol.*, 13:88–97, 2003.
- [1454] V. B. Zhurkin, Y. P. Lysov, and V. Ivanov. Anisotropic flexibility of DNA and the nucleosomal structure. *Nucleic Acids Res.*, 6:1081–1096, 1979.
- [1455] V. B. Zhurkin, N. B. Ulyanov, A. A. Gorin, and R. L. Jernigan. Static and statistical bending of DNA evaluated by Monte Carlo simulations. *Proc. Natl. Acad. Sci. USA*, 88:7046–7050, 1991.
- [1456] B. H. Zimm. Dynamics of polymer molecules in dilute solution: Viscoelasticity, flow birefringence and dielectric loss. *J. Chem. Phys.*, 24:269–278, 1956.
- [1457] O. Zimmerman and U. H. E. Hansmann. Understanding protein folding: small proteins in silico. *Biochim. Biophys. Acta*, 1784:252–258, 2008.
- [1458] G. Zou, R. D. Skeel, and S. Subramanian. Biased Brownian dynamics for rate constant calculation. *Biophys. J.*, 79:638–645, 2000.

- [1459] D. M. Zuckerman and E. Lyman. A second look at canonical sampling of biomolecules using replica exchange simulation. *J. Chem. Theo. Comp.*, 2:1200–1202, 2006.
- [1460] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids Res.*, 9:133–148, 1981.
- [1461] MFOLD: M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucl. Acids Res.*, 31:3406, 2003. <http://www.bioinfo.rpi.edu/~zukerm>.
- [1462] D. E. Shaw, R. O. Dror, J. K. Salmon, J. P. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, M. M. Deneroff, B. Batson, K. J. Bowers, E. Chow, M. P. Eastwood, D. J. Ierardi, J. L. Klepeis, J. S. Kuskin, R. H. Larson, K. Lindorff-Larsen, P. Maragakis, M. A. Moraes, S. Piana, Y. Shan, and B. Towles. Millisecond-scale molecular dynamics simulations on Anton. In *SC '09: Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, pages 1–11, San Diego, CA, 2009. ACM.
- [1463] J. Pogany, M. R. Fabian, K. A. White, and P. D. Nagy. A replication silencer element in a plus-strand RNA virus. *EMBO J.*, 22:5602–5611, 2003.
- [1464] T. Schlick, R. Colleparo-Guevara, L. Halvorsen, S. Jung, and X. Xiao. Biomolecular Modeling and Simulation: A Field Coming of Age. Submitted, 2010.
- [1465] C. Laing and T. Schlick. Computational approaches to RNA 3D Modeling. *J. Phys. Condens. Matter*, 22:283101–283119, 2010.

# Index

- anti*, 280f
- gauche*, 97f, 280f
- trans*, 97f, 280f
- ab initio* methods, 242–247
- actin, 126, 127
- Ada, Yonath, 4
- adiabatic map, 148, 148n, 149f
- agglutinin, 112f, 117
- AIDS
  - drugs, 59f, 58–62, 62b, 522b, 523
  - retrovirus, 59
- algorithm
  - definition, xvi
  - improvements, 2
- Allinger, Norman, 7t, 10, 262
- Altman, Sidney, 18
- AMBER, xvii, 262, 594h
  - Fortran language, 388
  - parameter examples, 286–297
- amino acids, 86f, 87f, 82–89
  - $\alpha$ -Carbon, 79f, 82
  - essential, 83b, 83–84
  - formula, 79f
  - nonessential, 83–84
- Anfinsen, Christian B., 103, 103n
- Anton, 11, 13, 239, 514
- apoptosis, 222
- aptamers, 220
- Arber, Werner, 7t, 22
- aspartame, 84f, 85
- ATP synthase B chain-like, 123f
- Avery, Oswald, 7t, 17
- backward error analysis, 466n
- bacteriorhodopsin, 106
- Baltimore, David, 31q
- barstar, 92, 96f
- base pair
  - stacking, 140, 157, 166–168, 168b, 206
- base pairing schemes, 206–216
  - Hoogsteen, 207, 209f
  - mismatches, 209–210
  - reverse Hoogsteen, 208
  - reverse Watson-Crick, 207
  - Watson-Crick, 206–207
  - wobbles, 209–210
- bending
  - in-plane
    - rocking, 271
    - scissoring, 271
  - out-of-plane
    - twisting, 271
    - wagging, 271

- Berendsen, H.J.C., 11  
 Berman, Helen, 572h  
 Bernal, J. D., 14  
 Bernanke, Ben , 346q  
 Berne, Bruce, 474  
 BFGS method, 371–372  
 biochips, 39b, 526, 530b  
 bioengineering, 70, 71b  
 bioinformatics, 43, 65  
 biology  
     dogma, 131  
     machinery, 251  
     symmetry, 53  
 biopharming, 70  
 Black, James W., 521  
 BLAST, 590h  
 Blue Gene, 49  
 Blue Waters, 50  
 Board, John, 323  
 Bohr, Niels, xix, 3, 520q  
 Boltzmann factor, 412  
 Boltzmann probability, 413f  
 Boltzmann statistics, 387  
 Born-Oppenheimer approximation, 242–247, 249  
 bovine pancreatic trypsin inhibitor (BPTI)  
     calculated spectra, 271  
 bovine papillomavirus E2 protein  
     DNA binding, 168b  
 bovine spongiform encephalopathy (BSE), *see* mad cow disease  
 Bragg's law, 19  
 Brahe, Tycho, 429  
 Brenner, Sydney, 7t, 26, 29  
 Brown, Robert, 489  
 Brownian dynamics, 337–338, 428t, 430f, 487–496  
     algorithms, 337, 491  
     complexity, 495b, 494–496, 496f  
     Ermak-McCammon, 491  
     inertial, 491  
     random-force computation, 495b, 494–496, 496f  
     biased, 509  
     diffusion constant, 337, 481, 490, 614h  
     diffusive limit, 336  
     DNA applications, 202f, 202, 335, 337  
     exercise, 614h  
     fluctuation/dissipation theorem, 336, 489  
     hydrodynamics, 337–338, 492b, 491–494  
     Kirkwood-Riseman approximation, 494b, 494  
     Oseen tensor, 492–494  
     Rotne-Prager tensor, 492–494  
     software, 343  
     theory and applications, 335  
 Brownian motion, 198, 337, 614h  
 Broyde, Suse, xv  
 bulge, 216, 218f  
 butane  
     torsional states, 280f  
 calcium-binding protein, 95f  
 calmodulin, 94, 115  
 Calugareanu, Grigore, 195  
 Carboxypeptidase inhibitor, 124f  
 catabolite gene-activating protein (CAP)  
     DNA binding, 168b, 172f  
 Cech, Thomas, 18  
 Celera Genomics, 26–27, 30–33, 33b  
 cellomics, 42  
 cellulase celA, 108f, 115  
 central limit theorem, 409–410, 438  
 CFF, 262  
 chaos, 441f, 440–441  
     Lyapunov exponent, 440n  
 chaperones, 51f, 50–51, 52b, 252b  
 Chargaff, Erwin, 7t, 17  
 CHARMM, xvii, 255, 256t, 257f, 262, 594h  
     Fortran language, 388  
     Monte Carlo, 416  
     parameter examples, 286–297  
 Chase, Martha, 7t, 17  
**Chemical Design**, 519–552  
 chemical libraries, 526–550  
     bioactivity, 529–530, 535–536  
     compound descriptors, 534–535  
     data compression and analysis, 540–550  
         PCA, 540–542  
         SVD, 542–544  
     problems, 526–530, 532–539  
         diversity, 527–528, 538–539  
         similarity, 527–528, 538–539  
 QSAR, 65

