Advances in Experimental Medicine and Biology 827

Dongqing Wei
Qin Xu
Tangzhen Zhao
Hao Dai *Editors*

Advance in Structural Bioinformatics

上海交通大學出版社
Shanghai Jiao Tong University Press

Springer

Advances in Experimental Medicine and Biology

Volume 827

More information about this series at http://www.springer.com/series/5584

Dongqing Wei · Qin Xu
Tangzhen Zhao · Hao Dai
Editors

Advance in Structural Bioinformatics

上海交通大學出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

Springer

Editors
Dongqing Wei
Qin Xu
Tangzhen Zhao
Hao Dai
Shanghai Jiao Tong University
Shanghai
China

ISSN 0065-2598 ISSN 2214-8019 (electronic)
ISBN 978-94-017-9244-8 ISBN 978-94-017-9245-5 (eBook)
DOI 10.1007/978-94-017-9245-5

Library of Congress Control Number: 2014951346

Springer Dordrecht Heidelberg New York London

Jointly published with Shanghai Jiao Tong University Press
ISBN: 978-7-313-11079-4, Shanghai Jiao Tong University Press

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

Structural bioinformatics, one of the hot spots of bioinformatics, is experiencing a rapid development in recent years. In the genome era, proteomics, genomics, and other data increase dramatically, providing a basis to clarify the problem of essential physiological functions of nucleic acids, proteins, and other biological macromolecules. Relative to the traditional sequence-based bioinformatics, structural bioinformatics focuses mainly on the exploration of the structure and function of biological macromolecules and their dynamic properties. Many human serious diseases are generally associated with some of the key enzymes, ion channels, or associated regulatory proteins. So, most of the new drug research is designed targeting on these proteins. Compared to the previous experimental approaches and sequence analysis, a more comprehensive knowledge of the physiological and pathological mechanism of the drug and the target protein could be obtained from the view of the spatial three-dimensional structure of these molecules and their dynamic structural changes.

The primary problem structural bioinformatics has been trying to solve is that we can build a protein model to fully reveal the nature of its structure and function through the extraction and analysis of the current high-throughput data of biological macromolecules, combining with structural biology knowledge and bioinformatics methods. Besides, to deduce and predict the unknown molecular structure and function based on the known one, and further to realize computer-aided the design and customization of the structure of protein complexes is a long-term goal.

This book represents comprehensive introduction and latest progresses in various aspects of structural bioinformatics. It covers not only the knowledge of mathematical and physical modeling theory, but also the computational methods and its applications in structural bioinformatics. More important, it takes the latest research achievements from the leading groups in this field as examples to illustrate the basic molecular dynamic theory. The content of this book mainly includes the basic knowledge of structural bioinformatics, genomics and proteomics sequence acquiring and analysis, structures of protein, DNA and RNA, basic methods of

molecular dynamic simulations and conformation search, the application examples of computing simulation methods and the structure-based drug design, recent research progress, and future prospects.

We are most grateful to professors and students in the class of "Structural Bioinformatics" at Shanghai Jiao Tong University, where the main contents of this book are accumulated.

Minhang, Shanghai, January 2014 Dongqing Wei

Acknowledgments

In the process of putting this book together, we are much indebted to many people who gave generous support. I would like to express deepest gratitude to the many friends who saw me through this book; to all those who provided support, talked things over, read, wrote, offered comments, allowed me to quote their remarks and assisted in the editing, proofreading and design.

I would like to especially thank the following authors, who were invited to contribute some chapters to this book. They are all from leading research groups in the field of structural bioinformatics in the world, with some of whom I have had the honor to collaborate, i.e., authors of the following chapters

Chapter 2 JVM: Java Visual Mapping Tool for Next Generation Sequencing Read. Ye Yang, Juan Liu*.

Chapter 3 Advancement of Polarizable Force Field and Its Use for Molecular Modeling and Design. Peijun Xu, Huiying Chu, Beibei Li, Yingchen Mao, Yang Ding, Guohui Li*.

Chapter 4 Systematic Methods for Defining Coarse-Grained Maps in Large Biomolecules. Zhiyong Zhang*.

Chapter 5 Quantum Calculation of Protein NMR Chemical Shifts Based on the Automated Fragmentation Method. Tong Zhu, John Z.H. Zhang, Xiao He*.

Chapter 7 Extended Structure of Rat Islet Amyloid Polypeptide in Solution. Lei Wei, Ping Jiang, Malathy Sony Subramanian Manimekalai, Cornelia Hunke, Gerhard Grüber, Konstantin Pervushin, Yuguang Mu*.

Chapter 8 Folding Mechanisms of Trefoil Knot Proteins Studied by Molecular Dynamics Simulations and Go-model. Xue Wu, Ting Fu, Zhilong Xiu, Guohui Li*.

Chapter 9 Binding Induced Intrinsically Disordered Protein Folding with Molecular Dynamics Simulation. Haifeng Chen*.

Chapter 10 Theoretical Studies on the Folding Mechanisms for Different DNA G-quadruplexes. Xue Wu, Ting Fu, Hujun Shen, Zhilong Xiu, Guohui Li*.

Chapter 11 RNA Folding: Structure Prediction, Folding Kinetics and Ion Electrostatics. Zhijie Tan*, Wenbing Zhang*, Yazhou Shi, Fenghua Wang.

Chapter 12 Binding Modes and Interaction Mechanism Between Different Base Pairs and Methylene Blue Trihydrate: A Quantum Mechanics Study. Huiying Chu, Jinguang Wang, Yong Xu, Hujun Shen, Guohui Li*.

Chapter 15 Evolutionary Optimization of Transcription Factor Binding Motif Detection. Zhao Zhang, Ze Wang, Guoqin Mai, Youxi Luo, Miaomiao Zhao, Fengfeng Zhou*.

Chapter 16 Prediction of Serine/Threonine Phosphorylation Sites in Bacteria Proteins. Zhengpeng Li, Ping Wu, Yuanyuan Zhao, Zexian Liu*, Wei Zhao*.

Chapter 22 Bayesian Analysis of Complex Interacting Mutations in HIV Drug Resistance and Cross-Resistance. Ivan Kozyryev, Jing Zhang*.

I would like to thank Springer and Shanghai Jiao Tong University Press for providing me with the opportunity to edit this book.

Contents

Chapter 1
Introduction to Structural Bioinformatics

Qin Xu, Hao Dai, Tangzhen Zhao and Dongqing Wei

Abstract Structural Bioinformatics is one of the hot spots of interdisciplinary sciences and obtained amazing advances in recent years. The first chapter overviews the concept of structural bioinformatics, and briefly describe the contents of this book. The interdisciplinary corporations make it difficult to further divide structural bioinformatics, so the chapters in this book are roughly separated according to the different fields of their applications. That is, fundamental developments in methods of structural bioinformatics, tertiary structure prediction and folding mechanism analysis, the binding mechanism and the interactions between biological macromolecules and ligands, structure-based functional analysis of biological macromolecules, as well as the applications in drug design.

Keywords Structural bioinformatics · Structure of macromolecules · Structure-based drug design

1.1 What Is Structural Bioinformatics

Structural Bioinformatics is generally looked as a branch of bioinformatics mainly about problems of structural biology, which the word "structural" is referred to here. In the early days, it was also named as "computational structural biology", using the distinctive techniques of computational molecular simulations. And the

Q. Xu · H. Dai · T. Zhao · D. Wei (✉)
State Key Laboratory of Microbial Metabolism, College of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China
e-mail: dqwei@sjtu.edu.cn

D. Wei et al. (eds.), *Advance in Structural Bioinformatics*, Advances in Experimental Medicine and Biology 827, DOI 10.1007/978-94-017-9245-5_1

research interests were mainly focused in analysis and prediction of the three-dimensional structures and related functions of biological macromolecules such as proteins, RNA, and DNA.

However, the fast developments in technologies and combinations with other fields make structural bioinformatics more and more diverse and interdisciplinary. Mathematics, statistics, informational sciences, bioinformatics, biophysics, computational chemistry, structural biology, enzymology, medical engineering, pharmaceutical sciences, and much more other disciplines are making contributions to structural bioinformatics. In the meanwhile, its applications are expanding into much more fields, like comparisons of overall folds and local motifs of both primary, secondary and tertiary structures, structural and functional predictions, molecular mechanism of folding/unfolding of macromolecules, evolution and bioengineering, binding interactions in the macromolecules complexes like drug-target complex, molecular mechanism of enzymatic catalysis, as well as other structure-function relationships. In addition to its wide application in the researches of biological sciences, it is showing more power in the industries of bioengineering and drug developments.

The award of 2013 Nobel Prize in Chemistry to Martin Karplus, Michael Levitt, and Arieh Warshel "for the development of multiscale models for complex chemical systems" is, in a way, recognition of the importance of computational techniques in chemistry and biology. However, the computational methods are not opposite to the experimental ones, but complimentary and embedded into them, boosting the developments of more new and advanced techniques and methods to be used. Finally, these new methods might result into a new field of technologies or sciences. Here, structural bioinformatics is a successful example: the advances in this interdisciplinary science have gradually made it an unignorable discipline.

1.2 What Is in This Book

The fast developments in structural bioinformatics attracted more research interests, brought more collaboration from different scientific scopes, and resulted into more advances, both in methodology and in applications. In this book, some of these new advances in structural bioinformatics are introduced, so that the researcher interested in this new field could get some new idea in the scientific developments or interdisciplinary collaborations from these successful examples.

The diverse interdisciplinary combinations make it difficult to trace the development of structural bioinformatics in a single line or divide it into sub-disciplines. But the emergence of structural bioinformatics could be somewhat simply explained as the application of new bioinformatic technologies into the research of structural biology. Therefore, in this book the chapters are organized roughly according to the different applications of the new techniques, additional to those advances with more emphasis on methodology, which are described briefly in the sections below.

1.2.1 Part I: Advances in Methods for Structural Bioinformatics

In Part I, we first introduced several new advances to improve the methodology of structural bioinformatics in different fields, like sequencing, molecular simulation and *in silico* computational chemistry.

Chapter 2 is about program JVM, a powerful tool for mapping next generation sequencing read to reference sequence. It can deal with millions of short read generated by sequence alignment using the Illumina sequencing technology, employing seed index strategy and octal encoding operations for sequence alignments. It is implemented in Java and designed as a desktop application, which supports reads capacity from 1MB to 10 GB. JVM is useful for DNA-Seq, RNA-Seq when dealing with single-end resequencing.

Molecular simulation is always one of the major methods of structural bioinformatics. The contribution of molecular simulation to the developments of chemistry was recently recognized by the 2013 Nobel Prize in Chemistry. The various methods of simulations have covered a diversity of biological scales now. The most popular method, the classical molecular dynamics is fully depended on the force field used. One of the current hot spots of force fields is how to deal with the influence of the electrostatic polarization. In Chap. 3, we review the history of the classical force fields and polarizable force fields, together with its application on small molecules and biological macromolecules simulation, as well as molecular design. In the meantime, various coarse-grained (CG) approaches have also attracted rapidly growing interest in this field of research, because they enable simulations of large biomolecules over longer effective timescales than all-atom molecular dynamics (MD) simulations. Chapter 4 reviews the recent development of a novel and systematic method for constructing CG representations of arbitrary biomolecules, which preserves large-scale and functionally relevant essential dynamics (ED) at the CG level. This method may serve as a very useful tool for the identification of functional dynamics of large biomolecules at the CG level. In Chap. 5, techniques of rare event dynamics are reviewed, followed by further discussion on the intrinsic difficulties to calculate free energy of rare events and the introduction of several well-developed free energy calculation methods. Then several examples of free energy calculations are illustrated, like the calculations on the drug binding in the M2 proton channel, as well as the insertion and association of membrane proteins and membrane active peptides.

In Chap. 6, the automatic fragmentation quantum mechanics/molecular mechanics (AF-QM/MM) is introduced to calculate the *ab initio* NMR chemical shifts so as to improve protein structure determination and refinement. Using the Poisson-Boltzmann (PB) model and first solvation water molecules, the influence of solvent effect is also discussed. Benefit from the fragmentation algorithm, the AF-QM/MM approach is computationally efficient, linear-scaling with a low pre-factor, and massively parallel.

1.2.2 Part II: 3D-Structure Prediction and Folding Mechanism of Biological Macromolecules

Part II focuses on one of the main applications of structural bioinformatics since its early days, that is, the structural prediction and analysis on the mechanisms of folding/unfolding of biological macromolecules. Without good understanding of the structure of the research objects, any in-depth study is questionable.

The case in Chap. 7 is about the research of the extend structure of human islet amyloid polypeptide (hIAPP). The human IAPP aggregates easily, so it is difficult to characterize its structural features by standard biophysical tools. The problem was solved by using rat version of IAPP (rIAPP) as substitute which differs from human IAPP by six amino acids and is not prone to aggregation and does not form amyloid fibrils and similar to human IAPP, it demonstrates random-coiled nature. However, the overall shape of it in solution still remains elusive. Using small angle X-ray scattering (SAXS) measurements combined with nuclear magnetic resonance (NMR) and molecular dynamics simulations (MD) the solution structure of rIAPP was studied and an overall random-coiled feature with residual helical propensity in the N-terminus was confirmed eventually.

The application of structural bioinformatics on the analysis of protein folding mechanisms is illustrated by two examples in Chaps. 8 and 9. In Chap. 8, the folding mechanism of two trefoil knot proteins was simulated under high temperature using all-atom Gō-model. Similar results of the folding process were obtained for the two proteins. That is, the contacts in β-sheet are important to the formation of knot protein. Without these contacts, the knot protein would be easy to untie. In Chap. 9, the folding mechanism of intrinsically disordered proteins upon partner binding was simulated under room temperature as well as high-temperature. The former suggests both nonspecific and specific interactions between the intrinsically disordered proteins and the partner, while the latter shows the kinetics of a two-state process for both the unfolding of apo-states and the unbinding of the bound states. Based on the results of the unfolding processes, the folding pathway of bound intrinsically disordered protein was proposed as: unfolded state, secondary structure folding, tertiary folding, partner binding, and finally to the folded state. In addition, induced-fit mechanism was suggested for the specific recognition between intrinsically disordered protein and its partner using Kolmogorov-Smirnov (KS) P test analysis.

In the rest part of Part II, we presented applications of structural bioinformatics in the studies of DNA and RNA folding. Chapter 10 discusses the folding mechanisms of different DNA G-quadruplexes, which could be a promising anticancer target. In this study, the folding of the thrombin aptamer, Form1 and Form3 G-quadruplexes were simulated with all-atom Gō-model and analyzed by the energy landscape theory, and all were suggested to be a two-state mechanism: the compact structures are formed in the initial stage of the folding process, then they are folded to the native states through the formation of G-triplex structures. The free energy barrier to fold Form 3 G-quadruplex is higher than those to fold thrombin aptamer and Form1, suggesting higher stability of Form 3 G-quadruplex

than those of the other two G-quadruplexes. In Chap. 11, we review the recent experimental and theoretical progress, especially the theoretical modeling of the three major problems in RNA folding: structure prediction, folding kinetics and influence of ion electrostatics.

1.2.3 Part III: The Interactions Between Biological Macromolecules and Ligands

Part III emphasizes on the interactions between macromolecules like protein or DNA/RNA and small ligand molecules, especially possible drug like compounds.

Chapter 12 studies the interactions between DNA base pairs and methylene blue trihydrate, a dye and therapeutic agent possibly to be inserted into two adjacent DNA base pairs. Thus it is called a DNA intercalator. Its binding mode with different base pairs was evaluated and compared using a series of quantum mechanical methods, including various semi-empirical methods, DFT methods and *ab initio* methods. The results showed that the DFT method WB97XD with 6-311+G* basis set best reproduced the result of the expensive *ab initio* method MP2 and determined that the best binding mode was into the AA-TT base pair according to the binding energies and charge density analyses.

Chapter 13 is about the influenza A virus matrix protein 2 (M2 protein), a pH-regulated proton channel crucial to the viral infection and replication. In this chapter, the experimental and computational studies of the two possible drug binding sites on the M2 protein were reviewed to explain the mechanisms for inhibitors to prevent proton conduction, the recent molecular dynamics simulations of the interactions between amantadine and drug-resistant mutant channels were summarized to propose mechanisms for drug resistance, and two proton conduction mechanisms in debate were discussed to further illustrate the applications of structural bioinformatics to understand the structure and functions of this interesting membrane protein.

In Chap. 14, the studies of protein-ligand interactions are in a totally different way, in which massive information about ligand bioactivity and the target protein structures were summarized into the ligand-protein networks so as to elucidate possible "multi-component—multi-target" mechanism of the traditional Chinese medicine (TCM) from its complex composition and unclear pharmacology.

1.2.4 Part IV: Functional Analysis of Biological Macromolecules

It is generally believed that the functions of biological macromolecules are in some ways determined by their structures, including primary structures, secondary structures and tertiary structures. Therefore, one of the major applications of structural bioinformatics is to analyze or predict the functions, activities,

specificities, binding affinities, etc., of protein, DNA/RNA, their complexes or some domains of these biological macromolecules according to their sequences or 3D structures. In this chapter, several examples are illustrated.

In Chap. 15, the primary structure of a DNA fragment is used to predict the possible binding sites (BSs) of a given transcription factor (TF). Based on the hypothesis that positions contribute differently to the motif scoring according to their nucleotide frequency patterns, this method formulated the position contribution as a weight for the position, randomly mutated the weights of different positions in the binding motif by an evolutionary algorithm, and optimized the overall TFBS prediction accuracy. It obtained better or similar performance in sensitivity, specificity, accuracy and Matthews correlation coefficient as the classical algorithm, Position Weight Matrix (PWM), and suggested the widely used assumption of independence between motif positions to be invalid.

Similarly, in Chap. 16 a new predictor named as cPhosBac is introduced to predict serine/threonine phosphorylation sites in bacteria proteins based on their primary structures. The predictor used the composition of k-spaced amino acid pairs (CKSAAP) method to encode the sequence context surrounding the phosphorylation sites, the motif length selection algorithm to optimize the length of the surrounding sequence, and the support vector machine (SVM) algorithm to classify the positive sites from negative sites. This method achieved promising performance and supports online services at http://netalign.ustc.edu.cn/cphosbac/.

In Chap. 17, we present a review on the available resources and methods for discovery and analysis of the single nucleotide polymorphisms (SNPs) in human cytochrome P450. A new method is illustrated as the example of computational SNPs prediction, which uses DNA sequence-based features like nucleotide composition, neighboring SNPs, and CpG dinucleotides occurrence. In addition, the current progress in the methods of annotation and prediction of functional SNPs are summarized.

In the last part of Part IV, the tertiary structure of cytochrome P450cam was used for QM/MM simulations to analyze the mechanism of the second protonation of P450cam. In this example, in order to explore the key factors for the coupling and uncoupling reactions, five 3D models suggesting five possible proton transfer pathways were build, in which two of them led to the coupling reaction while the other three resulted in uncoupling products. Analyses on the simulation results suggested the key factors for the high coupling rate of this enzyme is the Asp251–Thr252 channel, through which the second proton is transferred to the ideal position for coupling reaction.

1.2.5 Part V: Application of Structural Bioinformatics in Drug Design

Drug design is always one of the focuses of applications in structural biology and biochemistry. In recent decades, the burst of computational power boost the emergence of a diversity of new sciences and technologies, including structural

bioinformatics. Logically, it is quickly applied into the discovery and design of new drugs, such as the *in silico* structural or functional analyses on the target proteins, the virtual screening of drug candidates, constructions of databases and drug-target interaction networks, and so on. The applications of the new methods of structural bioinformatics are often surprising and interesting. Here only limited examples are introduced in this chapter.

Cytochrome P450 (CYP) families have been one of the hot spots in drug discovery and development for a long time, because of their critical role in human drug metabolisms. In Chap. 19, several structural bioinformatic studies on CYP are described, including the long-range effects of peripheral mutations on the catalytic activity of CYP1A2, the pharmacophore model for the active site of CYP1A2 and the preliminary prediction of functional consequences of single residue mutation in CYP. The impact of these results on the drug development, especially on the metabolic profile of the drug candidates is also discussed. On the other hand, Chap. 20 focuses on how the structural bioinformatic studies on SNPs of human cytochrome P450 could contribute to personal drug design and optimization of clinic therapies, such as to identify most possible genes associated with the therapeutic targets of given human diseases, to predict the drug efficacy and adverse drug response, to explore individual gene specific properties, etc. The application of diverse structural bioinformatic methods reviewed in this section is expected to greatly improve the current 30–40 % of drug efficacy and lower the possible adverse drug responses of specific patients with personalized medicines and treatments.

In Chap. 21, we introduced a study respect to nicotinic acetylcholine receptors (nAChRs), an ion channel in the central or peripheral nervous system that might be a possible target for Alzheimer's disease. Here the structure of the agonist binding site of $\alpha 7$ nAChR is analyzed to propose its interaction with the agonists, a pharmocophore model of the agonists is designed to explain their selectivity for $\alpha 7$ nAChR, and a brief review of the agonists discovered by far is summarized to confirm the proposed model. Another case of the drug design for resistant HIV is illustrated in Chap. 22, where the current challenges in the experimental and bioinformatic researches for anti-HIV therapy is reviewed, and a series of new Bayesian statistical modeling method are described as a powerful tool complementary to biochemical analysis and molecular simulations to understand the HIV drug resistance and to help the drug development for HIV.

Part I
Advances in Methods for Structural Bioinformatics

Chapter 2
JVM: Java Visual Mapping Tool for Next Generation Sequencing Read

Ye Yang and Juan Liu

Abstract We developed a program JVM (Java Visual Mapping) for mapping next generation sequencing read to reference sequence. The program is implemented in Java and is designed to deal with millions of short read generated by sequence alignment using the Illumina sequencing technology. It employs seed index strategy and octal encoding operations for sequence alignments. JVM is useful for DNA-Seq, RNA-Seq when dealing with single-end resequencing. JVM is a desktop application, which supports reads capacity from 1 MB to 10 GB.

Keywords Mapping · Reads · Algorithms · Next generation sequencing · Program

2.1 Introduction

Over the past 5 years, tens of read mapping programs were published to copy with Illumina sequencing data. But there are some problems have to be pointed out. The first is the limitation of the operating system (OS). Most of programs is designed by C++ language and only can be used on Unix/Linux OS. The biologist is boring by using Unix/Linux OS. Based on a survey on OS user, more than 90 % of users are used to apply "Windows" OS; almost 7 % of users are willing to use "Mac" OS provided by Apple Inc. So the program meeting the need of Multi-OS is required. The second is the restriction of the memory usage. Traditional sequence alignment softwares like MAQ [1], BWA [2], SOAP [3] are high memory consumption programs, and it is difficult to run these programs on the laptop normally. Therefore

Y. Yang · J. Liu (✉)
School of Computer, Wuhan University, Wuhan 430072, Hubei, China
e-mail: liujuanjp@163.com; liujuan@whu.edu.cn

Y. Yang
Military Economy Academy, Wuhan, Hubei, China

D. Wei et al. (eds.), *Advance in Structural Bioinformatics*, Advances in Experimental Medicine and Biology 827, DOI 10.1007/978-94-017-9245-5_2

the program with low memory consumption is needed. The last is the confusion of the parameter settings. There are so many parameters in most of the existed tools that it is difficult for a user to know how to set parameters to finish the alignment. In this work we present a new program JVM (Java Visual Mapping), trying to address to above three problems.

JVM is a desktop application program implemented with Java language, by which the user only needs mouse actions to fulfill the alignment. The best hit of each read which has zero number of sequence mismatch or gap will be reported. The read has multiple hits will be reported in the final list. JVM can handle reads around 11–1000 bp long, and can deal with single-end reads of FASTQ format files which produced by Illumina sequencing platform. JVM supports file sizes ranging from 1 MB to 10 GB. In order to run the program successfully, a Java Runtime Environment version 6.0 or later is required.

2.2 Problem Statement

Read document is a FASTQ format file with four lines per sequence. Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line). Line 2 is the raw sequence letters. Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again. Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in Line 2 [4].

The genome sequence document is a series of characters, each character is either a nucleic acid represented as A, G, C, or T, or an unknown character, named N [5]. This document contains the genome chromosome information.

Read alignment (mapping) is the course of sequence mapping. JVM takes read query sequences with equal length and a database of reference genome sequence as input. Read alignment is just to locate the right places where reads have a perfect alignment to reference genomes. JVM finds all valid alignment that satisfied the constraint on zero error in the set of query sequence.

2.3 Preprocessing

To addresses the problem of too much of the memory spending, we adopt the following preprocessing strategies in JVM.

2.3.1 File Block

The size of a human reference genome document has around 2–3 GB. A read document generated from Illumina platform has a size from 2 to 4 GB. JVM first intelligently separate the read document into several parts, separate the reference document base on the information of chromosome name, and write each part into the disk. Then it maps each read block to a specified chromosome. The Fig. 2.1 illustrates the mapping strategy of choosing read and reference part.

Through this way, we can reduce the peak memory consumption when reading the large capacity files.

2.3.2 Octal Encoding and Sequences Compressing

Problem 1: octal encoding

JVM uses five octal digits to represent each base in read and reference document. The symbols A, C, G, T and N are encoded as 0, 1, 2, 3, 4 respectively. Take a string 'GGGANAACAT' as an example, this string is encoded as octal string: $(2220400103)_8$ (see Table 2.1).

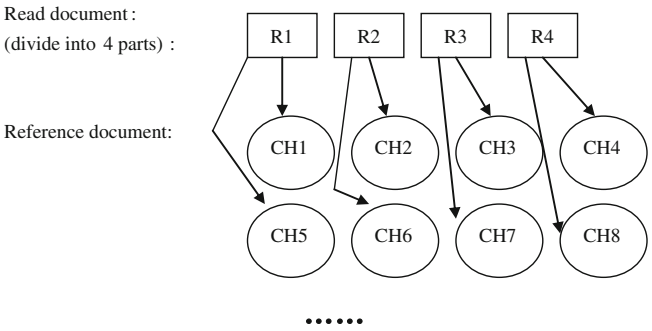


Fig. 2.1 A simulated image of the mapping strategy

Table 2.1 An example of octal encoding and string compressing

String S	GGGANAACAT
Octal encode	2220400103
Compressed value	306315331

Problem 2: sequences compressing

We denote a reference genome sequence as R = R(1, 2,…, m), the query read sequences as the set Q(q1, q2, …, qn), where m is the total bases of reference genome, n is the number of short read; we also let each sequence length be L(10<L<1,000). Alignment progress is the problem of mapping Q to R.

Reference compressing: We construct a new string P = P(1, 2,…, m) by using octal encoding. We partition the r into a set of factor F1, F2, …Fm − L + 1, where Fi = r(i, i + 1,…, i + L − 1), for 0 < i <=m − L + 1. In view of the range of integer type, we transform every 10 octal encoded number into a decimal value and store this numeric into an integer array. In this way, we can reduce memory cost five times when load a long sequences set into main memory. Table 2.1 give an example of the encoding and compressing progress.

Read compressing: Similarly, we can deal with the read document in the same way. We abstract the base sequences from the read and then encoding and compressing the base sequences into an integer array. We define the read array set as Q(1, 2,…n).

2.4 Method

JVM is a visualization tool that by using the mouse operation to complete read mapping and then create a file of ".SAM" format as the output result. In order to accelerate alignment, we take various measures to speed up the efficiency of JVM.

In the section of preprocessing, both reads and the reference sequences are converted to numeric data type using octal encoding for each base. We set the numeric reference as F(i)(0 < i<= m, and i > 108), and numeric read as Q(j)(0 < j<= n, and j > 107). As the progress of read alignment, we use the non-recursion quick sort algorithm combine with seed index strategy to complete ascending sort of F(i). Then using the seed index to quickly position the Q(j) to F(i). At the end of this section, we would give the time complexity of JVM.

The steps of our method are as follows.

Step 1.

Save the reference and read documents into memory. We get a numeric reference set F(i) and store each F(i) into an integer array Ai[t], for t = ⌈L/10⌉. Then we add Ai[t] to a list T. In this way, reference document is compressed into list T, and reduces the size of reference document for five times. We use the same strategy to copy with the read document Q(n).

Step 2.

Build the seed index. We extract every Ai[0] (the first element of Ai[]) from list T. We call Ai[0] as seed index. Then we load Ai[0] into a hash index array B[m − L + 1]. The number of element in B[m − L + 1] almost equal with the number of bases in reference document. In other words, B[m − L + 1] has an big order of magnitude. Considering the memory overflow, we use non-recursive quick sort algorithm to sort the B[m − L + 1]. At the same time, we save the original position of every F(i) so as to keep the exact location of F(i). Non-recursive quick sort algorithm is based on the divide-and-conquer strategy and takes O(m log m) time to sort B[m − L + 1].

Step 3.

Index search and read alignment. We get the read sequence element array from Q(n). To find the best hit of Q(j)(0 <=j < n), we get the first element of each array that is Q j (0), then map it to reference array set B[m − L + 1]. To improve the ability of searching speed, we search the B[m − L + 1] base on the divide-and-conquer strategy. And this take O(log m) time to find the best hit.

For three steps, in step 1 can be done in O(mL + nL) time. Step 2 should be done in O(m log m) time. Step 3 runs in O(n log m). So the overall time complexity is O(m log m + n log m).

2.5 Result

2.5.1 Test by Simulated Dataset

To evaluate speed and accuracy of JVM, we compared JVM with MAPNEXT [6] and WHAM [7]. We had mapped a simulated dataset of 246,558 49 bp-long Illumina single-end resequencing reads. Our reference genome is a dataset simulated the structure of the zebra fish genome NCBI Zv9. To guarantee a fair comparison, we ran the three programs on a same virtual machine and set the mismatch parameter as 0. The OS of this machine is Linux CentOS.5.4. The configuration of this machine includes 2G of main memory, dual 2.00 GHz AMD Turion 64 2-core CPUs. We also test JVM on the Windows OS with the Java Virtual Machine memory with 1.6 G. Table 2.2 shows the performance of each program.

As the result in Table 2.2, JVM has the similar performance on both Linux and Windows OS. Although WHAM is much faster and has more numbers of read mapped to reference, it needs to write parameters and build an index on the reference genome. In addition, WHAM ignores the memory limitation from personal computer during the mapping progress. JVM has a better performance than MAPNEXT on both speed and mapping rate.

Table 2.2 Mapping 246,558 49 bp-long simulated reads to simulated the structure of the zebra fish genome NCBI Zv9

Program	Total time (s)	Read aligned
JVM (on Linux)	47	224
WHAM (on Linux)	0.57	953
MAPNEXT (on Linux)	53	184
JVM (on Windows)	54	224

2.5.2 Test by Real Datasets

We evaluate three programs on a computer with dual 2.00 GHz AMD Turion 64 2-core CPUs, and 4G of DDR2 main memory, running Linux OS. We choose two real datasets containing 20,099,013 and 17,680,937 Illumina single-end resequencing reads (length 49 bp), which were generated from mRNA-Seq of zebra fish. We call the two datasets as dataset 1 and 2. Two read files are two different growth stages of zebra fish. Concerning the time consumption and feature of JVM, we get the same part from dataset 1 and mapping to the zebra fish chromosome 25, the result of three programs show in Table 2.3. In the same way, we take out part of dataset 2 and mapping to the chromosome 22. It gives the performance of each program. We finally run JVM on Windows 7 OS with the same computer configuration in the previous tests.

In the article of WHAM, the author claimed and verified that WHAM was a very fast alignment method. It is often orders of magnitude faster than BOWTIE [8] and RBSA [9]. From the results shown in Table 2.3, total time consumption of WHAM is much less than JVM and MAPNEXT. That is, we also confirmed its conclusion by our experiment. Although JVM is not as fast as WHAM, JVM has a better performance on mapping number. JVM has mapped nearly 20 % more reads than WHAM.

JVM has great advantage over MAPNEXT on time consumption. JVM finished alignment in 235.670 s, while MAPNEXT done in 480.000 s. In terms of mapped reads, MAPNEXT has only 1001 reads mapping to chromosome 25, but JVM has found 37009 reads, it is dozens of times to MAPNEXT.

Table 2.3 Mapping 20,099,013 49 bp-long real reads to the zebra fish chromosome 25 (38,499,472 bp)

Program	Total time (s)	Read aligned
JVM (on Linux)	235.670	37009
WHAM (on Linux)	25.685	31571
MAPNEXT (on Linux)	480.000	1001
JVM (on Windows)	243.890	37009

Table 2.4 Mapping 17,680,937 49 bp-long real reads to the zebra fish chromosome 22 (42,261,000 bp)

Program	Total time (s)	Read aligned
JVM (on Linux)	246.32	53369
WHAM (on Linux)	28.213	44627
MAPNEXT (on Linux)	520.000	1407
JVM (on Windows)	254.762	53369

We also run JVM on Windows 7 OS, we get the same result as that on the Linux OS.

In order to valid the effectiveness of the results in Table 2.3, we not only adjust the read and reference document, but also reset parameters on indexing and mapping progress of WHAM and MAPNEXT. As it is indicated by Table 2.4, the same result can be concluded.

2.5.3 Conclusion and Discussion

As it is demonstrated by above analysis, our developed JVM does a better overall performance than MAPNEXT. And JVM can find more hit reads than WHAM. Based on the efficiency and sensitivity on alignment, we believe that further development and functionality research is not only necessary but also feasible.

We have to admit that JVM is still in the process of improving. As a feature of JVM different from other software, file block is the first target which should be deal with. Now that the order of read part document aligned to reference chromosome document is defined, as the example showing in Fig. 2.1, we can take parallel processing method to accelerate the alignment speed. We can regulate the number of parallel threats based on the total number of physical cores in test machine. So application of parallelization processing mechanism in JVM is the next work we should to do.

In addition, alignment is just the first step in analysing and processing the next generation sequencing data. Further researches based on JVM such as gene fusion and gene expression profiles will be launched.

References

1. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18:1851–1858
2. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760
3. Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: short oligonucleotide alignment program. Bioinformatics 24:713–715

4. Cock P, Fields C, Goto N et al (2010) The Sanger FASTQ file format for sequences with quality scores and the Solexa/Illumina FASTQ variants. Nucl Acids Res 38:1767–1771
5. Frousios K, Iliopoulos CS, Mouchard L, Pissis SP, Tischler G (2010) REAL: An efficient REad ALigner for next generation sequencing reads. ACM, New York, pp 154–159
6. Bao H, Xiong Y, Guo H et al (2009) MapNext: a software tool for spliced and unspliced alignments and SNP detection of short sequence reads. BMC Genomics 10:S13
7. Li Y, Terrell A, Patel J (2011) WHAM: a high-throughput sequence alignment method. SIGMOD Conf 11:445–456
8. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3):R25
9. Papapetrou P, Athitsos V, Kollios G, Gunopulos D (2009) Reference-based alignment in large sequence databases. PVLDB 2(1):205–216

Chapter 3
Advancement of Polarizable Force Field and Its Use for Molecular Modeling and Design

Peijun Xu, Jinguang Wang, Yong Xu, Huiying Chu, Jiahui Liu, Meixia Zhao, Depeng Zhang, Yingchen Mao, Beibei Li, Yang Ding and Guohui Li

Abstract The most important requirement of biomolecular modeling is to deal with electrostatic energies. The electrostatic polarizability is an important part of electrostatic interaction for simulation systems. However, AMBER, CHARMM, OPLS, GROMOS, MMFF force fields etc. used in the past mostly apply fixed atomic center point charge to describe electrostatic energies, and are not sufficient for considering the influence of the electrostatic polarization. The emergence of polarizable force fields has solved this problem. In recent years, quickly developed polarizable force fields have involved a lot of fields. The chapter relating to polarizable force fields spread over several aspects. Firstly, we reviewed the history of the classical force fields and compared with polarizable force fields to elucidate the advancements of polarizable force fields. Secondly, it is introduced

Peijun Xu, Huiying Chu and Jinguang Wang have been contributed equally to this paper.

P. Xu · H. Chu · G. Li (✉)
Laboratory of Molecular Modeling and Design, State Key Laboratory of Molecular Reaction Dynamics, Dalian Institute of Chemical Physics, Chinese Academy of Science, Dalian, Liaoning, China
e-mail: ghli@dicp.ac.cn

J. Wang
The First Affiliated Hospital, Dalian Medical University, Dalian, China

Y. Xu
Guangzhou Institute of Biomedicine and Health, Guangzhou, China

P. Xu · J. Liu · M. Zhao · D. Zhang · Y. Mao · B. Li · Y. Ding
School of Physics and Electronic Technology, Liaoning Normal University, Dalian, Liaoning, China

D. Wei et al. (eds.), *Advance in Structural Bioinformatics*, Advances in Experimental Medicine and Biology 827, DOI 10.1007/978-94-017-9245-5_3

that the application of polarizable force fields to small molecules and biological macromolecules simulation, including molecular design. Finally, a brief development trend and perspective is given on rapidly growing polarizable force fields.

Keywords Polarizable force field · AMOEBA · Fluctuating charge model · Induced point dipole model · Molecular modeling

3.1 History of Classical Force Fields

Molecular force fields are mainly based on a kind of potential energy descriptions at different atomic and molecular levels, and can describe the topological structures and dynamic behaviors of molecules. Molecular force fields are usually adopted to calculate the energies of molecules by using positions of atoms, and greatly speed up calculations compared to quantum mechanics, thus it can be used to study the systems that contain tens of thousands of atoms. Many researches have shown that many physical problems of molecular systems can be explained based on molecular force fields.

In 1930, Andrews [1] first proposed the basic conception of molecular force fields, a bead-spring model was applied to describe the bond length and bond angle, and compute the interactions of non-bonded atoms by using van der Waals interaction expressions. Hill subsequently used process of molecular deformation under Van der Waals interaction to optimize the energies of systems and obtain a reasonable structure in 1946. Then Lifson et al. [2] described consistent force field (CFF) called empirical function force field in 1960s, and it should belong to the modern molecular force field. With the rapid development of molecular mechanics, so far, molecular force fields have been developed many dozens. There are CFF [2], MM1, MM2 [3], AMBER [4], CHARMM [5] etc. in early molecular force fields, these force fields are only limit to several kinds of atom types and some atoms of orbital hybridization, and they are mostly applied to simulations of organic molecular system.

There are two different types in the later development of the force field, such as the accurate type and complete type. The MM3/MM4 [6, 7], SHARP [8], MMFF [9], and OPLS [10] etc. are belonged to the accurate type force field. These force fields adopt simple functional form, and the number of force field parameters range from 10 to 100. They are generally applied to simple system and local specific system. The complete type force fields include DREIDING [11], UFF [12], COMPASS [13], AUA [14] etc., their parameters of force field are more than 100. This kind of force fields cover almost the whole periodic table of elements, many metal organic compounds, and a number of complicated ring compounds involving the orbital hybridization atoms, thus they are more universal.

Among the above molecular force fields, the AMBER, OPLS, CHARMM, MMFF force fields etc. are mostly used in the simulations of the biomolecules, while the MM4, DREIDING, UFF, COMPASS etc. are applied to the simulations

of the material science. Currently, three force fields AMBER, OPLS and CHARMM are also used for the modeling of the ionic liquids. In addition, the MM series force fields and CFF are suitable for the system of organic compounds. In the 1980s, molecular force fields such as AMBER, CHARMM, OPLS and GROMOS produce a positive impact on the research of life science, and promote the development of the molecular force fields targeting life science.

Assisted model building with energy refinement (AMBER) force field. AMBER force field is one of the earliest molecular force field used for the research of biological macromolecules, and covers the simulations of proteins, DNA, monosaccharide and polysaccharide. In this force field, $–CH_2–$ and $–CH_3$ are regard as united atom and used to treat hydrogen bonding interactions. The simulation results show that the AMBER force field can obtain reasonable molecular geometry, conformation energy, vibration frequency and solvation free energy. The parameters of the AMBER force field are obtained as follow, the parameters of equilibrium bonds length and angles are from the experimental data of microwave, neutron scattering and molecular mechanics calculations, the distorted constants are built by microwave, NMR and molecular mechanics calculations, the non-bonded parameters are obtained through the unit cell calculations, and the parameters of atomic charges are given by the calculations of local charge model and ab initio quantum mechanics. For non-bonded interactions within neighboring four atoms in the AMBER force field, the electrostatic interactions are reduced to1/1.2 of other atoms, while the van der Waals interactions reduced to 1/2 of other atoms. The bond stretching and angle bending energies in the AMBER force field are calculated using the harmonic oscillator model, dihedral angle torsion energy is described by Fourier series form, Lennard-Jones potential is chosen to represent the van der Waals force, and the Coulomb formula is applied to estimate the electrostatic interactions. The functional form of AMBER force field is shown as follows

$$\epsilon_{ij} = \frac{4\varepsilon_{ii}\varepsilon_{jj}}{\left(\varepsilon_{ii}^{1/2} + \varepsilon_{jj}^{1/2}\right)^2}$$

where r, θ, ϕ are the bond length, angle and dihedral angle, respectively. The forth term represents the sum of the van der Waals and the electrostatic interactions, and the fifth term is the hydrogen bonding interactions.

Optimized potentials for liquid simulations (OPLS) force field. The OPLS force field includes united-atom model (OPLS-UA) and all-atom model (OPLS-AA), and it is suitable for the simulations of organic molecules and peptides [15]. The bond stretching and bending parameters of OPLS force field are obtained based on the modifications of the AMBER force field. This force field is committed to calculate conformation energies of gas-phase organic molecules, solvation free energies of pure organic liquids and other thermodynamic properties. The OPLS force field is represent as follows

Chemistry at Harvard molecular mechanics (CHARMM) force field. The CHARMM force field is developed by Harvard University, and the force field

parameters are not only from the experimental results, but also involve many results of quantum chemical calculations. This force field is mostly used to study multi-molecular systems including small organic molecules, solutions, polymers, biochemical molecules etc. [16] it can also be used to perform energy minimization, molecular dynamics (MD) and Monte Carlo (MC) simulations. The form of CHARMM force fields is as follows

In the CHARMM force, hydrogen bonding interaction energies are computed by the expression form as follow where *sw* is defined as a switching function, and it is used to control the range of the hydrogen bonding interaction. The subscripts *on* and *off* indicate the start and termination point to calculate the bond lengths and angle values relating to hydrogen bonds in this function.

Force fields in themselves are not correct forms. If the performance of one force field is better than another one, it should be desirable. According to selected different simulation unit, the force field can be divided into all-atom models such as OPLS-AA and united-atom models such as OPLS-UA model. In all-atom model, one atom is regarded as a motion unit, while a alkyl group is took as an imaginary motion unit in the united-atom model. In present, the application of the classical force fields becomes more and more universal [17–25].

3.2 Advancement of Polarizable Force Fields

Due to the smaller amount of calculations and relatively accuracy, the classical force fields gradually become an important tool of biomolecules simulations, and of course, the results of molecular simulations depend on the quality of the force field. So far, the most of current force fields such as OPLS and CHARMM are limited to the theoretical model. In the calculations of electrostatic interactions of the biomolecules, the classical force fields are based on the model of fixed point charge focusing on the atom center, and ignore the electrostatic polarization and the intermolecular and intramolecular charge transfer. If solutes are put into the water-like solvent with a large dielectric constant, or when ions with a larger charge close to a neutral molecule, it will lead to the strong electrostatic polarization phenomena. Therefore, the polarizations can produce an important impact on the energies and structures in the process of the molecular recognition; and significantly reduce the partial electrostatic interactions between the atomic charges. The appearance of polarization force field can solve this problem. 20 years ago, the polarized force field was first introduced to elaborate the change of charge distributions in the dielectric environment. In the past 5 years, the polarized force fields are quickly developed, and it has been applied to many systems ranging from waters to metal enzymes. It should be noticed that the polarization model of water is of a good understanding in the advantages or limitations, furthermore effective insights into the polarization model of the polypeptide and protein-specific parameters are also obtained.

