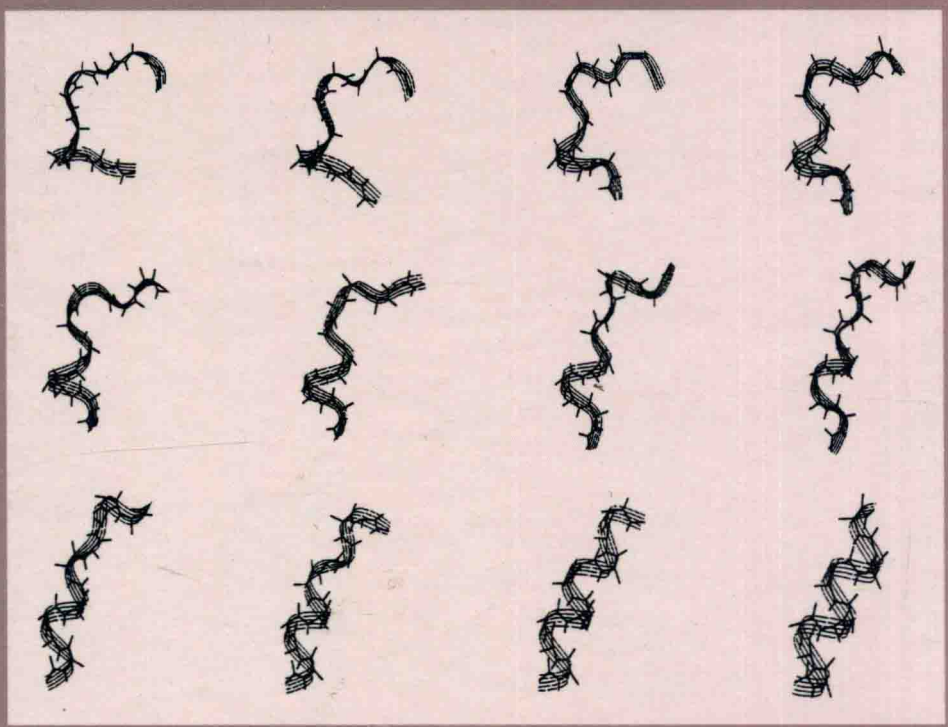
COMPUTATIONAL BIOCHEMISTRY AND BIOPHYSICS



edited by
Oren M. Becker
Alexander D. MacKerell, Jr.
Benoît Roux
Masakatsu Watanabe

COMPUTATIONAL BIOCHEMISTRY AND BIOPHYSICS

edited by

Oren M. Becker
Tel Aviv University
Tel Aviv, Israel

Alexander D. MacKerell, Jr.
University of Maryland
Baltimore, Maryland

Benoît Roux
Weill Medical College of Cornell University
New York, New York

Masakatsu Watanabe
Wavefunction, Inc.
Irvine, California

MARCEL DEKKER, INC. NEW YORK • BASEL

ISBN: 0-8247-0455-X

This book is printed on acid-free paper.

Headquarters
Marcel Dekker, Inc.
270 Madison Avenue, New York, NY 10016
tel: 212-696-9000; fax: 212-685-4540

Eastern Hemisphere Distribution
Marcel Dekker AG
Hutgasse 4, Postfach 812, CH-4001 Basel, Switzerland
tel: 41-61-261-8482; fax: 41-61-261-8896

World Wide Web
http://www.dekker.com

The publisher offers discounts on this book when ordered in bulk quantities. For more information, write to Special Sales/Professional Marketing at the headquarters address above.