- SAR, 536
- schematic illustration, 533f
- chemogenomics, 523
- chemoinformatics, 43, 65, 521
- chimeroplast, 68b
- chromatin, 184–185, 187f
  - model, 192f
  - simulation, 193b, 195f
- chromosome, 182–193
- classical mechanics
  - limitations, 432–435
- collagen, 78, 91f, 91, 110
- combinatorial chemistry
  - future, 551–552
  - history, 520–526
- combinatorial optimization, 527, 538
- computational biology, 41
- computing
  - progress, 2, 239
- Comte de Buffon, 387
- configuration, 258
  - definition, 258
  - relation to conformation, 258
- conjugate-gradient methods, 244, 372–374
- consilience, 2
- constrained dynamics, 453–455
  - RATTLE, 454
  - SHAKE, 453–454
    - in MTS, 482, 485b
- constraints
  - in minimization, 428t
- continuum solvation, 333–342, 428t
- Poisson-Boltzmann approach, 338–343
  - potential of mean force, 334–335
- convergence (of a sequence), 352b
  - linear, 352b
  - quadratic, 352b
  - superlinear, 352b
- Copernicus, Nicolas, 429
- Corey, Robert, 7t, 15
- Coronavirus NSP10-like, 124f
- cosmeceuticals, 74
- counterion condensation theory, 177, 180b
- Course Syllabus, 557–558**
- Creutzfeld-Jacob disease (CJD), *see* mad cow disease
- Crick, Francis, 7t, 17–18, 130, 135, 150, 155
- Crigler-Najjar (CN) syndrome, 67, 68b
- Critical Assessment of Techniques for Protein Structure Prediction (CASP), 48–49
- Crothers, Don, 170
- Crowfoot-Hodgkin, Dorothy, 14
- cruciform, 218f
- cryoelectron microscopy (cryo-EM), 23
  - illustration, 5f
- crystallography
  - direct lattice, 315b
  - electron density, 314, 315b
  - indices of reflection, 315b
  - reciprocal lattice, 315b
  - structure factor, 314, 315b, 322
  - unit cell, 314b
- cube, *see* periodic boundary conditions
- cutoff schemes, 302–310
  - force switch, 308–309
  - general formulation, 306–307
    - in multiple timesteps, 305, 308, 309f, 323
    - potential switch, 307–308, 309f
    - shift functions, 309f, 309–310
- CVFF, 262, 287
- cystic fibrosis, 67
- cytochrome, 92, 96f, 114, 115
- databases
  - analysis, 41–44
  - growth, 44f, 45t, 572h
  - NDB, 133, 135, 165, 206, 572h, 582h, 584h, 589, 590h
  - NRPR, 44f, 45t
  - PDB, 43, 45t, 165, 206, 572h, 589, 590h
  - PIR, 44f, 45t
  - RCSB, 44, 45t, 165, 206, 572h, 590h, 592h
  - sequence, 44f, 45t, 572h
  - structure, 44f, 45t, 572h
- Debye-Hückel theory, 177, 193b, 341f, 340–342
  - Debye length, 340, 341f
- Delbrück, Max, 17
- DelPhi, 342, 343
- density functional theory, 243
- derivatives
  - gradient vector, 352, 435–436