Recently, with the rapid development of the polarized force fields, a dozen different polarizable force fields come out. (1) A polarized force field base on a fragment-based electronic structure method is introduced by Gao [26, 27] at the University of Minnesota. This force field adopts the theory of the electronic structure of the explicit polarization (X-Pol) theory, and it can be used at any level of theory such as the ab initio Hartree-Fock (HF), semiempirical molecular orbital theory, correlated wave function theory, and Kohn-Sham (KS) density functional theory (DFT). In 2008, it is capable of performing more than 3,200 steps (3.2 ps) of MD simulations of a fully solvated protein in water with periodic boundary conditions, consisting of about 15,000 atoms and 30,000 basis functions on a single processor in 24 h, with a full quantum mechanical representation of the entire system [28]. Note that the first MD simulation of a protein by Gelin, McCammon and Karplus in 1979 lasted just over 9 ps using a United-Atom force field without solvent [29]. (2) Based on the induced dipole polarization force field, CFF/ind and ENZYMIX, which is the first polarizable force fields [30], have been applied to many biological systems, DRF90 is developed by Van Duijnen et al. and PIPF force field developed by Gao [31, 32], which is induced point dipole force field targeting organic liquids and biopolymers. (3) Many polarizable force fields, such as the CHARMM polarized force field [33–36] and AMBER polarization force field, are on the basis of the fix point charge polarization. (4) Sum of Interactions between fragments ab initio computed (SIBFA) force field [37], atomic Multipole Optimized Energetics for Biomolecular Applications (AMOEBA) force field [38], ORIENT procedure, Non-Empirical Molecular Orbital (NEMO) procedure [39], etc. belong to the polarizable force fields of the multipole distributions. Of course, there are also some polarized force fields based on density or bond polarization theory, such as the atom-bond electronegativity equalization method developed by Yang et al. [40] at Liaoning Normal University.

3.2.1 Methods Used to Account for Polarization

The methods, which used in the energy calculation of polarization during the simulation, are using different model. The widely used models are briefly introduced below.

3.2.1.1 Induced Point Dipole Model

In this method, a point dipole (PD) P_{ind} is induced at each contributing center in response to the total electric field according to:

$$\mathbf{P}_{ind} = \alpha(\mathbf{E}^0 + \mathbf{E}^p)$$

where E^0 is the field due to the permanent atomic charges and E^p is the field due to the (other) induced dipoles. The total field is determined self-consistently via an iterative procedure that minimizes the polarization energy or by means of the extended Lagrangian method [41]. The contribution of the polarization energy to the total non-bonded energy is then given by:

$$\mathbf{E}_{pol} = -1/2\mathbf{\Sigma}\mathbf{p}_i \cdot \mathbf{E}_i^0$$

where the summation is over polarizable centers *i*.

3.2.1.2 Fluctuating Charge Model

In this method, the atomic charges fluctuate in response to the environment is according to the principle of electronegativity equalization, which states that charge flows between atoms until the instantaneous electronegativities of the atoms are equal. In this approach, the fluctuating charges (FQs) are assigned fictitious masses and treated as additional degrees of freedom in the equations of motion. In the context of molecular dynamics, the equation is efficiently solved by using the extended Lagrangian method [42] at a computational cost little greater than that required for a fixed-charge, pairwise-additive force field. This model has also been implemented, though less efficiently, for use in Monte Carlo simulations [43].

Because liquid water holds a very good network of hydrogen bonds, it plays an important role in most biological processes. Firstly, liquid water serves as a good system to test the polarized force field. In the gas and condensed phases, the properties are accurately simulated and described by the polarized water models. Induced dipole model developed by Caldwell et al. [44] define that the molecular polarization is equal to simple addition of each atomic polarization degree. Bernardo et al. define that the polarization region is limited within the range of the atom in 1,2 or 1,3 bond. Some force fields more directly point the experimental values 1.444 $Å^3$ of unipolar degree on the oxygen atom or HOH angle bisector [45]. The research of Jedlovsky and Richardi compare three water models, and show that the divergence constant obtained through the polarized water models is closer to the experimental value compared with the water model of TIP4P and SPC, [45–48]. The polarization parameters of water in floating charge polarization force field are obtained by fitting the interaction of water dimer, trimer clusters to results of ab initio quantum mechanics. The biggest advantage of this water polarization model is high efficient, and drawback of one is that the polarization effect only limits on the planes of the water molecules. Although the calculated permittivity is very reasonable, their polarization degree is obtained by fitting, not set in advance.

The AMOEBA water model, proposed by Ren and Ponder in 2003, also gives excellent cluster and liquid phase results. This model uses a polarizable atomic multipole description of electrostatic interactions. Multipoles through the quadrupole are assigned to each atomic center based on a distributed multipole analysis

(DMA) derived from large basis set molecular orbital calculations on the water monomer. Polarization is treated via self-consistent induced atomic dipoles. A modified version of Thole's interaction model is used to damp induction at short range. Repulsion-dispersion (vdW) effects are computed from a buffered 14-7 potential. The new potential is fully flexible and has been tested versus a variety of experimental data and quantum calculations for small clusters, liquid water, and ice. Overall, excellent agreement with experimental and high level ab initio results is obtained for numerous properties, including cluster structures and energies, bulk thermodynamic and structural measures. The results of water potential should provide a useful explicit solvent model for organic solutes and biopolymer modeling.

Many models of molecular interactions have been improved by using electronic polarization, and the models including ion solvation [49–51], ion-pair interactions in micellar systems [52], a variety of small molecules condensate nature [53, 54], cation-л interactions [55], as well as interface system etc. [56].

The polarization force fields of small molecules have been reported a lot. Hermida-Ramon, Rios [57], and Krimm et al. take the formaldehyde as the study object, Levy et al. [58] perform studies on small fatty amines and amides, Kollman et al. [59] have studied the amines, Krimm uses polarization force fields to study the N-methylacetamide etc. These polarization force fields upon small molecules utilize induced point dipole method, but the ideas on the polarization model of spectroscopically derived force field (SDFF) of Krimm is more complex [60], SDFF model not only involves polarization simulations, but also the floating charge model. Krimm et al. use the polarized and non-polarized SDD model to calculate dipole moment and electrostatic potential of small molecule dimer as a function of the orientation. The results show that the charge distribution including polarization would significantly improve the consistency of the results with ab initio quantum chemistry calculation. Therefore, we conclude that the simulations ignoring the polarization will lead to some system errors, which include the hydrogen bonding electrostatic interactions.

3.3 Use for (Bio)Molecular Modeling

The main defect of protein force field, such as Amber, CHARMM, OPLS, GROMOS, MMFF, and most of the force fields, is not to fully consider electrostatic polarization effects, but this factor is essential for the protein and its solution system. Because of the existence of a large number of polar groups and the hydrogen bond network, and mutual polarization between the solvent molecules and the solution of protein, thus the polarization effects should be considered in the system of protein and solution.

Although the polarization force field has successfully used in the applications of small molecules, there are also some limitations in biological macromolecules. The first report used the polarization force field to perform the simulation on

protein macromolecules without solvents, and the length of simulation is only 2 ps [61]. Next, simulation time by the application of polarized force field to several small solvent and protein reach nano-second level, and the simulations of the DNA in the dissolved state has also been reported. Therefore, the development of biological macromolecules polarization force field is in a task of top priority.

Polarization force field used in the simulation of protein system can be divided into three types, the first one is the induced dipole (multi-pole) model, the second is a floating charge electrostatic potential model with electronegativity equalization method (fluctuating charge the model, FQ), and the third is combination of electronegativity equalization principle and molecular field.

Induced dipole (multi-pole) model. According to Applequist model, Kollman et al. first added the point polarization to the AMBER force field, and the method is applied effectively to organic molecules. Ponder et al. have treated the polypeptide polarization of the intramolecular and intermolecular by induction of a multi-polar model, and applied the polarization force field on the alanine dipeptids model. AMOEBA force field, which is based on multipole and considers the induced dipole effects, has been proposed for many years. The AMOEBA force field is in fact a significant improvement over fixed charge models for small molecule structural and thermodynamic observables in particular, although further fine-tuning is necessary to describe solvation free energies of drug-like small molecules, dynamical properties away from ambient conditions, and possible improvements in aromatic interactions. State of the art electronic structure calculations reveal generally very good agreement with AMOEBA for demanding problems, such as relative conformational energies of the alanine tetrapeptide and isomers of water sulfate complexes. AMOEBA is shown to be especially successful on protein-ligand binding and computational X-ray crystallography where polarization and accurate electrostatics are critical. In the calculation of the binding between protein and ligands, the AMOEBA force field has been utilized in calculating the binding free energy between trypsinization and a series of six benzamidine-like ligands [62–64], and the series of the calculation results are in good agreement with the experimental values. The polarizable atomic multipole is able to capture the chemical details of the substituted benzamidine ligands. Gresh et al. have studied the protein system using the SIBFA point dipole potential model. The polarization electrostatic interaction potential energy have been reflected and calculated by using the atomic and molecular dipole or multipole polarization and computing systems, achieved improved results in recent years, and it has been applied to study a number of small molecular clusters, peptide molecules and solution.

Floating charge electrostatic potential model with electronegativity equalization method. The electronegativity equalization method can obtain a lot of information in the molecular ground state, such as the electric dipole and multipole moments, polarization, dissociation energy, electron affinity energy. According to electronegativity equalization method, Rick et al. have performed a simulation on the liquid water and NMA aqueous systems taking advantage of fluctuating charge force field method [65] to calculate physical quantities of the

structure and thermodynamics of liquid water. The results are good consistent with the experimental results, and are better than ones of using the first fixed charge force field. On the basis of the electronegativity equalization principle, Banks et al. have established a new floating charge force field [66] and floating charge and induced dipole combined force field, which combined linear response model with OPLS-AA force field. Chelli and Tabacchi have developed a fluctuating charge force field based on the polarization of atomic orbitals, which is based on the chemical potential equalization method of York and Yang.

The combination of electronegativity equalization method and the molecular mechanics. The calculation results by using the combination of electronegativity equalization method and molecular mechanics show that the conformation of lowest energy is inconsistent with the description of force field, which the main defect is the simple combination electronegativity equalization method and the molecular mechanics. Another defect is that the fluctuations of energy there is greater in the molecular dynamics simulations, it shows that the system is certain lack of stability. Consistent implementation of electronegativity equalization method (CIEEM) can make up for these deficiencies.

Polarized force field has been used in a more in-depth study of small organic molecules and water, and also has been present some progress on biological macromolecules such as peptides and proteins. AMBER force field added a point degree of polarization according to Applequist model, which is created by Kollman and his collaborators, and that can offset and reduce certain fixed charge. This method is applied to the organic molecular formula is very convenient, but to some extent, if directly used the other parameters of the original force field, it will affect the calculations of conformational energy [67]. According to the polarization force field method introduced by Banks and Stern [66], Friesner et al. have developed the polarized force field which is applied to the polypeptides and proteins. And they have performed the simulation of the dipeptide models of 20 amino acids, the results show that the conformation of the dipeptides are good to repeat the results of ab initio calculations. They provided a complete polarization force field of protein, and provided a good tool to describe many-body impact and calculate the many-body energy [68]. Gresh et al. develop the polarization force field of proteins and peptides [69], and they make use of SIBFA method to study the formamide nitrogen methylacetamide dimer and the hydrogen bonds energies of alanine, glycine residues. The results show that the addition of lone pair electrons or polar moment is better than the simple atoms central fix charge model in description and calculation the orientation and energy of hydrogen bond between the amide groups. In 2004, Patel et al. have developed the first generation of CHARMM fluctuating charge force field [70–74], which is mainly used to study the parameters of electrostatic model in proteins and peptides. The dimer binding energy and bond length which is calculated using upon model are good agreement with the ab initio calculation results. Patel et al. then perform the molecular simulation on six smaller protein under the conditions of constant temperature and pressure using the fluctuating charge force field and polarization water model TIP-4P, and the length of the simulation reach to a few nanoseconds, while it is the longest time of simulation by

using the solvent and solute protein in the polarization potential field simulation in present.

Polarized force field, which is applied to protein systems, have provided a more effective model to calculate and described biological protein molecule and its electrostatic polarization in solution, and explained and calculated the phenomenon and the physical quantity that the fixed charge force establishments usually can not describe and calculate. So far, the application of polarization force field in molecular design is not widespread, and our laboratory is carrying out molecular design research by using AMOEBA polarizable force field.

3.4 Conclusion and Perspectives

We have briefly introduced the development of the molecular field from classical force field and now the rapid development of polarizable force field. The classical force field has achieved very good results in the study of biological molecules, either AMBER, CHARMM, OPLS and more early molecular field, but the fixed atoms center point charge method still exist limitations, which is that they ignore the intermolecular and intramolecular charge transfer and electrostatic polarization, and it should be noticed that this factor is critical.

So far, polarized force fields have been developed for over 30 years, only in recent years the development of it is better. The polarized force fields have relatively good development and application, whether it is used in simulation of small organic molecules, solution, peptides, proteins, metal enzyme or other biological macromolecules systems. Compared with classical force field, the fixed point charge model does not consider the polarization effect, and there are a lot of advantages of the polarization force field especially in the system where the polarization and accurate electrostatics are critical. The different polarization force fields have different shortcomings and limitations, which are required constantly improving to refine these force field methods, to achieve the purpose of the polarized force field requirements: (1) The charge is changed with the changes in the environment. For example, the polarization in the biological macromolecules is of great degrees of freedom, and the polarization between molecules is strongly dependent on the conformation of the molecule and spatial configuration. (2) The calculation of the many-body interactions is able to correctly, such as the intermolecular polarization effects of gaseous molecule clusters and solution. (3) The polarized force field parameters are of good portability. (4) The calculation of long-range interaction energy of the electrostatic is accurate.

The development of polarization force field is still very long, but we have been working hard on it. We are currently developing AMOEBA force field for lipid membrane system, and testing the binding energy data for large protein-ligand complex systems by using all-atom molecular dynamics simulation of AMOEBA polarizable force field.

References

1. Andrews DH (1930) The relation between the Raman spectra and the structure of organic molecules. Phys Rev 36:544–554
2. Lifson A, Warshel S (1968) Consistent force field for calculations of conformations, vibrational spectra, and enthalpies of cycloalkane and n-alkane molecules. J Chem Phys 49:5116–5229
3. Allinger NL (1977) Conformational analysis. 130. MM2. A hydrocarbon force field utilizing VI and V2 torsional terms. J Am Chem Soc 99(25):8127–8134
4. Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G et al (1984) A new force field for molecular mechanical simulation of nucleic acids and proteins. J Am Chem Soc 106:765–784
5. MacKerell AD Jr, Bashfor D, Bellott M, Dunbrack RL, Evanseck JD et al (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 102:3586–3616
6. Allinger NL, Yuh YH, Lii JH (1989) Molecular machanics. The MM3 force field for hydrocarbon. J Am Chem Soc 111(23):8551–8566
7. Allinger NL, Chen KH, Lii JH, Durkin KA (2003) Alcohols, ethers, carbohydrates, and related compounds. I. The MM4 force field for simple compounds. J Comput Chem 24:1447–1472
8. Allured VS, Kelly CM, Landis CR (1991) SHAPES empirical force field: new treatment of angular potentials and its application to square-planar transition-metal complexes. J Am Chem Soc 113(1):1–12
9. Halgren TA (1996) Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules. J Comput Chem 17:616–641
10. Jorgensen WL, Maxwell DS, Julian TR (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 118:11225–11236
11. Mayo SL, Olafson BD, Goddard WA (1990) DREIDING: a generic force field for molecular simulations. J Phys Chem 94(26):8897–8909
12. Rappe AK, Casewit CJ, Colwell KS, Goddard WA, Skid WM (1992) UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. J Am Chem Soc 114(25):10024–10035
13. Sun H (1998) COMPASS: an ab initio force-field optimized for condensed-phase applications-overview with details on alkane and benzene compounds. J Phys Chem B 102(38):7338–7364
14. Toxvaerd S (1990) Molecular dynamics calculation of the equation of state of alkanes. J Chem Phys 93:4290–4295
15. Gu RX, Liu LA, Wei DQ (2011) Free energy calculations on the two drug binding sites in the M2 proton channel. J Am Chem Soc 133(28):10817–10825
16. Lian P, Wei DQ, Wang JF, Chou KC (2011) An allosteric mechanism inferred from molecular dynamics simulations on phospholamban pentamer in lipid membranes. PLoS ONE 6:e18587
17. Arias HR, Gu RX, Feuerbach D, Guo BB, Ye Y, Wei DQ (2011) Novel positive allosteric modulators of the human α7 nicotinic acetylcholine receptor. Biochemistry 50:5263–5278
18. Zhang T, Liu L, Lewis D, Wei DQ (2011) Long-range effects of a surface mutation on the enzymatic activity of cytochrome P450 1A2. J Chem Info Model 51:1336–1346
19. Arias HR, Gu RX, Feuerbach D, Wei DQ (2010) Different interaction between the agonist JN403 and the competitive antagonist methyllycaconitine with the human alpha7 nicotinic acetylcholine receptor. Biochemistry 49:4169–4180
20. Xu BS, Shen HJ, Zhu X, Li GH (2011) Fast and accurate computation scheme for vibrational entropy of proteins. J Comp Chem 32(15):3188–3193

21. Wu J, Xia Z, Shen HJ, Li GH, Ren PY (2011) Gay-Berne and electrostatic multipole based coarse grained model and application with polyalanine in implicit solvent. J Chem Phys 135:155104
22. Xu BS, Dustin E, Wang YM, Liang HJ, Li GH (2013) A structural-based strategy for recognition of transcription factor binding sites. PLOS ONE, available online at http://dx.plos.org/10.1371/journal.pone.0052460
23. Wang JA, Zhu WL, Li GH, Hansmann UH (2011) Velocity-scaling optimized replica exchange molecular dynamics of proteins in a hybrid explicit/implicit solvent. J Chem Phys 135(8):084115
24. Zhang YX, Shen HJ, Zhang MB, Li GH (2013) Exploring the proton conductance and drug resistance of BM2 channel through molecular dynamics simulations and free energy calculations at different pH conditions. J Phys Chem B 117(4):982–988
25. Shen HJ, Sun H, Li GH (2012) What is the role of motif D in the nucleotide incorporation catalyzed by the RNA-dependent RNA polymerase from poliovirus? PLOS Comp Biol 8(12):e1002851
26. Gao J (1997) Toward a molecular orbital derived empirical potential for liquid simulations. J Phys Chem B 101:657–663
27. Xie W, Gao J (2007) Design of a next generation force field: the X-POL potential. J Chem Theory Comput 3:1890–1900
28. Xie W, Orozco M, Truhlar DG, Gao J (2009) X-Pol potential: an electronic structure-based force field for molecular dynamics simulation of a solvated protein in water. J Chem Theory Comput 5:459–467
29. McCammon JA, Gelin BR, Karplus M (1977) Dynamics of folded proteins. Nature 267:585–590
30. Warshel A, Levitt M (1976) Theoretical studies of enzymatic reactions: dielectric electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. J Mol Biol 103:227–249
31. Gao J, Habibollahzadeh D, Shao L (1995) A polarizable intermolecular potential functions for simulations of liquid alcohols. J Phys Chem 99:16460–16467
32. Xie W, Pu J, MacKerell AD Jr, Gao J (2007) Development of a polarizable intermolecular potential function (PIPF) for liquid amides and alkanes. J Chem Theory Comput 3:1878–1889
33. Patel S, Brooks CL (2004) CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations. J Comput Chem 25:1–16
34. Patel S, MacKerell AD Jr, Brooks CL (2004) CHARMM fluctuating charge force field for proteins: II Protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model. J Comput Chem 25:1504–1514
35. Anisimov VM, Lamoureux G, Vorobyov IV, Huang N, Roux B, MacKerell AD et al (2005) Determination of electrostatic parameters for a polarizable force field based on the classical drude oscillator. J Chem Theory Comput 1:153–168
36. Yu H, Whitfield TW, Harder E, Lamoureux G, Vorobyov I, Anisimov VM et al (2010) Simulating monovalent and divalent ions in aqueous solution using a drude polarizable force field. J Chem Theory Comput 6:774–786
37. Gresh N, Cisneros GA, Darden TA, Piquemal JP (2007) Anisotropic, polarizable molecular mechanics studies of inter-and intramolecular interactions, and ligand-macromolecule complexes. A bottom-up strategy. J Chem Theory Comput 3:1960–1986
38. Ponder JW, Wu CJ, Ren PY, Pande VS, Chodera JD, Schnieders MJ et al (2010) Current status of the AMOEBA polarizable force field. J Phys Chem B 114:2549–2564
39. Engkvist O, Astrand PO, Karlstrom G (2000) Accurate intermolecular potentials obtained from molecular wave functions: bridging the gap between quantum chemistry and molecular simulations. Chem Rev 100:4087–4108
40. Yang ZZ, Wang CS (1997) Atom-bond electronegativity equalization method, I. calculation of the charge distribution in large molecules. J Phys Chem A 101:6315–6321
41. Van Belle D, Froeyen M, Lippens G, Wodak SJ (1992) Molecular dynamics simulation of polarizable water by extended Lagrangian method. Mol Phys 77:239–255

42. Rick SW, Stuart SJ, Berne BJ (1994) Dynamical fluctuating charge force fields: application to liquid water. J Chem Phys 101:6141–6156
43. Martin MG, Chen B, Siepmann JI (1998) A novel Monte Carlo algorithm for polarizable force fields: application to a fluctuating charge model for water. J Chem Phys 108:3383–3385
44. Hem S, Bassler H (1994) Fluorescence spectroscopy of oligo. J Phys Chem 98(30):7355–7358
45. Shah S, Concolino T, Rheingold AL, Protasiewicz JD (2000) Sterically encumbered systems for two low-coordinate phosphorus centers. Inorg Chem 39(17):3860–3867
46. Spiliopoulos IK, Mikroyannidis JA (2002) Blue-light-emitting poly(phenylenevinylene)s with alkoxyphenyl substituents: synthesis and optical properties. Macromolecules 35:2149–2156
47. Lee SH, Jang BB, Tsutsui T (2002) Sterically hindered fluorenyl-substituted poly(p-phenylenevinylenes) for light-emitting diodes. Macromolecules 35:1356–1364
48. Roncali J (2000) Oligothienylenevinylenes as a new class of multinanometer linear л-conjugated systems for micro- and nanoelectronics. Acc Chem Res 33:147–156
49. Caldwell J, Dang LX, Kollman PA (1990) Implementation of nonadditive intermolecular potentials by use of molecular dynamics: Development of a water–water potential and water–ion cluster interactions. J Am Chem Soc 112:9144–9147
50. Stuart SJ, Berne BJ (1996) Effects of polarizability on the hydration of the chloride ion. J Phys Chem 100:11934–11943
51. Grossfield A, Ren P, Ponder JW (2003) Ion solvation thermodynamics from simulation with a polarizable force field. J Am Chem Soc 125:15671–15682
52. Shelley JC, Sprik M, Klein ML (1993) Molecular dynamics simulation of an aqueous sodium octanoate micelle using polarizable surfactant molecules. Langmuir 9(4):916–926
53. Rick SW, Berne BJ (1996) Dynamical fluctuating charge force fields: the aqueous solvation of amides. J Am Chem Soc 118:672–679
54. Gao J, Dariush H, Shao L (1995) A polarizable intermolecular potential function for simulation of liquid alcohols. J Phys Chem 99(44):16460–16467
55. Caldwell JW, Kollman PA (1995) Structure and properties of neat liquids using nonadditive molecular dynamics: Water, methanol, and N-methylacetamide. J Phys Chem 99(16):6208–6219
56. Dang LX (1999) Computer simulation studies of ion transport across a liquid/liquid interface. J Phys Chem B 103(39):8195–8200
57. Hermida-Ramon JM, Rios MA (1998) A new intermolecular polarizable potential for a formaldehyde dimer. Application to liquid simulations. J Phys Chem A 102:10818–10827
58. Ding Y, Bernardo DN, Krogh-Jespersen K, Levy RM (1995) Solvation free energies of small amides and amines from molecular dynamics/free energy perturbation simulations using pairwise additive and many-body polarizable potentials. J Phys Chem 99:11575–11583
59. Meng EC, Caldwell JW, Kollman PA (1996) Investigating the anomalous solvation free energies of amines with a polarizable potential. J Phys Chem 100:2367–2371
60. Mannfors B, Palmo K, Krimm S (2000) A new electrostatic model for molecular mechanics force fields. J Mol Struct 556:1–22
61. Kaminski GA, Stern HA, Berne BJ, Friesner RA, Cao YX, Murphy RB et al (2002) 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
62. Jiao D, Golubkov PA, Darden TA, Ren P (2008) Calculation of protein-ligand binding free energy by using a polarizable potential. Proc Natl Acad Sci USA 105:6290–6295
63. Jiao D, Zhang J, Duke RE, Li G, Schnieders MJ, Ren P (2009) Trypsin-ligand binding free energies from explicit and implicit solvent simulations with polarizable potential. J Comput Chem 30:1701–1711
64. Shi Y, Jiao D, Schnieders MJ, Ren P (2009) Trypsin-ligand binding free energy calculation with AMOEBA. In: IEEE Engineering in Medicine and Biology Society, EMBC proceedings, pp 2328–2331

65. Scherlis DA, Marzari N (2004) π-Stacking in charged thiophene oligomers. J Phys Chem B 108(46):17791–17795
66. Reichardt C (2003) Solvents and solvent effects in organic chemistry. Wiley-VCH, Weinheim
67. Leach AR (2001) Molecular modelling: principles and applications. Pearson Education Limited, England
68. Reichardt C (1994) Solvatochromic dyes as solvent polarity indicators. Chem Rev 94(8):2319–2358
69. Bendikov M, Duong HM, Starkey K, Houk KN, Carter EA, Wudl F (2004) Oligoacenes: theoretical prediction of open-shell singlet diradical ground states. J Am Chem Soc 126(24):7416–7417
70. Koshida N, Matsumoto N (2003) Fabrication and quantum properties of nanostructured silicon. Mater Sci Eng R 40:169–205
71. Miller RD, Michl J (1989) Polysilane high polymers. Chem Rev 89:1359–1410
72. Zeng XB, Liao XB, Wang B, Dai ST, Xu YY, Xiang XB et al (2004) Optical properties of boron-doped Si nanowires. J Cryst Growth 265:94–98
73. Cui Y, Lieber CM (2001) Functional nanoscale electronic devices assembled using silicon nanowire building blocks. Science 291:851–853
74. Holmes JD, Johnston KP, Doty RC, Korgel BR (2000) Control of thickness and orientation of solution-grown silicon nanowires. Science 287:1471–1473

Chapter 4
Systematic Methods for Defining Coarse-Grained Maps in Large Biomolecules

Zhiyong Zhang

Abstract Large biomolecules are involved in many important biological processes. It would be difficult to use large-scale atomistic molecular dynamics (MD) simulations to study the functional motions of these systems because of the computational expense. Therefore various coarse-grained (CG) approaches have attracted rapidly growing interest, which enable simulations of large biomolecules over longer effective timescales than all-atom MD simulations. The first issue in CG modeling is to construct CG maps from atomic structures. In this chapter, we review the recent development of a novel and systematic method for constructing CG representations of arbitrarily complex biomolecules, in order to preserve large-scale and functionally relevant essential dynamics (ED) at the CG level. In this ED-CG scheme, the essential dynamics can be characterized by principal component analysis (PCA) on a structural ensemble, or elastic network model (ENM) of a single atomic structure. Validation and applications of the method cover various biological systems, such as multi-domain proteins, protein complexes, and even biomolecular machines. The results demonstrate that the ED-CG method may serve as a very useful tool for identifying functional dynamics of large biomolecules at the CG level.

Keywords CG modeling · Principal component analysis · Elastic network model

4.1 Introdcution

With dramatic recent improvements in computer power, atomistic molecular dynamics (MD) simulations can be performed on timescales from μs to ms, which enable us to study key biological processes such as protein folding, ligand binding,

Z. Zhang (✉)
Hefei National Laboratory for Physical Sciences at Microscale and School of Life Sciences, University of Science and Technology of China, Hefei, Anhui, China
e-mail: zzyzhang@ustc.edu.cn

D. Wei et al. (eds.), *Advance in Structural Bioinformatics*, Advances in Experimental Medicine and Biology 827, DOI 10.1007/978-94-017-9245-5_4

and functional conformational changes [1]. However, many biochemical events in cells still take place on much longer timescales than ms. On the other hand, large spatial scales of biomolecular complexes involved in these processes make atomic-level MD simulations computationally expensive. Coarse-grained (CG) models, which reduce the large number of degrees of freedom in an atomic structure into a much smaller set of CG sites, allow us to simulate larger biomolecules over longer effective timescales than atomistic MD simulations. CG modeling may overcome the gap between computational capabilities and biological processes. Therefore various CG approaches have been widely developed with rapid-growing interest in many research groups [2–6].

In order to start CG modeling for a given biomolecular system, the first important issue is to establish a reasonable mapping between its atomistic and CG resolution. That is to say, one needs to determine the number of CG sites in the biomolecule and where to place them. Once the proper CG mapping is obtained, interactions among the CG sites (CG force field) are then defined before performing CG simulations via MD or Monte Carlo methods.

The scheme of CG mapping varies among different CG methodologies. In the MARTINI model [7], four heavy atoms are grouped into a single CG site (four-to-one mapping on average), but those ring-like structures are mapped with higher resolution (up to two-to-one). Therefore amino-acid residues are represented by one to five CG sites, respectively [8]. The United reside (UNRES) model simplifies each residue by two CG sites, which are its C_α atom and side-chain center [9]. Elastic network models (ENMs), which have been very popular in the study of protein functional dynamics [10–12], use a one-site per residue (usually at the position of its C_α atom) CG mapping. The above methods represent each residue by one or more CG sites, that is, the CG and atomic models have similar resolution. In this case, the CG mapping is straightforward and can be constructed by chemical intuition. However, it is more challenging to build CG maps with lower resolutions than one-site per residue, thus systematic methodologies have been developed, in order to define relatively few CG sites in a large biomolecule. Some of these methods are neural network-based approaches [13, 14], whereas another class of such methods uses dynamic information from atomic models to build CG maps [15–23].

We have developed a systematic and quantitative method to define a CG map beyond the resolution of individual residues, by using the information of essential dynamics (ED) in the biomolecule [18, 22, 23]. Essential dynamics [24], which can be characterized by principal component analysis (PCA) on a structure ensemble or ENM of a single structure, usually represent those functionally important collective domain motions in the biomolecule [25, 26]. In particular, this essential dynamics coarse-graining (ED-CG) approach designs a residual to variationally optimize the CG map, in order to preserve the essential dynamic domain motions.

In the subsequent sections, the theory and technical details of ED-CG method will be described. The resulting method is validated by applying to two proteins, which are the HIV-1 CA protein dimer and globular actin (G-actin), respectively.

Then an ED-CG map of the *E. coli* 70S ribosome is constructed from its MD data, and those bridge interactions between the small and large ribosomal subunits are analyzed at the CG level. Concluding remarks are provided at the end.

4.2 Theory and Methods

4.2.1 Essential Dynamics of Biomolecules

Internal dynamics is essential for a biomolecule to function, while it is a non-trivial issue to extract large-scale and functionally-relevant motions of the biomolecule from those small and uninteresting fluctuations. One solution to this problem is to use collective coordinates [24–26]. It has been well studied that motions occurring along the directions of a small number of properly-defined collective coordinates may dominantly contribute to internal dynamics of the biomolecule [25, 26]. Therefore a low-dimensional subspace can be defined by these essential collective coordinates, also named as essential subspace. Motions within the essential subspace, called essential dynamics [24], are usually functionally important. Collective coordinates are actually a set of eigenvectors obtained by diagonalizing a second moment matrix. Various computational techniques for determining the essential dynamics have been established.

Principal component analysis (PCA) on a structure ensemble. For a given biomolecule, if a sufficiently large number of experimental structures are available, PCA can be applied on them directly. Otherwise the structure ensemble is usually generated by MD or Monte Carlo (MC) simulations using the atomic force field. As an alternative, the CONCOORD method quickly produces a set of conformations around a know structure based on atomic distance constraints [27], which can then be used for PCA without doing CPU-intensive MD simulations. PCA needs to diagonalize a covariance matrix of atomic fluctuations, which may require a huge amount of memory storage and is computationally demanding for a large biomolecule. Fortunately, Amadei et al. have demonstrated that PCA by using only the coordinates of C_α atoms can preserve the essential dynamics obtained from the all-atom PCA quite well [24]. Therefore, only the C_α atoms (P atoms if there are nucleotides in the system) will be considered in the following.

For a biomolecule with n residues (that is, n C_α atoms), after removing the translational and rotational motion by least-square fitting each structure in the ensemble to a reference, a $3n \times 3n$ covariance matrix $\mathbf{C}$ is constructed, with each element defined by

$$C(i_x, j_y) = \frac{1}{n_t} \sum_{t=1}^{n_t} \Delta r_{i_x}(t) \Delta r_{j_y}(t), \tag{4.1}$$

where i_x and j_y are one of the three components of the atom i and j in Cartesian space, respectively, and n_t is the total number of conformations in the ensemble. $\Delta r_{i_x}(t)$ and $\Delta r_{j_y}(t)$ are atomic fluctuations calculated by

$$\begin{aligned} \Delta r_{i_x}(t) &= r_{i_x}(t) - \langle r_{i_x} \rangle \\ \langle r_{i_x} \rangle &= \frac{1}{n_t} \sum_{t=1}^{n_t} r_{i_x}(t). \end{aligned} \tag{4.2}$$

A $3n \times 3n$ matrix of eigenvectors and corresponding eigenvalues are obtained by diagonalizing the covariance matrix (Eq. 4.1)

$$C(i_x, j_y) = \sum_{q=1}^{3n} \Psi_q^{i_x} \lambda_q \Psi_q^{j_y}. \tag{4.3}$$

Here one column of the matrix $\mathbf{\Psi}$ represents a $3n$-dimensional eigenvector $\mathbf{\Psi}_q$ (also called a PCA mode), in which each atom i has three components. The eigenvalue λ_q is the mean square fluctuation of the corresponding PCA mode. If those PCA modes are sorted by decreasing order of their eigenvalues, the majority of the motions in the biomolecule can be described by first few PCA modes with the largest eigenvalues. Therefore this small subset of the PCA modes are essential modes, with corresponding motions called the essential dynamics.

Normal mode analysis (NMA) of a single structure. NMA determines independent normal modes of a biomolecule based on the assumption that the energy surface of the biomolecule can be characterized by the harmonic approximation with a single energy minimum [28]. Despite this crude approximation, NMA is very useful in determining functional motions in biomolecules [29]. Since a single structure is sufficient for NMA, one does not need to spend any CPU time on MD simulations. However, standard NMA uses an all-atom structure that needs to be energy minimized to a local minimum, and then a second moment Hessian matrix is constructed and diagonalized. These calculations are still computational demanding for large biomolecules. Elastic network models (ENMs) can significantly reduce the computational task by using residue-based CG models to perform NMA, and energy minimization is not necessary [10–12]. Therefore, we use an ENM called the anisotropic network model [10] in our ED-CG method.

Frequently in ENM, each residue is represented by the position of its C_α atom, and their interactions are defined by effective bonds. Therefore the harmonic potential of the ENM is written as

$$V = \sum_{i,j>i} \frac{1}{2} k_{ij} \Delta r_{ij}^2. \tag{4.4}$$

Here, $\Delta r_{ij} = r_{ij} - r_{ij}^0$ is the bond fluctuation connecting the atoms i and j. r_{ij}^0 is the equilibrium bond length. Those spring constants, k_{ij}, can be determined by

different rules. The most popular method is to pre-determine a cutoff distance, and only those atoms within the cutoff are connected, then a uniform spring constant is placed for all pairs of connected atoms [10]. Another strategy is to use distance-weighted spring constants, which may be physically better motivated [30]. Since there is no cutoff distance, this ENM is 'parameter-free' [31].

In NMA, the second moment Hessian matrix is a matrix of the second derivatives of the overall potential at an energy minimum

$$H(i_x, j_y) = \partial^2 V / \partial r_{i_x} \partial r_{j_y}. \tag{4.5}$$

Since the entire equilibrium bond lengths are taken from the single structure that means it locates at the minimum point on the energy surface under the NMA approximation, no energy minimization is needed. The potential energy function is harmonic (Eq. 4.4) in ENM, so the components of the matrix **H** can be solved analytically (more details can be found in Ref. [22]). Similar to the covariance matrix in PCA (Eq. 4.3), the Hessian matrix (Eq. 4.5) can be diagonalized to yield a $3n \times 3n$ matrix of eigenvectors (each column is called a normal mode) and $3n$ eigenvalues (each reflects the frequency of the corresponding normal mode). The normal modes are usually sorted by the increasing order of their eigenvalues, of which the first six modes have zero eigenvalues (frequencies) because they are associated with the overall translational and rotational motions of the biomolecule. Many studies have indicated that the first few low-frequency normal modes capture the essential dynamics of the biomolecules [12], which may describe nearly the same functional motions, as those revealed by large-amplitude PCA modes (Fig. 4.1).

4.2.2 The ED-CG Method

The ED-CG methodology has been described in two papers [18, 22] with more details. The atomic fluctuation of atom i in the essential subspace is denoted as $\Delta \mathbf{r}_i^{\mathrm{ED}}$. For those atoms that move in a highly correlated fashion with the atom i, their fluctuation differences are small. They can be grouped as a dynamic domain [32] and defined as a single CG site I by using the center-of-mass (COM) of the domain. Therefore, in order to define N CG sites in a biomolecule, the basic idea of ED-CG is to decompose the whole molecule into N dynamics domains, which is achieved by variationally minimizing the following residual

$$\begin{aligned} \chi^2 &= \frac{1}{3N} \sum_{I=1}^{N} \sum_{i \in I} \sum_{j \geq i \in I} \left\langle \left(\Delta \mathbf{r}_i^{ED} - \Delta \mathbf{r}_j^{ED} \right)^2 \right\rangle \\ &\equiv \frac{1}{3N} \sum_{I=1}^{N} \sum_{i \in I} \sum_{j \geq i \in I} \left\langle \left(\Delta \mathbf{r}_i^{ED} \right)^2 - 2 \Delta \mathbf{r}_i^{ED} \cdot \Delta \mathbf{r}_i^{ED} + \left(\Delta \mathbf{r}_i^{ED} \right)^2 \right\rangle, \end{aligned} \tag{4.6}$$

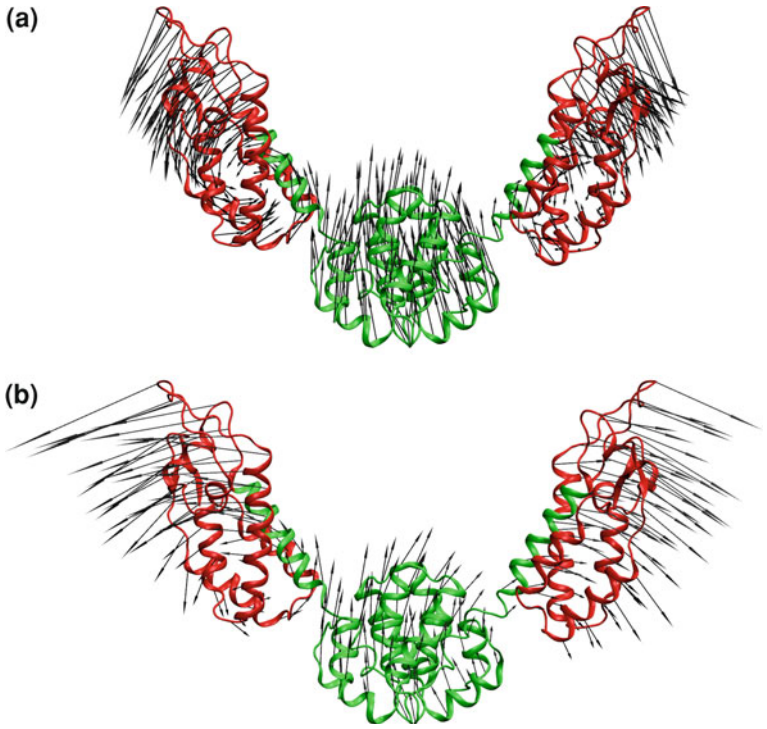


Fig. 4.1 Essential dynamics in the HIV-1 CA protein dimer. **a** The first PCA mode with the largest eigenvalue. PCA was performed on a 20 ns MD trajectory of the CA dimer. **b** The lowest-frequency normal mode calculated by ENM based on a single structure. Bothe the PCA and ENM mode describes the collective motion between the NTD and CTD. All Figures in this chapter were created using VMD [55]. Reprinted from the reference [22], Copyright (2009), with permission from Elsevier

where $\left\langle \left(\Delta \mathbf{r}_i^{ED}\right)^2 \right\rangle$ is the mean-square fluctuation of atom i in the essential subspace. Since the ED-CG method optimally preserve the dynamic domains and the essential dynamics describes the collective motions among them, the CG map defined by this algorithm can potentially capture the essential dynamics of the biomolecule at the CG level.

In PCA, the essential subspace is spanned by the first a few PCA modes ($n_{\mathrm{ED}} \ll 3n$) with dominant fluctuations. According to Eqs. (4.1 and 4.3),

$$\left\langle \left(\Delta \mathbf{r}_i^{ED}\right)^2 \right\rangle = \sum_{x=1}^{3} \sum_{q=1}^{n_{ED}} \Psi_q^{i_x} \lambda_q \Psi_q^{i_x} \equiv tr\left[\left(\mathbf{c}^{ED}\right)_{ii}\right]. \tag{4.7}$$

The 3×3 matrix $(\mathbf{c}^{ED})_{ii}$ is the *ith* super-element of the covariance matrix in the essential subspace (denoted as $\mathbf{C}^{ED}$), and *tr*[] refers to the trace of the matrix. Therefore, Eq. (4.6) is equivalent to the following form:

$$\chi^2 = \frac{1}{3N}\sum_{I=1}^{N}\sum_{i\in I}\sum_{j\geq i\in I}\left(tr\left[\left(\mathbf{c}^{ED}\right)_{ii}\right] - 2tr\left[\left(\mathbf{c}^{ED}\right)_{ij}\right] + tr\left[\left(\mathbf{c}^{ED}\right)_{jj}\right]\right). \tag{4.8}$$

Compared to Eq. (4.6) and (4.8) is computationally more convenient.

In ENM, the essential subspace consists of the first n_{ED} non-zero low-frequency normal modes. According to the classical theory of networks [33],

$$\left\langle \left(\Delta\mathbf{r}_i^{ED}\right)^2 \right\rangle = k_B T \,\mathrm{tr}\left[\left(\mathbf{h}^{ED}\right)_{ii}^{-1}\right], \tag{4.9}$$

where k_B is the Boltzmann constant, and T is the absolute temperature. The 3×3 matrix $\left(\mathbf{h}^{ED}\right)_{ii}^{-1}$ is the *ith* super-element of $(\mathbf{H}^{ED})^{-1}$, which is the inverse matrix of $\mathbf{H}$ in the essential subspace. Therefore, in the case of ENM, Eq. (4.6) can be rewritten as:

$$\chi^2 = \frac{k_B T}{3N}\sum_{I=1}^{N}\sum_{i\in I}\sum_{j\geq i\in I}\left(\mathrm{tr}\left[\left(\mathbf{h}^{ED}\right)_{ii}^{-1}\right] - 2\mathrm{tr}\left[\left(\mathbf{h}^{ED}\right)_{ij}^{-1}\right] + \mathrm{tr}\left[\left(\mathbf{h}^{ED}\right)_{jj}^{-1}\right]\right). \tag{4.10}$$

4.2.3 The Search Algorithms

In this section, we are going to introduce numerical algorithms to search the CG mapping with the minimal residual (Eq. 4.8 or 4.10), which is a non-trivial problem because the search space becomes extremely large with the increasing number of residues n in the biomolecule and CG sites N to be defined. Two algorithms are presented here, which are sequence-based [18] and space-based [23], respectively.

A sequence-based search algorithm. A restriction is employed to make the search more tractable, that is, the dynamic domains are assumed to be contiguous in the primary sequence. Therefore, if there are N dynamic domains (CG sites) to be defined, one just needs to determine $N - 1$ boundary atoms, and each of them corresponds to the last atom of one domain (note that the C-terminus is always a boundary). Under this restriction, the search space of CG mappings would be greatly reduced, and at the same time, it is a rather reasonable assumption because a group of sequentially-contiguous atoms may have a good chance to move collectively.