Current printing (last digit):
10 9 8 7 6 5 4 3 2 1

PRINTED IN THE UNITED STATES OF AMERICA

COMPUTATIONAL BIOCHEMISTRY AND BIOPHYSICS

Foreword

The long-range goal of molecular approaches to biology is to describe living systems in terms of chemistry and physics. Over the last 70 years great progress has been made in applying the quantum mechanical equations representing the underlying physical laws to chemical problems involving the structures and reactions of small molecules. This work was recognized in the awarding of the Nobel Prize in Chemistry to Walter Kohn and John Pople in 1998. Computational studies of mesoscopic systems of biological interest have been attempted only more recently. Classical mechanics is adequate for describing most of the properties of these systems, and the molecular dynamics simulation method is the most important theoretical approach used in such studies. The first molecular dynamics simulation of a protein, the bovine pancreatic trypsin inhibitor (BPTI), was published more than 20 years ago [1]. Although the simulation was ''crude'' by present standards, it was important because it introduced an important conceptual change in our view of biomolecules. The classic view of biopolymers, like proteins and nucleic acids, had been static in character. The remarkable detail evident in the protein crystal structures available at that time led to an image of ''rigid'' biomolecules with every atom fixed in place [2]. The molecular dynamics simulation of BPTI was instrumental in changing the static view of the structure of biomolecules to a dynamic picture. It is now recognized that the atoms of which biopolymers are composed are in a state of constant motion at ordinary temperatures. The X-ray structure of a protein provides the average atomic positions, but the atoms exhibit fluidlike motions of sizable amplitudes about these averages. The new understanding of protein dynamics subsumed the static picture in that the average positions are still useful for the discussion of many aspects of biomolecule function in the language of structural chemistry. The recognition of the importance of fluctuations opened the way for more sophisticated and accurate interpretations of functional properties.

In the intervening years, molecular dynamics simulations of biomolecules have undergone an explosive development and been applied to a wide range of problems [3,4]. Two attributes of molecular dynamics simulations have played an essential role in their increasing use. The first is that simulations provide individual particle motions as a function of time so they can answer detailed questions about the properties of a system, often more easily than experiments. For many aspects of biomolecule function, it is these details

that are of interest (e.g., by what pathways does oxygen get into and exit the heme pocket in myoglobin? How does the conformational change that triggers activity of ras p21 take place?). The second attribute is that, although the potential used in the simulations is approximate, it is completely under the user's control, so that by removing or altering specific contributions to the potential, their role in determining a given property can be examined. This is most graphically demonstrated in the calculation of free energy differences by "computer alchemy" in which the potential is transmuted reversibly from that representing one system to another during a simulation [5].

There are three types of applications of molecular dynamics simulation methods in the study of macromolecules of biological interest, as in other areas that use such simulations. The first uses the simulation simply as a means of sampling configuration space. This is involved in the utilization of molecular dynamics, often with simulated annealing protocols, to determine or refine structures with data obtained from experiments, such as X-ray diffraction. The second uses simulations to determine equilibrium averages, including structural and motional properties (e.g., atomic mean-square fluctuation amplitudes) and the thermodynamics of the system. For such applications, it is necessary that the simulations adequately sample configuration space, as in the first application, with the additional condition that each point be weighted by the appropriate Boltzmann factor. The third area employs simulations to examine the actual dynamics. Here not only is adequate sampling of configuration space with appropriate Boltzmann weighting required, but it must be done so as to properly represent the time development of the system. For the first two areas, Monte Carlo simulations, as well as molecular dynamics, can be utilized. By contrast, in the third area where the motions and their development are of interest, only molecular dynamics can provide the necessary information. The three types of applications, all of which are considered in the present volume, make increasing demands on the simulation methodology in terms of the accuracy that is required.

In the early years of molecular dynamics simulations of biomolecules, almost all scientists working in the field received specialized training (as graduate students and/or postdoctoral fellows) that provided a detailed understanding of the power and limitations of the approach. Now that the methodology is becoming more accessible (in terms of ease of application of generally distributed programs and the availability of the required computational resources) and better validated (in terms of published results), many people are beginning to use simulation technology without training in the area. Molecular dynamics simulations are becoming part of the "tool kit" used by everyone, even experimentalists, who wish to obtain an understanding of the structure and function of biomolecules. To be able to do this effectively, a person must have access to sources from which he or she can obtain the background required for meaningful applications of the simulation methodology. This volume has an important role to play in the transition of the field from one limited to specialists (although they will continue to be needed to improve the methodology and extend its applicability) to the mainstream of molecular biology. The emphasis on an in-depth description of the computational methodology will make the volume useful as an introduction to the field for many people who are doing simulations for the first time. They will find it helpful also to look at two earlier volumes on macromolecular simulations [3,4], as well as the classic general text on molecular dynamics [6]. Equally important in the volume is the connection made with X-ray, neutron scattering, and nuclear magnetic resonance experiments, areas in which molecular dynamics simulations are playing an essential role. A number of well-chosen "special topics" involving applications of simulation methods are described. Also, several chapters broaden

the perspective of the book by introducing approaches other than molecular dynamics for modeling proteins and their interactions. They make the connection with what many people regard—mistakenly, in my view—as "computational biology." Certainly with the announced completion of a description of the human genome in a coarse-grained sense, the part of computational biology concerned with the prediction of the structure and function of gene products from a knowledge of the polypeptide sequence is an important endeavor. However, equally important, and probably more so in the long run, is the biophysical aspect of computational biology. The first set of Investigators in Computational Biology chosen this year demonstrates that the Howard Hughes Foundation recognized the importance of such biophysical studies to which this volume serves as an excellent introduction.