- Hessian matrix, 353
- positive-definite, 353
- descent direction, 359–360
- detailed balance, 411–412
- dielectric constant, 316, 317, 334, 338–340, 342, 343f
- dielectric function, 338–339
- diffusion constant, 337, 481, 490
- dihedral angle
  - anti*, 280f
  - gauche*, 97f, 280f
  - trans*, 97f, 280f
  - definition, 97b, 98f
  - relation to torsion angle, 97b
- Dirac, Paul, 241
- Dirac, Paul Adrien Maurice, 431q
- directed molecular evolution, 68
- disulfide bonds, 89
- divergence operator, 319b
- Djerassi, Carl, 521
- DNA
  - London Times* quote, 130q
  - A-form, 158f, 156–158, 159, 160f
  - A-tracts, 169–173, 173f, 174b, 178, 201b
  - adducts, xv
  - B-form, 155–157, 158–160f
  - base-pair step analysis, 167t
  - bending, 166, 171f, 169–173, 173f, 174b
  - canonical forms, 155–161, 582h, 584h
  - cellular organization, 187f, 182–193
  - chips, 39b, 526, 530b
  - chromosome, 182–193
  - computing, 134, 165, 240
  - double helix, 136f
    - diameter, 150
    - grooves, 150
    - helix sense, 150, 150n
    - pitch, 136f, 150
    - residues per turn, 150
    - rise, 150
    - twist, 136f, 150
  - dynamics, 9t, 12, 15, 16f, 133–134, 179f, 193b, 195f, 198, 202b, 202f, 203f, 200–204, 215b
  - fundamental processes, 132–133
  - genome content, 184f, 182–184, 185t
  - geometry, 195–197
  - histone, 187f
  - hydration, 172, 174–180
  - ion patterns, 174–180
  - junk, 28
  - linking number, 196f, 195–196
  - Maddox quote, 240q
  - microarrays, 39b, 42
  - mimics, 212f, 212–213
  - model forms, 151, 152t, 158–160f
  - nucleosome, 189f, 186–193
  - overstretched, 215b, 214–216
  - persistence length, 187f, 198–202
  - polymer model, 198–202
  - quadruplexes, 211–212
  - S-DNA, 214
  - sequence effects, 165–174
    - Ascona Consortium, 169
    - site juxtaposition, 201, 202b, 202, 203f
  - supercoil, 156
  - supercoiling, 200b, 184–204
    - elastic approximation, 200b, 201b, 199–202
    - simulation, 197–204
  - TATA-box, 169b, 172f, 179f, 182, 207
  - telomeres, 211–212
  - topology, 195–197
  - triplexes, 211
  - twist, 196f, 196–197
  - understretched, 214–216
  - Watson and Crick's seminal paper, 164q
  - writhe, 196f, 197
  - Z-form, 156, 158–160f, 160–161
  - DNA topoisomerase I, 121f
  - DNA-binding leucine zipper protein, 114
  - DNA/protein interactions, 181t, 180–182
    - helix-turn-helix (HTH) motif, 584h
  - Down's syndrome, 31
  - Doyle, Arthur Conan, 346q
  - drugs, 58–67
    - abacavir, 531b
    - AIDS, 59f
    - biologics, 524
    - chlorpromazine, 527f
    - codeine, 531b
    - COX-2 inhibitors, 529b
    - cyclosporin, 58, 522b
    - for chronic pain, 529b
    - future, 552
    - herceptin, 530, 531b

- lead generation, 530
- lead optimization, 530
- mevacor, 522b
- Natural, 522b
- neuropeptide inhibitors, 523
- PDE-5 inhibitors, 523
- prilosec, 58
- quinacrine, 527f
- raloxifene, 527f
- rational, 58, 521–526
- tamoxifen, 527f
- taxol, 522b
- thrombin inhibitors, 523
- viagra, 58, 525f
- viracept, 525f
- warfarin, 531b
- zomig, 525f
- dyad, 152, 154
- ECEPP, xvii, 262, 594h
- Einstein, Albert, 490
- Elber, Ron, 517
- Elion, Gertrude, 58
- elongated dodecahedron, *see* periodic boundary conditions
- Engels, Friedrich, 77q
- Englund, Paul, 170
- ensembles, 412–413, 439, 455–461
  - canonical (NVT), 412–418, 455–457, 459–460
  - generalized, 421
  - isothermal-isobaric (NPT), 412, 455, 457–458, 460–461
  - microcanonical (NVE), 412, 439, 455
  - weak barostat coupling, 457–458
  - weak thermostat coupling, 456–457
- epigenetics, 222
- ergodicity, 411–412
- error bars, 388, 409–410
- ester, 138n
- Ewald method, 314b, 311–324
  - complexity, 320–321
  - direct space term, 321–322
  - Ewald's trick, 316–319
  - finite-dielectric correction, 320–322
  - Gaussian masking, 319f, 317–319
  - nonbonded-exclusion correction, 321–322
  - reciprocal space term, 321–322
- roots in crystallography, 314–316
- self-interactions correction, 321–322
- experimental progress, 14–23, 239
  - cryoelectron microscopy, 5f, 23
  - NMR, 20–23
  - X-ray crystallography, 14–20, 22–23
- experimental triumphs, 4
- face-centered cube, *see* periodic boundary conditions
- fatty acid binding protein, 111f, 116, 117
- ferritin, 114
- Fibonacci series, 396n
- fibronectin, 111f, 116, 117
- Finch, John, 189
- Fixman, Marshall, 338, 494, 495b
- flagellin, 126
- flavodoxin, 118, 120f
- fluctuation/dissipation theorem, 336, 489
- Fokker-Planck equation, 337, 489
- foods
  - amino acid sources, 83, 83b
  - designed, 71b, 69–72
  - vegetarian diet, 83b
- force field
  - AMBER, xvii, 262, 388, 594h
    - parameter examples, 286–297
  - approximations, xiv, 594h
    - polarizability, 263
  - assessment, 594h
  - CFF, 262
  - CHARMM, xvii, 255, 257f, 262, 388, 594h
    - cutoff functions examples, 306f, 307, 310
    - Monte Carlo, 416
    - nonbonded CPU examples, 302f, 303t
    - parameter examples, 286–297
  - criticisms, 248, 594h
  - CVFF, 262, 287
  - debate, xvii
  - early successes, 248
  - ECEPP, xvii, 262, 594h
  - first-generation, 10–11
  - GROMOS, 262
  - limitations, 262–264
  - MMFF, 262