Initially the boundary atoms are placed on the primary sequence randomly, then the residual (Eq. 4.8 or 4.10) is minimized by adjusting the locations of these boundary atoms, using a global simulated annealing (SA) [34] followed by a local search. At each step of SA, a boundary atom is randomly picked and its position on the sequence is changed randomly, to obtain a new CG map. This new map is accepted or rejected according to the Metropolis criterion [35]. If the residual of the new map is lower than its predecessor ($\Delta\chi^2 < 0$), it is certainly accepted as the start of the next SA step. However, if $\Delta\chi^2 > 0$, the new map can only be accepted

with a probability of exp $(-\Delta\chi^2/T)$, where T is the 'temperature'. At the beginning T is high and then gradually decreased during the SA process, which allows the boundary atoms to move widely and escape from local minima, and finally settle into the global minimum. The boundary-atom set after SA will be treated with a local search algorithm to see if it can be further optimized. One at a time, each boundary atom moves forward (+1) and backward (−1) on the primary sequence, and any change that decreases the residual is accepted. This procedure continues until the residual cannot be minimized anymore. To assure the convergence of the results, multiple minimizations starting with different initial boundary-atom sets are carried out. The boundary-atom set with the lowest residual is chosen, and the COM of each dynamic domain is calculated as a CG site.

A space-based search algorithm. This is a more general method than the sequence-based algorithm. Here we summarize the algorithm briefly, and more details can be found in the paper [23]. In order to define N CG sites from the atomic structure of the biomolecule, N 'seeds' are generated firstly. The position of each seed is determined by the coordinates of a randomly selected atom, plus a small random offset value between [−1, 1]. Then the atomic structure is decomposed into N domains according to the N seeds, such that each domain includes all the atoms that are closest to the corresponding seed. The N CG sites are computed as the COM of these domains, respectively, and the residual of this CG map is calculated by (Eq. 4.8 or 4.10). In the next step, a seed is picked out and its position is updated randomly. By repeating the above procedure, a new CG map is obtained, which is then accepted or rejected based on the Metropolis criterion. As in the sequence-based algorithm, SA is used to minimize the residual. The same calculations are performed beginning with different initial sets of seeds, and the CG model with the lowest residual is finally taken. Without the restriction of sequentially-contiguous domains, the space-based algorithm can search a much larger space of CG maps than the sequence-based algorithm, which may lead to a good side and a bad side. The good side is that, the space-based algorithm can obtain a CG map with lower residual than the sequence-based one at the same resolution, but the bad side is the poorer convergence of the former. Computational cost of the space-based algorithm is also larger since all the distances between seeds and atoms have to be calculated at each step.

Biomolecular complexes: a divide-and-conquer strategy. For a biomolecular complex with multiple subunits, the convergence of ED-CG results is usually not good due to its large size. Such a problem can be improved by a divide-and-conquer strategy. Here we will introduce how to employ it in the sequence-based algorithm. In a complex with N_s subunits, all the C-terminal atoms are defined as boundary atoms, that is to say, a dynamic domain is not allowed to cross the subunits. The sequence-based algorithm is then applied to the whole complex to define the leftover $N - N_s$ boundary atoms. We have observed that although multiple initial boundary-atom sets do not converge to the same ED-CG map, the CG-site distribution among the subunits is much better converged. Therefore, after fixing the number of CG sites in each subunit by the first round of minimization to the whole complex, the positions of CG sites on the primary sequence of each

subunit are then optimized, separately. Through this divide-and-conquer procedure, the convergence of the algorithm in each subunit is far superior, and thus a robust CG map to the whole biomolecular complex can be obtained.

4.3 Validation of the Methods

4.3.1 ED-CG Maps of the HIV-1 CA Protein Dimer

The assembly/disassembly of the HIV-1 viral capsid is a critical process during its infection to a host cell. The capsid shell consists of about 1,500 monomers of the CA protein, and each of them has 220 amino acid residues. Therefore, it is valuable to define a CG map of the CA protein with a much lower resolution than one-site per residue, in order to model the process of capsid assembly/disassembly efficiently [36].

The CA protein includes an N-terminal domain (NTD) and a C-terminal domain (CTD), which are connected by a α-helix (Fig. 4.1). From chemical intuition, each domain can be coarse-grained to a single site, so we tried to build a four-site map for the CA dimer using the sequence-based ED-CG method. PCA was carried out on a 20 ns MD trajectory of the CA dimer by considering only the coordinates of the C_α atoms, and thus the essential modes were obtained. Our study has suggested that it is reasonable to enforce symmetry between the two monomers [18], so only one boundary atom needs to be determined in the sequence-based four-site map of the CA dimer. We moved the boundary atom along the primary sequence of the monomer from the C_α atom 1–219, and found a global minimum of the residual Eq. (4.8) at the atom 131 that is located within the linking α-helix between NTD and CTD (Fig. 4.2). The ED-CG method can therefore automatically find the NTD and CTD, and place one CG site in each of them in the symmetric four-site map of the CA dimer (Fig. 4.2a), which is consistent to the chemical intuition. When increasing the number of CG sites to six, two boundary atoms needs to be defined in the CA monomer. They are the C_α atom 72 and 134, respectively, when the residual Eq. (4.8) is reached to the global minimum. The NTD is decomposed into two dynamic domains, but the CTD remains intact, in the symmetric six-site map of the CA dimer (Fig. 4.2c). By looking at the RMSF values of these boundary atoms, all of them are closely located at the 'local minima' in the RMSF curve of the CA dimer (Fig. 4.2b, d). These rigid hinge regions can be regarded as 'natural boundaries' between dynamics domains.

4.3.2 ED-CG Maps of G-Actin

G-actin is the basic component of the actin filament (F-actin). CG models of G-actin have proved to be very useful in the study of mesoscopic properties of F-actin, thus revealing how protein conformational changes affect the elastic properties of cytoskeleton [37–39]. From the work of Kabsch et al. [40], G-actin can be divided

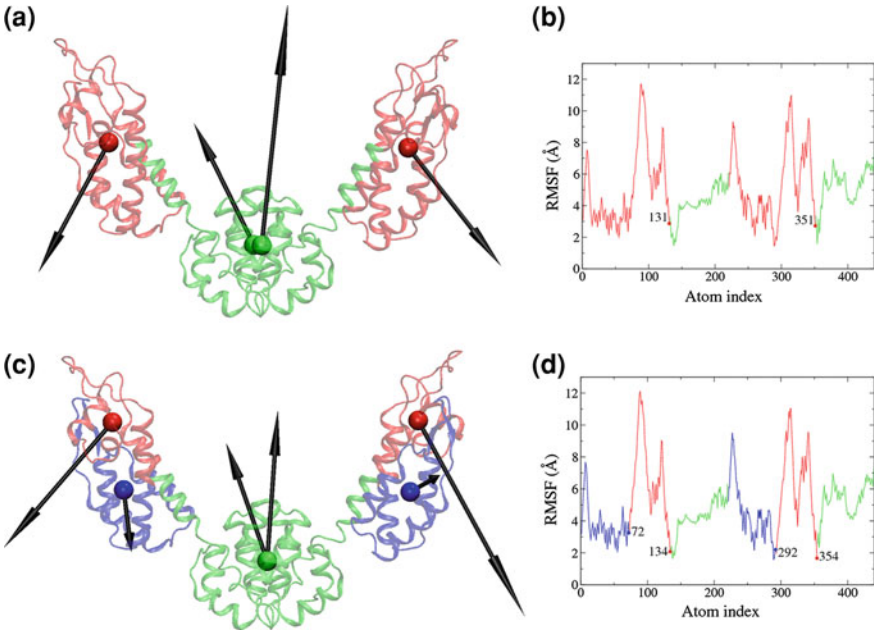


Fig. 4.2 ED-CG maps of the CA dimer. **a** The symmetric four-site map. Each CG site locates at the COM of its corresponding dynamic domain, and the arrows on the sites indicate the first PCA mode calculated from a CG trajectory that was constructed from the atomic MD trajectory based on the four-site CG mapping. **b** The RMSF values of all the C_α atoms, which are colored according to the dynamic domains in the panel a, and the boundary atoms are labeled. **c** The symmetric six-site map. **d** The RMSF values of all the C_α atoms, which are colored according to the dynamic domains in the panel **c**, and the boundary atoms are labeled. Reprinted from the reference [18], Copyright (2008), with permission from Elsevier

into four domains, which are D1 (residue numbers 1–32, 70–144, and 338–375), D2 (33–69), D3 (145–180, and 270–337), and D4 (181–269). Based on these domains, an intuitive four-site CG map of G-actin is defined (Fig. 4.3a).

The G-actin structure was taken from its ATP-bound state (PDB entry 1NWK) [41]. Low-frequency normal modes were calculated from a 'parameter-free' ENM [31], in which the spring constants between pairs of C_α atoms Eq. (4.4) are proportional to the inverse square of their distance. The sequence-based ED-CG four-site map of the G-actin (Fig. 4.3b) is very different to the intuitive four-site CG map, which is not surprising because the domains D1 and D3 are not contiguous on the primary sequence. That is to say, the intuitive four-site map is unreachable by the sequence-based ED-CG algorithm. In this case we need to switch to the space-based ED-CG algorithm instead. The space-based ED-CG four-site map of the G-actin (Fig. 4.3c) is very similar to the intuitive map with only minor differences, and the former also has a lower residual than the latter. This result supports that the ED-CG method can obtain CG maps that show agreement with these 'natural' domains.

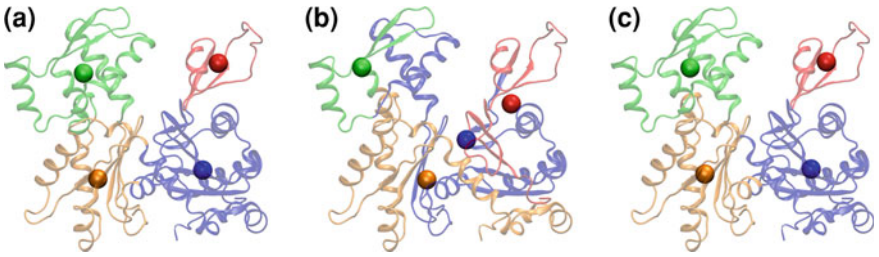


Fig. 4.3 Four-site CG models of the G-actin. **a** The intuitive model: D1 (1–32, 70–144, 338–375) *blue*; D2 (33–69) *red*; D3 (145–180, 270–337) *orange*; and D4 (181–269) *green*. **b** The sequence-based ED-CG model: (1–66) *red*; (67–219) *blue*; (220–256) *green*; and (257–375) *orange*. **c** The space-based ED-CG model: D1 (1–33, 70–140, 337–375) *blue*; D2 (34–69) *red*; D3 (141–181, 261–336) *orange*; and D4 (182–260). Adapted with permission from the reference [23]. Copyright 2010 American Chemical Society

4.4 Application: The *E. Coli* 70S Ribosome

The ribosome is a biomolecular machine, which is responsible for protein biosynthesis in the cell [42–44]. The structure of the *E. coli* 70S ribosome is a highly dynamic RNA-protein assembly, which consists of a 30S small subunit and a 50S large subunit. To start peptide synthesis, the two subunits need to associate through a network of 'bridge interactions' [45, 46]. Computationally expensive MD simulations of the ribosome have been performed [47, 48] and provided valuable information about these interactions in atomic details. However, the picture of the ribosomal bridge interactions at the atomic level is overly complex. We have tried to investigate all the bridge interactions at the CG level instead [49].

From an atomic MD trajectory of the ribosome, we calculated the COM of each residue (amino acids or nucleotides). Thus a COM trajectory of the ribosome was constructed, which was used to carry out PCA. These obtained essential PCA modes were then utilized to build a sequence-based ED-CG map with 480 sites. Detailed analysis has indicated that the CG map does capture those functionally important regions in the ribosome [49]. The interactions between CG sites can be approximated by a 'fluctuation matching' method [50], in which the CG sites within a cutoff distance are connected by effective harmonic bonds. In this network of the CG sites, mean-squared bond-length fluctuations will depend on these spring constants. Note that a CG trajectory of the ribosome can be constructed from its COM trajectory according to the ED-CG map, those bond-length fluctuations may also be calculated from the CG trajectory. Therefore, the spring constants are optimized by matching the fluctuations to the MD data of the ribosome. A large spring constant may indicate a strong interaction between two CG sites (e.g., two dynamic domains in the ribosome).

Table 4.1 Inter-subunit bridges between the small and large subunits based on the ribosome ED-CG model

Bridge	30S subunit	50S subunit	CG interactions (largest k)
B1a	S13[a]	H38[b]	3[c] (0.3[d])
B1b	S13	L5[e]	8 (1.0)
B2a	h44[f]	H69	4 (0.5)
B2b	h24	H67, H69	2 (1.4)
	h45	H69, H71	2 (2.2)
B2c	h24	H67	1 (0.8)
	h27	H67	1 (3.6)
B3	h44	H71	1 (13.3)
B4	h20	H34	1 (0.06)
	S15	H34	2 (1.8)
B5	h44	H64	2 (6.0)
	h44	L14	12 (3.7)
	h44	H62	2 (13.3)
B6	h44	H62	2 (13.3)
	h44	L19	8 (4.3)
B7a	h23	H68	0 (0.04)
B7b	h23	L2	5 (0.5)
	h24	L2	3 (1.4)
B8	h14	L14	5 (2.0)

[a] S13 means the S protein 13. [b] H38 means the helix 38 in the large subunit. [c] Number of CG interactions. [d] The largest spring constant (in $k_BT/\mathring{A}^2$) in CG interactions that belong to a certain bridge. [e] L5 means the L protein 5. [f] h44 means the helix 44 in the small subunit. Reprinted with permission from the reference [49]. Copyright 2011 American Chemical Society

From a static structure of the ribosome, one can visualize intersubunit bridges (from B1 to B8) and identify ribosomal components involved in the bridge interactions [46]. However, it is not straightforward to investigate how strong these bridges are by either a single structure or an atomic MD trajectory. Here we map the bridge interactions onto the CG sites and estimate their strength based on the number of CG interactions and their spring constants. Almost all the intersubunit bridges are preserved by one or more CG interactions (Table 4.1 and Fig. 4.4). The head of the small subunit and the top of the large subunit are connected by bridges B1a and B1b, which have significantly smaller spring constants than those in the ribosome body (B2 to B8). The head also has the minimal CG interactions with the rest of the small subunit. These results support the notion that the head is mobile and can easily move relative to the body, which is consistent to the functional motions of the head during the translocation process of the ribosome [51].

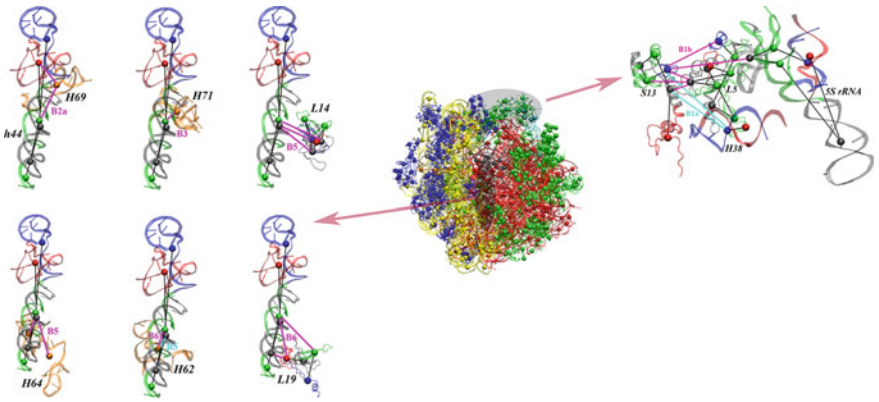


Fig. 4.4 Inter-subunit bridges between the small and large subunits, which are described by interactions between the CG sites. The bridge interactions are divided into two groups, which are the interactions between the head of the 30S subunit and the top of the 50S subunit, and the interactions located at the ribosome body, respectively. Adapted with permission from the reference [49]. Copyright 2011 American Chemical Society

4.5 Conclusions

This chapter introduces a systematic method to define CG maps of biomolecules when the CG sites are coarser than the resolution of one-site per residue. In this method, the atomic structure of the biomolecule is divided into a specified number of dynamic domains, which are optimized to capture the essential dynamics of the biomolecule. That is to say, the method has a functionality to identify domains, as the other methods like the DynDom algorithm [52]. In terms of coarse-graining, the COM of each domain is taken as a CG site. The ED-CG method has been applied to various biological systems, such as multi-domain proteins [18], protein complexes [22], and even biomolecular machines [23, 49]. The decomposition of dynamic domains is consistent to chemical intuition (Fig. 4.2 and 4.3). Those CG sites do identify functionally important regions in the biomolecule.

In the ED-CG scheme, the essential dynamics can be obtained by either PCA on a structural ensemble [18] or ENM of a single structure [22]. It is important to include enough number of essential modes (PCA or normal modes) in the essential subspace to constitute a relatively stable basis set [53]. In practice, one can use the first $3N$-6 essential modes since an N-site CG model has $3N$-6 internal DOF. A simpler but safe choice is to take those essential modes that contribute about 95 % of the total fluctuation as the essential subspace, in this case the number of essential modes is usually less than 5 % of the total number of collective coordinates [25]. The number of CG sites needs to be determined before the ED-CG calculations, which would certainly depend on what properties of the biomolecule to be investigated at the CG level. Since the residual (Eq. 4.8 or 4.10) is always decreasing when N increases (note that χ^2 is naturally 0 when N equals to the

number of residues), the residual itself may not be a good criterion to determine the optimal number of CG sites. This issue has recently been addressed by Voth and co-workers [54], to optimize the number of CG sites in different structural components of a biomolecular complex. An ED-CG map can be sequence- or space-based, but the former is recommended unless the space-based map is really necessary (Fig. 4.3c). With the same number of CG sites, the sequence-base algorithm is computationally more efficient than the space-based algorithm. The sequence-based ED-CG map always has a better convergence than the space-based one. Another advantage of the sequence-based CG map is that, it is straightforward to define effective bonded interactions between CG sites (bonds, angles, dihedrals, et al.) since they preserve the underlying primary sequence. The resulting CG model may potentially be used to explore large conformational changes, and meanwhile the CG map does not need to be altered.

The ED-CG method itself does not deal with the interactions between the CG sites, such information may be derived from all-atom MD data. In a simple case, any two CG sites within a cutoff distance are connected by a harmonic bond. Those spring constants can be determined by matching the computed bond-length fluctuations to those from the MD trajectory. Then it is possible for us to investigate the interactions between structural components in the biomolecule at the CG level, according to the strength of the harmonic springs. Therefore, ED-CG in combination with the fluctuation matching method provides a useful tool to build a CG interaction network in any biomolecular complex, and those functional couplings among the structural components in the complex may be reasonably represented by effective interactions between the CG sites. By comparing CG interaction networks of the biomolecule in different functional states, we can identify key changes during the conformational transition. CG simulations of F-actin based on these harmonic interactions have yielded insights on the heterogeneity in actin filaments [39]. More sophisticated CG interactions have also been developed in order to simulate the assembly of HIV-1 viral capsid [36].

References

1. Dror RO, Dirks RM, Grossman JP, Xu HF, Shaw DE (2012) Biomolecular simulation: a computational microscope for molecular biology. Annu Rev Biophys 41:429–452
2. Tozzini V (2005) Coarse-grained models of proteins. Curr Opin Struct Biol 15:144–150
3. Ayton GS, Noid WG, Voth GA (2007) Multiscale modeling of biomolecular systems: in serial and in parallel. Curr Opin Struct Biol 17:192–198
4. Murtola T, Bunker A, Vattulainen I, Deserno M, Karttunen M (2009) Multiscale modeling of emergent materials: biological and soft matter. Phys Chem Chem Phys 11:1869–1892
5. Saunders MG, Voth GA (2012) Coarse-graining of multiprotein assemblies. Curr Opin Struct Biol 22:144–150
6. Voth GA (2009) Coarse-graining of condensed phase and biomolecular systems. CRC Press-Taylor & Francis Group, Boca Raton 2009
7. Marrink SJ, Risselada HJ, Yefimov S, Tieleman DP, de Vries AH (2007) The MARTINI force field: Coarse grained model for biomolecular simulations. J Phys Chem B 111:7812–7824

8. Monticelli L, Kandasamy SK, Periole X, Larson RG, Tieleman DP, Marrink SJ (2008) The MARTINI coarse-grained force field: extension to proteins. J Chem Theory Comput 4:819–834
9. Liwo A, He Y, Scheraga HA (2011) Coarse-grained force field: general folding theory. Phys Chem Chem Phys 13:16890–16901
10. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J 80:505–515
11. Bahar I, Rader AJ (2005) Coarse-grained normal mode analysis in structural biology. Curr Opin Struct Biol 15:586–592
12. Bahar I, Lezon TR, Yang LW, Eyal E (2010) Global dynamics of proteins: bridging between structure and function. Annu Rev Biophys 39:23–42
13. Arkhipov A, Freddolino PL, Schulten K (2006) Stability and dynamics of virus capsids described by coarse-grained modeling. Structure 14:1767–1777
14. Murtola T, Kupiainen M, Falck E, Vattulainen I (2007) Conformational analysis of lipid molecules by self-organizing maps. J Chem Phys 126:17
15. Gohlke H, Thorpey MF (2006) A natural coarse graining for simulating large biomolecular motion. Biophys J 91:2115–2120
16. Stepanova M (2007) Dynamics of essential collective motions in proteins: theory. Phys Rev E 76:16
17. Gfeller D, De Los Rios P (2008) Spectral coarse graining and synchronization in oscillator networks. Phys Rev Lett 100:4
18. Zhang Z, Lu L, Noid WG, Krishna V, Pfaendtner J, Voth GA (2008) A systematic methodology for defining coarse-grained sites in large biomolecules. Biophys J 95:5073–5083
19. Zhang Z, Wriggers W (2008) Coarse-graining protein structures with local multivariate features from molecular dynamics. J Phys Chem B 112:14026–14035
20. Jang H, Na S, Eom K (2009) Multiscale network model for large protein dynamics. J Chem Phys 131:10
21. Potestio R, Pontiggia F, Micheletti C (2009) Coarse-grained description of protein internal dynamics: an optimal strategy for decomposing proteins in rigid subunits. Biophys J 96:4993–5002
22. Zhang Z, Pfaendtner J, Grafmüller A, Voth GA (2009) Defining coarse-grained representations of large biomolecules and biomolecular complexes from elastic network models. Biophys J 97:2327–2337
23. Zhang Z, Voth GA (2010) Coarse-grained representations of large biomolecular complexes from low-resolution structural data. J Chem Theory Comput 6:2990–3002
24. Amadei A, Linnsen ABM, Berendsen HJC (1993) Essential dynamics of proteins. Proteins 17:412–425
25. Kitao A, Go N (1999) Investigating protein dynamics in collective coordinate space. Curr Opin Struct Biol 9:164–169
26. Berendsen HJC, Hayward S (2000) Collective protein dynamics in relation to function. Curr Opin Struct Biol 10:165–169
27. de Groot BL, van Aalten DMF, Scheek RM, Amadei A, Vriend G, Berendsen HJC (1997) Prediction of protein conformational freedom from distance constraints. Proteins 29:240–251
28. Brooks B, Karplus M (1983) Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. Proc Natl Acad Sci USA 80:6571–6575
29. Cui Q, Bahar I (eds) (2006) Normal mode analysis: theory and applications to biological and chemical systems. Chapman & Hall/CRC, London
30. Hinsen K (2009) Physical arguments for distance-weighted interactions in elastic network models for proteins. Proc Natl Acad Sci USA 106:E128–E128
31. Yang L, Song G, Jernigan RL (2009) Protein elastic network models and the ranges of cooperativity. Proc Natl Acad Sci USA 106:12347–12352
32. Yesylevskyy SO, Kharkyanen VN, Demchenko AP (2006) Dynamic protein domains: Identification, interdependence, and stability. Biophys J 91:670–685

33. Flory PJ, Gordon M, McCrum NG (1976) Statistical thermodynamics of random networks [and discussion]. Proc R Soc Lond A Math Phys Sci 351:351–380
34. Kirkpatrick S, Gelatt Jr CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220:671–680
35. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. J Chem Phys 21:1087–1092
36. Krishna V, Ayton GS, Voth GA (2010) Role of protein interactions in defining HIV-1 viral capsid shape and stability: a coarse-grained analysis. Biophys J 98:18–26
37. Chu JW, Voth GA (2005) Allostery of actin filaments: molecular dynamics simulations and coarse-grained analysis. Proc Natl Acad Sci USA 102:13111–13116
38. Chu JW, Voth GA (2006) Coarse-grained modeling of the actin filament derived from atomistic-scale simulations. Biophys J 90:1572–1582
39. Fan J, Saunders MG, Voth GA (2012) Coarse-graining provides insights on the essential nature of heterogeneity in Actin filaments. Biophys J 103:1334–1342
40. Kabsch W, Mannherz HG, Suck D, Pai EF, Holmes KC (1990) Atomic structure of the actin: DNase I complex. Nature 347:37–44
41. Graceffa P, Dominguez R (2003) Crystal structure of monomeric actin in the ATP state. Structural basis of nucleotide-dependent actin dynamics. J Biol Chem 278:34172–34180
42. Steitz TA (2008) A structural understanding of the dynamic ribosome machine. Nat Rev Mol Cell Biol 9:242–253
43. Schmeing TM, Ramakrishnan V (2009) What recent ribosome structures have revealed about the mechanism of translation. Nature 461:1234–1242
44. Yonath A (2009) Large facilities and the evolving ribosome, the cellular machine for genetic-code translation. J R Soc Interface 6:S575–S585
45. Gabashvili IS, Agrawal RK, Spahn CMT, Grassucci RA, Svergun DI, Frank J, Penczek P (2000) Solution structure of the *E. coli* 70S ribosome at 11.5 Å resolution. Cell 100:537–549
46. Yusupov MM, Yusupova GZ, Baucom A, Lieberman K, Earnest TN, Cate JHD, Noller HF (2001) Crystal structure of the ribosome at 5.5 Å resolution. Science 292:883–896
47. Sanbonmatsu KY, Tung CS (2007) High performance computing in biology: multimillion atom simulations of nanoscale systems. J Struct Biol 157:470–480
48. Sanbonmatsu KY (2012) Computational studies of molecular machines: the ribosome. Curr Opin Struct Biol 22:168–174
49. Zhang Z, Sanbonmatsu KY, Voth GA (2011) Key intermolecular interactions in the *E. coli* 70S ribosome revealed by coarse-grained analysis. J Am Chem Soc 133:16828–16838
50. Lyman E, Pfaendtner J, Voth GA (2008) Systematic multiscale parameterization of heterogeneous elastic network models of proteins. Biophys J 95:4183–4192
51. Frank J, Agrawal RK (2000) A ratchet-like inter-subunit reorganization of the ribosome during translocation. Nature 406:318–322
52. Poornam GP, Matsumoto A, Ishida H, Hayward S (2009) A method for the analysis of domain movements in large biomolecular complexes. Proteins 76:201–212
53. Amadei A, Ceruso MA, Di Nola A (1999) On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of prtoeins' molecular dynamics simulations. Proteins Struc Func Genet 36:419–424
54. Sinitskiy AV, Saunders MG, Voth GA (2012) Optimal number of coarse-grained sites in different components of large biomolecular complexes. J Phys Chem B 116:8363–8374
55. Humphrey W, Dalke A, Schulten K (1996) VMD: Visual molecular dynamics. J Mol Graphics 14:33–38

Chapter 5
Quantum Calculation of Protein NMR Chemical Shifts Based on the Automated Fragmentation Method

Tong Zhu, John Z.H. Zhang and Xiao He

Abstract The performance of quantum mechanical methods on the calculation of protein NMR chemical shifts is reviewed based on the recently developed automatic fragmentation quantum mechanics/molecular mechanics (AF-QM/MM) approach. By using the Poisson-Boltzmann (PB) model and first solvation water molecules, the influence of solvent effect is also discussed. Benefiting from the fragmentation algorithm, the AF-QM/MM approach is computationally efficient, linear-scaling with a low pre-factor, and thus can be applied to routinely calculate the *ab initio* NMR chemical shifts for proteins of any size. The results calculated using Density Functional Theory (DFT) show that when the solvent effect is included, this method can accurately reproduce the experimental ^{1}H NMR chemical shifts, while the ^{13}C NMR chemical shifts are less affected by the solvent. However, although the inclusion of solvent effect shows significant improvement for ^{15}N chemical shifts, the calculated values still have large deviations from the experimental observations. Our study further demonstrates that AF-QM/MM calculated results accurately reflect the dependence of $^{13}C_{\alpha}$ NMR chemical shifts on the secondary structure of proteins, and the calculated ^{1}H chemical shift can be utilized to discriminate the native structure of proteins from decoys.

Keywords Automated fragmentation QM/MM · Linear scaling · Protein NMR chemical shift · Ab initio · Solvent effect

T. Zhu · J.Z.H. Zhang · X. He (✉)
State Key Laboratory of Precision Spectroscopy and Department of Physics, Institute of Theoretical and Computational Science, East China Normal University, Shanghai, China
e-mail: xiaohe@phy.ecnu.edu.cn

J.Z.H. Zhang
Department of Chemistry, New York University, New York, NY 10003, USA

D. Wei et al. (eds.), *Advance in Structural Bioinformatics*, Advances in Experimental Medicine and Biology 827, DOI 10.1007/978-94-017-9245-5_5

5.1 Introduction

Nuclear magnetic resonance (NMR) spectroscopy is an invaluable and widely used technique in areas of chemistry, biology and medicine [1, 2]. In proteins, the chemical shift tensors are key parameters in the NMR experiment, enabling signals from different nuclei of any given type in a molecule to be distinguished. During the past decades, there has been significant progress in using chemical shift information for characterization of protein structure and dynamics [3–11].

Although the chemical shifts are probably the most precise parameters that can be obtained for biomolecules, the inherently complex dependency on geometric, dynamic and electronic properties has made accurate calculation of chemical shifts of protein a significant challenge [12–14]. There are mainly two widely used methods to calculate protein NMR chemical shifts: the empirical approach based on the experimental database and the *ab initio* approach based on quantum mechanical (QM) calculations. Empirical methods rely on statistical data derived from a limited set of high-quality 3D structures and make use of empirical or semi-empirical equations to account for the non-sequential environment [15–19]. These methods are usually quite successful in predicting backbone chemical shifts, which are primarily determined by the local secondary structure, but they are not so well suited to handle proteins with nonstandard residues, metal cofactors, or protein-ligand complexes.

Over the past decade, QM methods have become increasingly useful for NMR chemical shift studies. Following the pioneering work of de Dios et al. [20–22], a number of quantum calculations have been carried out for chemical shifts in proteins and peptides [23–33]. However, due to the poor scaling of *ab initio* and DFT methods, it has not been practical to apply standard all-electron quantum chemistry methods to realistic macromolecules. In fact, full quantum mechanical computations on structures with 1,000 atoms or more are currently not routinely feasible. Fortunately, many previous studies have proven that there is no need to include all atoms in the QM NMR calculation because the nuclear shielding is fundamentally a local physical property. Cui and Karplus proposed a method for calculating chemical shifts in the QM/MM framework, and concluded that the QM/MM method can provide good descriptions of the environmental effect on chemical shifts [34]. Frank et al. calculated the chemical shifts using the fragment based adjustable density matrix assembler (ADMA) method [35–37]. Gao et al. also reported a fragment molecular orbital (FMO) method for NMR chemical shift calculations at the Hartree-Fock level [38, 39]. In our previous studies [40, 41, 49], a more efficient automated fragmentation quantum mechanics/molecular mechanics approach (AF-QM/MM) was shown to be applicable to routine *ab initio* NMR chemical shift calculation for proteins of any size. In this approach, the entire protein is divided into individual fragments, and residues within a certain buffer region surrounding each fragment are included in the QM calculation to preserve the chemical environment of the divided fragment. The remainder of the system outside the buffer regions is described by the MM method. The AF-QM/MM

calculated NMR chemical shifts of several proteins are in good agreement with the experimental measurement [40, 41].

Since most NMR measurements are performed on liquid samples, the NMR parameters (in particular NMR chemical shifts) are highly sensitive to the molecular environment, and especially the solvent effect. The effect of solvent on nuclear magnetic shielding parameters derived from NMR spectroscopy has been of great interest for a long time [42–47]. Several empirical approaches have been formulated to evaluate the solvent effects on nuclear shieldings, however, the development of *ab initio* calculation of NMR properties of proteins in solution has only recently received attention, and most of the studies were focused on small molecular structures or model peptides. In this review, we mainly discuss the influence of solvent effects on the QM calculation of protein NMR chemical shifts, by including both the implicit and explicit solvent model based on our previous works [40, 41, 48, 49].

5.2 Automated Fragmentation QM/MM Method

5.2.1 Fragmentation Criteria

The basic fragmentation scheme in the AF-QM/MM approach is shown in Fig. 5.1. In this approach, the entire protein system is divided into non-overlapping fragments termed core regions. The residues within a certain range from the core region are assigned as the buffer region. Both the core region and its buffer region are treated by QM, whereas the rest of the system is described by an empirical point-charge model. The purpose of the buffer area is to include the local QM

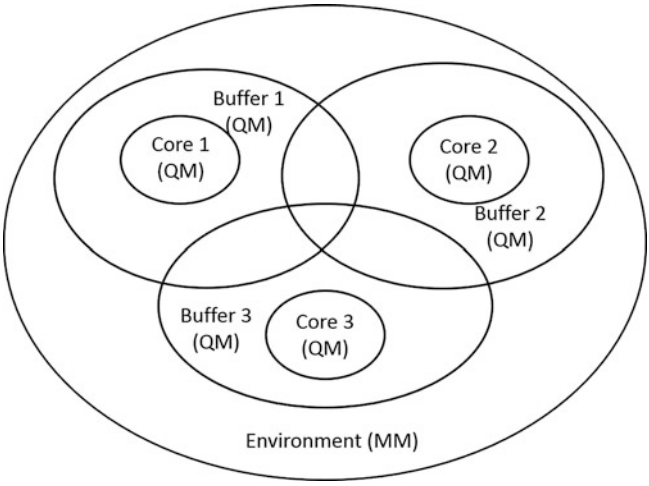


Fig. 5.1 Subsetting scheme for the AF-QM/MM approach

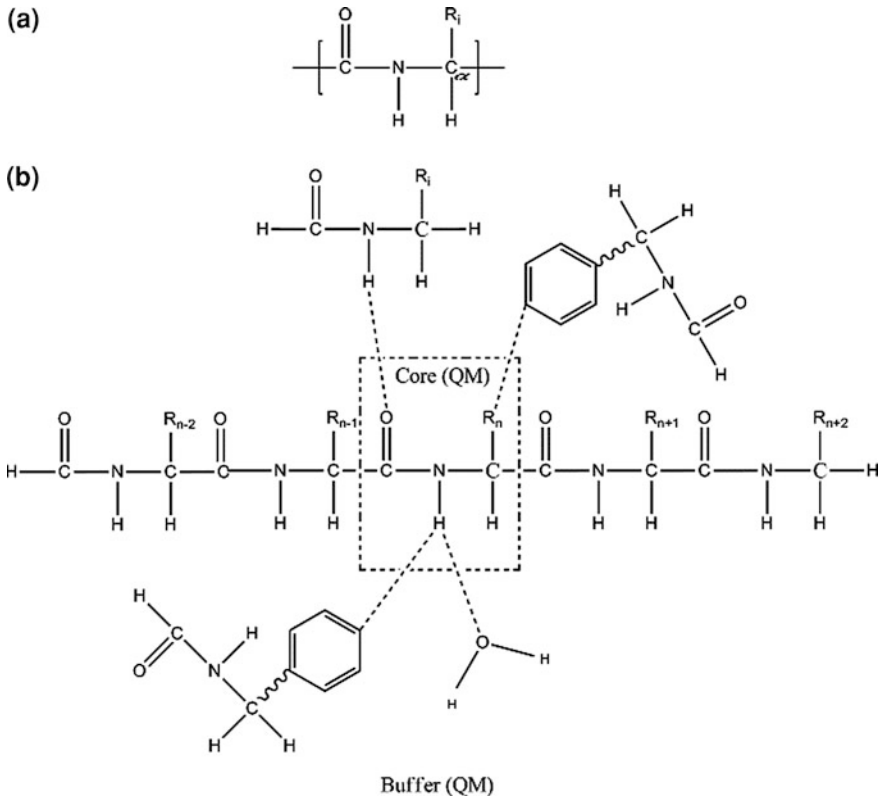


Fig. 5.2 **a** Definition of the residue unit used in this work. **b** *N*th amino acid is the core region. Sequentially connected $(n - 2)$th, $(n - 1)$th, $(n + 1)$th and $(n + 2)$th residues are included in the buffer region. In addition, the residues in spatial contact with the *n*th residue are also assigned to the buffer region (see text for further details)

effects on the chemical shifts. Each fragment-centric QM/MM calculation is carried out separately. Only the shielding constants of the atoms in the core region are extracted from the individual QM/MM calculations. A more detailed illustration of the automated fragmentation scheme is presented in Fig. 5.2.

For proteins discussed in this work, each residue is taken as the core region. A different definition of the residue that consists of the –CO–NH–CHR– is adopted to preserve the electron delocalization across the peptide bond (Fig. 5.2a). A generalized molecular cap was also introduced to take into account the QM polarization effect and charge transfer within the first shell from the residue of interest, as shown in Fig. 5.2b. In this and all our previous studies, we adopt the following distance-dependent criteria to include residues within the buffer region of each core residue: (1) if one atom of the residue outside the core region is less than 4 Å away from any atom in the core region and at least one of the two atoms is a non-hydrogen atom; (2) if the distance between one hydrogen atom in the core

region and the other hydrogen atom outside the core region is less than 3 Å, or (3) if a heavy atom on an aromatic ring is within 5 Å from any atom in the core region. Of course, other distance-dependent criteria could be used to further optimize the choice of the buffer region. The non-neighboring residues in the buffer region are simply capped by hydrogen atoms to construct the closed-shell fragment.

The remaining atoms beyond the buffer region are treated by MM method. A point-charge model is employed to account for the empirical electrostatic field outside the QM region. We use the full point charges for those junction atoms that are replaced by hydrogen atoms. Because a buffer region is added to smoothly link the core region and MM environment, atoms on the boundary between the QM and MM regions are relatively far from the core region and their influence is attenuated. By using a general criterion to assign a buffer zone to each residue, we can reduce the size of each fragment in order to make the QM calculation as small as possible until we strike a compromise between the desired accuracy and the computational cost. Although the total number of residue pairs is proportional to the square of the number of residues, the size of each fragment is independent of the overall protein size because each residue can have only a limited number of residues in its vicinity. Hence, the largest fragment normally contains less than 250 atoms consisting of C, H, O, N, and S, which is an affordable calculation at the HF and DFT levels. In this work, all the QM calculation were performed using Gaussian09 program [50].

5.2.2 Solvent Effects

The main obstacle of including solvent effects in QM/MM NMR calculation is the determination of solvent positions around the biomolecules. It is known that the interaction of the biomolecule with solvent is not well represented by the coordinates present in the experimental structures. There are no water molecules in the protein structure obtained by NMR experiment, and even some "crystallographic" waters are present in the X-ray structure, they represent only a fraction of the waters surrounding the biomolecule. In addition, the static positions of water molecules are probably not representative of the environment seen by the atoms of the solvated biomolecule. Therefore, in most of the calculations, the implicit continuum solvation model was used.

In continuum solvation model, the solute (protein) is represented by a charge distribution $\rho(r)$ embedded in a cavity surrounded by a polarizable medium with dielectric constant ε. The solute charge distribution polarizes the dielectric medium and creates a reaction field which acts back to polarize the solute until equilibrium is reached. The reaction field acting on the solute can be effectively represented by that of induced charges on the cavity surface according to the classical electrostatic theory. In the current approach, we use the DivCon [51] program which combines the linear-scaling divide-and-conquer semi-empirical algorithm with the Poisson-Boltzmann (PB) equation to perform the self-consistent reaction field (SCRF)

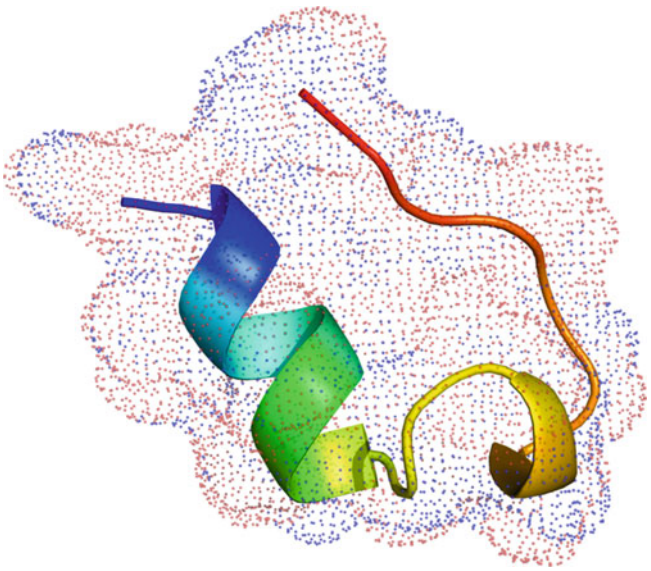


Fig. 5.3 NMR structure of Trp cage (PDB entry: 1L2Y) together with the surface charges calculated by DivCon (*red* and *blue dots* represent the positive and negative charges, respectively)

calculation. The CM2 charges for the atoms of proteins in conjunction with the PM3 methods were derived since the PM3/CM2 is one of the best polarizable charge models for NMR chemical shift calculations, as observed previously for HF/6-31G** and B3LYP/6-31G** calculations [41]. Then the set of point charges of the MM environment and on the molecular surface which represents the reaction field is used as the background charges in the QM calculation. The effective surface charges representing the solvent effects are shown in Fig. 5.3.

5.3 Applications

5.3.1 Comparison with the Full System Quantum Chemistry Calculations

Firstly, the AF-QM/MM method with the solvation model was used to compute the ^{1}H, ^{13}C and ^{15}N absolute chemical shielding tensors of a small protein Trp-cage (20 residues, PDB entry: 1L2Y). And the results are compared with the conventional full system calculations as shown in Fig. 5.4.

In the full system calculation, the protein is computed as an intact molecule with the presence of the same set of surface charges. As one can see from Fig. 5.4, the root mean square errors (RMSEs) for the ^{1}H, ^{13}C and ^{15}N are only 0.06, 0.22 and 0.55 ppm, respectively. All these errors are very small as all of them are less

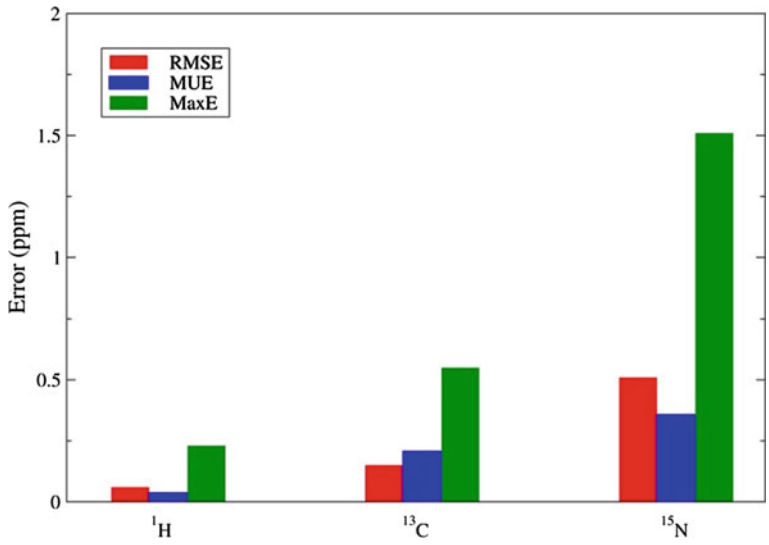


Fig. 5.4 Root mean square error (*RMSE*), mean unsigned error (*MUE*) and maximum error (*MaxE*) of AF-QM/MM with respect to the full system calculated ^{1}H, ^{13}C and ^{15}N chemical shifts in Trp-cage at the B3LYP/6-31G** level

than 1 % of the absolute chemical shielding tensor. The result clearly demonstrates that, as expected, the AF-QM/MM calculated chemical shifts can well reproduce the full quantum mechanical calculations for proteins.

5.3.2 ^{1}H Chemical Shifts in Proteins

Proton chemical shifts are the most important and most studied output of NMR experiments. In proteins, the proton atoms can be divided into two categories. One is called non-polar ^{1}H atoms which usually form covalent bonds with C atoms; the other group is called polar ^{1}H atoms which usually from covalent bonds with polar N, S or P atoms, and most of them are involved in hydrogen bonding interactions. The measured chemical shifts of ^{1}H atoms for these two groups are quite different. In this section, we first compare the non-polar ^{1}H chemical shift of Trp-cage calculated by the AF-QM/MM method with the experimental values. For the hydrogen atoms, calculations in both gas phase and solution phase give excellent agreement with the experimental value as shown in Fig. 5.5a.