I am very pleased to have been given the opportunity to contribute a Foreword to this very useful book. It is a particular pleasure for me to do so because all the editors and fifteen of the authors are alumni of my research group at Harvard, where molecular dynamics simulations of biomolecules originated.

REFERENCES

1. JA McCammon, BR Gelin, and M Karplus. Nature 267:585, 1977.
2. DC Phillips. In: RH Sarma, ed. Biomolecular Stereodynamics, II. Guilderland, New York: Adenine Press, 1981, p 497.
3. JA McCammon and S Harvey. Dynamics of Proteins and Nucleic Acids. Cambridge: Cambridge University Press, 1987.
4. CL Brooks III, M Karplus, and BM Pettitt. Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics. New York: John Wiley & Sons, 1988.
5. For an early example, see J Gao, K Kuczera, B Tidor, and M Karplus. Science 244:1069–1072, 1989.
6. MP Allen and DJ Tildesley. Computer Simulations of Liquids. Oxford: Clarendon Press, 1987.

Martin Karplus
Laboratoire de chimie Biophysique, ISIS
Université Louis Pasteur
Strasbourg, France
and
Department of Chemistry and Chemical Biology
Harvard University
Cambridge, Massachusetts

Preface

The first dynamical simulation of a protein based on a detailed atomic model was reported in 1977. Since then, the uses of various theoretical and computational approaches have contributed tremendously to our understanding of complex biomolecular systems such as proteins, nucleic acids, and bilayer membranes. By providing detailed information on biomolecular systems that is often experimentally inaccessible, computational approaches based on detailed atomic models can help in the current efforts to understand the relationship of the structure of biomolecules to their function. For that reason, they are now considered to be an integrated and essential component of research in modern biology, biochemistry, and biophysics.

A number of books and journal articles reviewing computational methods relevant to biophysical problems have been published in the last decade. Two of the most popular texts, however, were published more than ten years ago: those of McCammon and Harvey in 1987 and Brooks, Karplus, and Pettitt in 1988. There has been significant progress in theoretical and computational methodologies since the publication of these books. Therefore, we feel that there is a need for an updated, comprehensive text including the most recent developments and applications in the field.

In recent years the significant increase in computer power along with the implementation of a wide range of theoretical methods into sophisticated simulation programs have greatly expanded the applicability of computational approaches to biological systems. The expansion is such that interesting applications to important and complex biomolecular systems are now often carried out by researchers with no special training in computational methodologies. To successfully apply computational approaches to their systems of interest, these ''nonspecialists'' must make several important choices about the proper methods and techniques for the particular question that they are trying to address. We believe that a good understanding of the theory behind the myriad of computational methods and techniques can help in this process. Therefore, one of this book's aims is to provide readers with the required background to properly design and implement computational investigations of biomolecular systems. In addition, the book provides the needed information for calculating and interpreting experimentally observed properties on the basis of the results generated by computer simulations.

This book is organized so that nonspecialists as well as more advanced users can benefit. It can serve as both an introductory text to computational biology, making it useful for students, and a reference source for active researchers in the field. We have tried to compile a comprehensive but reasonably concise review of relevant theoretical and computational methods that is self-contained. Therefore, the chapters, particularly in Part I, are ordered so that the reader can easily follow from one topic to the next and be systematically introduced to the theoretical methods used in computational studies of biomolecular systems. The remainder of the book is designed so that the individual parts as well as their chapters can be read independently. Additional technical details can be found in the references listed in each chapter. Thus the book may also serve as a useful reference for both theoreticians and experimentalists in all areas of biophysics and biochemical research.

This volume thus presents a current and comprehensive account of computational methods and their application to biological macromolecules. We hope that it will serve as a useful tool to guide future investigations of proteins, nucleic acids, and biological membranes, so that the mysteries of biological molecules can continue to be revealed.

We are grateful to the many colleagues we have worked with, collaborated with, and grown with over the course of our research careers. The multidimensionality of those interactions has allowed us to grow in many facets of our lives. Special thanks to Professor Martin Karplus for contributing the Foreword of this book and, most important, for supplying the insights, knowledge, and environment that laid the foundation for our scientific pursuits in computational biochemistry and biophysics and led directly to the creation of this book. Finally, we wish to acknowledge the support of all our friends and family.