- nonbonded vs. bonded computations, 599h
- OPLS, 262
- potential
  - bond angle, 276–280
  - bond length, 272–276, 600h
  - bond length, harmonic, 272–274
  - bond length, Morse, 274–276
  - Buckingham, 288
  - Coulomb, 293b, 291–295, 599h
  - hydrogen bond, 263
  - Lennard, 599h
  - torsional, 280–288, 600h
  - van der Waals, 288–291
- second-generation, 267
- third-generation, 264, 267
- transferability, 284
- Force Fields**, 265–297
- force fields
  - CHARMM
    - atom types, 256t
  - Franklin, Rosalind, 7t, 17, 17n
  - Frauenfelder, Ulrich Hans, 443q
  - Fuller, F. Brock, 195
  - function
    - curvature, 353, 355, 365, 369, 370, 379
    - derivatives, 352
    - least-square, 350
    - non-differentiable, 350
    - nonlinear, 350
    - separable, 350
  - galactose oxidase, 112f, 117
  - Galilei, Galileo, 429
  - Gauss double integral, 197
  - Gauss' law, 338
  - Gaussian distribution, 392
  - Gaussian random variates, 403–405
    - Box/Muller/Marsaglia method, 405
    - Odeh/Evans method, 404–405
  - gene chips, 39b
  - gene therapy, 67–68, 68b
  - genetic code, 131, 132t
  - genetic engineering, 69–72
  - genome
    - content, 182–184, 185t
    - non-coding, 28
    - sequencing, 25–39
      - Arabidopsis thaliana*, 27–28
  - Bacillus subtilis*, 25
  - Haemophilus influenzae*, 7t, 25, 33
  - M. tuberculosis*, 25, 92
  - fruitfly (*Drosophila*), 8t, 26–27, 33b
  - human, 30–39
  - implications, 35–37
  - mouse, 28–29
  - pufferfish (*Fugu*), 29
  - rice, 29
  - roundworm, 7t, 25–26
  - strategies, 33b
  - yeast (*Saccharomyces cerevisiae*), 7t, 25
- genomics
  - comparative, 26, 29, 35, 42
  - functional, 42, 119
  - structural, 42, 119
- geometry optimization, *see* minimization
- Gerstmann-Straussler-Scheinker disease (GSS), *see* mad cow disease
- glycine amidinotransferase, 121f
- glycosyl rotation, 148–150
- Gosling, Raymond, 17
- gradient vector, 352, 435–436
- GRASP, 342
- Griffith, Fred, 17
- Griffith, J.S., 54
- GroEL/GroES, 51f, 52b
- GROMOS, 262
- guanine nucleotide-binding protein, 109f, 115
- hairpin, 218f
- harmonic oscillator, 619h
- Hartree-Fock, 242–244, 324
- Hauptman, Herbert, 19
- heat equation, 616h
- Heisenberg, Werner, 3
- helical parameters, 150–155
- helix-loop-helix-motif
  - EF-hand family, 94
- hemoglobin, 46, 78, 94, 106, 119
- hemophilia, 67
- Heraclitus of Ephesus, 41q
- Hershey, Alfred, 7t, 17
- Hessian matrix, 353
  - positive-definite, 353
- hexagonal prism, *see* periodic boundary conditions

- Hexameric HIV-1 CA, 122f  
 histone, 187f, 185–188  
 Hitchings, George, 58  
 HIV integrase, 523  
 HIV protease, 59f, 59, 521, 523  
 HIV reverse transcriptase, 59f, 59  
 Hoffman, Roald, xiii  
 Hofstadter, Douglas R., 301q  
 Holley, Robert, 7t  
**Homework Assignments, 571–621**  
 homework assignments  
     in syllabus, 557–558  
 homology, 589h  
     analysis tools  
         BLAST, 590h  
 Honig, Barry, 343  
 Hooke's law, 273  
 Horvitz, Robert, 26  
 Hugo, Victor, 238q  
 Human Genome Project, 7t, 30–39  
     sequencing strategies, 33b  
 human growth hormone, 114  
 human lysozyme, 106  
 Huxley, Thomas Henry, 105q  
 hybrid Monte Carlo, 420–422  
 HydB/Nqo4-like, 122f  
 hydrogen bond, 80b  
     acceptor, 80b  
     donor, 80b  
 Hoogsteen, 207, 209f  
     in DNA, 130, 135, 140, 206–210  
     reverse Hoogsteen, 208  
     reverse Watson-Crick, 207  
     strength, 80b  
     water, 81f  
 Watson-Crick, 136–137, 138f, 206–207  
 hydrophilic, 51n  
 hydrophobic, 51n  
 immunoglobulins, 96f  
 implicit integration, 496–503  
     computational time, 498–499  
     damping, 498  
     Euler, 498f, 497–499, 619h  
     midpoint, 499–500, 500b, 500f  
     resonance, 499–502  
     symplectic, 502b  
*in silico* biology, 239  
 inclination, 153f, 152–153  
 Ingenhausz, Jan, 489  
 Ingram, Vernon, 530b  
 Insight package, xvii, 574h, 577h  
     in syllabus, 557  
 integer programming, 349  
 interdisciplinary collaborations, 2, 240  
 ion channel proteins, 239  
 isothermal compressibility, 457, 458b  
 Janus, 50b  
 Joachim, Frank, 711  
 Jorgensen, William, 11  
 Kagay, Michael R., 386q  
 Karle, Jerome, 19  
 Karlin, Samuel, 266q  
 Karplus, Martin, xvii, 7t, 11, 262  
 Lord Kelvin, 387  
 Kendrew, John, 7t, 15, 16, 19  
 Kepler, Johannes, 429  
 Khorana, Gobind, 7t  
 Kirkwood-Riseman approximation, 494b,  
     494  
 Klug, Aaron, 189  
 knots, 103  
 Kohn, Walter, 238  
 Kollman, Peter, xvii, 11, 262  
 Kornberg, Roger, 186  
 L'Ecuyer, Pierre  
     combined generators, 399  
     expert advice, 402  
     long-period generators, 399  
     reviews, 389  
     website, 402  
 Langevin dynamics, 335–337, 479–484  
     diffusion constant, 337  
     diffusive limit, 337–338  
     discretization, 482, 484f  
     effect of  $\gamma$ , 336, 481f, 480–481,  
         486–488f  
     equilibration, 439  
     exercise, 619h  
     LN integrator, 484f, 485b, 486f,  
         482–486, 487, 488f  
     mollified method, 479  
     parameterization of  $\gamma$ , 336–337,  
         480–481  
     relation to thermostat coupling, 456