The RMSE, MUE, correlation coefficient and the fitted function are given in Table 5.1. Although the calculated results for trp-cage in the gas phase are pretty well, the inclusion of the solvent effects still improves the correlation between the theoretical and experimental values from 0.977 to 0.986. The RMSE also decreased from 0.39 to 0.29 ppm, and the slop of the correlation function is closer to 1.

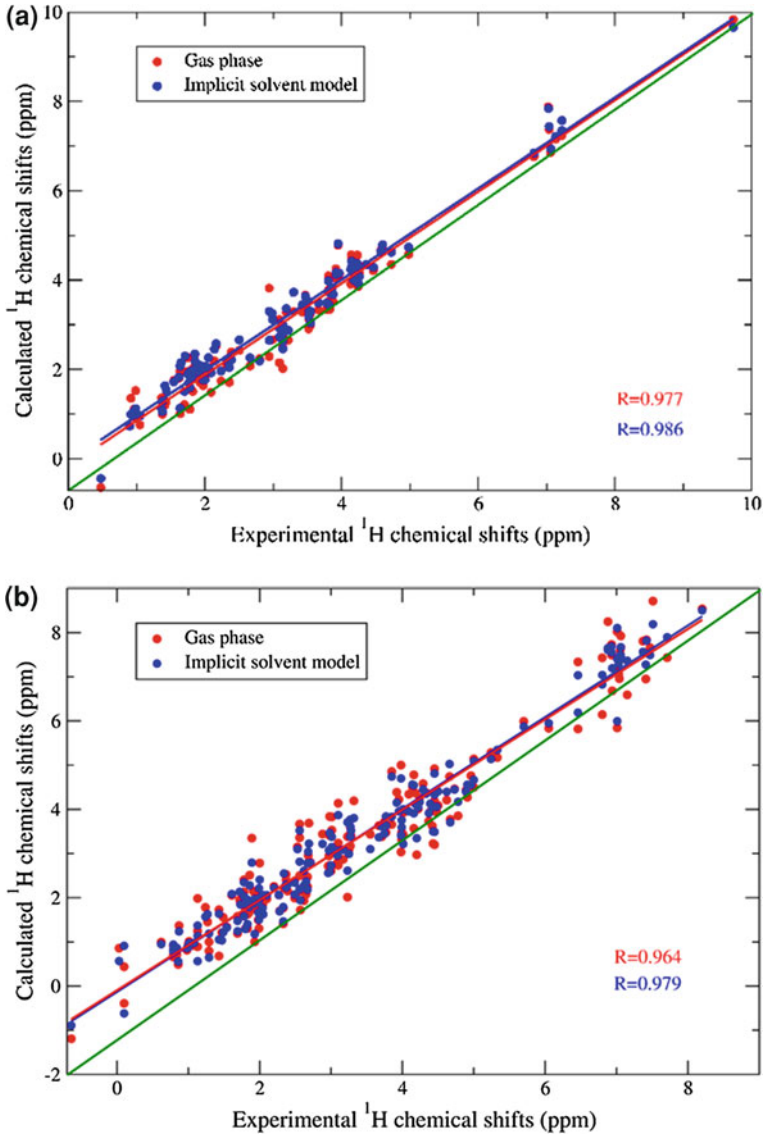


Fig. 5.5 Correlation between experimental and calculated ^{1}H NMR chemical shifts. **a** Trp-cage, **b** Pin1 WW domain. The exchangeable protons were excluded

The results here show that the solvent effects are important and calculated NMR chemical shifts with the solvation model for ^{1}H atom clearly improve the agreement between theory and experiment.

We also calculated the non-polar ^{1}H chemical shifts of Pin1 WW domain (PDB entry: 1PIN) which mainly consists of β-sheets. The comparison of our calculated

Table 5.1 Comparison of AF-QM/MM and experimental chemical shifts for the ^{1}H atoms in Trp-cage and Pin1 WW domain

		RMSE	MUE	R	Correlation function
Trp-Cage	G.	0.39	0.30	0.977	1.024 x −0.17
	S.	0.29	0.23	0.986	1.018 x −0.06
Pin1 WW domain	G.	0.57	0.44	0.964	1.036 x −0.13
	S.	0.43	0.33	0.979	1.023 x +0.09
GB3	G.	0.86	0.39	0.925	0.976 x −0.02
	S.	0.53	0.29	0.983	0.991 x −0.02

G. gas phase; *S.* in solution. The exchangeable protons were excluded

chemical shifts with the experimental values is given in Fig. 5.5b and Table 5.1. Here, the similar result as that of trp-cage is observed. The theoretical ^{1}H chemical shift in solvation is better correlated with the experimental values than results from gas-phase calculations. The RMSE using the solvent model is 0.42 ppm, which is smaller than the gas phase result of 0.57 ppm and the correlation between theoretical and experimental values also improved from 0.964 to 0.979. Thus, the inclusion of the solvent effects clearly improves the theoretical result. In our previous study [40], we also performed calculations on a large protein with mixed α-helical and β-sheet secondary structures, GB3 (PDB entry: 1IGD, 61 residues). The comparison between theoretical and experimental result is also shown in Table 5.1. Again, better correlation with experiment is seen for non-polar ^{1}H chemical shift with the inclusion of solvation.

Furthermore, we check the performance of the AF-QM/MM method on the polar hydrogen especially for protein amide H atoms. The $^{1}H_{N}$ chemical shift is one of the most precise NMR parameters that can be measured, which plays key roles in peak assignments. Thus, a QM model that can accurately predict their chemical shift is in demand. Previous studies have found that the main reason for the inaccuracy in computed amide H chemical shifts arises from the improper treatment of the solvation effect, especially the specific solvent-solute hydrogen bond effect. To include these effects in the calculation, explicit inclusion of solvent molecules is required. In our previous study [49], we used a 3D reference interaction site model (3D-RISM) to correct the distribution of explicit solvent molecules. The algorithm of 3D-RISM method is based on statistical mechanics and has been shown to accurately reproduce water distributions at a reduced computational cost. The PLACEVENT [52, 53] program developed by Hirata and co-workers was utilized to translate the continuous distributions to explicit water molecules. In the calculation, only the water molecules in the first and second solvation shell (within 6.0 Å from any atom in the protein) are regarded as part of the entire system. While the implicit solvent model was used to represent the bulk solvent effect beyond the second solvent shell as shown in Fig. 5.6 [49].

The protein GB3 is taken as the initial geometry. Besides the crystallographic water, 678 more water molecules were added by the PLACEVENT program to

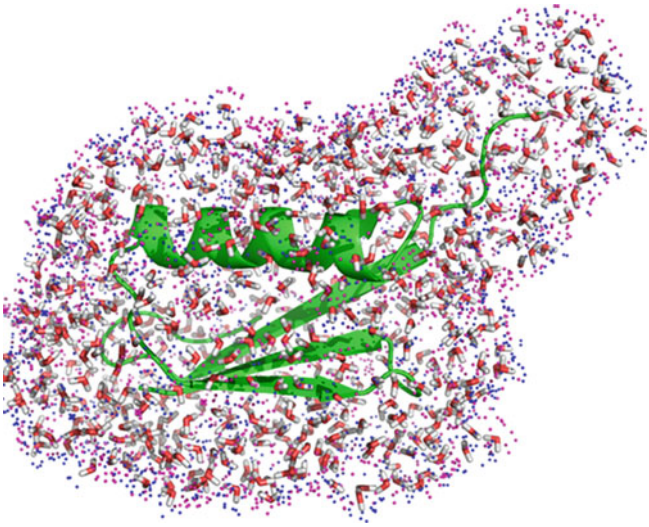


Fig. 5.6 Graphical representation of GB3 (PDB entry: 2IGD) together with the first, second solvation shells and surface charges calculated by DivCon program [51]. (*Red* and *blue dots* represent the positive and negative surface charges, respectively)

mimic the first and second solvent shell. Calculated $^{1}H_{N}$ chemical shifts using both the explicit and implicit solvent models are compared in Fig. 5.7. As can be seen, the inclusion of explicit water molecules gives considerably better agreement with experiment over the implicit solvent model. The correlation coefficient (R) between the theoretical and experimental values is improved from 0.673 to 0.835. The RMSE is also decreased from 1.19 to 0.86 ppm. Table 5.2 lists those residues which have amide protons forming hydrogen bonds (H-bonds) with water molecules. It can be seen that those calculated $^{1}H_{N}$ chemical shifts using the pure implicit solvent model show large upfield shifts as compared to experimental values.

When the explicit solvents were included in the fragment QM calculations, the results show significant improvement. It clearly indicates that hydrogen bonding has large electronic polarization effect on the $^{1}H_{N}$ chemical shift (up to 2–3 ppm). The water molecule which forms direct H-bond with the amide proton in proteins should be treated quantum mechanically to accurately reproduce the experimental $^{1}H_{N}$ chemical shifts.

As shown in Fig. 5.7, although the inclusion of explicit water molecules improves the results, the calculated $^{1}H_{N}$ chemical shifts with the explicit solvent model are systematically underestimated by about 0.5 ppm. Previous studies on some model systems have illustrated that the cooperative hydrogen bonding effect has a non-negligible influence on $^{1}H_{N}$ chemical shifts by affecting the primary hydrogen bond geometry and polarizing the electron density around the amide proton. Therefore, we further explored the cooperative hydrogen bond effect on the protein $^{1}H_{N}$ chemical shifts. For simplicity, we took the N-methylacetamide (NMA) as the central fragment, the cooperative hydrogen bonding effects caused

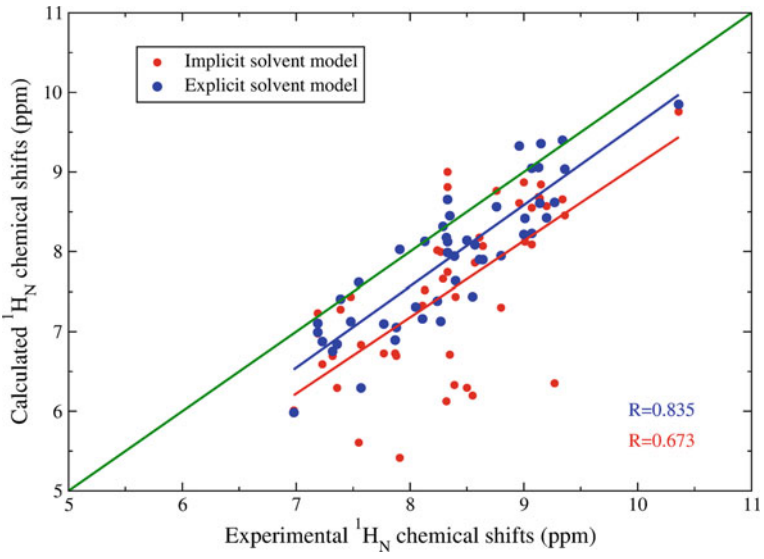


Fig. 5.7 Correlation between the experimental and calculated ^1H_N chemical shifts of GB3 using the AF-QM/MM method (the QM level is at B3LYP/6-31G**). (*red circle* ^1H_N chemical shifts calculated using the implicit solvent model; *blue circle* ^1H_N chemical shift calculated using the explicit solvent model.)

Table 5.2 Comparison of the experimental and AF-QM/MM calculated ^1H_N chemical shifts (in ppm) of GB3 for residues which form hydrogen bonds with water molecules using the explicit and implicit solvent models, respectively

Residue	LEU12	VAL21	ALA23	GLU24	GLY41	TRP43	THR45
Implicit solvation	5.61	6.30	6.13	6.71	5.42	6.35	6.20
Explicit solvation	7.62	8.14	8.18	8.45	8.03	8.62	7.94
Experiment	7.55	8.50	8.32	8.35	7.91	9.27	8.55

The QM level is at B3LYP/6-31G**

by both water and NMA molecules were investigated. As shown in Fig. 5.8, when the cooperative hydrogen bond was formed, the chemical shifts of the ^1H_N atom in the central residue are downfielded by around 0.3–0.5 ppm as opposed to the case of single H-bond. Therefore, we expand our definition of the buffer region to include the secondary hydrogen bond acceptor (the whole residue or water molecule) in the QM region. As depicted in Fig. 5.9, if the ^1H_N chemical shift in the core residue (A) is to be calculated and there is a cooperative hydrogen bond across the peptide bonds of residues: A, B (primary H-bond acceptor) and C (secondary H-bond acceptor), we also include residue C in the buffer region.

The other factors that may govern the accuracy of calculated ^1H_N chemical shifts include the density functional and the size of basis set chosen in our calculation. Previous studies on small organic molecules have demonstrated that, at

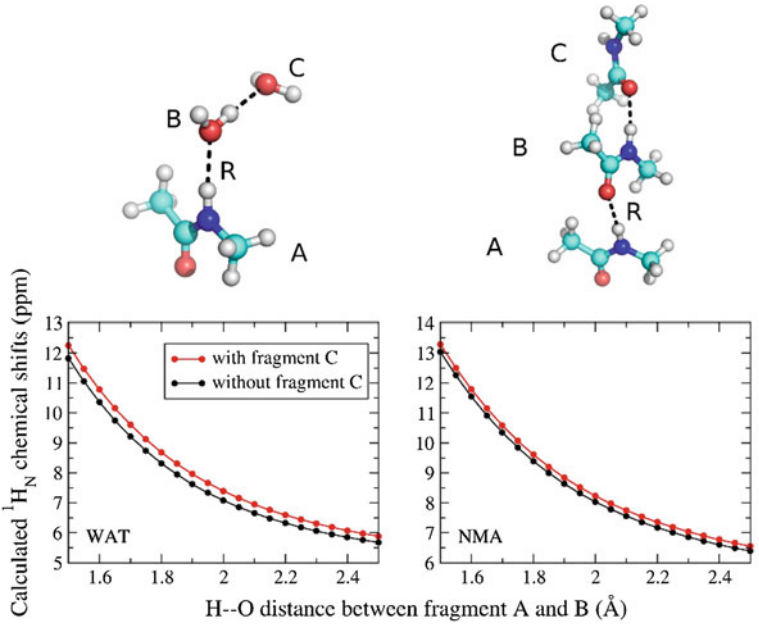


Fig. 5.8 The ^1H_N chemical shift of the central fragment (A) as a function of the ^1H_N-O distance between fragment A and B calculated at the B3LYP/6-311++G** level. *Left panel* both the primary and secondary hydrogen bond acceptors are water molecules; *right panel* both the primary and secondary hydrogen bond acceptors are N-methylacetamides (NMAs). The H-bond length between fragment B and C are fixed at the original optimized structure at the B3LYP/6-31G** level (1.98 Å for WAT-WAT and 2.09 Å for NMA-NMA, respectively.)

least a triple-zeta basis set with the diffuse basis function should be utilized to accurately reproduce the experimental amide hydrogen chemical shift. However, the computational cost is very demanding to apply large basis sets on the entire QM region consisting of normally 150–300 atoms, which is the normal size of each fragment (core + buffer region) using the current definition of the buffer region. Hence, the use of locally dense basis sets, i.e. the combination of two basis sets where the larger one is used for the atoms of interest and the smaller one for all the other atoms, is adopted. The 6-311++G** basis set was employed on the –CO–NH– atoms in both the core residue and other residues involved in the primary and secondary H-bonds (as illustrated in Fig. 5.9). If the H-bond acceptor is a water molecule, the entire water molecule is treated with the 6-311++G** basis set, while the rest atoms in the QM region are set to a smaller basis set. In this work, the 4-31G* basis set has been utilized and the result is shown in Fig. 5.10.

As can be seen, the inclusion of cooperative hydrogen bond effect and applying the locally dense basis set give remarkable improvement for the ^1H_N chemical shifts (compare Fig. 5.10 with Fig. 5.7). The calculation with the B3LYP/6-311++G**/4-31G* method decreases the RMSE from 0.86 to 0.49 ppm. In our previous study [49], we found that the increase of the lower basis set from 4-31G* to 6-31G*

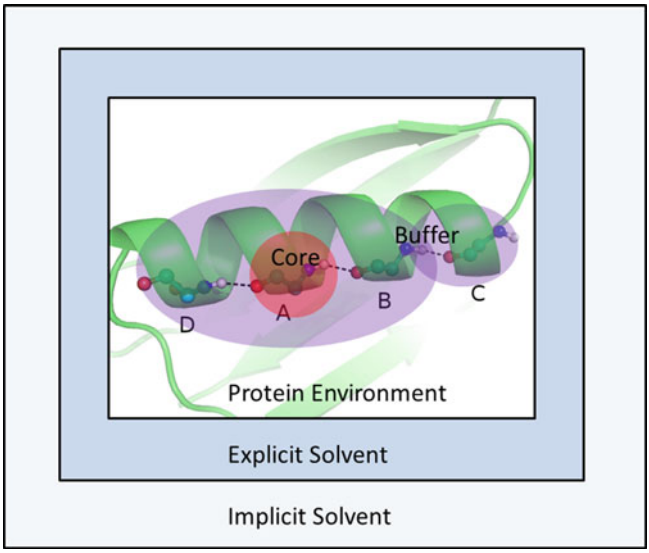


Fig. 5.9 Subsetting scheme for the AF-QM/MM-PB approach with the explicit solvent model. The *red* and *blue* region represents the core and buffer region, respectively. On top of the original definition of the buffer region described in Ref. [41], this study adds one additional criterion which is including the secondary hydrogen bond acceptor (residue C) in the buffer region to take cooperative hydrogen bonding effect into account. The rest of the protein and explicit solvent molecules are described by point charges. The bulk solvent effect is described by the classical electrostatic potential induced by the point charges on the cavity surface calculated using the PB model

or 6-311G** does not reduce the overall RMSE for GB3. Hence, we conclude that the B3LYP functional with the mixed basis set of 6-311++G**/4-31G* strikes a compromise between the computational cost and attained accuracy.

5.3.3 ^{13}C and ^{15}N Chemical Shifts in Proteins

Taking protein GB3 as an example, we also show the influence of solvation effects on the NMR chemical shifts of ^{13}C. The comparison between theoretical and experimental result is shown in Fig. 5.11.

It is not surprising that, as shown in Fig. 5.11, the chemical shift of ^{13}C atom is not significantly influenced by including the solvent effects, with the correlation coefficient of 0.994 in gas phase and 0.998 in implicit solvent. This is mainly because the ^{13}C chemical shifts span a large range from aliphatic region (15–35 ppm) to the carbonyl region (170–180 ppm). To further analyze the result in more details, we also plot the chemical shift of $^{13}C_\alpha$ in Fig. 5.12.

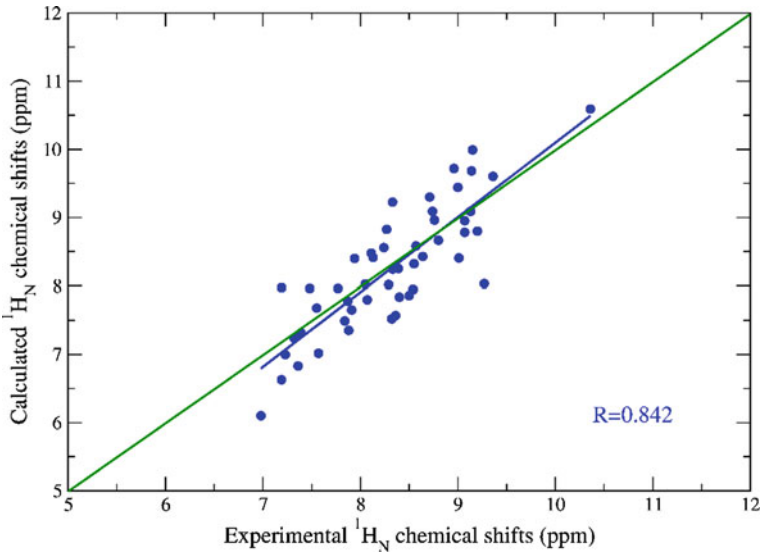


Fig. 5.10 Correlation between experimental and calculated $^{1}H_{N}$ chemical shifts of GB3 using the AF-QM/MM method (the QM level is at B3LYP/6-311++G**/4-31G*)

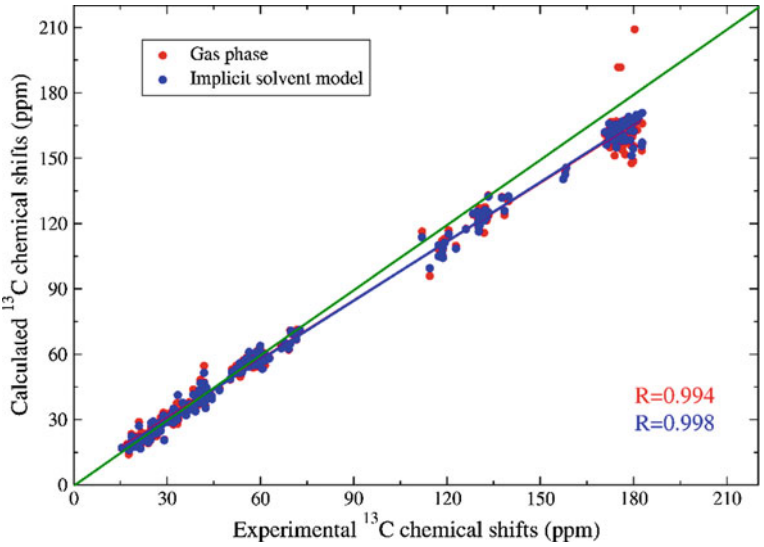


Fig. 5.11 Correlation between the experimental and calculated ^{13}C chemical shifts of GB3 using the AF-QM/MM method (the QM level is at B3LYP/6-31G**) (*red circle* ^{13}C chemical shifts calculated in the gas phase; *blue circle* ^{13}C chemical shifts calculated using the implicit solvent model)

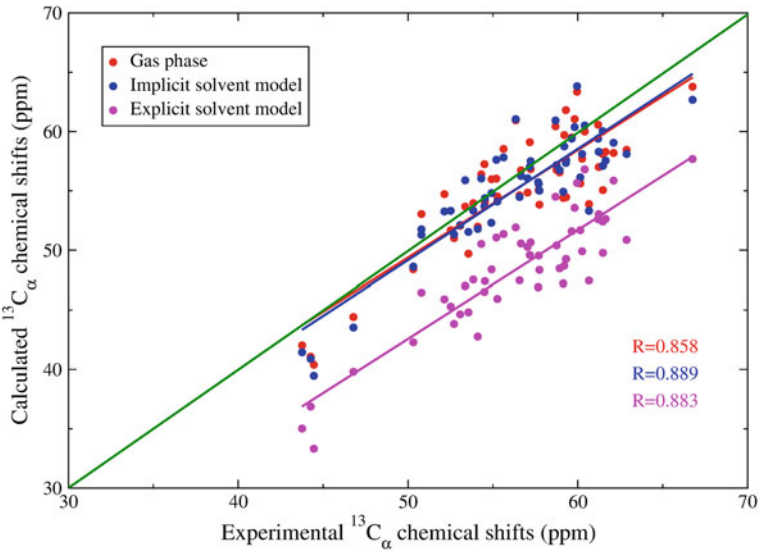


Fig. 5.12 Correlation between the experimental and calculated $^{13}C_\alpha$ chemical shifts of GB3 using the AF-QM/MM method (*red circle* $^{13}C_\alpha$ chemical shifts calculated in the gas phase at the B3LYP/6-31G** level; *blue circle* $^{13}C_\alpha$ chemical shifts calculated using the implicit solvent model at the B3LYP/6-31G** level; *magenta circle* $^{13}C_\alpha$ chemical shifts calculated using the explicit solvent model at the B3LYP/6-311++G**/4-31G* level)

As shown in Fig. 5.12, the inclusion of the implicit solvation model shows some improvement and the overall correlation with experiment increased from 0.858 to 0.889 for GB3, and the RMSE was reduced from 2.89 to 2.41 ppm, but the influence is not very significant. That is mainly because most of the C_α atoms are buried in the hydrophobic core region of the protein and are nonpolar, their chemical shifts are less affected by the solvent. The errors of the calculated $^{13}C_\alpha$ chemical shifts with respect to the experimental values are likely due to the insufficient sampling of the protein structure, since the experimental observed chemical shifts represent the ensemble-averaged values. When the explicit waters were included, the calculated results did not show any improvement and the calculated $^{13}C_\alpha$ chemical shifts using the 6-311++G** basis set are all clearly underestimated with respect to the experimental values. Similar findings have also been concluded in the previous study by Case and co-workers [28].

The ^{15}N results are summarized in Fig. 5.13. It has long been known that the ^{15}N chemical shift in protein presents a challenge for first principle prediction because they are very sensitive to the chemical environment and are influenced by numerous factors. To accurately predict the ^{15}N NMR chemical shifts usually requires high-level electron correlation methods beyond DFT.

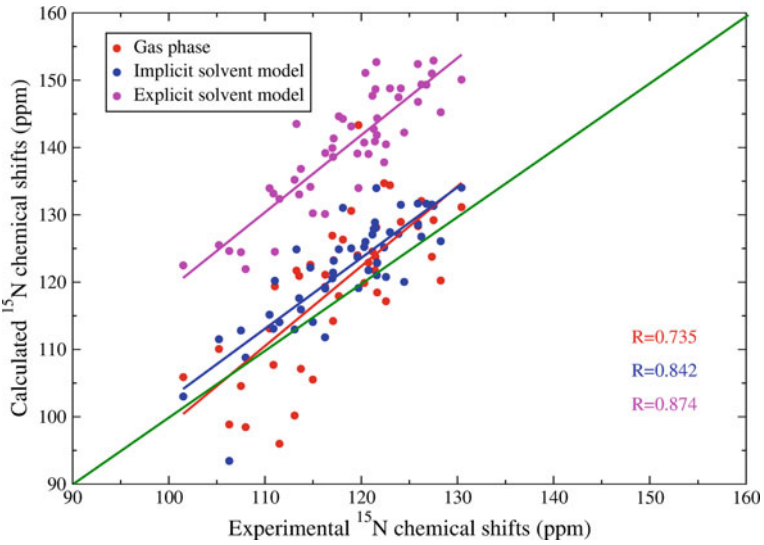


Fig. 5.13 Correlation between experimental and calculated ^{15}N chemical shifts of GB3 using the AF-QM/MM method (*red circle* ^{15}N chemical shifts calculated in the gas phase at the B3LYP/6-31G** level; *blue circle* ^{15}N chemical shifts calculated using the implicit solvent model at the B3LYP/6-31G** level; *magenta circle* ^{15}N chemical shifts calculated using the explicit solvent model at the B3LYP/6-311++G**/4-31G* level)

From the B3LYP/6-31G** calculation, the correlation (R) between the calculated and experimental ^{15}N chemical shifts is only 0.735 for GB3 in the gas phase. Although the inclusion of solvent effects shows significant improvement (with the correlation of 0.842 for GB3), it still has large deviations from the experimental values. As one can see from Fig. 5.13, the implicit solvation treatment on the nitrogen atoms improves more significantly than the nonpolar C_α atoms. However, as discussed in our previous study [40], there is a difference between backbone and side chain nitrogen atoms. For backbone amide nitrogen, which is buried in the core region of protein, the solvent effects on the ^{15}N chemical shift are relatively weak, and the calculated shifts are usually larger than the experimentally measured values [40]. In contrast, solvent effects on the nitrogen atoms from the side chain amine groups (mostly exposed to the solvent) are stronger. However, as shown in Fig. 5.13, including explicit water molecules did not give much improvement. The correlation coefficient is marginally increased from 0.842 to 0.874. Besides the solvent effect, there are other factors which may govern the accuracy of theoretical prediction on ^{15}N NMR chemical shifts, such as conformational sampling, the choice of DFT functionals, etc. Research along these lines is currently underway in our laboratory.

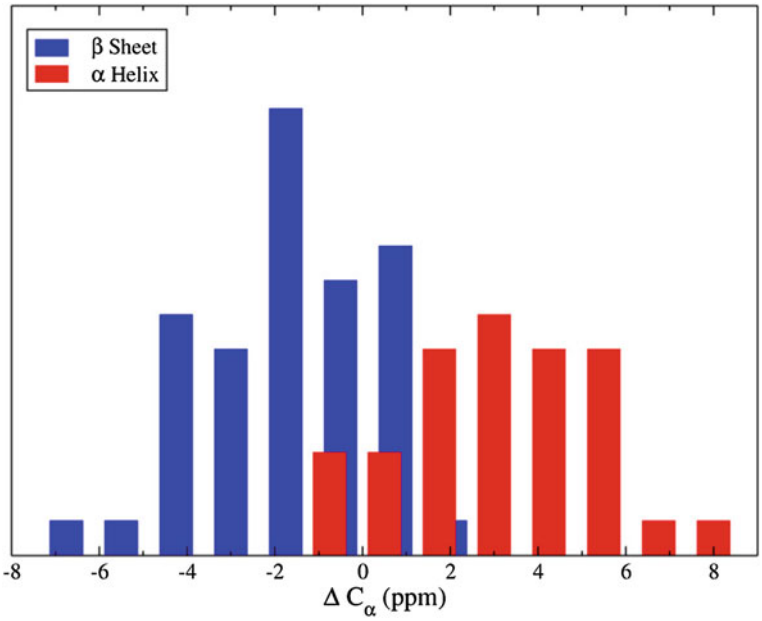


Fig. 5.14 Histograms of secondary shift (the calculated chemical shifts in the native structure minus the random coil values) distribution of α-helical and β-sheet chemical shifts for $^{13}C_\alpha$ in four proteins (Trp-cage, Pin1 WW domain, GB3 and ubiquitin)

5.3.4 Probing the Secondary Structures of Proteins by the AF-QM/MM Method

In structural biology, the protein chemical shift is a powerful tool for studying the structure and dynamics of the protein. They are often used to predict regions of secondary structure in native and nonnative states of proteins, to aid the refinement of complex structures and characterization of conformational changes. Here we validated the capability of using the $^{13}C_\alpha$ secondary chemical shifts (i.e. the calculated chemical shifts in the native structure minus the random coil values) calculated by AF-QM/MM approach to distinguish the α-helix and β-sheet structures. The $^{13}C_\alpha$ random coil chemical shifts are taken from the CamCoil module [54]. The calculated results are presented in Fig. 5.14. As expected, there is a clear separation between the shieldings of the two secondary structure types. The $^{13}C_\alpha$ chemical shift experiences a downfield shift with an average value of 2.55 ppm (with respect to the random coil value) when in a helical configuration and a comparable upfield shift of −2.38 ppm in average when in β-sheet configuration. It shows that the AF-QM/MM method accurately reflects the influence of the local geometry on the chemical shift calculation.

Recent studies [3, 6] have reported that, in combination with traditional molecular mechanical force field or de novo protein structure sampling techniques,

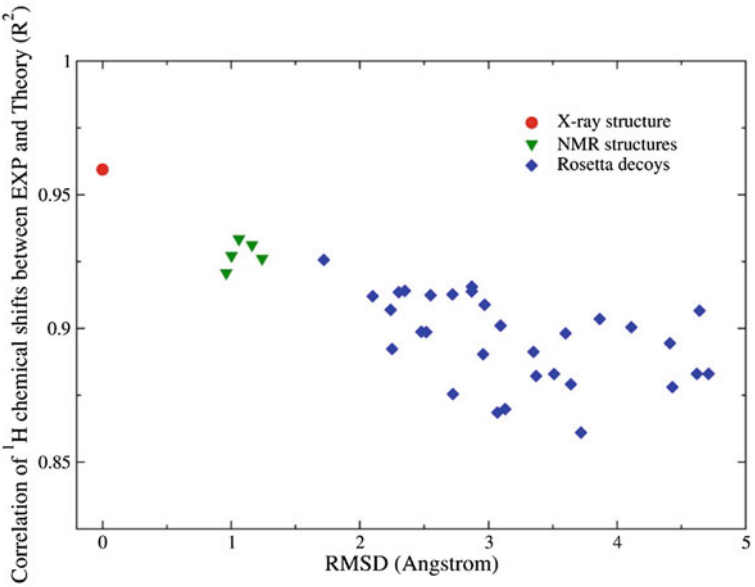


Fig. 5.15 Correlation between the experimental and calculated ^{1}H chemical shifts versus backbone RMSD for Pin1 WW domain (PDB entries for the X-ray structure and NMR structures are 1PIN and 1I6C, respectively)

protein structures can be derived using ^{1}H, ^{13}C and ^{15}N NMR chemical shifts. Hence, we also used the ^{1}H chemical shifts calculated by the AF-QM/MM method in detecting misfolded proteins relative to the natively folded target protein. The X-ray structure of Pin1 WW domain was taken as the native structure and a set of decoy structures for the same amino acid sequence was generated using Rosetta program. Figure 5.15 gives the correlation between calculated and experimental measured chemical shifts versus backbone RMSD with respect to the X-ray structure. As indicated, with the increase of the backbone RMSD, the NMR chemical shift correlation is declined. The RMSD values with the lowest correlations are far from the native structure. The results show that using the chemical shifts to detect the native structure from a collection of decoys is quite remarkable and would have significant potential in this regard.

5.4 Conclusions

In this review, we discussed the performance of QM methods on the calculation of protein chemical shifts based on the recently developed AF-QM/MM approach. By using the PB model and first solvation water molecules, the influence of solvent effect is also explored. Benefit from the fragment algorithm, the AF-QM/MM

approach is computationally efficient and linear-scaling with a low pre-factor. The calculation for each residue takes about 2–4 h of computer time using the current definition of the buffer region. The approach is massively parallel and can be applied to routinely calculate the *ab initio* NMR chemical shifts for proteins of any size.

The calculated results also indicate that when the solvent effect is included, the calculated ^{1}H and ^{15}N chemical shifts show remarkable improvement over those from the gas phase calculations, while the nonpolar ^{13}C chemical shifts are less affected by the solvent. In addition, to accurately calculate the $^{1}H_{N}$ chemical shifts, the explicit solvent method should be taken into account. However, although the inclusion of solvent effect shows significant improvement for ^{15}N chemical shifts, they still have large deviations from the experimental values.

Our study also demonstrated that the AF-QM/MM calculated result accurately reflects the dependence of $^{13}C_{\alpha}$ chemical shifts on the secondary structure of proteins, and the use of ^{1}H chemical shift to discriminate the native structure of proteins from decoys is quite remarkable as proton chemical shift is highly influenced by the local chemical environment. The use of *ab initio* calculated chemical shifts is capable of facilitating accurate protein structure refinement and determination.

The AF-QM/MM method can be further utilized to predict other local chemical properties, such as chemical shift tensor anisotropies and J coupling constants. The applications may also be extended to more general biological systems, such as proteins with nonstandard residues, metalloproteins, protein-ligand, protein-DNA/RNA and membrane protein-lipid complexes.

Acknowledgments This work was supported by the National Natural Science Foundation of China (Grants No. 10974054, 20933002 and 21303057) and Shanghai PuJiang program (09PJ1404000). X.H. is also supported by the Specialized Research Fund for Doctoral Program of Higher Education (Grant No. 20130076120019) and the Fundamental Research Funds for the Central Universities. We thank the Supercomputer Center of East China Normal University for providing us computational time. X.H. also gratefully acknowledges many helpful discussions with Kenneth Merz, Bing Wang, Ning Liao, David Case and Sishi Tang.

References

1. Bieri M, Kwan AH, Mobli M, King GF, Mackay JP, Gooley PR (2011) Macromolecular NMR spectroscopy for the non-spectroscopist: beyond macromolecular solution structure determination. FEBS J 278:704–715
2. Kwan AH, Mobli M, Gooley PR, King GF, Mackay JP (2011) Macromolecular NMR spectroscopy for the non-spectroscopist. FEBS J 278:687–703
3. Cavalli A, Salvatella X, Dobson CM, Vendruscolo M (2007) Protein structure determination from NMR chemical shifts. Proc Natl Acad Sci USA 104:9615–9620
4. Lang WH.; Coats JE, Majka J, Hura GL, Lin Y, Rasnik I, McMurray CT (2011) Conformational trapping of mismatch recognition complex MSH2/MSH3 on repair-resistant DNA loops. Proc Natl Acad Sci USA 108:1–8
5. Selvaratnam R, Chowdhury S, VanSchouwen B, Melacini G (2011) Mapping allostery through the covariance analysis of NMR chemical shifts. Proc Natl Acad Sci USA 108:6133–6138

6. Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D, Bax A (2008) Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci USA 105:4685–4690
7. Wylie BJ, Sperling LJ, Nieuwkoop AJ, Franks WT, Oldfield E, Rienstra CM (2011) Ultrahigh resolution protein structures using NMR chemical shift tensors. Proc Natl Acad Sci USA 108:16974–16979
8. Ulmer TS, Ramirez BE, Delaglio F, Bax A (2003) Evaluation of backbone proton positions and dynamics in a small protein by liquid crystal NMR spectroscopy. J Am Chem Soc 125:9179–9191
9. De Gortari I, Portella G, Salvatella X, Bajaj VS, van der Wel PC, Yates JR, Segall MD, Pickard CJ, Payne MC, Vendruscolo M (2010) Time averaging of NMR chemical shifts in the MLF peptide in the solid state. J Am Chem Soc 132:5993–6000
10. Sahakyan AB, Vranken WF, Cavalli A, Vendruscolo M (2011) Using side-chain aromatic proton chemical shifts for a quantitative analysis of protein structures. Angew Chem Int Ed 50:9620
11. Sahakyan AB, Vranken WF, Cavalli A, Vendruscolo M (2011) Structure-based prediction of methyl chemical shifts in proteins. J Biomol NMR 50:331–346
12. Helgaker T, Jaszunski M, Ruud K (1999) Ab initio methods for the calculation of NMR shielding and indirect spin-spin coupling constants. Chem Rev 99:293–352
13. Facelli JC (2011) Chemical shift tensors: theory and application to molecular structural problems. Prog Nucl Magn Reson Spectrosc 58:176–201
14. Saito H, Ando I, Ramamoorthy A (2010) Chemical shift tensor—the heart of NMR: insights into biological aspects of proteins. Prog Nucl Magn Reson Spectrosc 57:181–228
15. Moon S, Case DA (2007) A new model for chemical shifts of amide hydrogens in proteins. J Biomol NMR 38:139–150
16. Kohlhoff KJ, Robustelli P, Cavalli A, Salvatella X, Vendruscolo M (2009) Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. J Am Chem Soc 131:13894–13895
17. Baskaran K, Brunner K, Munte CE, Kalbitzer HR (2010) Mapping of protein structural ensembles by chemical shifts. J Biomol NMR 48:71–83
18. Han B, Liu Y, Ginzinger SW, Wishart DS (2011) SHIFTX2: significantly improved protein chemical shift prediction. J Biomol NMR 50:43–57
19. Sahakyan AB, Vranken WF, Cavalli A, Vendruscolo M (2011) Structure-based prediction of methyl chemical shifts in proteins. J Biomol NMR 50:331
20. De Dios AC, Pearson JG, Oldfield E (1993) Secondary and tertiary structural effects on protein NMR chemical shifts: an ab initio approach. Science 260:1491–1496
21. De Dios AC (1996) Ab initio calculations of the NMR chemical shift. Prog Nucl Magn Reson Spectrosc 29:229–278
22. De Dios AC, Pearson JG, Oldfield E (2008) Secondary and tertiary structural effects on protein NMR chemical shifts: an ab initio approach. J. Chem. Phys. 260:1491–1496
23. Ochsenfeld C, Kussmann J, Koziol F (2004) Ab initio NMR spectra for molecular systems with a thousand and more atoms: a linear scaling method. Angew Chem Int Ed 43:4485–4489
24. Oldfield E (2002) Chemical shifts in amino acids, peptides, and proteins: from quantum chemistry to drug design. Ann Rev Phys Chem 53:349–378
25. Abildgaard J, Hansen PE, Manalo MN, LiWang A (2009) Deuterium isotope effects on 15 N backbone chemical shifts in proteins. J Biomol NMR 44:119–126
26. Tang S, Case DA (2011) Calculation of chemical shift anisotropy in proteins. J Biomol NMR 51:303
27. Beer M, Kussmann J, Ochsenfeld C (2011) Nuclei-selected NMR shielding calculations: a sublinear-scaling quantum-chemical method. J Chem Phys 134:074102
28. Moon S, Case DA (2006) A comparison of quantum chemical models for calculating NMR shielding parameters in peptides: mixed basis set and ONIOM methods combined with a complete basis set extrapolation. J Comput Chem 27:825–836

29. Komin S, Gossens C, Tavernelli I, Rothlisberger U, Sebastiani D (2007) NMR solvent shifts of adenine in aqueous solution from hybrid QM/MM molecular dynamics simulations. J Phys Chem B 111:5225–5232
30. Hinton JF, Guthrie P, Pulay P, Wolinski K (1992) Ab initio quantum mechanical calculation of the chemical shift anisotropy of the hydrogen atom in the (H_2O) 17 cluster. J Am Chem Soc 114:1604
31. Vila JA, Aramini JM, Rossi P, Kuzin A, Su M, Seetharaman J, Xiao R, Tong L, Montelione GT, Scheraga HA (2008) Quantum chemical C-13(alpha) chemical shift calculations for protein NMR structure determination, refinement, and validation. Proc Natl Acad Sci USA 105:14389–14394
32. Buehl M, Mourik VT (2011) NMR spectroscopy: quantum-chemical calculations, *Wires Comput*. Mol. Sci. 1:634–647
33. Mourik VT (2006) Density functional theory reveals an increase in the amino (1)H chemical shift in guanine due to hydrogen bonding with water. J Chem Phys 125:191101
34. Cui Q, Karplus M (2000) Molecular properties from combined QM/MM methods. 2. Chemical shifts in large molecules. J Phys Chem B 104:3721–3743
35. Frank A, Onila I, Möller HM, Exner TE (2011) Toward the quantum chemical calculation of nuclear magnetic resonance chemical shifts of proteins. Proteins 79:2189–2202
36. Exner TE, Frank A, Onila I, Moeller HM (2012) Toward the quantum chemical calculation of nmr chemical shifts of proteins. 3. conformational sampling and explicit solvents model. J Chem Theory Comput 8:4818–4827
37. Frank A, Moeller HM, Exner TE (2012) Toward the quantum chemical calculation of NMR chemical shifts of proteins. 2. Level of theory, basis set, and solvents model dependence. J Chem Theory Comput 8:1480–1492
38. Gao Q, Yokojima S, Kohno T, Ishida T, Fedorov DG, Kitaura K, Fujihira M, Nakamura S (2007) Ab initio NMR chemical shift calculations on proteins using fragment molecular orbitals with electrostatic environment. Chem Phys Lett 445:331–339
39. Gao Q, Yokojima S, Fedorov DG, Kitaura K, Sakurai M, Nakamura S (2010) Fragment-molecular-orbital-method-based ab initio NMR chemical-shift calculations for large molecular systems. J Chem Theory Comput 6:1428–1444
40. Zhu T, He Xiao, Zhang JZH (2012) Fragment density functional theory calculation of NMR chemical shifts for proteins with implicit solvation. Phys Chem Chem Phys 14:7837–7845
41. He X, Wang B, Merz KM (2009) Protein NMR chemical shift calculations based on the automated fragmentation QM/MM approach. J Phys Chem B 113:10380–10388
42. Mogelhoj A, Aidas K, Mikkelsen KV, Kongsted J (2008) Solvent effects on the nitrogen NMR shielding and nuclear quadrupole coupling constants in 1-methyltriazoles. Chem Phys Lett 460:129–136
43. Kitevski-LeBlanc JL, Evanics F, Prosser RS (2009) Approaches for the measurement of solvent exposure in proteins by 19F NMR. J Biomol NMR 45:255–264
44. Dracinsky M, Bour P (2010) Computational analysis of solvent effects in NMR spectroscopy. J Chem Theory Comput 6:288–299
45. Witanowski M, Biedrzycka Z, Sicinska W, Grabowski Z (1998) A study of solvent polarity and hydrogen bonding effects on the nitrogen NMR shielding of isomeric tetrazoles and ab initio calculation of the nitrogen shielding of azole systems. J Magn Reson 131:54–60
46. Witanowski M, Sicinska W, Biedrzycka Z, Webb GA (1996) Solvent effects on the nitrogen NMR shieldings of cyanamide and N,N-dimethyl cyanamide. J Mol Struc 380:133
47. Mennucci B, Martinez JM, Tomasi J (2001) Solvent effects on nuclear shieldings: continuum or discrete solvation models to treat hydrogen bond and polarity effects? J Phys Chem A 105:7287–7296
48. Tang M, Sperling LJ, Berthold DA, Schwieters CD, Nesbitt AE, Nieuwkoop AJ, Gennis RB, Rienstra CM (2011) High-resolution membrane protein structure by joint calculations with solid-state NMR and X-ray experimental data. J Biomol NMR 51:227–233

49. Zhu T, Zhang JZH, He X (2013) Automated fragmentation QM/MM calculation of amide proton chemical shifts in proteins with explicit solvent model. J Chem Theory Comput 9:2104–2114
50. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JAJ, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Bakken V, Adamao C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski, JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL. Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople J (2010) Gausian09, revision B.01, Gaussian, Inc., Wallingford, CT
51. Dixon SL, van der Vaart A, Gogonea V, Vincent M, Brothers EN, Suarez D, Westerhoff LM, Jr. Merz KM (1999) DivCon. The Pennsylvania State University, University Park, PA
52. Imai T, Hiraoka R, Kovalenko A, Hirata F (2007) Locating missing water molecules in protein cavities by the three-dimensional reference interaction site model theory of molecular solvation. Proteins 66:804–813
53. Yoshida N, Phongphanphanee S, Maruyama Y, Imai T, Hirata F (2006) Selective ion-binding by protein probed with the 3D-RISM theory. J Am Chem Soc 128:12042–12043
54. De Simone A, Cavalli A, Hsu S-TD, Vranken W, Vendruscolo M (2009) Accurate random coil chemical shifts from an analysis of loop regions in native states of proteins. J Am Chem Soc 131:16332

Chapter 6
Applications of Rare Event Dynamics on the Free Energy Calculations for Membrane Protein Systems

Yukun Wang, Ruoxu Gu, Huaimeng Fan, Jakob Ulmschneider and Dongqing Wei

Abstract Techniques of rare event dynamics were reviewed including the string methods, which will be implemented with the biochemical simulation packages. The existing methods were applied to study biological systems with relevance to drug design and drug metabolism. The rare event dynamics simulations were performed to understand the kinetic and thermodynamic free energy information on the drug binding sites in the M2 proton channel, and also the free energy of insertion and association of membrane proteins and membrane active peptides. Results give a theoretical framework to interpret and reconcile existing and often conflicting opinions.