Oren M. Becker
Alexander D. MacKerell, Jr.
Benoît Roux
Masakatsu Watanabe

Contributors

Oren M. Becker Department of Chemical Physics, School of Chemistry, Tel Aviv University, Tel Aviv, Israel

Thomas A. Darden Laboratory of Structural Biology, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, North Carolina

Roland L. Dunbrack, Jr. Institute for Cancer Research, Fox Chase Cancer Center, Philadelphia, Pennsylvania

András Fiser Laboratories of Molecular Biophysics, The Rockefeller University, New York, New York

Steven Hayward School of Information Systems, University of East Anglia, Norwich, England

Fumio Hirata Department of Theoretical Study, Institute for Molecular Science, Okazaki National Research Institutes, Okazaki, Japan

Toshiko Ichiye School of Molecular Biosciences, Washington State University, Pullman, Washington

Shigeki Kato Department of Chemistry, Kyoto University, Kyoto, Japan

Paul D. Lyne Computer Aided Drug Design, Biogen, Inc., Cambridge, Massachusetts

Alexander D. MacKerell, Jr. School of Pharmacy, University of Maryland, Baltimore, Maryland

Alexey K. Mazur Institut de Biologie Physico-Chimique, CNRS, Paris, France

Francisco Melo Laboratories of Molecular Biophysics, The Rockefeller University, New York, New York

Michael Nilges Structural and Computational Biology Program, European Molecular Biology Laboratory, Heidelberg, Germany

Lennart Nilsson Department of Biosciences at NOVUM, Karolinska Institutet, Huddinge, Sweden

Benoît Roux Department of Biochemistry and Structural Biology, Weill Medical College of Cornell University, New York, New York

Andrej Šali Laboratories of Molecular Biophysics, The Rockefeller University, New York, New York

Roberto Sánchez Laboratories of Molecular Biophysics, The Rockefeller University, New York, New York

Hirofumi Sato Department of Theoretical Study, Institute for Molecular Science, Okazaki National Research Institutes, Okazaki, Japan

Thomas Simonson Laboratory for Structural Biology and Genomics, Centre National de la Recherche Scientifique, Strasbourg, France

Jeremy C. Smith Lehrstuhl für Biocomputing, Interdisziplinäres Zentrum für Wissenschaftliches Rechnen der Universität Heidelberg, Heidelberg, Germany

John E. Straub Department of Chemistry, Boston University, Boston, Massachusetts

Seiichiro Ten-no Graduate School of Information Science, Nagoya University, Nagoya, Japan

Douglas J. Tobias Department of Chemistry, University of California at Irvine, Irvine, California

Alexander Tropsha Laboratory for Molecular Modeling, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

Owen A. Walsh Physical and Theoretical Chemistry Laboratory, Oxford University, Oxford, England

Masakatsu Watanabe* Moldyn, Inc., Cambridge, Massachusetts

Weifan Zheng Laboratory for Molecular Modeling, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

* *Current affiliation*: Wavefunction, Inc., Irvine, California.

Contents

1

Introduction

Oren M. Becker
Tel Aviv University, Tel Aviv, Israel

Alexander D. MacKerell, Jr.
University of Maryland, Baltimore, Maryland

Benoît Roux
Weill Medical College of Cornell University, New York, New York

Masakatsu Watanabe*
Moldyn, Inc., Cambridge, Massachusetts

I. INTRODUCTION

The first hints of the chemical basis of life were noted approximately 150 years ago. Leading up to this initial awareness were a series of insights that living organisms comprise a hierarchy of structures: organs, which are composed of individual cells, which are themselves formed of organelles of different chemical compositions, and so on. From this realization and the observation that nonviable extracts from organisms such as yeast could by themselves catalyze chemical reactions, it became clear that life itself was the result of a complex combination of individual chemicals and chemical reactions. These advances stimulated investigations into the nature of the molecules responsible for biochemical reactions, culminating in the discovery of the genetic code and the molecular structure of deoxyribonucleic acid (DNA) in the early 1950s by Watson and Crick [1]. One of the most fascinating aspects of their discovery was that an understanding of the mechanism by which the genetic code functioned could not be achieved until knowledge of the three-dimensional (3D) structure of DNA was attained. The discovery of the structure of DNA and its relationship to DNA function had a tremendous impact on all subsequent biochemical investigations, basically defining the paradigm of modern biochemistry and molecular biology. This established the primary importance of molecular structure for an understanding of the function of biological molecules and the need to investigate the relationship between structure and function in order to advance our understanding of the fundamental processes of life.