- resonance, 477f
- versus Newtonian, 483f
- Langevin piston, 458
- de Laplace, Pierre Simon, 431, 432b, 432q, 464q
- Laplace's equation, 320b
- Laplacian operator, 319, 320b, 617h
- leapfrog integrator, *see* Verlet integrator
- Lehmer, D. H., 392
- Leibowitz Schmidt, Shira, xiii
- lentivirus, 68b
- Levinthal, Cyrus, 252b
- Levitt, Michael, 245f
- Lifson, Shneior, 7t, 10, 248
- linear programming, 350
- linking number, 196f, 195–196
- Liouville operator, 475
- Los Alamos pioneers, 387
- Luria, Salvador, 17
- Lyapunov exponent, 440n
- Lyapunov instability, 440
- lysozyme, 96f, 121f
- MacLeod, Colin, 17
- mad cow disease, 53–54, 55b
- Maddox, John, 240
- Maddox, Sir John Royden, 426q
- maltate dehydrogenase, 118, 120f
- Markov chain, 414
- Marsaglia, G., 395
- materials
  - designed, 74
- matrix
  - dense, 353
  - indefinite, 353
  - negative-definite, 353
  - positive semi-definite, 353
  - positive-definite, 353
  - sparse, 353
  - symmetric, 350
- McCammon, Andrew, 343
- McCarty, Maclyn, 17
- Mendel, Gregor, 7t
- metabolome, 43
- Metropolis algorithm, 411–416
- Mezei, Mihaly, 314
- MgtE membrane domain-like, 123f
- microarrays, 39b, 526, 530b
- minimization, 428t
- algorithm
  - conjugate gradient, 244, 372–374
  - convergence criteria, 362–363
  - descent algorithm, 359–363
  - descent direction, 359–360
  - interior methods, 349
  - line search, 359–363
  - Newton's method, 363–369, 618h
  - quasi-Newton, 370–372
  - truncated Newton, 374–376
  - trust-region methods, 361–362
- DNA/adduct, 430f
- exercise — advanced, 616h
- exercise — biphenyl structure, 606h
- exercise — pentapeptide global
  - minimum, 612h
- for large-scale functions, 369–383
- recommendations, 380–383
- software, 378t, 378–380
- with restraints, 608h
- minimum
  - global, 616h
  - local and global, 351–352
- MMFF, 262
- molecular biology
  - momentum, 130–132
- Molecular Dynamics**, 425–519
  - molecular dynamics
    - accuracy, 447
    - as refinement tool, 436
    - as sampling tool, 436
    - background reading, 427–428
    - butane simulation, 442f, 444f
    - chaos, 441f, 440–441
    - computational complexity, 443–444, 446–448
    - constrained, 447–448, 453–455
    - convergence of properties, 441–442, 442f, 444f
    - degrees of freedom, 438
  - Desmond, 11
  - ensembles, 439, 455–461
  - equilibration, 439, 439f
  - exercise — advanced, 619h
  - explicit-Euler, 619h
  - historical perspective, 12–14, 15, 16f
  - implicit integrators, 496–503
    - computational time, 498–499
    - damping, 498

- Euler, 498f, 497–499, 619h  
 midpoint, 500f, 499–500, 500b  
 symplectic, 501–502, 502b  
 initial conditions, 437–439  
 integration, 445–446, 448–461  
 Laplace’s vision, 429–432  
 limitations, 432–435  
 long simulations, 514–517  
 multiple timesteps (MTS), 429, 442f,  
     447–448, 476f, 472–478  
 multiple trajectories, 515  
 NAMD, 11  
 natural timescale, 619h  
 overview, 435–448  
 pathway generation, 517  
 RATTLE integrator, 454  
 resonance, 472f, 470–472, 477f,  
     475–479, 499–502  
 Respa integrator, 474  
 Runge-Kutta integrator, 446f  
 setup, 437–439  
 SHAKE integrator, 453–454  
     in MTS, 482, 485b  
 stability, 447  
 stability limit, 467–468, 468t, 471, 473t  
 symplectic methods, 429, 446f,  
     445–446, 454, 465–472  
 targeted, 428t, 430f, 508  
 TATA/TBP, 430f  
 temperature (kinetic), 438  
 timescales, 434t, 436–437  
 trajectory analysis, 445  
 Trotter factorization, 475  
 Verlet integrator, 439, 440, 446f, 446,  
     448–461, 467–479  
     effective frequency, 470  
     generalized, 482  
     linear stability, 467–468, 468t,  
         471–472  
     MTS variants, 473–479  
     phase-space rotation, 470f  
     resonance, 473t, 501  
     water simulation, 446f  
 molecular geometry, 260–262, 602h  
     biphenyl example, 606h  
 molecular mechanics, 247, 256–264  
     configuration space, 258–259  
     functional form, 259–260  
 molecular geometry, 260–262  
 pioneers, 8–11  
 underlying principles, 251–255  
     additivity, 252–254  
     thermodynamics, 251–252  
     transferability, 254–255  
 molecular modeling  
     assessment, 5  
     current progress, 573h  
     early literature, 573h  
     evolution, 9t  
     failures, 604h  
     in rational drug design, 523–524  
     Insight package, xvii, 574h, 577h  
     introduction, 2–5  
     successes, 604h  
 molecular replacement, 20  
 molten globule, 51, 52b  
 Monte Carlo, 428t  
     as sampling tool, 503  
     batch mean, 410–411  
     biased, 419–420  
     DNA applications, 195f  
     error bars, 388, 409–410  
     exercise — advanced, 616h  
     hybrid, 420–422  
     illustration, 430f  
     mean, 406–411, 613h  
 Monte Carlo distributions, 417f  
**Monte Carlo Methods**, 385–425  
 Monte Carlo moves, 416f, 415–418  
 Monte Carlo sampling, 417f, 411–418  
     detailed balance, 411–412  
     ergodicity, 411–412  
     importance sampling, 413–418  
 Morgan, Thomas , 7t  
 Mullis, Kary, 7t, 23  
 multi-wavelength anomalous diffraction  
     (MAD), 20  
 multidisciplinary collaborations, 2, 240  
 multiple isomorphous replacement (MIR),  
     15, 19  
 multiple timesteps (MTS), 429, 442f,  
     447–448, 476f, 472–478  
     extrapolation, 472–475  
     impulses, 472–475  
     resonance, 476f, 477f, 475–479  
 multipole method, 328b, 324–333  
     biomolecular applications, 324f, 332  
     comparison to PME, 324f, 332