Keywords Rare event dynamics · Drug design · Free energy

6.1 Introduction

Membrane proteins pay an important role in many cellular processes, energy transduction, active or passive molecules transport, transmembrane signaling, Endocytosis and exocytosis. As to our recent knowledge 20–30 % of protein encoding region of human genome encodes membrane proteins. Furthermore, as a lot of membrane proteins are the terminal or central functional parts of some cellular processes, nearly 50 % of these membrane proteins are considered to be putative drug targets. Study of their structures and their interaction could facilitate to get deep insight of those basic biological processes. Among tens of thousands of

Y. Wang · R. Gu · H. Fan · D. Wei (✉)
State Key Laboratory of Microbial Metabolism, College of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China
e-mail: dqwei@sjtu.edu.cn

J. Ulmschneider
Department of Physics and the Institute of Natural Sciences, Shanghai Jiao Tong University, Shanghai, China

D. Wei et al. (eds.), *Advance in Structural Bioinformatics*, Advances in Experimental Medicine and Biology 827, DOI 10.1007/978-94-017-9245-5_6

protein structures you can find in the protein data bank [1], only hundreds of proteins are membrane proteins, so it is easy to understand how difficult it is to research membrane proteins by experiments.

Computationally studying membrane protein systems can give some useful information which is hard to get by experiment. We can get dynamical behavior of membrane proteins by using molecular dynamics with atomic revolution. Access to massively parallel computational resources and great progress on developing linear scaling molecule dynamics algorithm has removed a lot of limitation of molecular simulations, enabling larger system (millions of atoms) and longer time (μs) to be simulated. But lots of bio-macromolecules work in the time-scale which is much longer than μs. Lots of membrane transport proteins assist the transmembrane movement of substances in millisecond to several seconds. Some Ligand-gated ion channels open themselves by conformational change induced by binding of ligands. Quaternary structure change of proteins occurs on microseconds to milliseconds. These membrane protein activities are very important, however currently can't be stimulated by standard molecular dynamics methods.

Activated processes such as nucleation events during protein folding, conformational changes of macromolecules, or chemical reactions usually occur on a time scale that is much larger than the micro-time scale in the system. The reason is that these processes require an unusually large thermal fluctuation to drive the system over some energy barrier separating the conformations. Because of the wide separation of time scales, it is impossible to study activated processes by conventional molecular dynamics simulations. Those activated processes are usually called rare events.

To understand in depth the molecular mechanism for these membrane proteins' function, the underlying free energy activity should be studied. Free energy gives the most measurable connect between experimental and computational investigation [2, 3]. A calculation of a priori free energy differences with a meaningful accuracy can check the quality of the designed models. The ability of getting an accurate free energy is reachable. A lot of researchers over the last 20 years developed many free energy calculation methods.

This review paper consists of three parts. First part is an introduction of rare event dynamics. The second part gives a review of the intrinsic difficulties to calculate free energy of rare events and some well-developed free energy calculation methods. The last part presents a few examples on membrane protein free energy calculations.

6.2 Rare Events of Proteins in Membrane Systems

Large-scale conformation change of protein happens rarely in the atomic resolution and specifically in the bio-membrane environment. Those rare events sometimes play a very important role in proteins' to function. For example, Voltage-gated potassium channel switched between open and closed channel state by

transmembrane movement of its S4-helix which corresponds to the transmembrane voltage change [4]. Those events can be understood to be a physical phase transition among two or more thermodynamically stable or meta-stable states separated by free energy barriers, which happens rarely because system has to wait for a long time to cross high energy region in phase space. If we know free energy of those states and free energy barrels between those states, we can generally describe membrane protein rare dynamics by this information.

$$\mathbf{A} = -\, k_B T \, \ln Q_{NVT} \tag{6.1}$$

$$\mathrm{A} = \; k_B T \, \ln \left\langle \exp \left[\frac{+H(p^{3N}, r^{3N})}{k_B T} \right] \right\rangle \tag{6.2}$$

($\langle\, \rangle$ represents canonical ensemble average)

6.3 Intrinsic Difficulties to Calculate Free Energy by MD Simulations

The free energy is usually expressed as the Helmholtz function (A) under NVT ensemble or the Gibbs function (G) under NPT ensemble. For the simplicity, take Helmholtz free energy for instance:

Helmholtz free energy A can be represented to be phase space integral as Eq. (6.1). Theoretically Helmholtz free energy can be calculated by using MD or MC sampling method according to Eq. (6.2), however in practice we can't get a correct result by implementing this for two main reasons. Firstly, exponentially increased function in integrand in Eq. (6.2) makes some rarely accessed regions in phase space have considerable contribution to the whole integral. Secondly, the phase space is a very huge and complex high dimensional space which makes the normal sampling method like MD and MC very hard to sample it ergodically. Due to the above reasons accurate calculation of absolute free energy is nearly impossible due to insufficient sampling in a finite length and time scale simulation.

Although it is very hard to calculate the absolute free energy, the relative free energy is easier to calculate. The transition state theory (TST) which was developed in 1935 by Henry Eyring and Michael Polanyi gives the simplest way how rare event happens. If a protein has two meta-stable states: A and B. According to ergodic hypothesis, the time average equals ensemble average, we can run a very long time simulation and get the probability of state A and state B, then we can get the free energy difference between state A and state B by Eq. (6.3). But if free energy barrel between A and B is too big, system has to wait a long time for transition from A to B. In practice we always can't afford for such a long time waiting, then may get problematic result: $P_B = 0$ and $\Delta \mathbf{G} = -\infty$ in the finite simulation time.

$$\Delta G = -k_B T \ln \frac{p_A}{p_B} \tag{6.3}$$

Umbrella sampling as general algorithm was developed in 1977 by G.M. Torrie and J.P. Valleau to solve above problem [7]. The basic idea is that if we can find an order parameter (for example certain function of system atom coordinates) which can monotonously distinguishes state A, state B and saddle point then we can apply an external potential to the order parameter and force system to sample specific phase region of system. The free energy of system projected into order parameter can be gotten by histogram analysis.

The above picture in which two stable states are separated by one high energy saddle point hardly works in practical protein conformational change situation. Those real systems have a rugged potential energy landscape and also have plenty of local minimal energy points between protein's stable state A and state B. When entropic (i.e., volume) effects matter (as they typically do in high dimensions), the saddle points do not necessarily play the role of transition states.

So a serial of transition path based theories and methods were developed in recent years. Transition Path Theory (TPT) was one the most mathematically rigorous theory framework of them [34]. Intuitively, for systems with rugged energy landscapes, TPT replaces the notion of a transition state by the notion of a transition-state ensemble and to replace the notion of most probable transition paths by that of transition tubes (inside which most of the flux of the transition paths is concentrated).

6.4 Well Developed Free Energy Calculation Methods

Although absolute free energy is difficult to calculate, considering two well-defined states X and Y, for example, X could be a hydrophobic peptides with cell membrane surface-bound helix state, Y a transmembrane-inserted helix (TM) states, the free energy difference between X and Y is more physical meaningful and is easy to calculate. There are a lot of methods to solve this problem. Broadly speaking these methods can be classified in three categories, according to their scope and range of applicability:

1. Methods aimed at reconstructing the probability distribution or enhancing the sampling as a function of one or a few predefined collective variables (CVs). There is a great deal of degrees of freedom for a typical molecule dynamics system. However, the number of the intrinsic slow degrees of freedom is usually not too large. One can define some collective variables (CVs) to capture those intrinsic slow degrees of freedom. The histogram of CVs are determined and transferred to free energy which projected to those CVs. If the free energy barrier is too high, the unbiased MD can sample the projected phase space badly. The whole reaction process can be partitioned to several adjacent

windows along CVs, and using some constraint algorithm one can force the system to be sampled nearly uniformly along CVs, so speeds up the convergence greatly. One would choose CVs, for the above case, as the distance between center of mass of the peptide and membrane, in the study of chemical reaction, the distance between reactive atoms and enhance the sampling as a function of these coordinates. Examples of these methods include thermodynamic integration [4, 5], free energy perturbation [6], umbrella sampling [7]. These approaches are very powerful but require a careful choice of the CVs that must provide an intrinsic description of the reaction coordinate. If an important variable is missed the calculation will suffer from quasi-nonergodicity or hysteresis and lack of convergence. Moreover, when the number of involved CVs increases linearly, the cost of CPU time grows exponentially. So for some complex chemical or conformational reaction with a very complicated free energy surface, this kind of methods can't work efficiently.

2. Methods aimed at exploring the transition mechanism and building reactive trajectories [8], such as finite-temperature string method [9, 10], transition path sampling [11–13], transition interface sampling [14], milestoning [15] and forward flux method [16]. These methods do not require in most of the cases the explicit definition of a reaction coordinate, but require an a priori knowledge of the initial and final states of the process that has to be simulated. Take string method for instance, it postulates that in the normal temperature the reaction is taken place following a smooth transition tube (inside which most of the flux of the transition paths is concentrated). Center of the tube is a curved string in the high dimension phase space.
3. Methods in which the phase space is explored simultaneously at different temperature, such as replica exchange [17], or as a function of the potential energy, such as multicanonical MD [18] and Wang–Landau [19]. These approaches are very general and powerful; however, they also meet the troubles from some of the limitations of the first category. These methods actually use potential energy as a generalized CV. In several cases, ordered and disordered states may correspond to the same value of potential energy, or be present in the thermal ensemble at the same temperature. This may lead to hysteresis and convergence problems [20].

6.5 Application of Free Energy Calculations to Membrane Protein Systems

6.5.1 Binding Free Energy for the Influenza a M2 Protein Channel

The influenza A M2 proton channel is critical for the viral life cycle. Two adamantane-based antiviral drugs, amantadine and rimantadine take M2 proton channel as the acting target. Understanding how these drugs bind to the M2

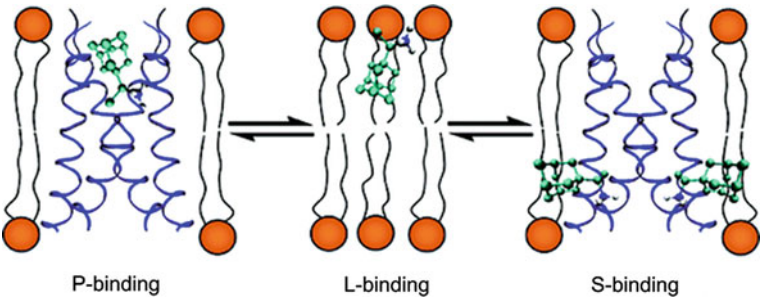


Fig. 6.1 Three possible rimantadine binding site for M2 proton channel membrane system. *P-binding* pore-binding site, *L-binding* lipid-binding site and *S-binding* surface binding site (adapted from [21] with permission from ACS Publications)

channel and block its proton conduction will help design new drug against the fact that virus has quickly obtained drug resistance. There are two alternative binding sites of amantadine and rimantadine in the M2 channel reported recently, with one amantadine molecule bound in the channel pore (pore binding or P-binding) and with four molecules of rimantadine bound at the C-terminal surface of the transmembrane domain of the M2 channel (surface binding or S-binding) in Fig. 6.1, until recently there are a lot controversy about which is the primary binding site.

Gu et al. [21] carried out molecular dynamics simulations and Potential of Mean Force calculations using umbrella sampling on the M2-rimantadine complex for two alternative drug binding models: pore binding and surface binding models. From the PMF calculations for the two drug binding models, the free energy profiles of two binding pattern were obtained in Fig. 6.2. Pore binding requires a high energy barrier to be overcome but is thermodynamically favorable, leading to stable drug binding and inhibition. In comparison, the less energetically stable surface binding site can be easily accessed by rimantadine molecules in the lipid-water environment. These results complement existing work, expand our understanding of these binding sites, and may help guide drug resign and screening studies.

6.5.2 *Free Energy of Insertion of Membrane Proteins and Membrane Active Peptides*

Recently, it has been observed that in some cases, it is possible to generate rare events via straight MD simulations, without need for free energy methods. One such example is the transfer free energy of the insertion of peptides into lipid bilayers from an interfacial state. The most fundamental stability principle of helix-bundle membrane proteins (MPs) is that the free energy of transfer of the constituent transmembrane (TM) helices must favor the membrane rather than the

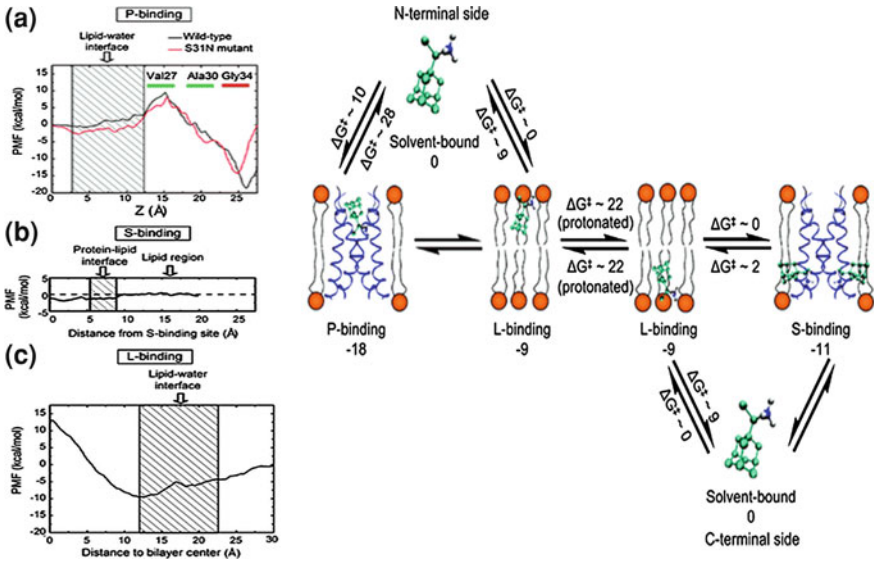


Fig. 6.2 The free energy of three binding site of rimantadine. **a** PMF from solvent-bound state to P-binding state. **b** PMF from L-binding state to S-binding state. **c** PMF from solvent-bound state to L-binding state. (*right figure*) Whole picture about the interaction rimantadine with M2 membrane system (adapted from [21] with permission from ACS Publications)

aqueous phase. This truism has resisted direct quantitation, because of the experimental challenges of measuring water-to-bilayer transfer free energies of hydrophobic peptides. Aggregation in the aqueous phase is the principal issue. Cells have conquered this problem by means of the translocon machinery, consisting primarily of the SecY complex of membrane proteins in bacteria and archaea and the highly homologous Sec61 complex in eukaryotes. To circumvent the experimental challenges of partitioning transmembrane segments across lipid membranes, Ulmschneider et al. have adopted a computational approach using molecular dynamics (MD) simulations carried out in the microsecond time regime [22, 23]. Because the simulations use the same TM segments used in a recent in vitro study of the translocon-assisted insertion of poly-leucine segments of various lengths, it was possible to compare direct peptide partitioning with translocon-to-bilayer partitioning. The strength of this partitioning approach is that all states populated at equilibrium are directly detected, and the free energy between them is obtained from their relative occupancies (Fig. 6.3). Key to the success of this method is the use of elevated temperatures to speed up rare event kinetics. Unfolding is not observed due to the high thermostability of hydrophobic peptide in membranes. How far such direct equilibrium approaches can be applied to related biophysical simulation studies has yet to be investigated, but they appear to be a promising alternative to free energy perturbation techniques, which are usually limited by large hysteresis errors for these type of transitions.

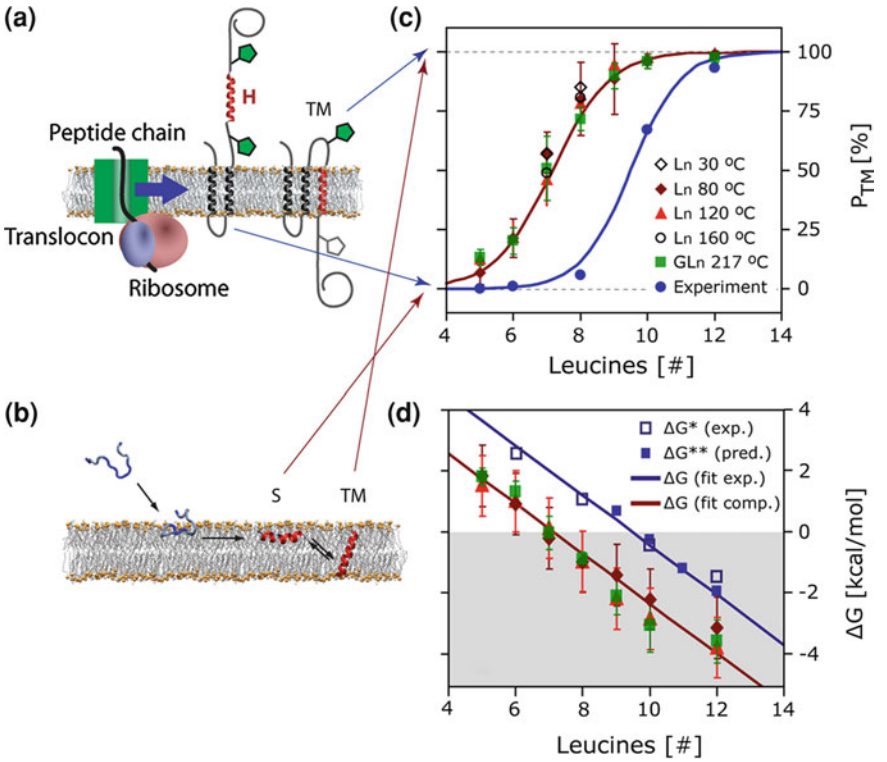


Fig. 6.3 Bilayer insertion efficiency and transfer free energy as a function of peptide length n. **a** The experimental values are for translocon mediated insertion into dog pancreas rough microsomes of GGPG-(L)n-GPGG constructs embedded into the leader peptidase carrier sequence. **b** The computed values are for spontaneous partitioning of Ln peptides into POPC lipid bilayers at 30–160 °C, and for GGPG-(L)n-GPGG at 217 °C. **c** Insertion propensity (**d**) free energy of insertion ΔG(n) as a function of peptide length n (insertion for negative ΔG—*shaded*). The *straight lines* indicate the two-state Boltzmann fit, while the data points show the computed (*red*, *green*) and experimental (*blue*) values for the individual peptides (*measured ΔG, peptide IDs: 43 and 380–383; **predicted ΔG, http://dgpred.cbr.su.se/) (adapted from [22, 23] with permission from ACS Publications)

6.5.3 Free Energy of Transmembrane Peptides Helix Tilting

Hydrophobic match or mismatch in transmembrane (TM) helices (or proteins) refers to the match or mismatch between the length of the hydrophobic core of the helix and the native thickness of the hydrocarbon region of the membrane. Hydrophobic mismatch is a fascinating and important example of mutual protein−membrane interaction. The tilt angle between TM helices and membrane is a direct response to TM helices membrane mismatch. At the positive mismatch, the TM helix tilts and the membrane swells to prevent the hydrophobic part of TM exposed into the hydrophilic environment. But experiments and MD simulation both find

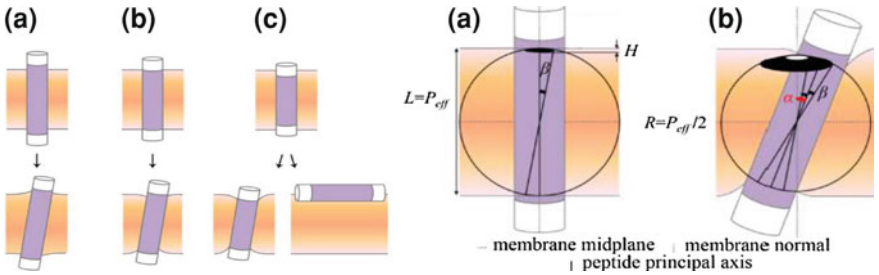


Fig. 6.4 *Left figure* helix-membrane configurations with **a** positive hydrophobic mismatch, **b** perfect match, and **c** negative hydrophobic mismatch. The helix is represented as a *cylinder*, with the hydrophobic core in *purple* and the hydrophilic termini in *white*. *Right figure* the precession entropy gain associated with TM helix tilting in the membrane. β is the maximum amplitude of around helix axis with tilt angle α. Assuming that β is independent on α. The helix's precession entropy is proportional to the *dark cap* or *ring*-like surface area. Helix with small tilt angle (*right A*) has smaller precession entropy than that with large tilt angle (*right B*) (adapted from [26] with permission from ACS Publications)

TM helix always tilts in a certain degree in perfect match and even negative mismatch situation (Fig. 6.4). Free energy calculation of transmembrane Helix Tilting can explore the microscopic forces governing the helix tilting in membranes. Lee et al. using umbrella sampling studied the potential of mean force (PMF) as a function of tilt angle τ of WALP19, a TM model peptide (hydrophobic length L = 19.5 A), in a dimyristoylphosphatidylcholine membrane (width of the effective hydrocarbon lipid region (P_{eff}) = 25.4 A) [24]. The PMF shows a wide range of thermally accessible tilt angles (5–22) with a minimum at $\tau = 12.5$. The free energy decomposition reveals that the helix tilting up to $\tau = 12.5$ is mostly driven by the entropy contribution arising from the helix precession around the membrane normal, whereas the PMF increase after $\tau = 12.5$ results from helical deformation due to the sequence specific helix-lipid interactions.

6.5.4 Free Energy of Transmembrane α-Helixes Self-assembly and Association

Many channel proteins contain a central pore lined by a bundle of approximately parallel **α**-helices. Such channels range in complexity from the M2 protein of influenza A (ca. 100 amino acids per subunit) to the nicotinic acetylcholine receptor (ca. 500 amino acids per subunit). Given the importance of this structural motif in a number of channel proteins, it is important to have a simple yet detailed model system for channels formed by α-helices. Given some membrane inserted **α**-helices, the first step to self-assemble into a functional structure is association of those discrete peptides. Because self-assemble of **α**-helices is usually a energy downhill process, when two peptides are associated into a dimer, then the process will be

followed that two dimer forms a tetramer or one dimer and one helix forms a trimer. Association of TM helix is very hard to be studied by experiments. Recently several theoretical computational studies have aimed to elucidate the detailed atomic interactions and driving forces of TM helix association [24–28]. Most notably, Hénin et al. recently calculated the dimerization free energy of the GpA TM region by calculating the potential of mean force (PMF) as a function of the distance between the centers of mass of the helices from MD simulations in a lipid membrane. The free energy was decomposed into helix-helix and helix-solvent contribution which has greatly improved our understanding of the recognition and association mechanism of the GpA TM domain [28]. Zhang et al. recently computed the standard association free energy of GpA with an implicit membrane model, then derived translational, rotational, and conformational entropy contributions from the total free energy. The gotten association free energy of GpA in micelles gives a good agreement with the experimental result [29].

A framework of reaction coordinates which describes helix-helix distance and crossing angle was developed by Lee and Im [30]. They applied external potential to those RCs to enhance sampling in MD simulations. Lee et al. using those RC to explore the role of hydrogen bonding and helix-lipid interactions in transmembrane helix pVNVV peptides association in DMPC membrane [31]. They found that the As n residues in the middle of the helices show the most significant per-residue contribution to the PMF with various hydrogen bonding patterns as a function of helix-helix distance. Release of lipid molecules between the helices into bulk lipid upon helix association makes the helix-lipid interaction enthalpically unfavorable but entropically favorable.

6.6 Methods to Solve the Difficult Convergence of Free Energy Calculations in Membrane Protein System

As mentioned above, free energy calculation is very hard to be gotten converged, for there is huge volume of rugged phase space to be sampled. Even for the simplest TM homo-to-dimer Association case, there are many degrees of freedom to be explored, such as helix to helix distance and crossing angle, tilt and rotation of each helix, and displacement of each helix along the membrane normal. Such high dimensionality in TM helix assembly also makes the computational studies challenging. Most recently, the method of window exchange umbrella sampling molecular dynamics (WE-USMD) with a preoptimized parameter set was recently used to obtain the most probable conformations and the energetics of transmembrane (TM) helix assembly of a generic TM sequence [32]. WEUSMD method with optimal parameter set acquires a significantly more efficient sampling of helix-helix interfaces than normal umbrella sampling method. Park et al. applied WEUSMD method furthermore into Two Dimensional RC space to explore glycophorin A TM domain Association problem [33]. The two-dimensional WEUSMD results demonstrate that the incomplete sampling in the one-dimensional WEUSMD arises from high barriers

along the crossing angle between the GpA-TM helices. Together with the faster convergence in both the assembled conformations and the potential of mean force, the 2D-WEUSMD can be a general and efficient approach in computational studies of TM helix assembly.

References

1. Sussman JL, Lin D, Jiang J, Manning NO, Prilusky J, Ritter O, Abola EE (1998) Protein Data Bank (PDB): database of threedimensional structural information of biological macromolecules. Acta Cryst D54:1078–1084
2. Kollman PA (1993) Free energy calculations: applications to chemical and biochemical phenomena. Chem Rev 93:2395–2417
3. Chipot C, Pohorille A (eds) (2007) Free energy calculations. Theory and applications in chemistry and biology. Springer, Heidelberg
4. Carter EA, Ciccotti G, Hynes JT, Kapral R (1989) Constrained reaction coordinate dynamics for the simulation of rare events. Chem Phys Lett 156:472–477
5. Sprik M, Ciccotti G (1998) Free energy from constrained molecular dynamics. J Chem Phys 109:7737–7744
6. Bash PA, Singh UC, Brown FK, Langridge R, Kollman PA (1987) Calculation of the relative energy of a protein-inhibitor complex. Science 235:574–576
7. Patey GN, Valleau JP (1975) Monte-Carlo method for obtaining interionic potential of mean force in ionic solution. J Chem Phys 63:2334–2339
8. Elber R, Karplus M (1987) A method for determining reaction paths in large molecules: application to myoglobin. Chem Phys Lett 139:375–380
9. EW, Ren W, Vanden-Eijnden E (2005) Finite-temperature string method for the study of rare events. J Phys Chem B 109:6688–6693
10. Maragliano L, Fischer A, Vanden-Eijnden E, Ciccotti G (2006) String method in collective variables: minimum free energy paths and isocommittor surfaces. J Chem Phys 125:024106
11. Dellago C, Bolhuis P, Csajka FS, Chandler D (1998) Transition path sampling and the calculation of rate constants. J Chem Phys 108:1964–1977
12. Dellago C, Bolhuis P, Geissler P (2002) Transition path sampling. Adv Chem Phys 123:1–78
13. Peters B, Trout BL (2006) Obtaining reaction coordinates by likelihood maximization. J Chem Phys 125:054108
14. Van Erp T, Moroni D, Bolhuis P (2003) A novel path sampling method for the sampling of rate constants. J Chem Phys 118:7762–7774
15. Faradjian A, Elber R (2004) Computing timescales from reaction coordinates by milestoning. J Chem Phys 120:10880–10889
16. Allen R, Warren P, Ten Wolde P (2005) Sampling rare switching events in biochemical networks. Phys Rev Lett 94:018104
17. Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. Chem Phys Lett 314:141–151
18. Nakajima N, Higo J, Kidera A, Nakamura H (1997) Flexible docking of a ligand peptide to a receptor protein by multicanonical molecular dynamics simulation. Chem Phys Lett 278:297–301
19. Wang F, Landau DP (2001) Efficient, multiple-range random walk algorithm to calculate the density of states. Phys Rev Lett 86:2050
20. Trebst S, Troyer M, Hansmann U (2006) Optimized parallel tempering simulations of proteins. J Chem Phys 124:174903
21. Gu RX, Liu LA, Wei DQ (2011) Equilibrium of four binding states of anti-viral drug rimantadine in M2-lipid bilayer system. J Am Chem Soc 133(28):10817–10825

22. Ulmschneider JP, Smith JC, White SH, Ulmschneider MB (2011) In silico partitioning and transmembrane insertion of hydrophobic peptides under equilibrium conditions. J Am Chem Soc 133(39):15487–15495
23. Ulmschneider MB, Doux JPF, Killian JA, Smith JC, Ulmschneider JP (2010) Mechanism and kinetics of peptide partitioning into membranes from all-atom simulations of thermostable peptides. J Am Chem Soc 132:3452–3460
24. Lee J, Im W (2008) Transmembrane helix tilting: insights from calculating the potential of mean force. Phys Rev Lett 100:018103
25. Lee J, Im W (2007) Restraint potential and free energy decomposition formalism for helical tilting. Chem Phys Lett 441:132–135
26. Yana G, Turkan H, Nir BT (2012) The transmembrane helix tilt may be determined by the balance between precession entropy and lipid perturbation. J Chem Theory Comput 8:2896–2904
27. Choma C, Gratkowski H, Lear JD, DeGrado WF (2000) Asparagine-mediated self-association of a model transmembrane helix. Nat Struct Biol 7:161–166
28. Hénin J, Pohorille A, Chipot C (2005) The free energy of r-helix dimerization in glycophorin A. J Am Chem Soc 127:8478–8484
29. Zhang J, Lazaridis T (2006) Calculating the free energy of association of transmembrane helices. Biophys J 91:1710–1723
30. Lee J, Im W (2007) Implementation and application of helix-helix distance and crossing angle restraint potentials. J Comput Chem 28:669–680
31. Lee J, Im W (2008) Role of hydrogen bonding and helix-lipid interactions in transmembrane helix association. J Am Chem Soc 130:6456–6462
32. Park S, Kim T, Im W (2012) Transmembrane helix assembly by window exchange umbrella sampling. Phys Rev Lett 108:108102
33. Park S, Im W (2013) Two dimensional window exchange umbrella sampling for transmembrane helix assembly. J Chem Theory Comput 9:13–17
34. WE, Vanden-Eijnden E (2010) Transition-path theory and path-finding algorithms for the study of rare events. Annu Rev Phys Chem 61:391–420

Part II
3D-Structure Prediction and Folding Mechanism of Biological Macromolecules

Chapter 7
Extended Structure of Rat Islet Amyloid Polypeptide in Solution

Lei Wei, Ping Jiang, Malathy Sony Subramanian Manimekalai, Cornelia Hunke, Gerhard Grüber, Konstantin Pervushin and Yuguang Mu

Abstract The process of islet amyloid polypeptide (IAPP) formation and the prefibrillar oligomers are supposed to be one of the pathogenic agents causing pancreatic β-cell dysfunction. The human IAPP (hIAPP) aggregates easily and therefore, it is difficult to characterize its structural features by standard biophysical tools. The rat version of IAPP (rIAPP) that differs by six amino acids when compared with hIAPP, is not prone to aggregation and does not form amyloid fibrils. Similar to hIAPP it also demonstrates random-coiled nature in solution. The structural propensity of rIAPP has been studied as a hIAPP mimic in recent works. However, the overall shape of it in solution still remains elusive. Using small angle X-ray scattering (SAXS) measurements combined with nuclear magnetic resonance (NMR) and molecular dynamics simulations (MD) the solution structure of rIAPP was studied. An unambiguously extended structural model with a radius of gyration of 1.83 nm was determined from SAXS data. Consistent with previous studies, an overall random-coiled feature with residual helical propensity in the N-terminus was confirmed. Combined efforts are necessary to unambiguously resolve the structural features of intrinsic disordered proteins.

Keywords IAPP · NMR · Molecular dynamics simulations

7.1 Introduction

Islet amyloid polypeptide (IAPP) is a peptide hormone secreted by the endocrine β-cells of the pancreas together with insulin [1]. It has 37 amino acids with a disulfide bond between residue 2 and 7 in the N-terminus. In solution, IAPP is characterized as a natively disordered protein [2–6]. In patients with type 2

L. Wei · P. Jiang · M.S.S. Manimekalai · C. Hunke · G. Grüber · K. Pervushin · Y. Mu (✉)
School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore, Singapore
e-mail: ygmu@ntu.edu.sg

D. Wei et al. (eds.), *Advance in Structural Bioinformatics*, Advances in Experimental Medicine and Biology 827, DOI 10.1007/978-94-017-9245-5_7

diabetes, IAPP changes its conformation to form amyloid fibers [7]. The process of IAPP amyloid formation and the prefibrillar oligomers are supposed to be one of the pathogenic agencies causing pancreatic β-cell dysfunction [8–11]. Thus the structural characterization of IAPP in the form of monomer or small oligomer states would be beneficial towards the full understanding of the toxicity mechanism of IAPP oligomers.

The human IAPP (hIAPP) aggregates easily and therefore, it is difficult to characterize its structural features by standard biophysical tools. Whereas the rat version of IAPP (rIAPP), that differ by six amino acids compared with hIAPP, is not prone to aggregation and does not form amyloid fibrils. But similar to hIAPP it also demonstrates random-coiled nature in solution. The structural propensity of rIAPP has been studied as a hIAPP mimic in several of the recent works [2–5, 12, 13]. However, the overall shape of it in solution still remains elusive. Here, we focus on resolving the low resolution structure of this peptide in solution by small angle X-ray scattering (SAXS) measurements and with complementary nuclear magnetic resonance (NMR) data. Further, these structural information were utilized to assess the ability of the three commonly used classic energy functions (force fields) for simulating the intrinsic disordered peptides/proteins.

7.2 Materials and Methods

7.2.1 Systems and NMR Spectroscopy

The ^{15}N uniformly labeled rIAPP sample preparation and NMR experiments have been described in details in our last publication [3]. Briefly, ^{15}N-HSQC, ^{15}N-TOCSY-HSQC (τmix = 120 ms) and ^{15}N-NOESY-HSQC (τmix = 200 ms) were performed using a Bruker Advance II 700 MHz spectrometer at 25 °C. The data were collected on a sample containing 50 μM ^{15}N uniformly labeled rIAPP (in 5 mM potassium phosphate buffer, 10 mM KCl, 3 % D_2O, pH 6). The spectra were analyzed and the chemical shifts were assigned with CARA software (www.nmr.ch). Peaks were picked manually from the 3D ^{15}N-NOESY-HSQC spectrum. The peak list, together with the chemical shift assignments were used as the input for structure calculations by CYANA 2.0 [14].

7.2.2 Small Angle X-ray Scattering Experiments

The synchrotron radiation X-ray scattering data for rIAPP were collected following standard procedures on the X33 SAXS camera of the EMBL Hamburg located on a bending magnet (sector D) on the storage ring DORIS III of the Deutsches Elektronen Synchrotron (DESY).

7.2.3 Molecular Dynamics Simulations

Classic molecular dynamics simulations were performed on rIAPP using three different force fields, AMBER03 [15] with recent modifications [16], OPLSAA [17] and CHARMM with CMAP [18, 19]. All simulations lasted 50 ns, which began with the NMR model 1 that has a Rg value of 1.78 nm. A dodecahedron box of size 7 nm was used with 7,869 water molecules (SPC model) and four chloride ions. The simulation was performed using Gromacs simulation package [20], during which all bonds involving hydrogen atoms were constrained in length according to LINCS protocol [21] with the integration step 2 fs. Non-bonded pair lists were updated every five integration steps. The protein and the water were separately coupled to the external heat bath with the relaxation time of 0.1 ps. The structure snapshots were output every 1 ps. Electrostatic interactions were treated with the particle mesh Ewald method [22] with a cutoff of 0.9 nm, while for the van der Waals interactions a cutoff of 1.4 nm was used. The simulations were repeated three times for each force field with different initial velocities. The helical structures of peptides were assessed by DSSP algorithm [23].

7.3 Results

7.3.1 Solution Structure Obtained from SAXS Measurements

Solution X-ray scattering experiments have been performed with the aim to determine the low resolution structure of rIAPP in solution. SAXS patterns from solutions of the peptide were recorded as described in "Materials and Methods" to yield the final composite scattering curve in Fig. 7.1a, showing that the peptide is monodispersed in the solution. Inspection of the Guinier plots at low angles indicated good data quality and no protein aggregation. The radius of gyration R_g of rIAPP is 1.83 ± 0.1 nm and the maximum dimension D_{max} of the peptide is 6.4 ± 0.4 nm (Fig. 7.1b). The solution shape of rIAPP was restored ab initio from the scattering pattern in Fig. 7.1a using the dummy residues modeling program DAMMIN [24], which fitted well to the experimental data in the entire scattering range (a typical fit displayed in Fig. 7.1a, red curve, has the discrepancy of $\chi^2 = 1.061$). Ten independent reconstructions yielded reproducible models and the average model is displayed in Fig. 7.2. rIAPP appears as an elongated molecule with a length of 6.4 nm and an overall spiral-like shape.

The solution structure of rIAPP has also been studied by us through NMR spectroscopy in our previous work [3]. Although the residual helical structure was relatively well-defined, the global structures generated by applying NMR constraints were quite heterogeneous, which was due to the lack of long-range constraints. Twenty structural models generated from NMR constrains are shown in Fig. 7.3. The radius of gyration (R_g) for this structure ensemble ranges from 1.1 to 1.78 nm.

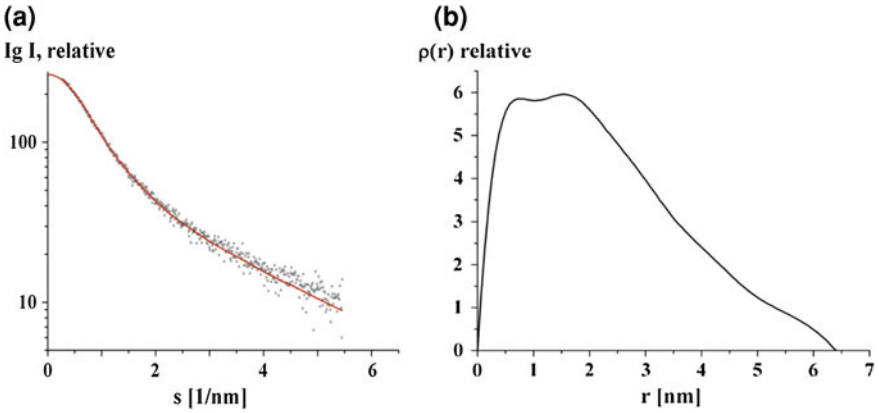


Fig. 7.1 Small-angle X-ray scattering data of rIAPP. **a** Experimental scattering data (*circle*) and the *fitting curves* (*line*; *green* experimental, *red* calculated from ab initio model) for rIAPP. **b** The distance distribution function of the same peptide

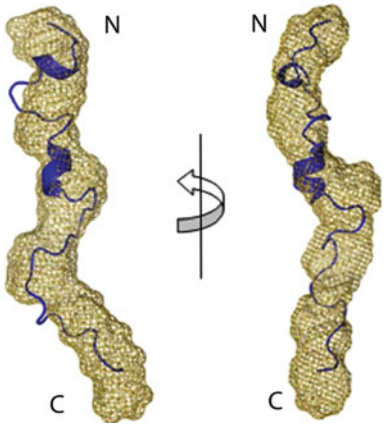


Fig. 7.2 Superposition of the DAMMIN model of rIAPP (*mashed shape*) with the NMR solution structure (*blue*) of the same peptide. The two models are rotated clockwise by around 90° along the Y-axis. The two helical regions are residue 9–12 and residue 15–17

Out of the 20 structural models we do found one model (NMR1) with an R_g of 1.78 nm. This NMR1 model and the solution shape, determined by SAXS ($R_g = 1.83$ nm), were superimposed with the program SUBCOMP [25] which showed good fitting with an r.m.s. deviation of 1.47 Å (Fig. 7.2). When we repeated the structural refinement processes using NMR constrains, only 5 % of the generated structural models have R_g larger than 1.73 nm. Clearly combing local structural information from NMR measurement with the global structural profile from SAXS can greatly narrow down the configuration space of the model structures. Thus the

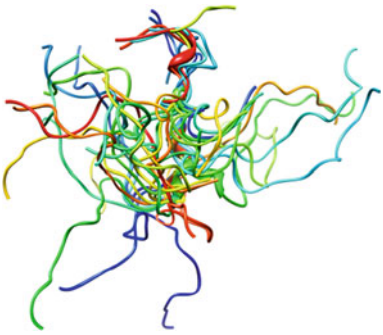


Fig. 7.3 The superimposing of 20 structure models from NMR constraints. The average helical residue number is 5.75

SAXS refined structural models of rIAPP are quite extended. Previously the dynamics of contact formation between the N- and C-termini in monomeric IAPP from human and rat were probed by triplet quenching technique [5]. This showed that the relaxation rates are approximately 2-fold faster for hIAPP than for rIAPP, which indicated that rIAPP is always more expanded than hIAPP.

7.3.2 Evaluation of Three Force Fields

Three all-atom force fields, AMBER03d [16], CHARMM [18] and OPLSAA [17], were employed to simulate this peptide in the presence of explicit water. The initial structure was taken from the elongated NMR1 model (Fig. 7.2 and 7.4a). The back-calculated scattering curve from the NMR1 model with CRYSOL program [26] gave a χ^2 value of 2.93 after superimposition onto the experimental data (Fig. 7.4c, solid line), indicating the consistency with both the experimental NMR- and SAXS data. Unfortunately, the configuration of the solution model cannot be maintained within the three force fields. In the AMBER03d and OPLSAA force field simulations the peptide collapses quickly in the first 10 ns from the initial R_g value of 1.78–1.1 nm (Fig. 7.5a). These compact conformations (Fig. 7.5b) were nearly unchanged during the following 40 ns simulations. The back-calculated scattering curve from such compact model (Fig. 7.5b) with CRYSOL resulted in a χ^2 value of 7.49 after superimposition with the experimental data (Fig. 7.4c, blue dashed line).