As the molecular structure of DNA was being elucidated, scientists made significant contributions to revealing the structures of proteins and enzymes. Sanger [2] resolved the

* *Current affiliation*: Wavefunction, Inc., Irvine, California.

primary sequence of insulin in 1953, followed by that of an enzyme, ribonuclease A, 10 years later. The late 1950s saw the first high resolution 3D structures of proteins, myoglobin and hemoglobin, as determined by Kendrew et al. [3] and Perutz et al. [4], respectively, followed by the first 3D structure of an enzyme, lysozyme, by Phillips and coworkers [5] in 1965. Since then, the structures of a very large number of proteins and other biological molecules have been determined. There are currently over 10,000 3D structures of proteins available [6] along with several hundred DNA and RNA structures [7] and a number of protein–nucleic acid complexes.

Prior to the elucidation of the 3D structure of proteins via experimental methods, theoretical approaches made significant inroads toward understanding protein structure. One of the most significant contributions was made by Pauling and Corey [8] in 1951, when they predicted the existence of the main elements of secondary structure in proteins, the α-helix and β-sheet. Their prediction was soon confirmed by Perutz [9], who made the first glimpse of the secondary structure at low resolution. This landmark work by Pauling and Corey marked the dawn of theoretical studies of biomolecules. It was followed by prediction of the allowed conformations of amino acids, the basic building block of proteins, in 1963 by Ramachandran et al. [10]. This work, which was based on simple hard-sphere models, indicated the potential of computational approaches as tools for understanding the atomic details of biomolecules. Energy minimization algorithms with an explicit potential energy function followed readily to assist in the refinement of model structures of peptides by Scheraga [11] and of crystal structures of proteins by Levitt and Lifson [12].

The availability of the first protein structures determined by X-ray crystallography led to the initial view that these molecules were very rigid, an idea consistent with the lock-and-key model of enzyme catalysis. Detailed analysis of protein structures, however, indicated that proteins had to be flexible in order to perform their biological functions. For example, in the case of myoglobin and hemoglobin, there is no path for the escape of O_2 from the heme-binding pocket in the crystal structure; the protein must change structure in order for the O_2 to be released. This and other realizations lead to a rethinking of the properties of proteins, which resulted in a more dynamic picture of protein structure. Experimental methods have been developed to investigate the dynamic properties of proteins; however, the information content from these studies is generally isotropic in nature, affording little insight into the atomic details of these fluctuations [13]. Atomic resolution information on the dynamics of proteins as well as other biomolecules and the relationship of dynamics to function is an area where computational studies can extend our knowledge beyond what is accessible to experimentalists.

The first detailed microscopic view of atomic motions in a protein was provided in 1977 via a molecular dynamics (MD) simulation of bovine pancreatic trypsin inhibitor by McCammon et al. [14]. This work, marking the beginning of modern computational biochemistry and biophysics, has been followed by a large number of theoretical investigations of many complex biomolecular systems. It is this large body of work, including the numerous methodological advances in computational studies of biomolecules over the last decade, that largely motivated the production of the present book.

II. OVERVIEW OF COMPUTATIONAL BIOCHEMISTRY AND BIOPHYSICS

Although the dynamic nature of biological molecules has been well accepted for over 20 years, the extent of that flexibility, as manifested in the large structural changes that

biomolecules can undergo, has recently become clearer due to the availability of experimentally determined structures of the same biological molecules in different environments. For example, the enzyme triosephosphate isomerase contains an 11 amino acid residue loop that moves by more than 7 Å following the binding of substrate, leading to a catalytically competent structure [15,16]. In the enzyme cytosine-5-methyltransferase, a loop containing one of the catalytically essential residues undergoes a large conformational change upon formation of the DNA–coenzyme–protein complex, leading to some residues changing position by over 20 Å [17]. DNA, typically envisioned in the canonical B form [18], has been shown to undergo significant distortions upon binding to proteins. Bending of 90° has been seen in the CAP–DNA complex [19], and binding of the TATA box binding protein to the TATAAAA consensus sequence leads to the DNA assuming a unique conformation referred to as the TA form [20]. Even though experimental studies can reveal the end points associated with these conformational transitions, these methods typically cannot access structural details of the pathway between the end points. Such information is directly accessible via computational approaches.