- complexity, 329–330
- domain decomposition, 324–326, 326f
- multipole expansion, 325, 327, 328b
- spherical coordinates, 330–332
- variants, 333
- Multivariate Minimization**, 345–385
  - muramidase, 115
  - myoglobin, 46, 78, 94, 96f, 106, 109f, 115
  - myohemerythrin, 108f, 114
  - myosin, 126
  - nanotechnology, 240
  - Nathans, Daniel, 7t, 22
  - Newton equations, 369
  - Newton’s method, 363–369
    - classic example, 366–367
    - historical perspective, 364
    - multivariate version, 368–377
    - one-dimensional version, 364–367
  - Newton, Sir Isaac, 429
  - Newtonian physics, 426–437
    - limitations, 432–435
  - Nirenberg, Marshall, 7t
- Nonbonded Computations**, 299–344
  - nonbonded computations
    - CPU time, 302f, 303t
    - normal modes, 267–272, 438, 444
    - NP-complete, 349, 538
    - nuclear magnetic resonance (NMR)
      - high resolution, 21
      - NOEs, 21
      - residual dipolar coupling, 22
      - technique details, 20–22
  - Nuclear Overhauser Effects (NOEs), 21
  - nucleic acid
    - atom labeling, 141–142
    - backbone torsion analysis, 147f
    - building blocks, 137f, 135–142
      - nitrogenous bases, 135
      - nucleoside, 137, 140
      - nucleotide, 137, 140
    - bulge, 210, 216, 218f
    - dihedral angle, 585h
    - flexibility, 142–155
      - backbone torsions, 145, 147f
      - base-pair, 153f, 155f, 151–155, 165–169
      - glycosyl rotation, 149f, 148–150
    - helical parameters, 150–155
    - sugar, 143–145, 146f, 149f
    - hairpin, 218f
    - loop, 210, 218f
    - polynucleotide chain, 139f, 137–142
    - pseudorotation, 143f, 144f, 145f, 143–145
    - stacking, 140, 157, 166–168, 168b, 206, 213b, 216, 224
    - sugar pucker analysis, 146f
    - torsion-angle labeling, 142t, 142
  - Nucleic Acid Database (NDB), 133, 135, 165, 206, 572h, 582h, 584h, 589, 590h
    - backbone torsion analysis, 147f
    - sugar pucker analysis, 146f
  - Nucleic Acids Structure**, 129–237
    - nucleosome, 4n, 189f, 186–193, 239
      - H3 K79 dimethylated nucleosome, 190f
      - poly(dA•dT) nucleosome, 190f
      - tetrarnucleosome, 190f
    - nutraceuticals, 71b, 91
    - nutrigenomics, 39b, 72–73, 530b
    - nutritional genomics, *see* nutrigenomics
  - OPLS, 262
  - optimization, *see* minimization
  - Overton, Michael, xv
  - p53* tumor suppressor protein
    - DNA binding, 181t
  - p53* gene
    - mutations, 24b
  - p53* tumor suppressor gene, 27
  - P-DNA, 215b
  - Paracelsus challenge, 47f, 50b
  - partition function, 412
  - Pauling, Linus, 7t, 15, 215, 530b
  - pectin lyase A, 112f, 117
  - peptide, 82f
    - aspartame, 84f, 85
    - Met-enkephalin, 577h, 612h
  - peptide nucleic acid (PNA), 212f, 212, 213b
  - periodic boundary conditions, 311–314
    - cube, 311
    - elongated dodecahedron, 313f
    - face-centered cube, *see* rhombic dodecahedron

- hexagonal prism, 311, 313f  
 minimum-image convention, 312  
 rhombic dodecahedron, 311, 313f  
 squashed dodecahedron, 313f  
 triangular prism, 313f  
 truncated octahedron, 311, 313f  
 persistence length, 187f, 198–202  
 personalized medicine, *see*  
     pharmacogenomics  
 Perutz, Max, 7t, 15, 16, 19  
 pharmacogenomics, 39b, 43, 74, 526, 530,  
     531b  
 phosphocarrier protein, 121f  
 photosynthetic assembly, 126  
 Photosystem I reaction center subunit XI,  
     PsaL, 123f  
 photosystem I, 126  
 $\pi$   
     Monte Carlo estimates, 387, 406–407,  
         407f, 408t  
 Picasso, Pablo, 3q  
 pix, 108f, 115  
 Plant proteinase inhibitors, 124f  
 Poincaré, Jules Henri, 432  
 Poisson Boltzmann  
     contours, 189f, 343f, 430f  
 Poisson’s ratio, 199n  
 Poisson-Boltzmann equation, 338–343  
     Debye-Hückel approximation, 341f,  
         340–342  
     DelPhi, 342, 343  
     GRASP, 342  
     UHBD, 342, 343  
 polymer  
     polyethylene, 601h  
 polymerase chain reaction (PCR), 22–23,  
     25b  
 polypeptide, 84f, 85  
 Pople, John, 238, 241  
 potential of mean force, 334–335  
 presidential elections, 555q  
 pressure tensor, 455, 458, 458b  
 principal component analysis (PCA), 40b,  
     540–542  
 prion, 53–54, 55b, 57f, 117  
 programming, xiii, xvii, 599h  
     in syllabus, 558  
 propeller-twisting, 140, 155f, 154–155,  
     165–166
- PROSITE, 126  
 protease inhibitors, *see* AIDS  
 protein  
     chaperones, 51f, 50–51, 52b  
     class, 113–114  
     dynamics, 9t, 12–14, 15, 16f, 246b, 246  
     fibrous, 78, 91f, 91–92  
         collagen, 91f, 91  
         silk, 88, 91f, 91–92  
     flexibility, 97–103  
          $\omega$  angle, 98, 99f  
          $\phi$  angle, 98, 99f  
          $\psi$  angle, 98, 99f  
         Ramachandran plot, 101f, 102f,  
             99–103, 575h, 580h  
     fold, 114–119  
          $\alpha$ , 108, 109f, 114–115  
          $\alpha+\beta$ , 117–118, 121f  
          $\alpha/\beta$ , 117–118, 120f  
          $\beta$ , 111, 112f, 115–117  
         membrane and cell-surface, 123f  
         multi-domain, 122f  
         number, 118–119  
     folding, 4, 6, 38, 46–53, 57, 252b  
     aggregates, 50, 51, 52b, 54, 56  
     CASP, 48–49  
     challenge, 48–49, 50b  
     compared to nucleic-acid folding,  
         133  
     intermediates, 46, 51, 52b  
     Levinthal paradox, 46–48, 252b,  
         621h  
     molten globule, 51, 52b  
     old and new views, 46–48, 252b  
     Paracelsus challenge, 47f, 50b  
     problem solvability, 49–50, 427,  
         621h  
     timescale, 49–50, 621h  
     function, 56–57  
     globular, 78, 90  
     Greek root, 77  
     knot, 103  
     machinery of life, 77–78  
     membrane, 90, 91, 552  
     misfolding, 6, 53–56  
     muscle, 126  
     sequence  
         growth, 44f  
     sequence databases, 44f, 45t