This compact feature of rIAPP has also been proposed from theoretical simulations and infrared spectroscopy data [4]. However, a compact model is not consistent with the experimental SAXS- and NMR data presented. The simulated conformations using CHARMM force field have larger R_g values, than the other two force fields (Fig. 7.5a), however, they are highly helical (Fig. 7.5b). The average number of helical residues is above 19 (more than half of the residues of the peptide) during the 50 ns trajectory. Such highly helical propensity is in contradiction with the overall random-coil nature of the peptide resolved by NMR

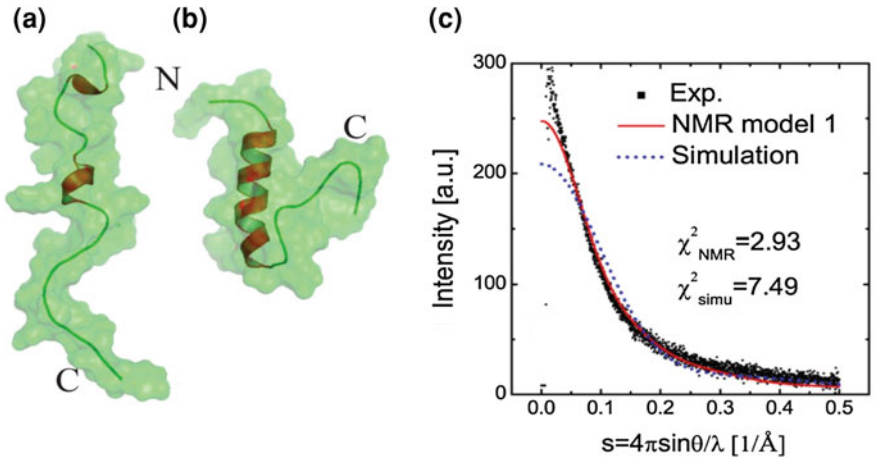


Fig. 7.4 Comparison of NMR1 model (**a**) and simulation model of AMBER03d (**b**) scattering intensities between experimental data and calculated from structural models (**c**)

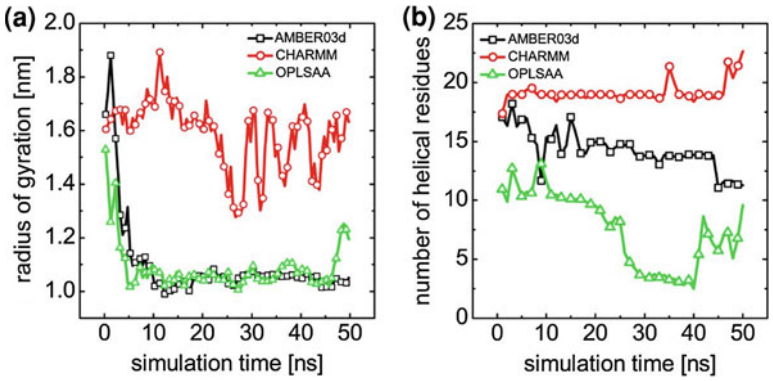


Fig. 7.5 Evolution of radius of gyration, Rg, **a** and the number of backbone hydrogen bonds **b** of rat IAPP from three different force fields simulations. Each data point is an averaged value during 1 ns simulation

measurement [2, 3]. The average helical residue number of 20 NMR structural models (Fig. 7.3) is only 5.75. Two more simulations in each type of force fields have been performed with different initial velocities in which similar results were obtained.

7.4 Conclusions

In summary we combined NMR measurements, which mainly provided local secondary structure information in this case, and SAXS data, which delivered a global structural profile, to resolve an extended, random-coiled structural model for the rat IAPP peptide. The presented structure will provide an invaluable reference to further study the conformational propensity of more disease related hIAPP.

References

1. Jaikaran E, Clark A (2001) Islet amyloid and type 2 diabetes: from molecular misfolding to islet pathophysiology. Biochim Biophys Acta-Mol Basis Dis 1537:179–203
2. Williamson JA, Miranker AD (2007) Direct detection of transient alpha-helical states in islet amyloid polypeptide. Protein Sci 16:110–117
3. Wei L, Jiang P, Yau YH, Summer H, Shochat SG, Mu Y, Pervushin K (2009) Residual structure in islet amyloid polypeptide mediates its interactions with soluble insulin. Biochemistry 48:2368–2376
4. Reddy AS, Wang L, Lin YS, Ling Y, Chopra M, Zanni MT, Skinner JL, De Pablo JJ (2010) Solution structures of rat amylin peptide: simulation, theory, and experiment. Biophys J 98:443–451
5. Vaiana SM, Best RB, Yau WM, Eaton WA, Hofrichter J (2009) Evidence for a partially structured state of the amylin monomer. Biophys J 97:2948–2957
6. Dupuis NF, Wu C, Shea JE, Bowers MT (2009) Human islet amyloid polypeptide monomers form ordered beta-hairpins: a possible direct amyloidogenic precursor. J Am Chem Soc 131:18283–18292
7. Haataja L, Gurlo T, Huang CJ, Butler PC (2008) Islet amyloid in type 2 diabetes, and the toxic oligomer hypothesis. Endocr Rev 29:303–316
8. Campioni S, Mannini B, Zampagni M, Pensalfini A, Parrini C, Evangelisti E, Relini A, Stefani M, Dobson CM, Cecchi C, Chiti F (2010) A causative link between the structure of aberrant protein oligomers and their toxicity. Nat Chem Biol 6:140–147
9. Hull RL, Westermark GT, Westermark P, Kahn SE (2004) Islet amyloid: a critical entity in the pathogenesis of type 2 diabetes. J Clin Endocrinol Metab 89:3629–3643
10. Gurlo T, Ryazantsev S, Huang CJ, Yeh MW, Reber HA, Hines OJ, O'Brien TD, Glabe CG, Butler PC (2010) Evidence for proteotoxicity in beta cells in type 2 diabetes toxic islet amyloid polypeptide oligomers form intracellularly in the secretory pathway. Am J Pathol 176:861–869
11. Janson J, Ashley RH, Harrison D, McIntyre S, Butler PC (1999) The mechanism of islet amyloid polypeptide toxicity is membrane disruption by intermediate-sized toxic amyloid particles. Diabetes 48:491–498
12. Soong R, Brender JR, Macdonald PM, Ramamoorthy A (2009) Association of highly compact type 2 diabetes related islet amyloid polypeptide intermediate species at physiological temperature revealed by diffusion nmr spectroscopy. J Am Chem Soc 131:7079–7085
13. Nanga RPR, Brender JR, Xu JD, Hartman K, Subramanian V, Ramamoorthy A (2009) Three-dimensional structure and orientation of rat islet amyloid polypeptide protein in a membrane environment by solution nmr spectroscopy. J Am Chem Soc 131:8252–8261

14. Güntert P (2003) Automated NMR protein structure calculation. Prog Nucl Magn Reson Spectrosc 43:105–125
15. Duan Y, Wu C, Chowdhury S, Lee MC, Xiong GM, Zhang W, Yang R, Cieplak P, Luo R, Lee T, Caldwell J, Wang JM, Kollman P (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J Comput Chem 24:1999–2012
16. Best RB, Hummer G (2009) Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. J Phys Chem B 113:9004–9015
17. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. J Phys Chem B 105:6474–6487
18. Mackerell AD, Feig M, Brooks CL (2004) 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
19. Bjelkmar P, Larsson P, Cuendet MA, Hess B, Lindahl E (2010) Implementation of the CHARMM force field in GROMACS: analysis of protein stability effects from correction maps, virtual interaction sites, and water models. J Chem Theory Comput 6:459–466
20. David Van Der Spoel EL, Hess B, Groenhof G, Mark AE, Berendsen HJC (2005) GROMACS: fast flexible and free. J Comput Chem 26:1701–1718
21. Hess B, Bekker H, Berendsen HJC, Fraaije J (1997) LINCS: a linear constraint solver for molecular simulations. J Comput Chem 18:1463–1472
22. Darden T, York D, Pedersen L (1993) Particle mesh Ewald: an N•log(N) method for Ewald sums in large systems. J Chem Phys 98:10089–10092
23. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637
24. Svergun DI (1999) Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. Biophys J 76:2879–2886
25. Kozin MB, Svergun DI (2001) Automated matching of high- and low-resolution structural models. J Appl Crystallogr 34:33–41
26. Svergun D, Barberato C, Koch MHJ (1995) CRYSOL—a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. J Appl Crystallogr 28:768–773

Chapter 8
Folding Mechanisms of Trefoil Knot Proteins Studied by Molecular Dynamics Simulations and Go-model

Xue Wu, Peijun Xu, Jinguang Wang, Yong Xu, Ting Fu, Depeng Zhang, Meixia Zhao, Jiahui Liu, Hujun Shen, Zhilong Xiu and Guohui Li

Abstract Most proteins need to avoid the complex topologies when folding into the native structures, but some proteins with nontrivial topologies have been found in nature. Here we used protein unfolding simulations under high temperature and all-atom Gō-model to investigate the folding mechanisms for two trefoil knot proteins. Results show that, the contacts in β-sheet are important to the formation of knot protein, and if these contacts disappeared, the knot protein would be easy to untie. In the Gō-model simulations, the folding processes of the two knot proteins are similar. The compact structures of the two knot proteins with the native contacts in β-sheet are formed in transition state, and the intermediate state has loose C-terminal. This model also reveals the detailed folding mechanisms for the two proteins.

Keywords Trefoil · Knot · High-temperature unfolding · Gō-model · Fold

Xue Wu, Ting Fu, Peijun Xu and Jinguang Wang have been contributed equally to this paper.

X. Wu · T. Fu · H. Shen · G. Li (✉)
Laboratory of Molecular Modeling and Design, State Key Laboratory of Molecular Reaction Dynamics, Dalian Institute of Chemical Physics, Chinese Academy of Science, Dalian, Liaoning, China
e-mail: ghli@dicp.ac.cn

P. Xu · D. Zhang · M. Zhao · J. Liu
School of Physics and Electronic Technology, Liaoning Normal University, Dalian, Liaoning, China

J. Wang
The First Affiliated Hospital, Dalian Medical University, Dalian, China

Y. Xu
Guangzhou Institute of Biomedicine and Health, Guangzhou, China

Z. Xiu
School of Life Science and Technology, Dalian University of Technology, Dalian, China

D. Wei et al. (eds.), *Advance in Structural Bioinformatics*, Advances in Experimental Medicine and Biology 827, DOI 10.1007/978-94-017-9245-5_8

8.1 Introduction

The protein molecule performs the biological function through folding into the compact structure. In the folding process, most proteins avoid complex topologies, but some proteins are able to fold into nontrivial topologies, especially the main chain fold into a knotted conformation [1–3], which is an evolutionary curiosity. If pulling the knotted protein from both the two ends, this structure can't be disengaged. So far, most of the discovered knotted proteins are belong to the 3_1 knot, and the others are belong to the 4_1, 5_2 or 6_1 knots [4–6]. Though the knotted proteins are existent, but how the knotted proteins overcoming the energy barrier fold into the complicated and intact topologies from the disordered linear polypeptide is still a mystery. The shape of the protein and the chain connectivity of its backbone may determine the folding routes of a well-designed protein sequence [7]. So the structure based protein models can capture the essential features of protein folding through separating from the effects of topology and eliminating all non-native energetic traps [8–10]. From the unfolded state to the native state of protein, the energy landscape directs this folding route of protein, and the diverse sizes and shapes of the free energy barriers are directed by the pattern of contacts especially the native contacts [11–13]. The knot protein with the complicated topology may not fold easily because of the emerging unlikely configurations in the folding process [14–16]. The folding of knot protein needs right crossing of polypeptide, otherwise may have an unknotted protein or a wrong chirality. Based on the structure-based model, the information of protein folding pathway is contained in the folded configuration, so it is a good model for studying the folding process of knot protein. The time scale of protein folding is incompatible with the time scale of molecular dynamics simulations, so studying the protein unfolding process under high temperature is another meaningful method for studying the folding process of protein. At high temperature, it is easy to cross the energy barrier of knot protein, so this knot could be untied in this condition, and through studying the unfolding progress of knot protein to get the information of the folding process. Here we used the high-temperature unfolding method, all-atom and Ca structure-based model to research the folding pathways of knot proteins. The all-atom model can supply more accurate thermodynamic information for the folding of knotted protein than the Ca structure-based model, which is good for uncovering the folding mechanisms for the simple knot proteins.

For studying the formation of knot protein, various biochemical and biophysical techniques have been employed, like chemical denaturants, single-molecule atomic force microscopy (AFM) measurements [17]. The studies for the folding of protein YibK from *H. influenzae* and YbeA from *E. coli* in experiment were through using the denaturant urea to get the unfolded structure reversibly which lacked secondary or tertiary structure, and then gave a detailed folding study for protein [18, 19]. The folding pathways of protein YibK have been extensively studied. The double-jump refolding experiment has been used to investigate the presence of multiple unfolded states of protein YibK [20]. The folding mechanism

of YibK has been probed by using single-site mutants, this folding process of protein YibK was from the denatured state, but this structure was not unfold completely [21]. The theoretical investigations of knot protein can be started from the wholly unfolded structure, which does not contain a knot, so the theoretical investigations can make up the defects of the experimental studies and give more information for the folding of knot protein. The atomistic simulations have been used for studying the unfolding of bovine carbonic anhydrase II [22]. On the coarse-grained level, the simulations also could be used for the studies of protein folding. The simulations of Gō-model for knot protein was applied to some studies, and this model could make the protein fold from the unfolding structure to the topologically frustrated, knotted structure. Generally, the Gō-model reduces the protein to its Ca-backbone. Wallin undertook the coarse-grained model on the knot protein YibK for studying the folding kinetics, and through introducing the attractive nonnative interactions on the knot protein, this protein could take the knotted mechanism of plug motion to form native structure [23]. The coarse-grained model has been used for probing the folding processes of protein YibK and YbeA, and succeeded in forming a native knot structure in 1–2 % of the simulations with native interactions through using this model [24]. In the folding processes of protein YibK and YbeA, an intermediate configuration with a slipknot was involved, and the appearance of this configuration was aimed at reducing the topological bottlenecks. The researches about slipknots of proteins also revealed that these slipknots could give contribution to the thermal stability for the slipknot feature [25]. The molecular dynamics simulations have been used widely in the studies of proteins [26–34], so here molecular dynamics simulation methods were used to study the folding of knot proteins. Two different approaches comparing with these previous studies have been used for probing the folding mechanisms of knot proteins. The method of protein unfolding under high temperature and all-atom Gō-model were applied to two 3_1 knot proteins for studying the folding mechanisms and thermodynamics of the two proteins.@@@@@

When folding to the correct native structure, the knot protein has to avoid the topological traps and kinetic traps on the landscape. In the folding process, the tying of knot protein is refer to a problem about the chain crossing, and a topological constraint may solve this problem, otherwise this process is not allowed. The geometric constraint of the native structure may dominate the knot protein through a subset of possible folding pathways. In theory, the protein with the minimally frustrated structure is supposed to have the energy landscape of funnel shape. In the folding process of protein, the shape of the landscape is dominated by the strong energetic bias, which could reduce traps caused by non-native interactions. Thus, this geometric constraint model is better for determining the folding mechanisms of proteins. Using this geometric constrain model also is necessary for guiding the chain to form knot, and the final folded structure of protein plays a major role in determining its foldability, so this model may make the protein have more chances to fold into the native state. Under high temperature, the knot protein has more probability to cross the energy barrier, so this protein has more chances to unfold. The all-atom model can make up the gap between coarse-grained

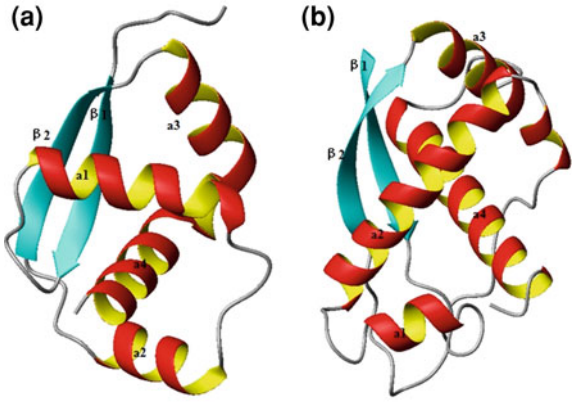


Fig. 8.1 Folded structures of the two knot proteins. **a** The crystal structure of protein MJ0366 (PDB ID code 2efv). **b** The crystal structure of protein VirC2 (PDB ID code 2rh3)

models and all-atom empirical forcefields. Hence, here we used the method of high-temperature unfolding and all-atom model to research two knot proteins in order to have more information about the folding mechanisms and topological constraint effects of these proteins.

In this study, the knot proteins are the smallest knot protein MJ0366, from *Methanocaldococcus jannaschii* [5, 19], and protein VirC2, the border-specific endonuclease, from *Agrobacterium tumefaciens* [25]. The two proteins have trefoil knot structures (Fig. 8.1). At high temperature, the protein MJ0366 could unfold. The conformational clustering method was used to find the transition state, and this state has the native contacts in β-sheet. The unfolding process has relation ship with the stability of this β-sheet. The all-atom model for the smallest knot protein shows the intermediate state has native contacts in β-sheet, and slipknot and plug knotting routes are found at folding temperature. The protein VirC2 is prone to have traps in the folding process and through backtracking to fold into the native state.

8.2 Results and Discussion

In this paper, we study two trefoil knot proteins which are two simple examples of nontrivial knots. The protein MJ0366 with 82 residues belongs to α/β protein. The trefoil knot is one end of the chain through into a loop. The C-terminal of protein MJ0366 threads into the loop consists of α1, α2 and their linkers, and the N-terminal threads into the loop which is comprised of β2, α3 and their linkers. The protein VirC2 with 121 residues has ribbon-helix-helix (RHH) fold. This protein has two β-strands, and four α-helices like protein MJ0366. The C-terminal of protein VirC2 threads into the loop which is created by α1, α2 and the linkers between α2 and β2, and the N-terminal threads into the loop formed by β2, α3 and their linkers.

8.2.1 Protein MJ0366 Unfolding Pathway

Here we used molecular dynamics simulations under high temperature to study the protein unfolding process. We selected 530 K for studying the protein unfolding process, and the molecular dynamics simulation of native state was in 298 K as a comparison. The Ca root-mean-square deviation (Ca RMSD) cluster method was used to find out the transition state. We took nine unfolding simulation trajectories for protein MJ0366, knot_1-knot_9. The transition states were identified at 8.175 ns in knot_1, 23.42 ns in knot_2, 14.076 ns in knot_3, 4.345 ns in knot_4, 14.073 ns in knot_5, 17.831 ns in knot_6, 8.431 ns in knot_7, 9.511 ns in knot_8, and the transition state in the last trajectory knot_9 was not found.

The number of native contact for protein MJ0366 as a function of time in a typical trajectory knot_6 is shown in Fig. 8.2. Under high temperature, the number of native contact was obviously changed as the time growth, and the change trend of the number of native contact for the whole protein was the same as the number of native contact in the β-sheet. The native contacts in β-sheet decreased along with the decreasing number of native contact of the whole protein, so the β-sheet unfolding may have significant impact on the whole system. After the β-sheet untied, the whole system may have low stability, so the knot would be easier to unfold. The unfolding process of protein MJ0366 in the typical trajectory knot_6 is shown in Fig. 8.3. Under high temperature, the α-helices especially the α2 unfold firstly. The native contacts between α2 and the other secondary structures were

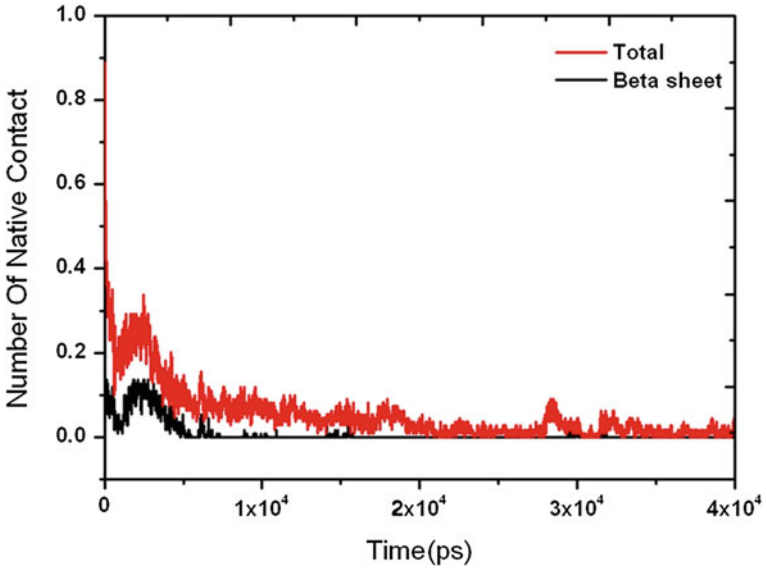


Fig. 8.2 The native contacts in β-sheet and the whole knot protein in a typical kinetic folding trajectory for protein MJ0366 under high temperature

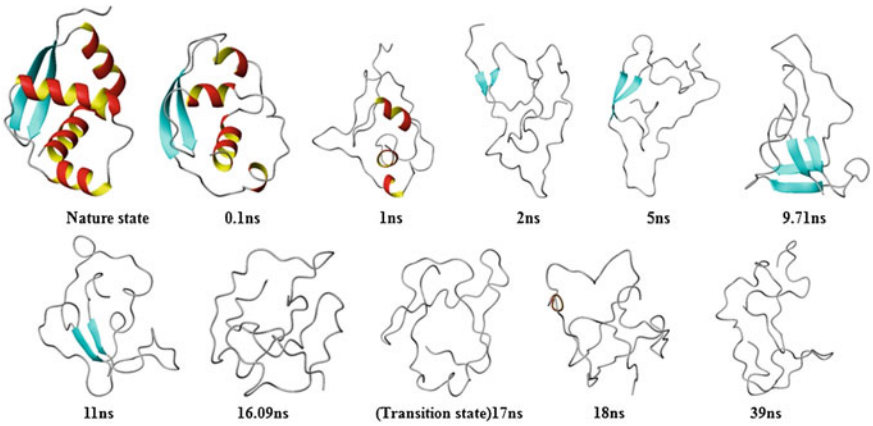


Fig. 8.3 The unfolding process of protein MJ0366 under 530 K. The transition state is at ~17 ns

few (Fig. 8.5b), which may effect the stability of this helix. In the trajectory knot_6, the α-helices were almost disappeared after 1 ns, and the native contacts in β-sheet were still existent. The α-helices disappeared after 2 ns, at this time the β-sheet still was stable, and the protein MJ0366 formed a compact structure. The β-sheet disappeared after 5 ns, and the new β-sheet between the position of α1 and α3 was appeared. Though the β-sheet was disappeared, the two β-strands were in close distance, and the loop controlled by this β-sheet was enlarged, so the C-terminal may have the chance to unfold. The untied protein was appeared for the first time at ~9.71 ns, in this time scale the β-sheet was reformed, and the two terminals formed a new β-sheet. This new β-sheet made the protein fluctuate around N-terminal, so the C-terminal could have the chance to unfold in short time. From this time, the knot protein entered a fluctuant stage lasting for ~7 ns, the knot protein was varied between the untied state and the knot state. In the fluctuant stage, the contacts in β-sheet were diminished, which made the protein change to a loose structure, and then made protein easier to untie. In this stage, the β-sheet was prone to form loops to make the knot untie. Though the C-terminal has formed loop, and it seems to be excluded from the loop formed by α1, α2 and their linkers, but the contacts in the β-sheet were still existent, which effected the unfolding of the C-terminal of knot protein. From the above, the β-sheet is important for the stability of knot protein. The β-sheet disappeared completely after ~11 ns. At 16.09 ns, the protein was untied, and did not form the knot again. The β-sheet between the two terminals was disappeared, and the terminal of β1 was prone to form a loop, which made the β1 exclude from the loop formed by β2, α3 and their linkers. At ~17 ns, the knot protein got to the transition state.

After the transition state the contacts between β-strands were disappeared, the N-terminal excluded from the loop formed by β2, α3 and their linkers, and then the C-terminal had the chance to exclude from the loop formed by α1, α2 and their linkers. Under high temperature, the β-sheet is prone to be destroyed first, and then the C-terminal may have the chance to exclude from the loop formed by α1, α2 and their linkers. The unfolding trajectories are considered in reverse as a description of the folding pathway. The α-helixes of protein MJ0366 have been disappeared in the early stage of the folding process, and then the β-sheet disappeared, so the β-sheet may be formed earlier than α-helixes. The process that C-terminal unfolds firstly may consume more energy for the knot protein, so this protein chooses the pathway that the N-terminal unfolds firstly. Hence, this protein may choose a pathway that a structure with the β-sheet is formed firstly, and then the C-terminal thread into this loop controlled by the β-sheet for folding into the native state.

8.2.2 *Transition States for Protein MJ0366 Under High Temperature*

In the protein unfolding process, the transition state was decided by the Ca-RMSD cluster method. The Ca-RMSD has been used as a crucial criterion for the convergence measure of the protein systems [35, 36]. The Ca-RMSD for protein MJ0366 in a typical trajectory knot_6 is shown in Fig. 8.4a. Before performing the unfolding simulations under high temperature, the dynamic behavior of protein MJ0366 was investigated at room temperature. Under room temperature, the knot

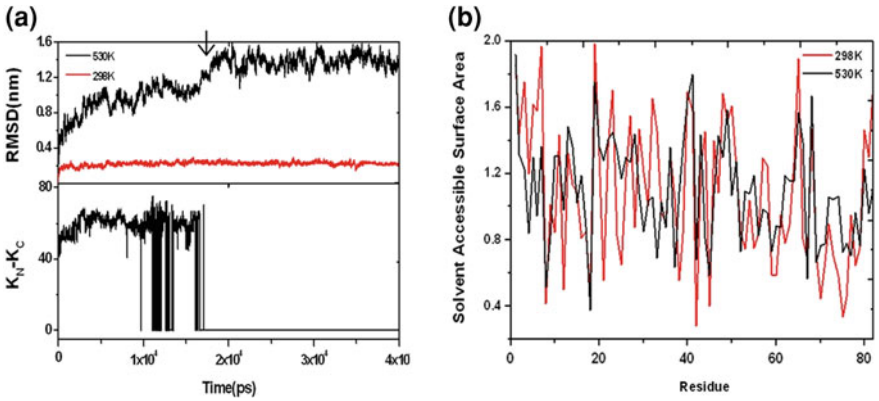


Fig. 8.4 Transition state for protein MJ0366 in the unfolding process. **a** The Ca-RMSD of the crystal structure as a function of time at 530 and 298 K. The protein unfolded at transition state. **b** The average solvent accessible surface area for the residues of protein MJ0366 in transition state

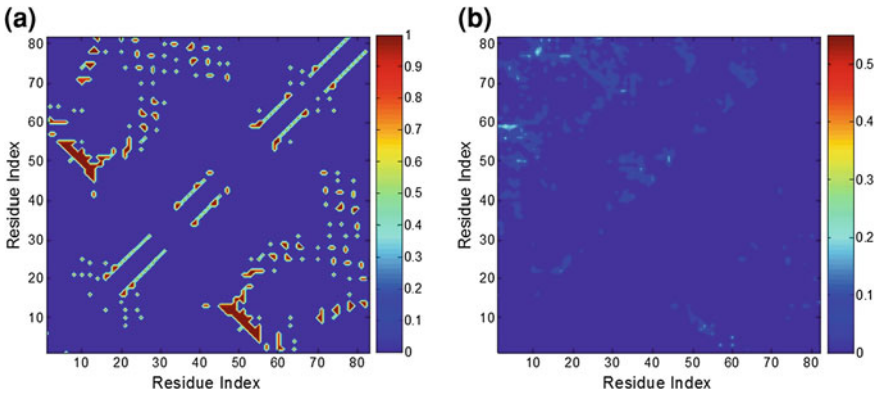


Fig. 8.5 The average native contact maps for protein MJ0366. **a** The contact map of the trajectory at 298 K. **b** The native contact map for the transition states of the nine simulation trajectories at 530 K. The *upper triangular* presents the nonnative contacts, and the *lower triangular* presents the native contacts

protein was stable, and the Ca-RMSD for this protein was remained at ~2.0 Å during the 40 ns simulation. Under the temperature of 530 K, the Ca-RMSD of protein MJ0366 had a rapid structural deviation comparing with the crystal structure in the native state at ~17 ns in the typical trajectory, and the transition state was found through the method of Ca-RMSD cluster at ~17 ns. The knot position can be characterized by its depth, the distance along the sequence from N-terminal and C-terminal of the knot [24]. Here we used the residues that form the knot to monitor this protein. The knot protein server was used for the detection of knot proteins [37]. The size of knot protein as a function of time under temperature of 530 K is shown in Fig. 8.4a. At transition state, the knot of protein MJ0366 was untied, and before reaching the transition state the protein fluctuated between folded and untied states. After transition state, the protein was untied and no longer formed a knot. The contact map for the knot protein in native state is shown in Fig. 8.5a. In transition state, some of the native contacts in β-sheet were existent, which implied the two β strands fluctuated in the close distance between each other. This state effected the excluding of C-terminal from the loop formed by α1, α2 and their linkers. The number of native contact of residues A8-I53, R7-S57 and T8-L60 were higher than 30 % in the transition state. Some of the native contacts between the loop of N-terminal and β2 were maintained at high level. The number of native contact K5-E60 was higher than 40 %. The residues K3-E57, K3-E65 had the number of native contact higher than 30 %. In the transition state, the non-native contacts for this knot protein were increased, especially the contacts in the β-sheet and between C-terminal and N-terminal. The non-native contacts between N-terminal and β2 were increased, which implied the β1 was prone to exclude from the loop formed by β2, α3 and their linkers, and the contacts in the β-sheet effected the unfolding of knot protein. The non-native contacts between C-terminal and the region around β1 were appeared. The decreasing native contacts in β-sheet made

the surrounding secondary structures of C-terminal become loose, so the C-terminal had more chances to have contacts with N-terminal. All the above, the contacts in the β-sheet is important for the protein stability, if breaking these contacts may promote protein untie. The molecular dynamics simulations for this knot protein were in water, and in transition state the solvent accessible surface area (SASA) was changed (Fig. 8.4b). The SASA values of C-terminal and N-terminal were decreased. The emerging non-native contacts between the C-terminal and N-terminal made the two regions eliminate the surrounding water molecules. The changes in the surrounding loop of C-terminal made the contacts among α-helices decreased, which may impact the SASA value of the C-terminal in α1, and this region had SASA decreased. In transition state, the whole system did not unfold, so the SASA of the knot protein was not changed very much.

8.2.3 Protein MJ0366 Folding Pathway in Gō-model

We performed constant temperature molecular dynamics simulations to obtain the free energy landscape for the monomer structure of knot protein at folding temperature. Each simulation of all-atom model included the folded/knotted state and unfolded/unknotted state. The folding process for protein MJ0366 was monitored by reaction coordinates. The free energy as a function of the number of native contact is shown in Fig. 8.6a. In the folding process, the knot protein had three states. The unfolded state was near the number of native contact 0.15, and then this protein folded into the intermediate state. This result is consistent with the investigation by Jeffrey K. Noel et al., and they made use of Gaussian-type contact potential to study knot protein [38]. After crossing the free energy barrier with the number of native contact ~ 0.4, the protein folded into the native state. The free energy as a function of two reaction coordinates, the number of native contact and RMSD, is shown in Fig. 8.6b. In the folding process, the RMSD of knot protein were changed with the increasing number of native contacts. The RMSD of the unfolded state for knot protein was near 20 Å. When the RMSD value decreased to ~ 3 Å, the knot protein folded into the native state. The two-dimensional free energy landscape as a function of the number of native contact and radius of gyration (Fig. 8.6c) was not shown an obvious L-shaped landscape, which indicated the whole system not aggregated rapidly. The radius of gyration of knot protein decreased with an increasing number of native contact. When the value of radius of gyration decreased to ~ 3 Å, the protein folded into the native state. The landscape as a function of the number of native contact and the number of native contact formed in the β-sheet was shown three states (Fig. 8.6d). The number of native contact in β-sheet increased rapidly with an increasing number of native contact of the whole protein, but stayed low and increased little further once the number of native contact in β-sheet of ~ 0.7 was formed. In the folding process, the intermediate state was appeared, which near the number of native contact of 0.7. After intermediate state the protein needed to cross the energy barrier to form a knot. In the folding process, protein must overcome an

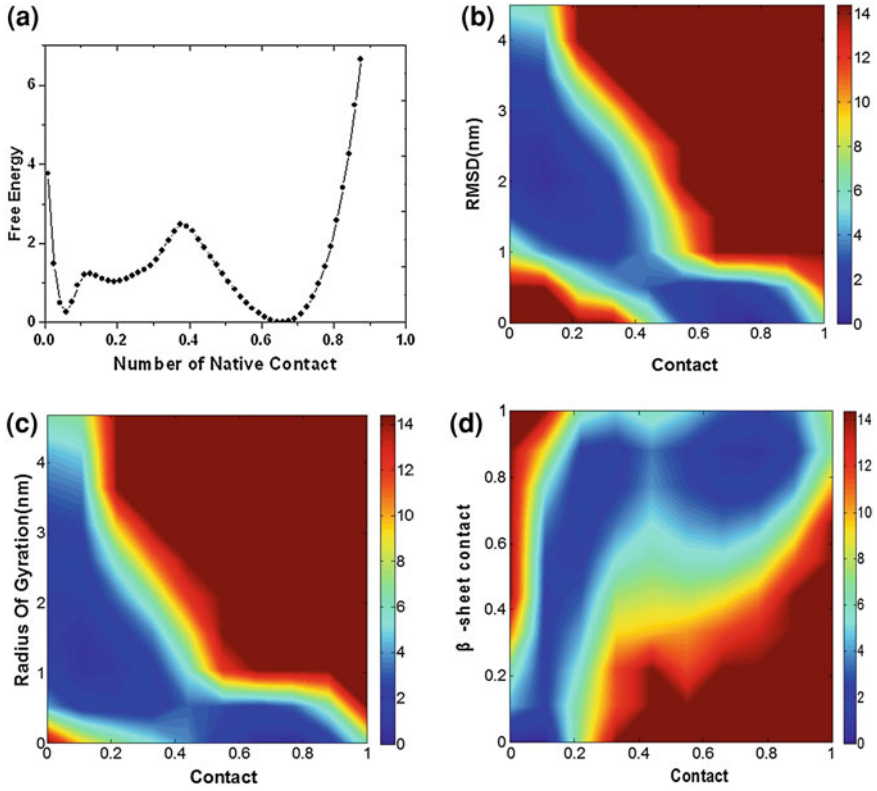


Fig. 8.6 The folding routes of knot protein MJ0366 from all-atom Gō-model at folding temperature T = 111. **a** The free energy as a function of the number of native contact. **b** The free energy as a function of the number of native contact and Ca-RMSD. **c** Two-dimensional free energy landscape as a function of the number of native contact and radius of gyration. **d** The free energy as a function of the number of native contact of the whole protein and the number of native contact in β-sheet

energy barrier to form the β-sheet, and this state could form a loop, which is necessary for the formation of the knot. This loop needs to twist correctly, otherwise the protein may form the topological trap structures like the results of the investigation by Jeffrey K. Noel et al. The C-terminal needed to thread into this loop for the formation of native structure, and this step required to cross the high energy barrier. The C-terminal may thread into this loop through plug or slipknot motion [38]. Here we found when the loop formed by α1, α2 and their linkers was loose, the C-terminal was prone to thread into this loop with plug motion, otherwise the C-terminal tended to adopt slipknot motion. From the above, the native contacts between C-terminal and the loop formed by α1, α2 and their linkers are stable, so more energy is needed

to destroy these contacts than the native contacts in β-sheet. Under high temperature, the protein chooses to untie the β-sheet firstly, which is because of the unstability of this region comparing with C-terminal of the knot protein.

8.2.4 *Intermediate and Transition States for Protein MJ0366 in Gō-model*

The native contact maps in intermediate state and transition state for protein MJ0366 are shown in Fig. 8.7. The intermediate and transition states were defined according to the free energy as a function of the number of native contact. The intermediate state had the number of native contact of ~ 0.2, and the transition state was located in the maximum free energy as the function of the number of native contact. In intermediate state, all the native contacts in β-sheet were almost appeared, but the C-terminal was loose. After the intermediate state the protein entered the transition state with high energy barrier. The transition state appeared some native contacts, such as the native contacts between C-terminal and the surrounding region of β1, and the C-terminal and α1. So the C-terminal was ready to thread into the loop formed by α1, α2 and their linkers in transition state. In addition, some structures in the transition state formed the native contacts between α2 and α4, which means the main native contacts needed by the formation knot protein were appeared, so some structures have formed knot and the formation knot protein maybe at the late transition state.

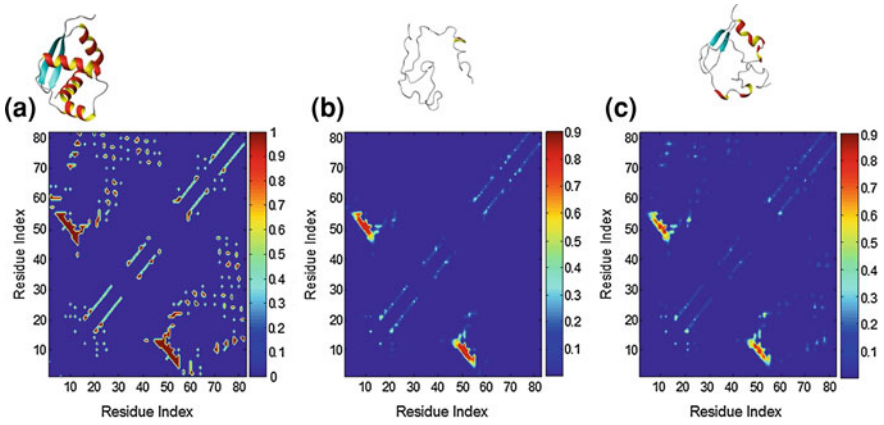
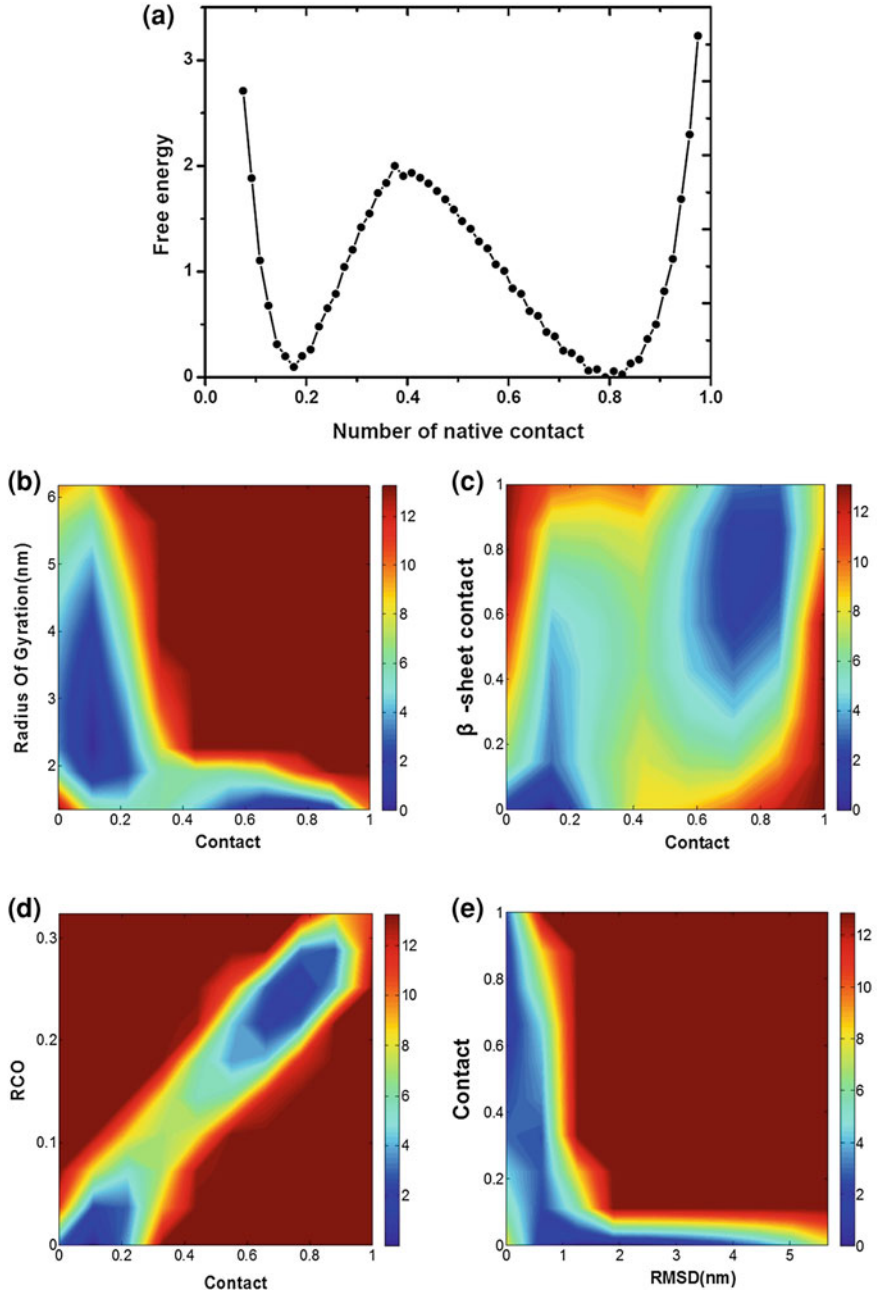


Fig. 8.7 The native contact maps for protein MJ0366 from all-atom Gō-model at folding temperature. **a** The contact map for protein in the native state. **b** The native contact map in the intermediate state. **c** The native contact map in the transition state. The typical structures in the three states are shown

8.2.5 Protein VirC2 Folding Pathway in Gō-model

We used constant temperature molecular dynamics simulations of Ca Gō-model to get the free energy landscape of the structure of protein VirC2 at folding temperature for better understanding the folding mechanism of trefoil protein. The free energy as a function of the number of native contact and radius of gyration is shown in Fig. 8.8b. The L-shaped landscape indicated the radius of gyration decreased rapidly with an increasing number of native contact, but once the number of native contact of $\sim$0.4 was formed, the radius of gyration decreased little further. The unfolded state had the radius of gyration of $\sim$0.1. After the radius of gyration reached $\sim$0.6, the protein folded to the nature state. The two states were separates by the area of transition state. The sharply decreased radius of gyration implied the system of knot protein had the initial collapse. The landscape of the free energy as a function of the number of native contact for the whole protein and the number of native contact in β-sheet (Fig. 8.8c) was showed the β-sheet formed first, subsequently the number of native contact increased until the protein folded into the native state. So the native contacts in β-sheet may promote the formation of compact structure. The free energy landscape was plotted as a function of the number of native contact and the relative contact order (RCO) parameter (Fig. 8.8d), which can be used to investigate more detail about the folding mechanism of this knot protein [39]. In the folding process of protein VirC2, the RCO increased with an increasing number of native contact. The change trend of RCO value coincided with the number of native contact. This implied that the local native contacts were formed in the initial stage of the folding process, and then the long-range native contacts were formed as an increasing number of native contact. In the Ca Gō-model, the β-sheet formed firstly, which promoted the compaction between N-terminal and the other parts of this knot protein. In this process, the local native contacts were important for the formation of β-sheet. After forming the native contacts in β-sheet, the knot protein needed to cross the transition state to fold into the native state. A typical folding process for this protein with all-atom Gō-model at T = 103 is shown in Fig. 8.9. In the folding process, the structure of knot protein formed the native contacts in β-sheet at the number of native contact of $\sim$0.2. When the number of native contact reached $\sim$0.5, the protein entered a state with a compact structure, but the C-terminal can not thread into the loop formed by α1, α2 and their linkers, it is likely that this process needed to adjust the conformation of this loop. When the number of native contact decreased to $\sim$0.2 again, this loop was readjusted, and the orientation of this loop was changed. After this process the C-terminal could thread into this loop. The folding process for this trefoil knot protein was similar to the protein MJ0366. The formation of the native contacts in β-sheet was important for the whole protein, after forming the native contacts in β-sheet, the N-terminal could have chances to thread into the loop formed by α1, α2 and their linkers. At the last stage of the folding process, the C-terminal was prone to adopt slipknot motion to thread into this loop.