Computational approaches can be used to investigate the energetics associated with changes in both conformation and chemical structure. An example is afforded by the conformational transitions discussed in the preceding paragraph. Conformational free energy differences and barriers can be calculated and then directly compared with experimental results. Overviews of these methods are included in Chapters 9 and 10. Recent advances in techniques that combine quantum mechanical (QM) approaches with molecular mechanics (MM) now allow for a detailed understanding of processes involving bond breaking and bond making and how enzymes can accelerate those reactions. Chapter 11 gives a detailed overview of the implementation and current status of QM/MM methods. The ability of computational biochemistry to reveal the microscopic events controlling reaction rates and equilibrium at the atomic level is one of its greatest strengths.

Biological membranes provide the essential barrier between cells and the organelles of which cells are composed. Cellular membranes are complicated extensive biomolecular sheetlike structures, mostly formed by lipid molecules held together by cooperative noncovalent interactions. A membrane is not a static structure, but rather a complex dynamical two-dimensional liquid crystalline fluid mosaic of oriented proteins and lipids. A number of experimental approaches can be used to investigate and characterize biological membranes. However, the complexity of membranes is such that experimental data remain very difficult to interpret at the microscopic level. In recent years, computational studies of membranes based on detailed atomic models, as summarized in Chapter 21, have greatly increased the ability to interpret experimental data, yielding a much-improved picture of the structure and dynamics of lipid bilayers and the relationship of those properties to membrane function [21].

Computational approaches are now being used to facilitate the experimental determination of macromolecular structures by aiding in structural refinement based on either nuclear magnetic resonance (NMR) or X-ray data. The current status of the application of computational methods to the determination of biomolecular structure and dynamics is presented in Chapters 12 and 13. Computational approaches can also be applied in situations where experimentally determined structures are not available. With the rapid advances in gene technology, including the human genome project, the ability of computational approaches to accurately predict 3D structures based on primary sequence represents an area that is expected to have a significant impact. Prediction of the 3D structures of proteins can be performed via homology modeling or threading methods; various approaches to this problem are presented in Chapters 14 and 15. Related to this is the area

of protein folding. As has been known since the seminal experimental refolding studies of ribonuclease A in the 1950s, the primary structure of many proteins dictates their 3D structure [22]. Accordingly, it should be possible ''in principle'' to compute the 3D structure of many proteins based on knowledge of just their primary sequences. Although this has yet to be achieved on a wide scale, considerable efforts are being made to attain this goal, as overviewed in Chapter 17.

Drug design and development is another area of research where computational biochemistry and biophysics are having an ever-increasing impact. Computational approaches can be used to aid in the refinement of drug candidates, systematically changing a drug's structure to improve its pharmacological properties, as well as in the identification of novel lead compounds. The latter can be performed via the identification of compounds with a high potential for activity from available databases of chemical compounds or via de novo drug design approaches, which build totally novel ligands into the binding sites of target molecules. Techniques used for these types of studies are presented in Chapter 16. In addition to aiding in the design of compounds that target specific molecules, computational approaches offer the possibility of being able to improve the ability of drugs to access their targets in the body. These gains will be made through an understanding of the energetics associated with the crossing of lipid membranes and using the information to rationally enhance drug absorption rates. As evidenced by the recent contribution of computational approaches in the development of inhibitors of the HIV protease, many of which are currently on the market, it can be expected that these methods will continue to have an increasing role in drug design and development.

Clearly, computational and theoretical studies of biological molecules have advanced significantly in recent years and will progress rapidly in the future. These advances have been partially fueled by the ever-increasing number of available structures of proteins, nucleic acids, and carbohydrates, but at the same time significant methodological improvements have been made in the area of physics relevant to biological molecules. These advances have allowed for computational studies of biochemical processes to be performed with greater accuracy and under conditions that allow for direct comparison with experimental studies. Examples include improved force fields, treatment of long-range atom–atom interactions, and a variety of algorithmic advances, as covered in Chapters 2 through 8. The combination of these advances with the exponential increases in computational resources has greatly extended and will continue to expand the applicability of computational approaches to biomolecules.

III. SCOPE OF THE BOOK

The overall scope of this book is the implementation and application of available theoretical and computational methods toward understanding the structure, dynamics, and function of biological molecules, namely proteins, nucleic acids, carbohydrates, and membranes. The large number of computational tools already available in computational chemistry preclude covering all topics, as Schleyer et al. are doing in *The Encyclopedia of Computational Chemistry* [23]. Instead, we have attempted to create a book that covers currently available theoretical methods applicable to biomolecular research along with the appropriate computational applications. We have designed it to focus on the area of biomolecular computations with emphasis on the special requirements associated with the treatment of macromolecules.