- sequence variability, 89f, 90t, 89–92
- small, 124f
- structure, 56, 90–103, 105–127
  - $\alpha$ -helix, 107f, 106–107, 109
  - $\beta$ -sheet, 107f, 110
  - challenge, 50b
  - classification, 126–127
  - collagen-helix, 110
  - helices, 106–110
  - homology, 589h
  - knot, 103
  - loops, 110–112
  - $\omega$  angle, 98, 99f
  - $\phi$  angle, 98, 99f
  - $\pi$ -helix, 107–108
  - primary, 103
  - $\psi$  angle, 98, 99f
  - quaternary, 103, 119–126
  - Ramachandran plot, 101f, 102f, 99–103, 575h, 580h
  - rotamer, 99f, 99, 100f, 577h
  - secondary, 103, 106–119
  - sheets, 110
  - similarity, 46, 89, 92–95, 95, 96f
  - supersecondary, 113–119
  - tertiary, 103, 113–119
    - $3_{10}$ -helix, 107–108
    - torsional rotation, 578h, 580h
    - turns, 110–112
- protein**, 77
  - unstructured, 52
  - variability, 78–79
- Protein Data Bank (PDB), 43, 45t, 165, 206, 572h, 589, 590h
  - file format, 575h
  - growth, 44f, 43–44, 45t
  - NMR-structure fraction, 44
- protein kinase, 114
- protein sequence
  - growth, 45t
- Protein Structure**, 77–127
- proteinase inhibitor PMP-C, 114
- Proteins**, 77–127
- proteomics, 42
  - Human Proteomics Project, 42
- Prusiner, Stanley, 53
- pseudoknot, 217, 219, 220f, 225
  - pseudorotation, 144f, 145f, 143–145
    - envelope, 143f
    - twist, 143f
  - pseudorotation path, 259
  - purine, 135
  - pyrimidine, 135
- quadratic programming, 350
- quantitative structure/activity relationships (QSAR), 65, 536
- quantum mechanics, 3, 241–247
  - ab initio* methods, 242–247
  - Born-Oppenheimer approximation, 242–247, 249
  - density functional theory, 243
  - DNA application, 245, 246b, 248f
  - DNA polymerase application, 250f
  - Hartree-Fock, 242–244, 324
  - QM/MM concept, 245f, 245–246
  - Schrödinger equation, 241–247
  - semi-empirical methods, 242–247
  - quasi-Newton methods, 370–372
    - BFGS, 371–372
    - QN condition, 370–371
- Rahman, Aneesur, 10
- Ramachandran plot, 101f, 102f, 99–103, 575h, 580h
- Ramachandran, G.N., 101
- Ramakrishnan, Venkatraman, 18
- random number generators, 388–405
  - artifacts, 400, 400b, 402b, 613h
  - combination, 398–399
  - drand48**, 394, 397f, 407, 408t
  - exercise, 613h
  - Fibonacci series, 396–398
  - Gaussian variates, 613h
  - independence, 389–390
  - lattice structure, 396f, 395–396, 397f, 613f, 613h
  - linear congruential, 392–396, 613h
  - period, 389–390
  - portability, 391
  - rand**, 394, 407, 408t
  - rand48**, 397f
  - random variates, 613h
  - RANDU**, 396, 613h
  - recommendations, 401–402
  - shift-register, 398

- spectral tests, 395
- speed, 391
- SPRNG, 402
- SURAND, 393–394, 397f, 402b
- timings, 408t
- uniformity, 389–390
- random variates
  - from other distributions, 390, 403
  - Gaussian, 403–405
    - Box/Muller/Marsaglia method, 405
    - Odeh/Evans method, 404–405
  - uniform, 388–402
- replication, 130
- Research Collaboratory for Structural Bioinformatics (RCSB), 44, 45t, 165, 206, 572h, 590h, 592h
- residual dipolar coupling, 22
- Resistin, 124
- resonance artifacts, 472f, 470–472
  - in implicit methods, 499–502
  - in impulse-MTS, 475–476
  - in MTS, 477f, 476–479
- Respa integrator, 474
- restraints
  - in minimization, 608h
- retrovirus, 59, 68b
- reverse transcriptase inhibitors, *see* AIDS
- rhombic dodecahedron, *see* periodic boundary conditions
- ribonuclease, 103
- ribonuclease inhibitor, 118, 120f
- ribosome, 4n, 123–125, 131, 239, 343f
- Rich, Alexander, 156
- ricin, 71
- rise, 154
- RNA
  - Tetrahymena* group I intron, 220f, 223, 224b, 227
  - in vitro* technology, 219–221
  - aptamers, 220
  - first genetic material, 213b
  - folding, 225–227
  - genes, 28
  - hairpin ribozyme, 218f
  - hammerhead ribozyme, 218f, 223b, 223
  - Hepatitis Delta virus (HDV) ribozyme, 220f, 223b
  - HIV-1 hairpin, 218f
  - interference (RNAi), 26, 221–222
  - micro (miRNA), 221–222
  - non-coding (ncRNA), 221–222
  - pseudoknot, 217, 219, 220f, 225
  - riboswitch, 222, 223
  - ribozyme, 219, 220f, 221, 223b
  - roles, 217t, 216–223
  - silencing, 221–222
  - small-interfering (siRNA), 221–222
  - structure, 223b, 224f, 226f, 227f, 216–227, 228, 229f
  - world, 213, 219n
- RNA structure
  - RNA Ontology Consortium, 210, 225
- roll, 154, 155f, 165–169
- Rop (repressor of primer), 93f, 92–95, 106, 114
- Rossmann, Michael, 118
- rotamer, 99f, 99, 100f, 577h
- rubredoxin, 114
- S-DNA, 214, 215b
- sampling
  - enhanced, 479, 498, 503–513, 515
  - limited, 470, 501
- Samuelson, Paul A., 386q
- sarcoplasmic calcium-binding protein, 94
- SARS, 30, 36
- satellite panicum mosaic virus, 111f, 116
- Schaefer, Henry, 238
- Schellman, John, 101
- Scheraga, Harold, xvii, 7t, 10, 262, 419, 596h, 616h
- Schrödinger equation, 241–247
- Schulten, Klaus, 474
- sea anemone toxin k, 114
- semi-empirical methods, 242–247
- sequence similarity, 46, 89, 92–95, 95, 96f
  - challenge, 50b
- severe combined immune deficiency (SCID), 67
- shift, 154
- sickle-cell anemia, 67, 94, 530b
- silk, 88, 91f, 91
- simulated annealing, 387, 415
  - exercise, 616h
- singular value decomposition (SVD), 542–544
- Slater-Kirkwood, 290
- slide, 154, 155f, 168b, 165–169