(a)
Free energy
Number of native contact
(b)
Radius Of Gyration(nm)
Contact
(c)
β -sheet contact
Contact
(d)
RCO
Contact
(e)
Contact
RMSD(nm)

◀ **Fig. 8.8** The folding routes of knot protein VirC2 from Ca Gō-model at folding temperature T = 146. **a** The free energy as a function of the number of native contact. **b** The free energy as a function of the number of native contact and radius of gyration. **c** Two-dimensional free energy landscape as a function of the number of native contact for the whole protein and the native contacts in β-sheet. **d** The free energy as a function of the number of native contact and RCO parameter. **e** Two-dimensional free energy landscape as a function of Ca-RMSD and the number of native contact in β-sheet

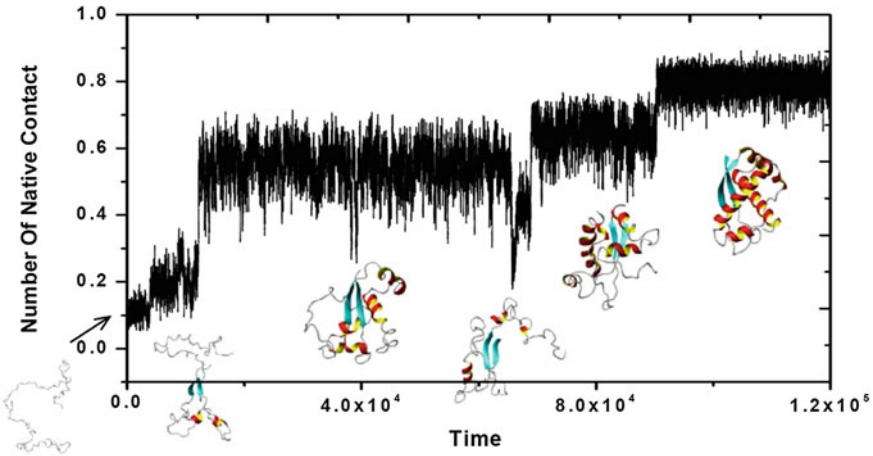


Fig. 8.9 A typical folding route for protein VirC2 from all-atom Gō-model at T = 103 close to the folding temperature. The typical conformations in this trajectory are shown below each states

8.2.6 Transition State for Protein VirC2 in Gō-model

The free energy as a function of the number of native contact is shown in Fig. 8.8a. The transition state has the number of native contact of ∼0.4. The native contact map of transition state for protein VirC2 in Ca Gō-model is shown in Fig. 8.10. In transition state, the native contacts in β-sheet were formed, some of the native contacts were formed between N-terminal and C-terminal, and between C-terminal and α2. Comparing with the native state, most of the native contacts have been formed for some structures in the transition state, which implied this knot may be formed at this stage. Hence, the protein VirC2 may be formed in the late transition state like the trefoil knot MJ0366. The free energy as a function of Ca-RMSD and the number of native contact in β-sheet (Fig. 8.8e) was presented a state with the number of native contacts of ∼0.3 in the β-sheet. This state formed the native contacts in the β-sheet, and the C-terminal was loose like the intermediate state of protein MJ0366. So the knot protein VirC2 may have the intermediate state, in this state the loop was controlled by the β-sheet which needed to be readjusted, and then the protein could fold into the native structure.

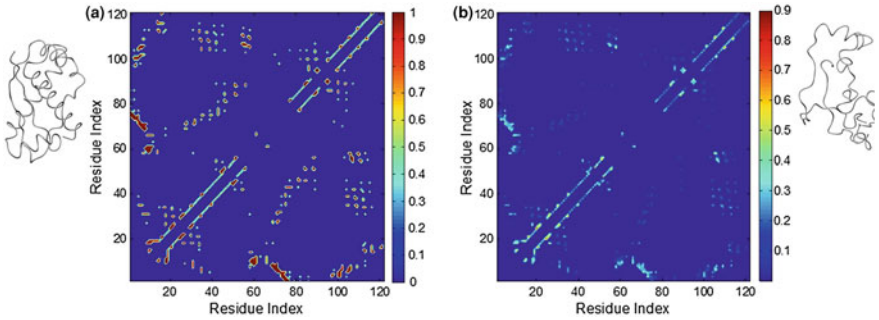


Fig. 8.10 The native contact maps for protein VirC2 from Ca Gō-model at folding temperature. **a** The contact map for protein in native state. **b** The native contact map for protein in transition state

8.3 Conclusions

We simulated two trefoil proteins with Gō-model, and high-temperature unfolding simulations was used for the study of trefoil protein MJ0366. The unfolding process of protein MJ0366 showed the contacts in β-sheet decreased firstly, and then the C-terminal of knot MJ0366 could thread out of the loop controlled by the contacts in β-sheet. In all-atom Gō-model, the native contacts in β-sheet promote the formation of a loop, and then the C-terminal threads into this loop to form the native state. The folding processes of the two trefoil knots were similar, and the formation of β-sheet was important for the two knot proteins. The C-terminal was prone to thread into the loop formed by secondary structures in correct size with slipknot motion, but when the loop was loose, the C-terminal was probably to thread into the loop with plug motion. In the intermediate state, the compact structure with the native contacts in the β-sheet was formed, but the C-terminal was loose. In transition state, the native contacts in β-sheet were formed, and the C-terminal was prone to thread into the loop.

8.4 Materials and Methods

High-temperature unfolding. The molecular dynamics simulations for protein MJ0366 were performed through using the software package GROMACS 4.0.7 with GROMOS force field [40]. The starting structure of protein MJ0366 was taken from the NMR structure of the Protein Data Bank. This protein had nine simulation trajectories at 530 K for 40 ns, and a molecular dynamics simulation in native state was performed under 298 K at neutral PH. For preparing the molecular dynamics simulations, the starting structure was solvated with SPC216 water, and then subjected to 20,000 steps of steepest descent minimization. The nearest

distance between solute and box was 1.2 nm. Following the minimization, the whole system was subjected to 500,000 steps molecular dynamics simulations under NVT canonical ensemble and NPT constant pressure and constant temperature ensemble, respectively. The initial velocities were assigned from the Maxwellian distribution. The time step for these molecular dynamics simulations was 2 fs, and the neighboring list was updated every 5 steps. The transition states in the high-temperature unfolding process were determined by the conformational cluster method which was based on the Ca root-mean-square deviation (Ca RMSD) among the structures taken from the molecular dynamics simulation trajectories. For the nine simulation trajectories, the Ca RMSD values of the whole trajectory were used to generate positive definite matrix. The Michael Levitt's projecting co-ordinate spaces method was used to project this positive definite matrix onto the best plane [41]. The structures in the last 5 ps of the first obvious cluster were regarded as the transition state.

All-Atom Model. The all-atom model has been described [42] and has an available web server [43]. In the all-atom model of protein, only the heavy atoms were included. The single bead with unit mass was used to represent each atom. The harmonic potentials were used to restrain the bond lengths, bond angles, improper dihedrals, and planar dihedrals. The attractive 6–12 interactions were used for the nonbonded atom pairs which formed the native contacts, and the repulsive interactions were given to the other nonlocal interactions. For the Ca coarse-grained protein model [44], the single bead was centered in the Ca position to represent each residue. The contact map was constructed by including all residue pairs that at least had one atom-atom contact between them. Here we used GROMACS 4.0.7 software package to perform the molecular dynamics simulations [40]. The constant temperature molecular dynamics simulations at folding temperature were used to get thermodynamics datas, and these datas were compiled through using weighted histogram analysis method [45].

Reaction Cordinates. We used Q_{AA} and Q_{CA} as the reaction coordinates. Q_{AA} is the fraction of native contact which is the probability of interactional atoms comparing with the native state. If any atom-atom interaction between two residues within 1.2 times the native distance σij are considered as the native contact. Q_{CA} is the fraction of native contact for the Ca coarse-grained model which includes the residue pairs whose Ca atoms within 1.2 times their native distance.

References

1. Virnau P, Mirny LA, Kardar M (2006) Intricate knots in proteins: function and evolution. PLoS Comput Biol 2:1074–1079
2. Taylor WR (2000) A deeply knotted protein structure and how it might fold. Nature 406:916–919
3. Taylor WR (2007) Protein knots and fold complexity: some new twists. Comput Biol Chem 31:151–162
4. Mansfield ML (1994) Are there knots in proteins? Nat Struct Biol 1:213–214

5. Bölinger D, Sułkowska JI, Hsu HP, Mirny LA, Kardar M, Onuchic JN, Virnau P (2010) A Stevedore's protein knot. PLoS Comput Biol 6:e1000731
6. King NP, Yeates EO, Yeates TO (2007) Identification of rare slipknots in proteins and their implications for stability and folding. J Mol Biol 373:153–166
7. Gosavi S, Chavez LL, Jennings PA, Onuchic JN (2006) Topological frustration and the folding of interleukin-1 beta. J Mol Biol 357:986–996
8. Go N (1983) Theoretical studies of protein folding. Annu Rev Biophys Bioeng 12:183–210
9. Nymeyer H, Garcia AE, Onuchic JN (1998) Folding funnels and frustration in off-lattice minimalist protein landscapes. Proc Natl Acad Sci USA 95:5921–5928
10. Koga N, Takada S (2001) Roles of native topology and chain-length scaling in protein folding: a simulation study with a go-like model. J Mol Biol 313:171–180
11. Bryngelson JD, Wolynes PG (1987) Spin glasses and the statistical mechanics of protein folding. Proc Natl Acad Sci USA 84:7524–7528
12. Leopold PE, Montal M, Onuchic JN (1992) Protein folding funnels—a kinetic approach to the sequence structure relationship. Proc Natl Acad Sci USA 89:8721–8725
13. Onuchic JN, Wolynes PG (2004) Theory of protein folding. Curr Opin Struct Biol 14:70–75
14. Bayro MJ, Mukhopadhyay J, Swapna GV, Huang JY, Ma LC, Sineva E. Dawson PE, Montelione GT, Ebright RH (2003) Structure of antibacterial peptide microcin J25: a 21-residue lariat protoknot. J Am Chem Soc 125:12382–12383
15. Duff AP, Cohen AE, Ellis PJ, Kuchar JA, Langley DB, Shepard EM, Dooley DM, Freeman HC, Guss JM (2003) The crystal structure of Pichia pastoris lysyl oxidase. Biochemistry 42:15148–15157
16. Wagner JR, Brunzelle JS, Forest KT, Vierstra RD (2005) A light-sensing knot revealed by the structure of the chromophore-binding domain of phytochrome. Nature 438:325–331
17. Vimau P, Mallam A, Jackson S (2011) Structures and folding pathways of topologically knotted proteins. J Phys: Condens Matter 23:1–17
18. Mallam AL, Jackson SE (2005) Folding studies on a knotted protein. J Mol Biol 346:1409–1421
19. Mallam AL, Jackson SE (2007) A comparison of the folding of two knotted proteins: YbeA and YibK. J Mol Biol 366:650–665
20. Mallam AL, Jackson SE (2006) Probing nature's knots: the folding pathway of a knotted homodimeric protein. J Mol Biol 359:1420–1436
21. Mallam AL, Morris ER, Jackson SE (2008) Exploring knotting mechanisms in protein folding. Proc Natl Acad Sci USA 105:18740–18745
22. Ohta S, Alam MT, Arakawa H, Ikai A (2004) Origin of mechanical strength of bovine carbonic anhydrase studied by molecular dynamics simulation. Biophys J 87:4007–4020
23. Wallin S, Zeldovich KB, Shakhnovich EI (2007) The folding mechanics of a knotted protein. J Mol Biol 368:884–893
24. Sulkowska JI, Sulkowski P, Onuchic J (2009) Dodging the crisis of folding proteins with knots. Proc Natl Acad Sci USA 106:3119–3124
25. Lu J, den Dulk-Ras A, Hooykaas PJ, Glover JN (2009) *Agrobacterium tumefaciens* VirC2 enhances T-DNA transfer and virulence through its C-terminal ribbon-helix-helix DNA-binding fold. Proc Natl Acad Sci U S A 106:9643–9648
26. Li J, Wei DQ, Wang JF, Yu ZT, Chou KC (2012) Molecular dynamics simulations of CYP2E1. Med Chem 8:208–221
27. Gu RX, Liu LA, Wei DQ, Du JG, Liu L, Liu H (2011) Free energy calculations on the two drug binding sites in the M2 proton channel. J Am Chem Soc 133:10817–10825
28. Wang Y, Wei DQ, Wang JF (2010) Molecular dynamics studies on T1 lipase: insight into a double-flap mechanism. J Chem Inf Model 50:875–878
29. Wang JF, Yan JY, Wei DQ, Chou KC (2009) Binding of CYP2C9 with diverse drugs and its implications for metabolic mechanism. Med Chem 5:263–270
30. Gong K, Li L, Wang JF, Cheng F, Wei DQ, Chou KC (2009) Binding mechanism of H5N1 influenza virus neuraminidase with ligands and its implication for drug design. Med Chem 5:242–249

31. Li L, Wei DQ, Wang JF, Chou KC (2007) Computational studies of the binding mechanism of calmodulin with chrysin. Biochem Biophys Res Commun 358:1102–1107
32. Arias HR, Gu RX, Feuerbach D, Guo BB, Ye Y, Wei DQ (2011) Novel positive allosteric modulators of the human α7 nicotinic acetylcholine receptor. Biochemistry 50:5263–5278
33. Zhang T, Liu LA, Lewis DF, Wei DQ (2011) Long-range effects of a peripheral mutation on the enzymatic activity of cytochrome P450 1A2. J Chem Inf Model 51:1336–1346
34. Wang JF, Gong K, Wei DQ, Li YX, Chou KC (2009) Molecular dynamics studies on the interactions of PTP1B with inhibitors: from the first phosphate-binding site to the second one. Protein Eng Des Sel 22:349–355
35. Lian P, Wei DQ, Wang JF, Chou KC (2011) An allosteric mechanism inferred from molecular dynamics simulations on phospholamban pentamer in lipid membranes. PLoS ONE 6:e18587
36. Li J, Wei DQ, Wang JF, Li YX (2011) A negative cooperativity mechanism of human CYP2E1 inferred from molecular dynamics simulations and free energy calculations. J Chem Inf Model 51:3217–3225
37. Kolesov G, Virnau P, Kardar M, Mirny LA (2007) Protein knot server: detection of knots in protein structures. Nucleic Acids Res 35:W425–W428
38. Noel JK, Sułkowska JI, Onuchic JN (2010) Slipknotting upon native-like loop formation in a trefoil knot protein. Proc Natl Acad Sci USA 107:15403–15408
39. Wallin S, Chan HS (2006) Conformational entropic barriers in topology-dependent protein folding: perspectives from a simple nativecentric polymer model. J Phys: Condens Matter 18:S307–S328
40. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. J Chem Theory Comput 4:435–447
41. Levitt M (1983) Molecular dynamics of native protein. II. Analysis and nature of motion. J Mol Biol 168:621–657
42. Whitford PC, Noel JK, Gosavi S, Schug A, Sanbonmatsu KY, Onuchic JN (2009) An all-atom structure-based potential for proteins: bridging minimal models with all-atom empirical forcefields. Proteins 75:430–441
43. Noel JK, Whitford PC, Sanbonmatsu KY, Onuchic JN (2010) SMOG@ctbp: simplified deployment of structure-based models in GROMACS. Nucleic Acids Res. doi:10.1093/nar/gkq498
44. Clementi C, Nymeyer H, Onuchic JN (2000) Topological and energetic factors: what determines the structural details of the transition state ensemble and "on-route" intermediates for protein folding? An investigation for small globular proteins. J Mol Biol 298:937–953
45. Kumar S, Bouzida D, Swendsen RH, Kollman PA, Rosenberg JM (1992) The weighted histogram analysis method for free-energy calculations on biomolecules I. The method. J Comput Chem 13:1011–1021

Chapter 9
Binding Induced Intrinsically Disordered Protein Folding with Molecular Dynamics Simulation

Haifeng Chen

Abstract Intrinsically disordered proteins lack stable tertiary and/or secondary structures under physiological conditions in vitro. Intrinsically disordered proteins undergo significant conformational transitions to well folded forms only on binding to partner. Molecular dynamics simulations are used to research the mechanism of folding for intrinsically disordered protein upon partner binding. Room-temperature MD simulations suggest that the intrinsically disordered proteins have non-specific and specific interactions with the partner. Kinetic analysis of high-temperature MD simulations shows that bound and apo-states unfold via a two-state process, respectively. Φ-value analysis can identify the key residues of intrinsically disordered proteins. Kolmogorov-Smirnov (KS) *P* test analysis illustrates that the specific recognition between intrinsically disordered protein and partner might follow induced-fit mechanism. Furthermore, these methods can be widely used for the research of the binding induced folding for intrinsically disordered proteins.

Keywords IDPs · Molecular dynamics simulations · Induced-fit mechanism

9.1 Introduction

Intrinsically disordered proteins lack stable tertiary and/or secondary structures under physiological conditions in vitro [1]. A large number of proteins (between 25 and 41 %) are intrinsically disordered. If the dogma dedicates that proteins need a structure to function, why do so many proteins live in the disorder state? [2] However, these intrinsically disordered proteins also play key function in regulation, signaling, and control upon binding with multiple interaction partners [3].

H. Chen (✉)
State Key Laboratory of Microbial Metabolism, Department of Bioinformatics and Biostatistics, College of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China
e-mail: haifengchen@sjtu.edu.cn

D. Wei et al. (eds.), *Advance in Structural Bioinformatics*, Advances in Experimental Medicine and Biology 827, DOI 10.1007/978-94-017-9245-5_9

These proteins have many names, like rheomorphic, flexible or highly flexible, natively denatured, natively unfolded, intrinsically unstructured, intrinsically disordered. These proteins composed of an ensemble of highly heterogeneous conformations. After statistics of disordered protein database, IDPs include significantly higher levels of polar amino acids for Glu, Lys, Arg, Gln, Ser, Asp and Pro, and lower levels of hydrophobic residues for Ile, Leu, Val, Trp, Phe, Tyr, Thr, Met, Cys, His and Asn [4].

Furthermore, regions of disorder are found to be abundant in proteins associated with signaling, cancer, cardiovascular disease, amyloidoses, neurodegenerative diseases, and diabetes [5]. Different from structural protein as drug target, IDPs as drug target can bring low binding affinity and low side effect. There are two strategies for drug design targeting IDPs. Firstly, drug is binding to structured partner, thereby preventing the binding of the disordered partner. Secondly, drug is binding directly to the disordered partner, thereby preventing the association of two proteins. For this approach both partners were disordered, but small molecules bound to one of the two partners only. For example, c-Myc-Max inhibitors bind to distinct ID regions of c-Myc [6, 7]. These binding sites are composed of short contiguous stretches of amino acids that can selectively and independently bind small molecules. Inhibitor binding induces only local conformational changes, preserves the overall disorder of c-Myc, and inhibits dimerization with Max.

Furthermore, many intrinsically disordered proteins undergo significant conformational transitions to well folded forms only on binding to target ligands [8–11]. These experimental observations raise a set of interesting questions if these intrinsic disordered proteins obey an induced fit upon binding.

Coarse-grained modeling simulation [12] and all-atomic model with high temperature simulation [13] were used in intrinsically disordered protein folding coupled partner binding. So far the folding time scales of all atomic MD simulations are restricted to microsecond magnitude at room temperature (298 K), which is significant shorter than the folding half times of most proteins [14, 15]. In order to reveal the conformational changes within reasonable time, all MD simulations in explicit solvent at high temperature have been widely used to monitor the unfolding pathways of proteins. The unfolding timescales could be nanosecond at 498 K [14, 16]. Moreover, according to the principle of microscopic reversibility, experiments have demonstrated that the transition state for folding and unfolding is supposed to be same [14]. Therefore, MD simulations high temperatures are widely used in the folding of intrinsically disordered proteins coupled partner binding.

9.2 Materials and Method

The atomic coordinates of intrinsically disordered proteins were obtained from pdb data bank. Point mutants were modeled with SCWRL3 [17]. All hydrogen atoms were added using the LEAP module of AMBER [18]. Counter-ions were used to

maintain system neutrality. All systems were solvated in a truncated octahedron box of TIP3P waters with a buffer of 10 Å [19]. Particle Mesh Ewald (PME) [20] was applied to handle long-range electrostatic interactions with default setting in AMBER [18]. The parm99 force filed was used to compute the interactions within protein [21]. The SHAKE algorithm [22] was employed to constrain bonds including hydrogen atoms. All solvated systems were first minimized by steepest descent to remove any structural clash, followed by heating up and brief equilibration in the NPT ensembles at 298 K. The time step was 2 fs with a friction constant of 1 ps^{-1} using in Langevin dynamics. To study the folded state of each solvated system, multiple independent trajectories in the NPT ensemble at 298 K were simulated with PMEMD of AMBER. Then multiple independent unfolding trajectories were performed to investigate unfolding pathways for each solvated system in the NVT ensemble.

9.3 Results

TIS11d, KID, LEF, p53, CBP, and Brinker are partially or fully intrinsically disordered proteins. [13, 23–27] As transcription factor, they play key roles in signal transduction. Upon binding with DNA, RNA, or other transcription factors, they can well fold and will be introduced in this book. Their complex structures are illustrated in Fig. 9.1.

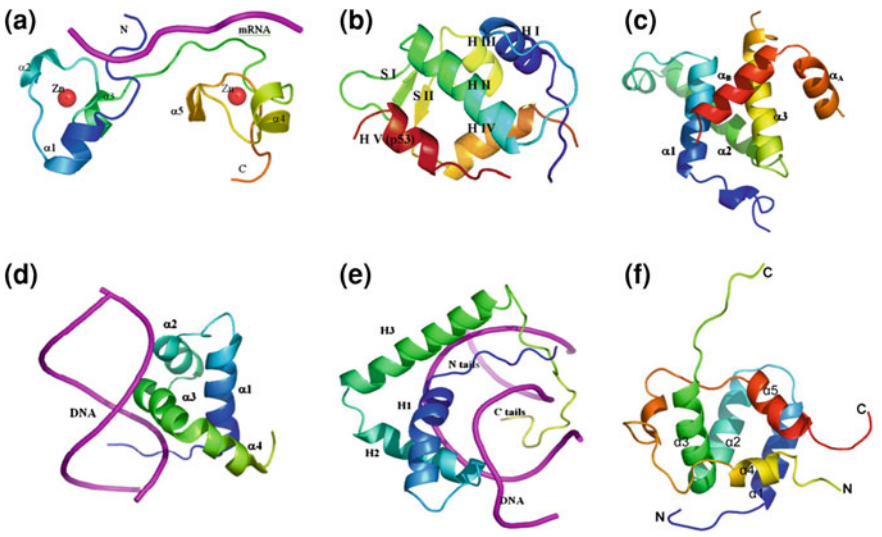


Fig. 9.1 The complex structure of intrinsically disordered proteins. **a** TIS11d/mRNA. **b** p53/MDM2. **c** pKID/KIX. **d** Brinker/DNA. **e** LEF/DNA. **f** p53/CBP

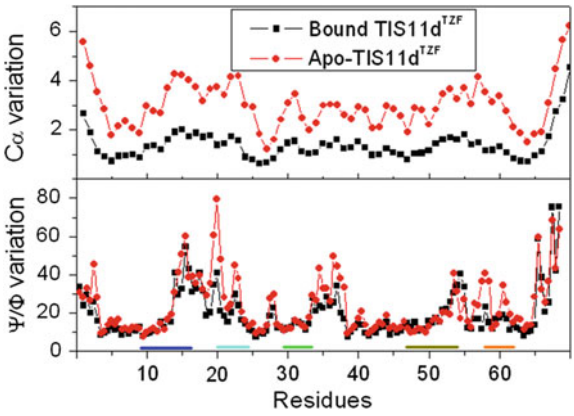


Fig. 9.2 Cα and Φ/Ψ variations for TIS11d

To capture the average properties of proteins, multiple trajectories for MD simulations (5–10) are necessary [28]. To study the recognition for intrinsically disordered proteins, multiple independent trajectories for apo-states and their complex were simulated at room temperature (298 K), respectively. Cα and Φ/ψ fluctuations for apo and bound states are researched. In general, the Cα variations of bound state are significant smaller than those of apo-state, especially in the region of the binding site. The results of apo and bound TIS11d are shown in Fig. 9.2 [26]. The Cα fluctuation of bound TIS11d^{TZF} is much smaller than that of apo-TIS11d^{TZF}, especially in the binding site of mRNA and zinc. This suggests that bound TIS11d^{TZF} become less flexible and more stable upon mRNA and zinc binding, which is consistent with the experiment. However, the Φ/ψ variation of bound TIS11d^{TZF} is similar to that of apo-TIS11d^{TZF}, suggesting that the secondary structure of bound TIS11d^{TZF} does not significantly change upon mRNA and zinc binding. Indeed, the helices of α1, α3 and α4 are already stable within apo-TIS11d^{TZF}.

To clearly illustrate the conformational difference, the landscapes of distance difference between the average pairwise intra-molecular distance of bound states and corresponding average pairwise intra-molecular distance of apo states for intrinsically disordered protein are shown in Fig. 9.3 [24]. The landscapes can reflect the relative conformational change of DNA and LEF backbone. The deep red area indicates that the distance difference for bases 5–8 and 23–26 is positive value. These bases are corresponding to the minor groove. This suggests that the minor groove is widened upon LEF-binding. Furthermore, disordered C-tail of LEF is located at the minor groove. This suggests that the disordered C-tail of LEF has interactions with DNA and open the minor groove of DNA. The deep blue area represents that the distance difference is negative value. It suggests that the major groove is contracted. That is consistent with the experimental observation that DNA is bended upon LEF-binding [29, 30]. For LEF, the deep red and blue areas are locked at disordered C-tail. This suggests that C-tail of LEF has significant conformational change.

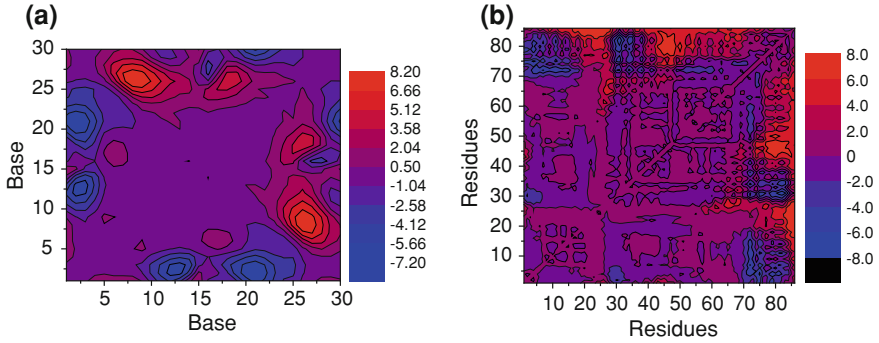


Fig. 9.3 Distance difference landscapes for DNA and LEF. **a** DNA. **b** LEF

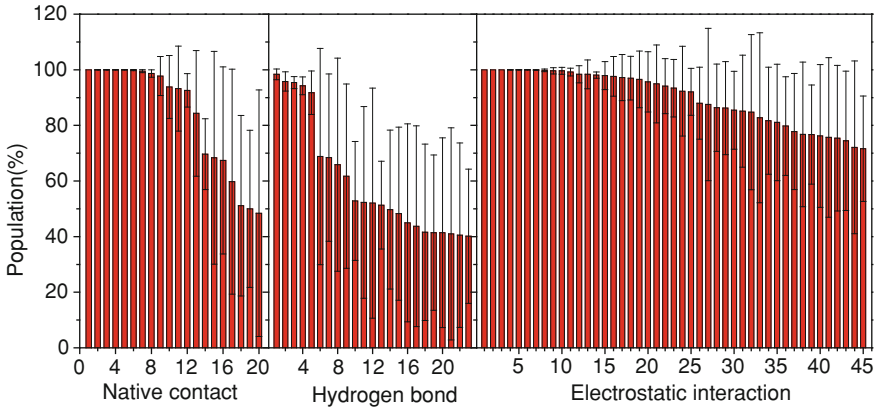


Fig. 9.4 Interactions between LEF and DNA

To study the driving force for these conformational adjustments, the electrostatic, hydrophobic, and hydrogen-binding interactions between intrinsically disordered protein and partner were analyzed and shown in Fig. 9.4. From this figure, stable electrostatic interactions, hydrogen bonds, and hydrophobic interactions can be calculated. In general, partner binding will introduce more electrostatic interactions, native contacts and hydrogen bonds at the interface which are responsible for the higher stability for intrinsically disordered proteins.

9.3.1 Unfolding Kinetics

High temperature simulation was used to research the unfolding kinetics of intrinsically disordered proteins with the parameters of the fraction of native tertiary contact (Qf) and the fraction of native binding contact (Qb). Time evolutions

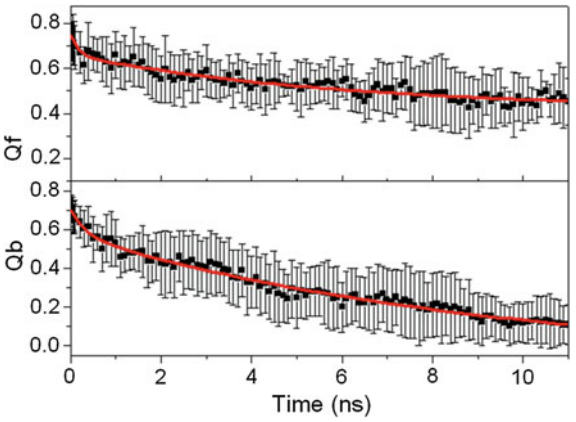


Fig. 9.5 Unfolding kinetics for bound pKID

of Qb and Qf for apo and bound states are shown in Fig. 9.5 [23]. The tertiary unfolding and unbinding can be fitted well by a single exponential function, indicating first order kinetics in the NVT ensemble at high temperature (498 K). This suggests that the binding of partner significantly postpones the tertiary unfolding of intrinsically disordered proteins. This is in agreement with the experimental observations [8, 31].

9.3.2 Φ-Value Prediction

Φ values have been widely used by theoretical and experimental works to identify the key residues for protein folding [32–34]. The Φ values of pKID were predicted and shown in Fig. 9.6. Note also that the highest Φ values are found for Asn139, Asp140 and Leu141, suggesting these residues play key role in the folding of

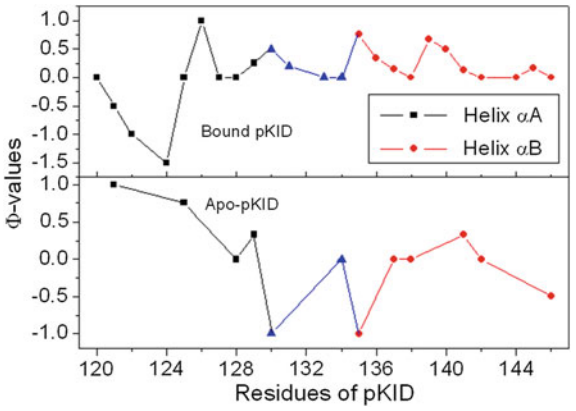


Fig. 9.6 Φ-values for bound and apo pKID

pKID [23]. A critical role of Leu141, which deeply extends into the hydrophobic groove of KIX, forms three hydrophobic contacts with KIX. All predicted Φ values can be confirmed by experiments.

9.3.3 Unfolding Pathway

According to the unfolding kinetics analysis, the unfolding orders of bound intrinsically disordered protein are shown in Fig. 9.7 [13]. If we assumed folding is reverse of unfolding, the proposed folding pathway of bound intrinsically disordered protein is from the unfolded state, then secondary structure folding, tertiary folding, partner binding, then to the folded state.

9.3.4 Recognition Mechanism

Conformational selection and induced fit are two widely used models to interpret the recognition between intrinsic disordered proteins [35]. According to the conformational selection paradigm, various conformational ensembles explore the free energy landscapes corresponding to diverse stable unbound states in equilibrium. During the binding process, the favorable conformation compatible with binding selectively stabilize, and the populations of conformational ensembles shift towards stabilizing state [36–39]. However, the induced fit scenario proposes that the favorable conformation results from significant changes of unbound ensembles upon allosteric binding [40–43]. It is worthy to point out that conformational selection and induced fit models cannot be distinguished absolutely [44]. Indeed, some systems involve kinetic elements of both mechanisms [45, 46].

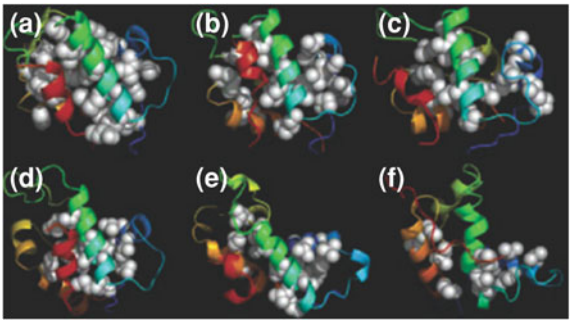


Fig. 9.7 Unfolding pathway for bound p53. **a** fold state. **b** unbinding. **c** tertiary unfolding. **d** helix 3/5 unfolding. **e** helix 1/2/4 unfolding. **f** unfolded state

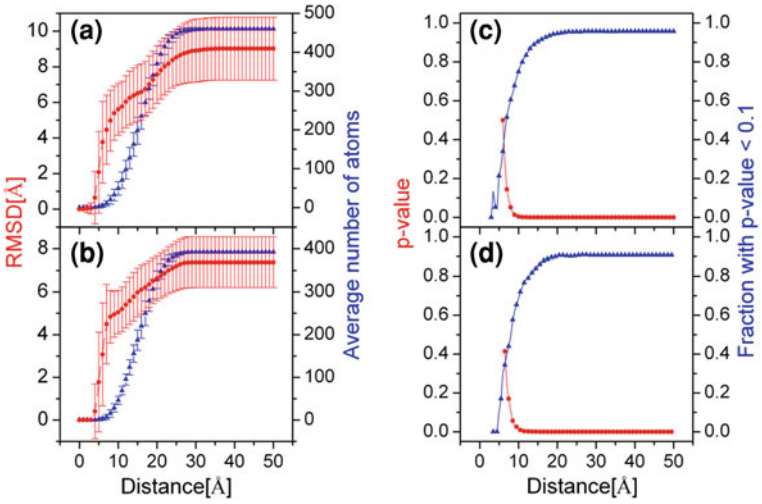


Fig. 9.8 Local conformational RMSD differences between bound and apo conformations as a function of distance from the centroid of binding partner and statistical significance of conformational selection in p53 and CBP binding. Average local RMSD for 10 pairs of bound conformations and the most similar apo conformation and for 90 pairs of bound NCBD and the other apo conformations, as a function of distance from the centroid of binding partner. **a** CBP. **b** p53. **c** CBP. **d** p53

The possible magnitudes of conformational selection and induced fit [47] are calculated to reveal the recognition mechanism. To explore the recognition mechanism, the average RMSD deviations of bound conformation and apo conformations are analyzed as a function of distance from the centroid of binding partner and shown in Fig. 9.8 [27]. This figure illustrates that the RMSD variation gradually increases until to the global level. This suggests that there is an induced fit far away for the binding site.

To address the statistical significance for differences of deviations between these two systems, two sample Kolmogorov-Smirnov test [48] is used to calculate the P value for each distance group. Figure 9.8c illustrates the median of P values and the fraction with $P < 0.1$ for all 100 pairs of CBP conformations in each distance group. It is found that the median P values are typically smaller than 0.1 in most distance group, especially in some larger distance group with median P values approximates 0. The conformations with $P < 0.1$ exceed 50 % in most distance groups. This suggests that the bound CBP is significant different from the apo conformation far away from the binding site and the differences are statistically significant. In summary, the recognition between intrinsic disordered CBP and p53 might obey an induced fit based on the RMSD and P-value analysis.

9.4 Conclusion and Remark

Intrinsically disordered proteins lack stable tertiary and/or secondary structures under physiological conditions in vitro. Intrinsically disordered proteins undergo significant conformational transitions to well folded forms only on binding to partner. Molecular dynamics simulations are used to research the mechanism of folding for intrinsically disordered protein upon partner binding. Room-temperature MD simulations suggest that the intrinsically disordered proteins have non-specific and specific interactions with the partner. Kinetic analysis of high-temperature MD simulations shows that bound and apo-states unfold via a two-state process, respectively. Φ-value analysis can identify the key residues of intrinsically disordered proteins. Kolmogorov-Smirnov (KS) *P* test analysis illustrates that the specific recognition between intrinsically disordered protein and partner might follow induced-fit mechanism. Furthermore, these methods can be widely used for the research of the binding induced folding for intrinsically disordered proteins.

References

1. Uversky VN, Oldfield CJ, Dunker AK (2008) Intrinsically disordered proteins in human diseases: introducing the D2 concept. Annu Rev Biophys 37:215–246
2. Chouard T (2011) Structural biology: breaking the protein rules. Nature 471:151–153
3. Liu J, Faeder JR, Camacho CJ (2009) Toward a quantitative theory of intrinsically disordered proteins and their function. Proc Natl Acad Sci USA 106:19819–19823
4. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z (2001) Intrinsically disordered protein. J Mol Graph Model 19:26–59
5. Metallo SJ (2010) Intrinsically disordered proteins are potential drug targets. Curr Opin Chem Biol 14:481–488
6. Wang H, Hammoudeh DI, Follis AV, Reese BE, Lazo JS, Metallo SJ, Prochownik EV (2007) Improved low molecular weight Myc-Max inhibitors. Mol Cancer Ther 6:2399–2408
7. Hammoudeh DI, Follis AV, Prochownik EV, Metallo SJ (2009) Multiple independent binding sites for small-molecule inhibitors on the oncoprotein c-Myc. J Am Chem Soc 131:7390–7401
8. Sugase K, Dyson HJ, Wright PE (2007) Mechanism of coupled folding and binding of an intrinsically disordered protein. Nature 447:1021–1025
9. Canon F, Ballivian R, Chirot F, Antoine R, Sarni-Manchado P, Lemoine J, Dugourd P (2011) Folding of a salivary intrinsically disordered protein upon binding to tannins. J Am Chem Soc 133:7847–7852
10. Wang J, Wang Y, Chu X, Hagen SJ, Han W, Wang E (2011) Multi-scaled explorations of binding-induced folding of intrinsically disordered protein inhibitor IA3 to its target enzyme. PLoS Comput Biol 7:e1001118
11. Zhang W, Ganguly D, Chen J (2012) Residual structures, conformational fluctuations, and electrostatic interactions in the synergistic folding of two intrinsically disordered proteins. PLoS Comput Biol 8:e1002353

12. Chen J (2009) Intrinsically disordered p53 extreme C-terminus binds to S100B(betabeta) through "fly-casting". J Am Chem Soc 131:2088–2089
13. Chen HF, Luo R (2007) Binding induced folding in p53-MDM2 complex. J Am Chem Soc 129:2930–2937
14. Fersht AR, Daggett V (2002) Protein folding and unfolding at atomic resolution. Cell 108:573–582
15. Klepeis JL, Lindorff-Larsen K, Dror RO, Shaw DE (2009) Long-timescale molecular dynamics simulations of protein structure and function. Curr Opin Struct Biol 19:120–127
16. Shea JE, Brooks CL 3rd (2001) 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
17. Canutescu AA, Shelenkov AA, Dunbrack RL Jr (2003) A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci 12:2001–2014
18. Case DA, Darden TA, Cheatham TE, Simmerling ICL, Wang J, Duke RE, Luo R, Walker RC, Zhang W, Merz KM, Roberts B, Wang B, Hayik S, Roitberg A, Seabra G, Kolossváry I, Wong KF, Paesani F, Vanicek J, Liu J, Wu X, Brozell SR, Steinbrecher T, Gohlke H, Cai Q, Ye X, Wang J, Hsieh M-J, Cui G, Roe DR, Mathews DH, Seetin MG, Sagui C, Babin V, Luchko T, Gusarov S, Kovalenko A, Kollman PA (2010) Amber 11, University of California, San Francisco
19. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926
20. Darden T, York D, Pedersen L (1993) Particle mesh Ewald: an N·log(N) method for Ewald sums in large systems. J Chem Phys 98:10089
21. Wang JM, Cieplak P, Kollman PA (2000) 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
22. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys 23:327–341
23. Chen HF (2009) Molecular dynamics simulation of phosphorylated KID post-translational modification. PLoS ONE 4:e6516
24. Qin F, Ye W, Chen Y, Chen X, Li Y, Zhang J, Chen HF (2012) Specific recognition between intrinsically disordered LEF and DNA. Phys Chem Chem Phys 14:538–545
25. Qin F, Jiang Y, Chen Y, Wu M, Yan G, Ye W, Li Y, Zhang J, Chen HF (2011) Conformational selection or induced fit for Brinker and DNA recognition. Phys Chem Chem Phys 13:1407–1412
26. Qin F, Chen Y, Li YX, Chen HF (2009) Induced fit for mRNA/TIS11d complex. J Chem Phys 131:115103
27. Yu Q, Ye W, Wang W, Chen HF (2013) Global conformational selection and local induced fit for the recognition between intrinsic disordered p53 and CBP. PLoS ONE 8:e59627
28. Day R, Daggett V (2005) Ensemble versus single-molecule protein unfolding. Proc Natl Acad Sci USA 102:13445–13450
29. Love JJ, Li X, Case DA, Giese K, Grosschedl R, Wright PE (1995) Structural basis for DNA bending by the architectural transcription factor LEF-1. Nature 376:791–795
30. Love JJ, Li X, Chung J, Dyson HJ, Wright PE (2004) The LEF-1 high-mobility group domain undergoes a disorder-to-order transition upon formation of a complex with cognate DNA. Biochemistry 43:8725–8734
31. Radhakrishnan I, Perez-Alvarado GC, Parker D, Dyson HJ, Montminy MR, Wright PE (1997) Solution structure of the KIX domain of CBP bound to the transactivation domain of CREB: a model for activator: coactivator interactions. Cell 91:741–752
32. Fersht AR (2000) Transition-state structure as a unifying basis in protein-folding mechanisms: contact order, chain topology, stability, and the extended nucleus mechanism. Proc Natl Acad Sci USA 97:1525–1529

33. Fernandez-Escamilla AM, Cheung MS, Vega MC, Wilmanns M, Onuchic JN, Serrano L (2004) Solvation in protein folding analysis: combination of theoretical and experimental approaches. Proc Natl Acad Sci USA 101:2834–2839
34. Fersht AR, Sato S (2004) Phi-value analysis and the nature of protein-folding transition states. Proc Natl Acad Sci USA 101:7976–7981
35. Boehr DD, Wright PE (2008) Biochemistry. How do proteins interact? Science 320:1429–1430
36. Frauenfelder H, Sligar SG, Wolynes PG (1991) The energy landscapes and motions of proteins. Science 254:1598–1603
37. Tsai CJ, Ma B, Nussinov R (1999) Folding and binding cascades: shifts in energy landscapes. Proc Natl Acad Sci USA 96:9970–9972
38. Boehr DD, McElheny D, Dyson HJ, Wright PE (2006) The dynamic energy landscape of dihydrofolate reductase catalysis. Science 313:1638–1642
39. Weikl TR, von Deuster C (2009) Selected-fit versus induced-fit protein binding: kinetic differences and mutational analysis. Proteins 75:104–110
40. Koshland DE (1958) Application of a theory of enzyme specificity to protein synthesis. Proc Natl Acad Sci USA 44:98–104
41. Rini JM, Schulze-Gahmen U, Wilson IA (1992) Structural evidence for induced fit as a mechanism for antibody-antigen recognition. Science 255:959–965
42. Turjanski AG, Gutkind JS, Best RB, Hummer G (2008) Binding-induced folding of a natively unstructured transcription factor. PLoS Comput Biol 4:e1000060
43. Schrank TP, Bolen DW, Hilser VJ (2009) Rational modulation of conformational fluctuations in adenylate kinase reveals a local unfolding mechanism for allostery and functional adaptation in proteins. Proc Natl Acad Sci USA 106:16984–16989
44. Csermely P, Palotai R, Nussinov R (2010) Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. Trends Biochem Sci 35:539–546
45. James LC, Tawfik DS (2003) Conformational diversity and protein evolution–a 60-year-old hypothesis revisited. Trends Biochem Sci 28:361–368
46. Okazaki K, Takada S (2008) Dynamic energy landscape view of coupled binding and protein conformational change: induced-fit versus population-shift mechanisms. Proc Natl Acad Sci USA 105:11182–11187
47. Wlodarski T, Zagrovic B (2009) Conformational selection and induced fit mechanism underlie specificity in noncovalent interactions with ubiquitin. Proc Natl Acad Sci USA 106:19346–19351
48. Massey Jr, FJ (1951) The Kolmogorov-Smirnov test for goodness of fit. J Am stat Assoc, 68–78

Chapter 10
Theoretical Studies on the Folding Mechanisms for Different DNA G-quadruplexes

Xue Wu, Peijun Xu, Jinguang Wang, Yong Xu, Ting Fu, Meixia Zhao, Depeng Zhang, Jiahui Liu, Hujun Shen, Zhilong Xiu and Guohui Li

Abstract The G-quadruplex DNA formed by the stack of guanines in human telomere sequence is a promising anticancer target. In this study we used the energy landscape theory to elucidate the folding mechanisms for the thrombin aptamer, Form 1 and Form 3 G-quadruplexes. The three G-quadruplexes were simulated with all-atom Gō-model. Results show that, the three G-quadruplexes fold through a two-state mechanism. In the initial stage of the folding process, the compact structures are formed. The G-quadruplexes need to form G-triplex structures on the basis of the compact structures before folding to the native states. The folding free energy barrier of Form 3 G-quadruplex is higher than thrombin aptamer and Form 1, which shows that the structure of Form 3 G-quadruplex has more stability than the other two G-quadruplexes.