- Smith, Hamilton, 7t, 22  
 Smith, Michael, 7t, 23  
 SNPs (single nucleotide polymorphisms), 39b, 74, 530, 531b  
 soluble lytic transglycosylate, 109f, 115  
 squashed dodecahedron, *see* periodic boundary conditions  
 statistical ensembles, 412–413  
 Steitz, Thomas, 18  
 Stillinger, Frank, 7t, 10  
 stochastic dynamics, 335–338, *see* Langevin and Brownian dynamics  
 stochastic path approach, 428t  
 Stokes' law, 336–337, 480–481, 489, 494b, 494, 615h  
 stress, *see* pressure tensor  
 structural biology, 41  
 chronology, 9t  
 Structural Classification of Proteins (SCOP), 126–127, 589, 590h  
 structure/activity relationships (SAR), 536  
 Sulston, John, 26  
 Sumner, James, 14  
 symplectic methods, 429, 446f, 445–446, 454, 465–472  
 effective rotation, 469, 470f  
 phase-space transformation, 467–469
- T4 lysozyme, 95, 106  
 TATA-box, 169b, 172f, 182, 207  
 hydration, 179f  
 TATA-box binding protein (TBP)  
 DNA binding, 172f, 175b, 182, 215  
 DNA binding and mutation fragility, 57  
 DNA unwinding, 169b, 215  
 transcription activity, 182  
 tautomeric forms, 210  
 tautomerization, 18n, 135n, 209  
 technology advances, 22–23  
 NMR, 22–23  
 PCR, 22–23, 25b  
 recombinant DNA, 22  
 site-directed mutagenesis, 23  
 X-ray crystallography, 14–20, 22–23  
 telomeres, 211–212  
 Thomas, Jean, 186  
 tilt, 154, 155f, 165–169  
 timescales  
 in biomolecules, 434t, 436–437, 468  
 tip, 153f, 152–153  
 titin, 78  
 tobacco mosaic virus, 114  
 Tolstoy, Lev, 464q  
 tomato bushy stunt virus, 125f  
 torsion angle, *see* dihedral angle  
 transcription, 130, 131n  
 translation, 131n  
 triangular prism, *see* periodic boundary conditions  
 triosephosphate isomerase (TIM), 117, 120f  
 Trp repressor, 106  
 truncated Newton methods, 374–376  
 truncated octahedron, *see* periodic boundary conditions  
 trust-region methods, 361–362  
 twist, 155f, 165–169, 196f, 196–197
- UHBD, 342, 343  
 uniform random variates, 388–402
- V-type ATP synthase subunit C, 123f  
 vegetarian diet, 83, 83b  
 Venter, Craig, 30, 33b  
 Verlet integrator, 439, 448–461, 467–479  
 effective frequency, 470  
 generalized, 482  
 leapfrog variant, 449–453  
 linear stability, 467–468, 468t, 471–472  
 MTS variants, 473–479  
 phase-space rotation, 470f  
 position variant, 449–453  
 resonance, 472f, 473t, 501  
 velocity variant, 449–453  
 vibrational frequencies, *see* normal modes, 269f, 270t, 271f, 271n, 277, 282, 295, 434t, 468, 473t  
 characteristic period, 473t  
 coupling, 497  
 harmonic oscillator, 433–434  
 Langevin damping, 480  
 spectral densities, 485b, 487, 488f  
 virial, 458b, 460  
 virus, 119–122, 125f  
 bluetongue, 122  
 polio, 120  
 tobacco mosaic, 120  
 tomato bushy stunt, 121

- Wang, Andrew, 156  
Warshel, Arieh, 245f  
water  
  dynamics, 79b  
  hydration shells, 79b, 176  
  hydrogen bonds, 79b, 81f  
  SPC potential, 619h  
  spine of hydration, 176, 178  
  structure, 79b, 81f  
Watson, James, 7t, 17–18, 130, 135, 150, 155  
Westheimer, Frank, 7t  
Wheeler, John Archibald, 426q  
White, Jim, 195  
Wilkins, Maurice, 7t, 17, 130  
Wilson angle, 286  
Wilson, Edward, 2q, 3  
Wolff algorithm, 402  
Wolynes, Peter, 443q  
writhe, 196f, 197  
writhing number, *see* writhe  
X-ray crystallography, 18–20  
  diffraction pattern, 19  
  historical perspective, 14–18  
  MAD, 20  
  MIR, 19  
  molecular replacement, 20  
  phase problem, 19–20, 20b  
  technique details, 18–20  
  time-resolved, 20  
  unit cell, 20b  
Yang, Weitao, 246b, 249f  
Yonath, Ada, 18  
York, Darren, 246b, 248f