Keywords G-quadruplex · Gō-model · Fold

Xue Wu, Ting Fu, Peijun Xu and Jinguang Wang have been contributed equally to this paper.

X. Wu · T. Fu · H. Shen · G. Li (✉)
Laboratory of Molecular Modeling and Design, State Key Laboratory of Molecular Reaction Dynamics, Dalian Institute of Chemical Physics, Chinese Academy of Science, Dalian, Liaoning, China
e-mail: ghli@dicp.ac.cn

P. Xu · M. Zhao · D. Zhang · J. Liu
School of Physics and Electronic Technology, Liaoning Normal University, Dalian, Liaoning, China

J. Wang
The First Affiliated Hospital, Dalian Medical University, Dalian, China

Y. Xu
Guangzhou Institute of Biomedicine and Health, Guangzhou, China

Z. Xiu
School of Life Science and Technology, Dalian University of Technology, Dalian, China

D. Wei et al. (eds.), *Advance in Structural Bioinformatics*, Advances in Experimental Medicine and Biology 827, DOI 10.1007/978-94-017-9245-5_10

10.1 Introduction

Telomere, a nucleoprotein complex, is located at the ends of linear eukaryotic chromosomes. It is important for maintaining the chromosomal stability and the integrity of the genome [1, 2]. In the process of the replication of eukaryotic chromosomes packed by the DNA, the ends of telomeric DNA cannot be copied by DNA polymerase, which is due to lack of a template strand in the extreme 3′ end of a DNA sequence [3, 4]. As a result, the 3′ end of telomeric DNA is eroded, and the telomeric DNA is shortened without compensation mechanism. Nevertheless, the telomeric DNA in the tumour cell is not shortened during the replication. The telomerase is important for maintaining the stability and integrity of the telomeric DNA in most of proliferating tumour cells [5]. It is interesting to study the telomere and telomerase because of the difference between the somatic and tumour cells for the maintenance of telomere. Human telomeric DNA is composed of thousands of tandem repeats of the guanine-rich sequences, and the 3′-end overhangs 100–200 nt [6]. The G-quadruplex structures can be built from the vertical stacking of planar G · G · G · G tetrads in the G-rich DNA sequences in vitro, and this structures have been found in the telomeric sequences and telomeres [7–11]. The activity of telomerase is inhibited by these structures, so the G-quadruplexes in human telomere sequences are the promising anticancer targets. It is meaningful to study these structures as the promising anticancer targets, and the research about the folding of nucleic acids is useful to understand the biological natures, so here we study the folding dynamics for DNA tetraplex.

To date, there have been reported various G-quadruplex structures [12, 13]. The NMR structure of intramolecular quadruplex formed by four repeats of d(TTAGGG) in the Na(+)-containing solution have been reported [9]. In this structure, three G-quartets are held together by strands in the alternating orientations, and two lateral loops and a central diagonal loop connect these G-quartets. Under approximately physiological ionic conditions, the G-quadruplex has a different conformation in present of the K+ solution comparing with the structure in Na+ solution. In the K+ solution, this crystal structure is consisted of all four parallel strands [12]. The thrombin aptamer sequence d(GGTTGGTGTGGTTGG) could form a G-quadruplex structure in the K+ solution, and this structure is composed of two guanine quartets connected by two T-T loops and a T-G-T loop [14]. The human telomeric DNA in physiologically was observed to form the (3 + 1) G-quadruplex topology (Form 1 and Form 2) in K+ solution, and this structure is consisted of three strands oriented in one direction and the fourth in the opposite direction [15, 16]. Both Form 1 and Form 2 contain one double-chain-reversal and two edgewise T-T-A loops, but the two structures differ in loop arrangement. Furthermore, a novel G-quadruplex fold (Form 3) has been found in K+ solution, and this structure is a basket-type G-quadruplex with two G-tetrad layers [17].

For better understanding the structure and biological nature, the investigation of the folding dynamics for G-quadruplex structures is needed. The stopped-flow mixing coupled with rapid wavelength scanning method was used to study the

folding dynamics for G-quadruplexes [18]. The folding of G-quadruplex structures from the G-rich oligonucleotides may be via intermediates. Some reports about the G-quadruplexes were shown a triplex structure which was important for the formation of G-quadruplex [19, 20]. The theoretical research has been used to study the folding dynamics of G-quadruplexes. A thrombin-bind DNA aptamer was investigated by replica exchange molecular dynamics simulation method at the all-atom level for giving more insight into its fold in atomic detail [21]. Though the structure and the folding of G-quadruplexes have been investigated, how this molecule overcoming the energy barrier folds from the G-rich oligonucleotide to the correct G-quadruplex topology still needs to be investigated. Here we used a different approach from these studies to discuss the folding mechanisms of G-quadruplexes. The folding routes of a well-designed sequence with the reduced effect of the local trap may be determined by the shape of this molecule and the chain connectivity of its backbone [22]. So the structure-based model can capture the essential folding features through isolating the effect of topology and removing all non-native energetic trap [23–25]. In the folding process, the energy landscape directs the folding of protein from unfolded state to the native state, the pattern of contacts directs the diverse sizes and shapes of the free energy barriers, and the native contacts may be more favorable than the nonnative contacts [26–28]. The DNA molecules have been studied widely [29–31], and this structure-based model has been used for the folding of nucleic acid in theory [32–34]. So here we used the all-atom structure-based model to study the folding pathways of thrombin-bind DNA aptamer, Form 1 and Form 3 G-quadruplexes. The PDB entries for thrombin-bind DNA aptamer, Form 1 and Form 3 G-quadruplexes are 148d, 2jsm and 2kf8, respectively. The study for the folding of the three G-quadruplexes demonstrated that all the three G-quadruplexes had a two-state folding behavior. In the folding process, the thrombin aptamer formed the two T-T loops firstly, and then the two G-quartets stacked together by the native contacts between the two ends. The Form 1 and Form 3 G-quadruplexes needed to form the compact structures first, and then through forming the G-triplex structures to fold into the native states. The energy barrier of Form 3 was higher than the other two G-quadruplexes, which may explain the reason that the stability of Form 3 G-quadruplex is higher than the other two G-quadruplexes.

10.2 Results and Discussion

The thrombin aptamer is consisted of two stacked G-quartets (Fig. 10.1a). Two T-T loops at the two ends and one T-G-T loop link the two G-quartets. The guanines in the two G-quartets have the alternative glycoside orientations. The two G-quartets compose of G1syn · G6anti · G10syn · G15anti and G2anti · G5syn · G11anti · G14syn [14]. The Form 1, a (3 + 1) quadruplex,is consisted of one anti · syn · syn · syn and two syn · anti · anti · anti G-tetrads (Fig. 10.1b). This G-quadruplex has one narrow, one wide and two medium grooves, and the double-chain-reversal loop is located in a medium groove

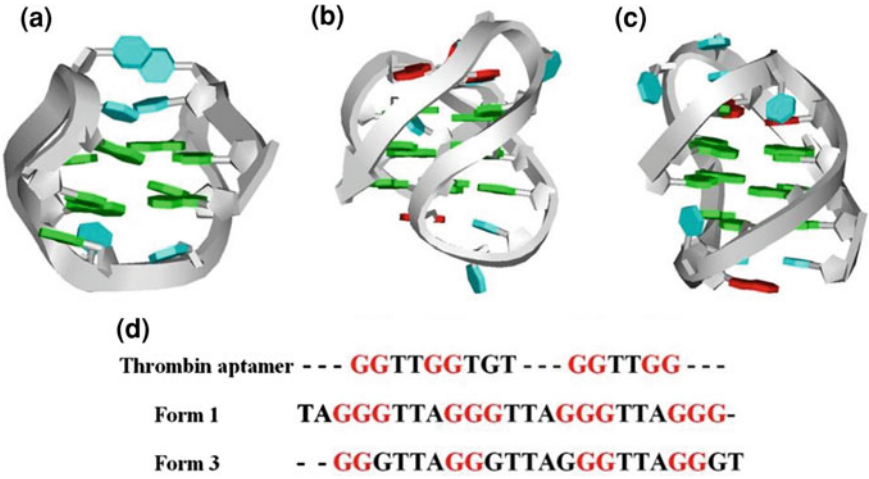


Fig. 10.1 The secondary structures and sequences alignment for thrombin aptamer, Form 1 and Form 3 G-quadruplexes. **a** Thrombin aptamer ribbon diagram. **b** The ribbon diagram for Form 1 G-quadruplex. **c** Form 3 G-quadruplex ribbon diagram. **d** Sequences alignment for thrombin aptamer, Form 1 and Form 3 G-quadruplexes. These guanines form the G-quartets in the three G-quadruplexes marked in *red*

[16]. The Form 3 G-quadruplex structure has two layers of G-tetrads like the thrombin aptamer, but forms antiparallel-stranded basket-type structure (Fig. 10.1c). The glycosidic conformations of guanines in the two G-tetrads are G1syn · G14syn · G20anti · G8anti and G2anti · G15anti · G19syn · G7syn. Form 3 has one diagonal and two edgewise loops.

10.2.1 Transition States in All-Atom Gō-Model

Thrombin aptamer. The free energy as a function of the number of native contact at folding temperature is shown in Fig. 10.3a. There are two basins corresponding to the folded and denatured states. The two states are separated by a free energy barrier as the transition state near the number of native contact of ~ 0.5. The transition state is defined as the state that near the maximum free energy as a function of the number of native contact [35]. The temperature wasn't defined as the usual Boltzmann constant but an arbitrary chosen constant in this study. The native contact maps for thrombin aptamer is shown in Fig. 10.2a. Comparing with the native state, the structures in the transition state have been formed some of the native contacts. In the transition state, the native contacts between G2 and G5, G1 and G6, G11 and G14, and G10 and G15 at the two ends were formed. Between G5 · G6 and G10 · G11, the native contacts G5-G11 and G6-G10 were appeared. In this state the native contacts in the T7-G8-T9 loop were formed. Some

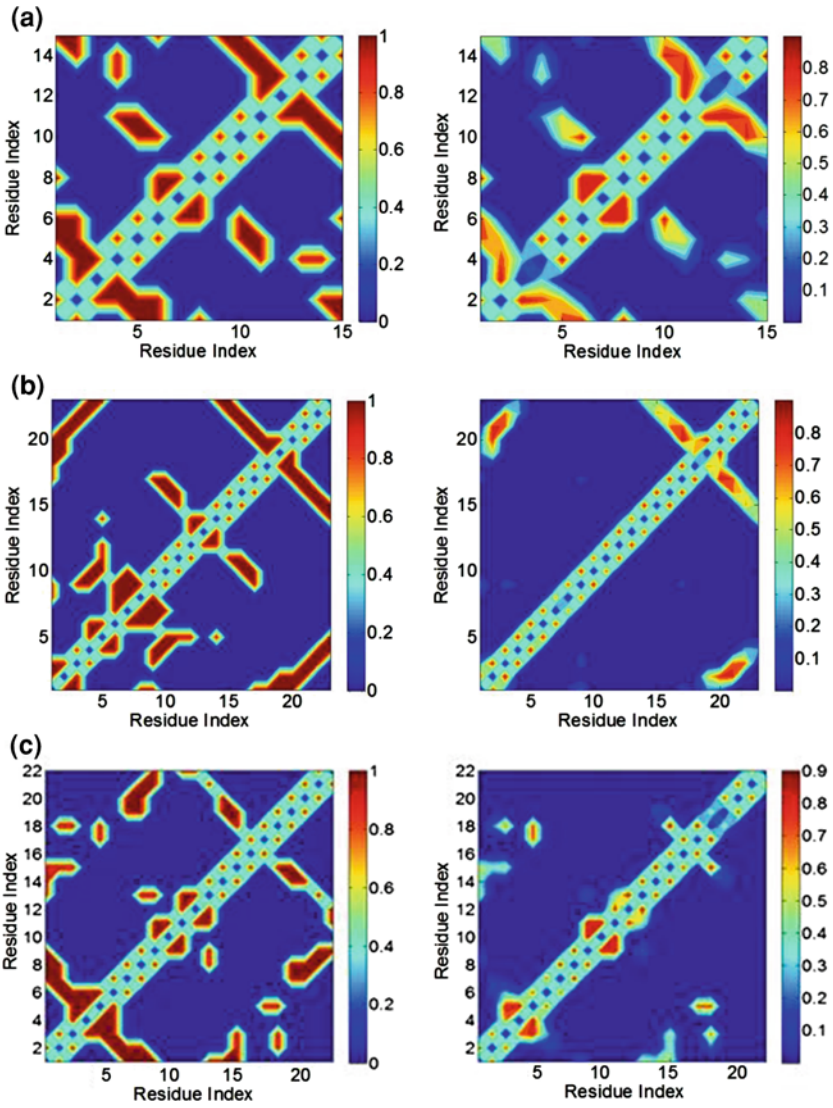


Fig. 10.2 Native contact maps for **a** thrombin aptamer, **b** Form 1 G-quadruplex and **c** Form 3 G-quadruplex. The *left maps* show the native contacts in native states, and the *right maps* present the native contacts in transition states for the three G-quadruplexes

structures formed the native contacts G2-G14, G1-G15, and T4-T13 between the loops of the two ends, but these native contacts had low probability than the other native contacts in the two G-tetrads. So in the transition state part of structures formed most of the native contacts, but most of the structures only presented the native contacts G2-G5, G1-G6, G11-G14, and G10-G15 at the two ends. Before

reaching to the native state, the G-triplex structures were found, which were formed by the native contacts G5-G11 and G6-G10 and the native contacts at the two ends.

10.2.2 Form 1 G-quadruplex

The free energy as a function of the number of native contact at folding temperature is shown in Fig. 10.4a. Two basins corresponding to the folded and denatured states are presented like the thrombin aptamer. The denatured states have the number of native contact of $\sim$0.15, and the number of native contact in the folded state is near 0.65. The transition state with the number of native contact of $\sim$0.35 separates the folded state and denatured state. The native contact maps of the structures in the native and transition states are shown in Fig. 10.2b. The contact map of the transition state had only two regions, so some of the native contacts weren't formed. One of the regions in the contact map had the native contacts G3-G21 and G4-G22 at the two ends of Form 1 G-quadruplex. At one end of the G-quadruplex, the native contacts G17-G21, G16-G22, and G15-G23 were appeared at the transition state. In the transitions state, the Form 1 G-quadruplex folded into the compact structure through forming the native contacts between the two ends. The native contacts A2-T19 and A2-A20 stabilized this compact structure. The native contact T18-A20 also stabilized this structure. The G9-G10-G11 strand did not stack with the other three strands. So the G-triplex structures were formed in the transition state, and these structures included the formed native contacts G3-G21, G4-G22, G17-G21, G16-G22, and G15-G23.

10.2.3 Form 3 G-quadruplex

The free energy as a function of native contact at the folding temperature is shown in Fig. 10.5a. There are two basins corresponding to the folded and denatured states like thrombin aptamer and form 1 G-quadruplex. The folded state has the number of native contact of $\sim$0.7, and the number of native contact in the denatured state is around 0.15. The transition state with the number of native contact of $\sim$0.3 separates the folded and denatured states. The native contact maps of the folded and transition states for Form 3 G-quadruplex are shown in Fig. 10.2c. Comparing with the native state, the some of the native contacts were formed at the transition state. The native contacts G1-G14 and G2-G15 were appeared at the transition state. Few structures in the transition state had native contacts G1-G8 and G2-G7. The native contact between G3 and T5 was appeared in the transition state. This native contact may stabilize the loop at one end, and promote the formation of the native contacts G1-G14 and G2-G15. The native contact G3-A18 was formed in the loop at one end, but the high probability of this native contact did not promote the formation of native contacts G14-G20 and G15-G19. The structures were compact at transition

state, and the native contacts T5-T17 and T5-A18 were good for the formation of these compact structures. The native contact between G9 and T11 was appeared, and this contact was benefit to the formation of native contacts G1-G14, G2-G15, G1-G8 and G2-G7. All the above, few structures in the transition state showed the G-triplex conformations formed by the native contacts G1-G14, G2-G15, G1-G8 and G2-G7, but most of the structures had native contacts G1-G14 and G2-G15 between the loops of the two ends.

10.2.4 The Folding Pathway of G-quadruplexes

Thrombin aptamer. The constant temperature simulations of thrombin aptamer were performed at folding temperature. The landscape for free energy as a function of the number of native contact and the Ca root-mean-square deviation (RMSD) is shown in Fig. 10.3b. In the folding process, the RMSD decreased as an increasing number of native contact, and the RMSD didn't have the sharp change. The transition state with the native contact of $\sim$0.5 separates the two folded and denatured regions. The free energy landscape as a function of the number of native contact and the radius of gyration is shown in Fig. 10.3c. The L-shaped landscape indicated that the radius of gyration decreased sharply in the initial stage of the folding process, when the structure reached the transition state with the number of native contact of $\sim$0.5, the radius of gyration decreased little further. So in the initial stage of the folding process, the compact structures were formed. The Fig. 10.3d is shown for the free energy landscape as a function of the number of native contact and the distance between 5′-end and 3′-end. In the initial stage of the folding process, the distance between 5′-end and 3′-end decreased sharply, so the 5′-end and the 3′-end was in close distance, and this G-rich oligonucleotide has formed a compact structure. After transition state the number of native contact increased rapidly as the slowly decreasing distance between the 5′-end and the 3′-end, and the two ends was in short-distance as the native state. So the unfolded state were formed a compact structure first through the stack between the 5′ and 3′ ends. More details on the folding mechanism can be derived from the free energy landscape as a function of the number of native contact and the relative contact order (RCO) parameter (Fig. 10.3e). The relative contact order was defined as a function of the distance between two residues that formed native contact [36]. Two regions as the denatured and folded states were shown in this plot. The unfolded region had the RCO value of $\sim$0.3, the number of native contact of the unfolded region was inferior to 0.4, and the folded region with the RCO of $\sim$0.5 presented the number of native contact higher than 0.6. In the folding process, the RCO increased as the increasing number of native contact, so the local native contacts were important for the initial stage of the folding process. With the increasing number of native contact the non-local native contacts were increased, so the local native contacts may promote the stack of the local structures, and then stabilized the whole structure with non-local native contacts.

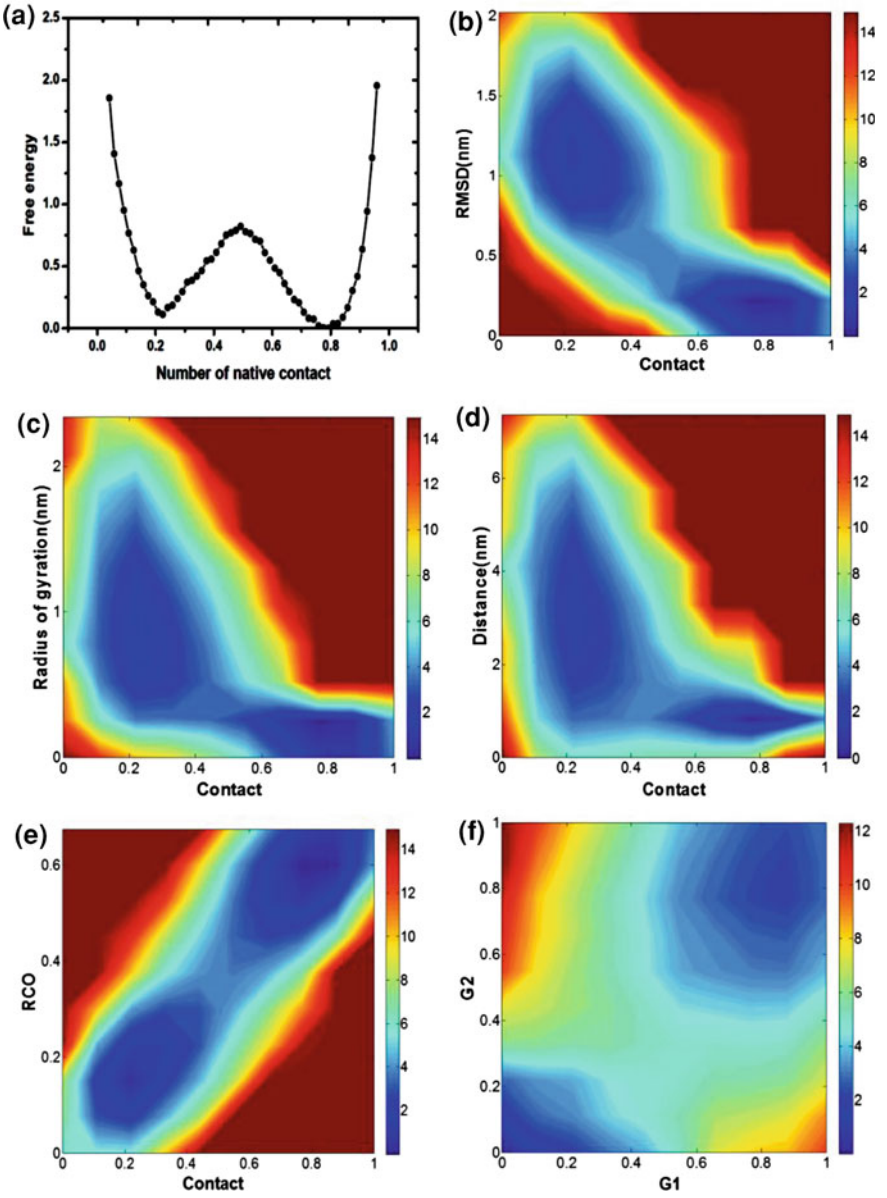


Fig. 10.3 Folding routes for thrombin aptamer. **a** Free energy is plotted as a function of the number of native contact. **b–f** The free energy landscapes as a function of various quantities at folding temperature: T = 115

The typical structures for the folding process are shown in Fig. 10.6a. The two ends of the unfolded structure folded to the native state firstly, but the native contacts G5-G11 and G6-G10 were not formed. The native contacts in the region of loop T7-G8-T9 were formed in the denatured state, and were stable in the whole folding process. The native contacts in the two ends were not formed, but the compact structures were appeared through forming the native contacts in the T7-G8-T9 loop. So the local native contacts had the main contribution for the compact structures. After the native contacts at the two ends formed, the native contacts G5-G11 and G6-G10 were appeared, at this time the G-triplex was formed by the native contacts G1-G6, G2-G5, G6-G10, G5-G11, G11-G14, and G10-G15. The native contacts G2-G14 and G1-G15 were the last formed in the G-tetrads, and these native contacts made the two ends stack together. In the experimental study, two-state folding process was found, the two ends folded firstly, and G-triplex structures were needed for the formation of the native state [37]. This folding process studied by the all-atom Gō-model is consistent with the conclusion of the experimental research. In the folding process, the native contacts in the loops contributed to the formation of the G-tetrads. The free energy landscape as the function of the number of native contact in the two G-triplexes is shown in Fig. 10.3(F). The G1 and G2 represent the native contacts of the G-quartets G1 · G6 · G10 · G15 and G2 · G5 · G11 · G14, respectively. The number of native contact in G1 increased rapidly as an increasing number of native contact of G2 before folding to the native state, so the native contacts in G1 may be formed prior to G2. The local native contacts formed in these loops also promoted the formation of the compact structure, and the native contact T4-T13 between the loops at the two ends stabilized the compact structure. After the formation of G-triplex structure the two ends stacked together and folded to the native state.

10.2.5 Form 1 G-quadruplex

We used constant temperature simulations for Form 1 G-quadruplex at folding temperature. The free energy landscape as a function of the number of native contact and the RMSD is shown in Fig. 10.4b. The RMSD decreased sharply with slowly increasing the number of native contacts, and after the number of native contact reached $\sim$0.4, the RMSD decreased little further. The transition area with the number of native contact of $\sim$0.3 separates the two regions of folded and denatured states. The Fig. 10.4c is shown for the free energy as a function of the number of native contact and the radius of gyration. The L-shaped landscape showed the radius of gyration for the integral G-quadruplex was decreased rapidly with an increasing number of native contact, but once the number of native contact reached $\sim$0.3 the radius of gyration decreased little further. In the initial stage of the folding process, the sharply decreased radius of gyration implied the unfolded G-quadruplex formed a compact structure, which was regarded as a basis for folding to the native state. The landscape of the free energy as a function of the number of native contact and the

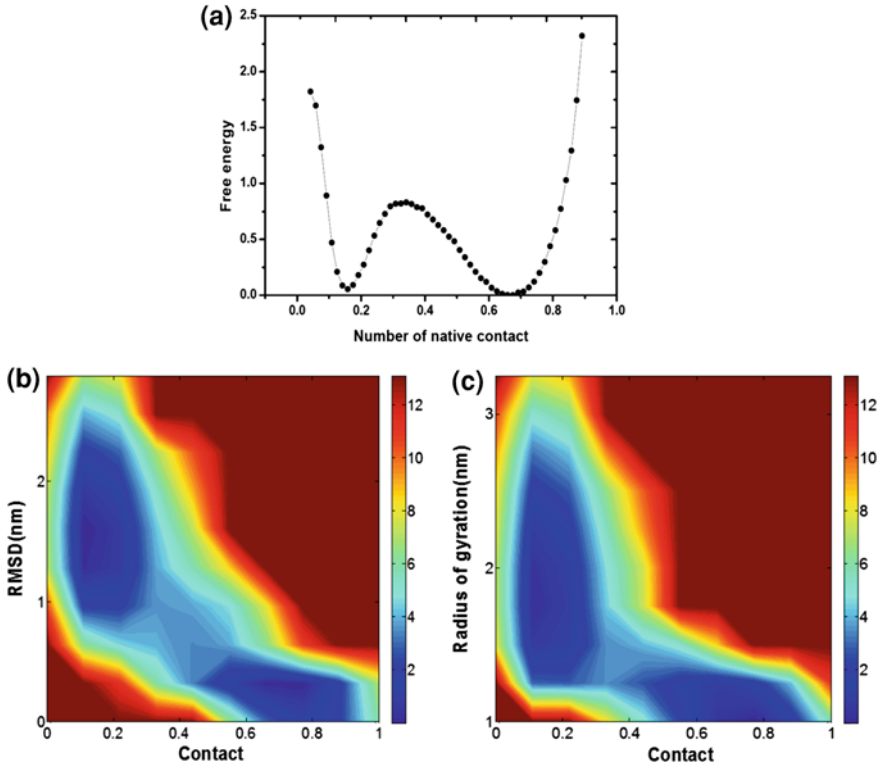


Fig. 10.4 Folding routes for Form 1 G-quadruplex. **a** The free energy as a function of the number of native contact. **b–g** Free energy as a function of two coordinates at folding temperature: T = 105

distance between 5′-end and 3′-end is shown in Fig. 10.4d. In this plot, the distance between 5′-end and 3′-end was decreased sharply in the initial stage of the folding process, and the closeness between the two ends made the unfolded state have a compact structure. After the formation of the compact structure with the number of native contact ∼0.3 the distance between the two ends decreased little further. The transition state with the distance between 5′-end and 3′-end of ∼2 nm separates the two regions of the folded and denatured states. The distance between the two ends in the folded state was lower than 2 nm, and the distance in denatured state was higher than 2.5 nm. More detailed about the folding mechanism can be derived from the free energy landscape of a function of the number of native contact and RCO parameter (Fig. 10.4e). In the folding process, the RCO increased with an increasing number of native contact, but this plot didn't show the sharply increasing RCO. The transition state with the RCO of ∼0.1 separates the folded and denatured states. The RCO in the folded state was higher than 0.2, and the number of native contact was higher than 0.6. The denatured state had the RCO of ∼0.01 and the number of native contact ∼0.2. In the denatured state, the RCO increased with an increasing the

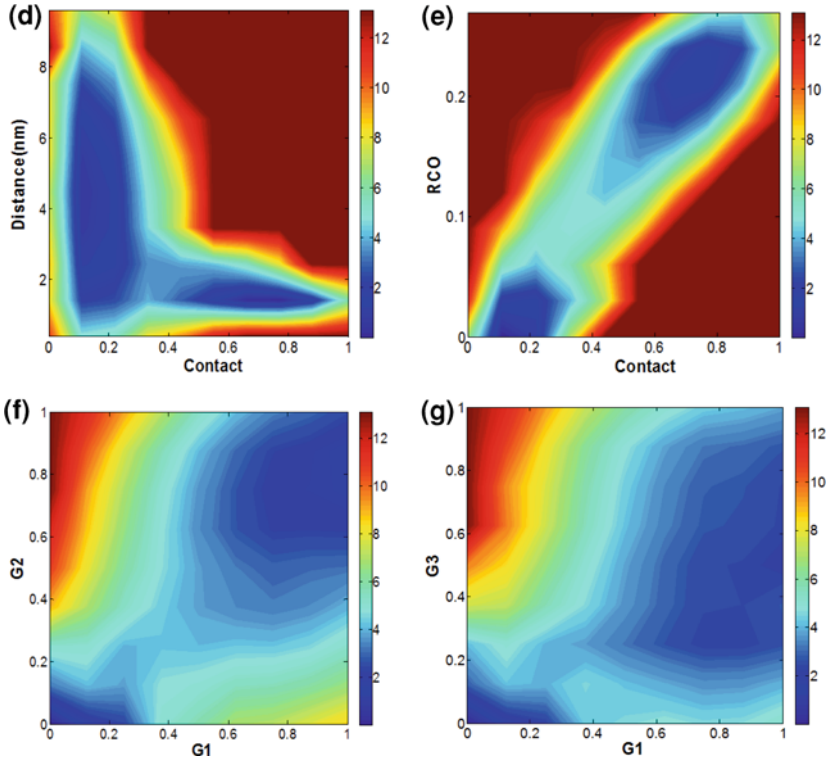


Fig. 10.4 (continued)

number of native contact, so the local native contacts had the main contribution for the formation of the compact structure in the initial stage of the folding process. After the compact structure formed, the RCO was increased because of the increasing non-local native contacts. Hence the local native contacts promoted the formation of the compact structure like thrombin aptamer G-quadruplex, and the non-local native contacts further contributed to the formation of the native structure.

The folding process with the typical structures is shown in Fig. 10.6b. The starting structure was unfolded. In the initial stage of the folding process a compact structure was formed. The native contacts in the 3′-end were formed firstly, and the native contacts G17-G21, G16-G20, and G15-G23 were appeared. The loop T18-T19-A20 had fewer native contacts in the denatured state, and the loops in the regions T6-T7-A8 and T12-T13-A14 were formed in the denatured state and stable in the folding process. The native contacts in the loop T18-T19-A20 were increased with the formation of the G-triplex at 3′-end. After forming the native contacts at the 3′-end, the native contacts between 5′-end and 3′-end were formed, and the compact structure was appeared. In the process of forming the compact structure, the local native contacts were important, which implied the native contacts in loop regions may have the main contribution for this structure.

The formation of the native contacts in these loops drew the 5′-end and the 3′-end toward each other, so the two ends could have the chance to form the native contacts. The G1, G2 and G3 represent the three G-quartets G3 · G9 · G17 · G21, G4 · G10 · G16 · G22 and G5 · G11 · G15 · G23, respectively. The landscape of the number of native contact of G1 and G2 is shown in Fig. 10.4f, and the landscape for the number of native contact of G1 and G3 is shown in Fig. 10.4g. The number of native contact of G1 increased as the increasing number of native contact of G2, and this change trend is the same as the landscape of the number of native contact of G1 and G3. Hence, the three G-quartets may be formed in the same time. The compact structure had the native contacts at the two ends, and the strand of G9-G10-G11 didn't stack with the other strands. The native contacts G3-G21, G4-G22, G17-G21, G16-G22, and G15-G23 formed a G-triplex structure, and then this structure folded to the native state through stacking the strand G9-G10-G11 with the other three strands.

10.2.6 Form 3 G-quadruplex

The constant temperature simulations at the folding temperature were used for Form 3 G-quadruplex. The free energy landscape as a function of the number of native contact and RMSD is shown in Fig. 10.5b. In the initial stage of the folding process, the RMSD decreased sharply as an increasing number of native contact, but once the number of native contact of ~ 0.3 was fromed the RMSD decreased little further. The transition state with the number of native contact of ~ 0.3 separates two regions of the folded and denatured states, and the folded state had the number of native contact of ~ 0.8. In Fig. 10.5c, the free energy landscape as a function of the number of native contact and radius of gyration of the whole molecule for Form 3 G-quadruplex is shown. The L-shaped landscape implied that the radius of gyration decreased rapidly with an increasing number of native contact, after the number of native contact reached ~ 0.3, the radius of gyration decreased little further. The transition state with the radius of gyration of ~ 1.2 nm separates the folded and denatured states in this landscape. Hence, the unfolded state of this G-quadruplex formed a compact structure in the initial stage of the folding process, and then through adjusting this conformation to fold into the native state. The landscape as a function of the number of native contact and the distance between 5′-end and 3′-end at the folding temperature is presented in Fig. 10.5d. In this L-shaped landscape, the distance between the two ends decreased sharply with an increasing number of native contact, after the number of native contact got up to ~ 0.3 the distance decreased little further. The transition state with the distance between the two ends of ~ 2 nm separates folded and denatured states in this L-shaped landscape. In the initial stage of the folding process, the 5′-end was closed to 3′-end for the formation of a compact structure. The Fig. 10.5e presents the free energy landscape as a function of the number of native contact and RCO parameter at the folding temperature. The RCO increased sharply with an increasing number of native contact

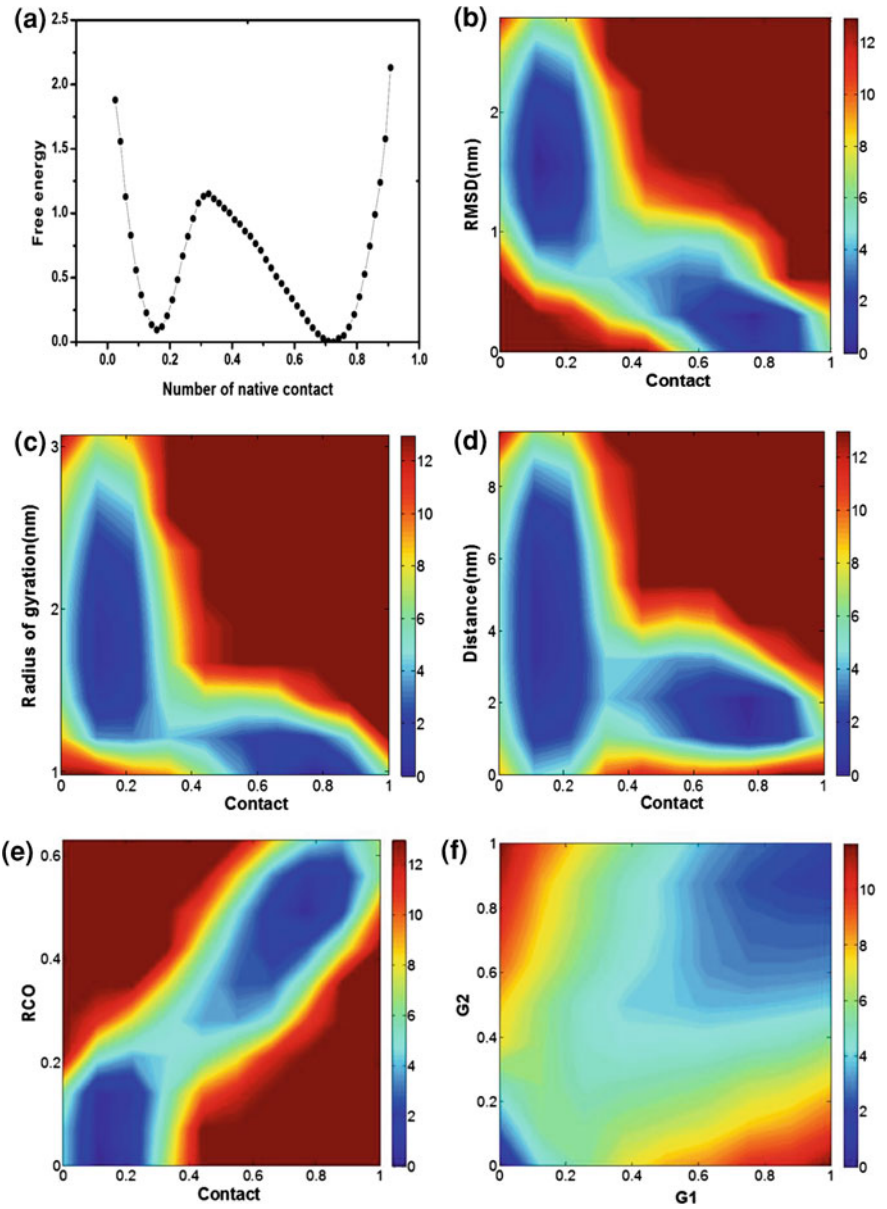


Fig. 10.5 Folding routes for Form 3 G-quadruplex. **a** Free energy as a function of the number of native contact is plotted at folding temperature: T = 105. **b–f** The free energy landscapes are shown as a function of various quantities

until the number of native contact reached ∼0.3 in the initial stage of the folding process. The transition state with the RCO of ∼0.25 divided the two regions of folded and denatured states in this landscape. In the initial stage of the folding

process, the compact structure was formed, so the local native contacts had the main contribution for this process. The RCO increased sharply in the denatured state, so the contribution for the formation of the compact structure can't eliminate the non-local native contacts. These non-local native contacts may be mainly come from the native contacts between the 5′-end and one strand.

The snapshots for the folding process are shown in Fig. 10.6c. The starting structure was unfolded. The compact structure was formed in the initial stage of the folding process like Form 1 G-quadruplex. The free energy landscape as a function of the number of native contact of G1 (G1 · G8 · G20 · G14) and G2 (G2 · G7 · G19 · G15) is shown in Fig. 10.5f. The number of native contact of G1 increased as an increasing number of native contact of G2. So the two G-quartets may be formed in the same time. The native contacts in loop G9-T10-T11-A12-G13 were formed in the denatured state with high probability. The loops G3-T4-T5-A6 and T16-T17-A18 formed fewer native contacts comparing with the loop G9-T10-T11-A12-G13. The formed native contacts especially the native contacts in loop G9-T10-T11-A12-G13 contributed to the formation of the compact structure in the initial

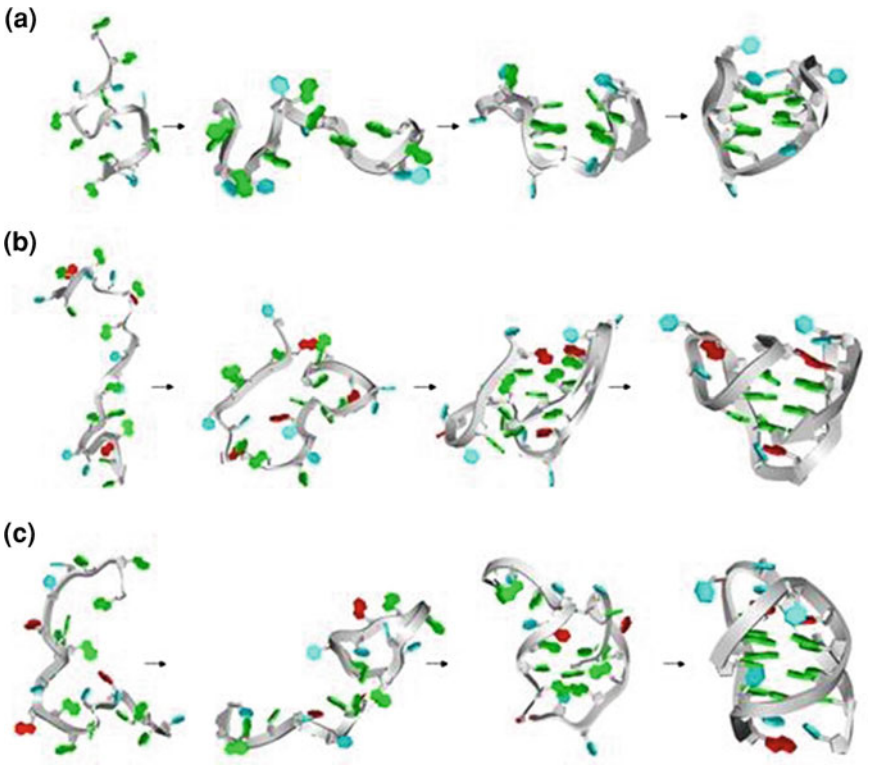


Fig. 10.6 The folding pathways for **a** thrombin aptamer, **b** Form 1 G-quadruplex and **c** Form 3 G-quadruplex. The unfolded G-quadruplexes fold into the compact structures, and then through folding into the G-triplex structures to form the native structures

stage of the folding process. The native contacts G1-G14 and G2-G15 constituted the compact structure, and the native contacts in loop G9-T10-T11-A12-G13 could stabilize this structure. After forming the compact structure, the native contacts G1-G8 and G2-G7 were formed at the 5'-end. Hence, in this stage the G-triplex was formed by native contacts G1-G8, G2-G7, G1-G14 and G2-G15. The 3′-end of the Form 3 G-quadruplex was loose comparing with the other three strands. After the 3′-end stacking with the other three strands, the native structure was formed.

10.2.7 Comparing the Folding Mechanisms of Thrombin Aptamer, Form 1 and Form 3 G-quadruplexes

The thrombin aptamer has two G-quartets like the Form 3 G-quadruplex. The Form 1 G-quadruplex has three G-quartets and folds to the (3 + 1) G-quadruplex. The Form 3 G-quadruplex only has three different bases in the 5′ and 3′ ends comparing with Form 1 G-quadruplex but folds to the basket form. Though the structures of the three G-quadruplexes are obviously different, the folding processes for the three G-quadruplexes are similar. A compact structure formed firstly, and then the G-triplex structure was appeared for folding into the native state. The compact structure of the thrombin aptamer was formed by the native contacts in the two ends and the loop between the two ends. The compact structure was formed by the native contacts between the two ends for Form 1 G-quadruplex. The Form 3 G-quadruplex had the compact structure formed by the native contacts between one strand and the 5′-end. The sequences alignment for the three G-quadruplexes is shown in Fig. 10.1d. In the folding process of thrombin aptamer, the native contact between T4 and T13 stabilized the G-triplex, and comparing with the Form 3 G-quadruplex the native contact T5-T17 stabilized the loops at the two ends. The native contact G3-T5 in the G-triplex of Form 3 determined the loop, and the native contact between G9 and T11 gave the contribution for the formation of the loop. The native contacts G1-G14 and G2-G15 were formed in the G-triplex of Form 3, but these native contacts were formed at the last stage in the folding process of thrombin aptamer. The native contacts in loop T7-G8-T9 were formed and stable in denatured state, and the native contacts in the loops of the two ends were existent, so these loops were determined in the initial stage of the folding process of thrombin aptamer for defining the stacking way of G-quartets, which was different from Form 3. Some of the native contacts in loops G9-T10-T11-A12-G13 and T16-T17-A18 of Form 3 were formed in the denatured state, and most of the native contacts in these loops were formed in folded state. These native contacts made the 3′-end and one strand in a close distance, so the compact structure of Form 3 could be formed. Hence, the local native contacts in these loops have made a difference between thrombin aptamer and Form 3 for the stacking ways of G-quartets in the denature state. The native contacts in loop T18-T19-A20 of Form 1 G-quadruplex were appeared, and then

these local native contacts promoted the formation of the native contacts in the 3′-end. The native contacts in loop G9-T10-T11-A12-G13 of Form 3 for stabilizing this region were higher than the number of the native contact in the corresponding loop T12-T13-A14 of Form 1, and these native contacts in Form 3 were formed and stable in the denatured state. The formed native contacts in the loops of Form 1 made the two ends stack together, however the stable native contacts in the loop G9-T10-T11-A12-G13 of Form 3 promoted the stack between the 3'-end and one strand of G-triplex. So the local native contacts formed in the denatured state may impact the folding pathways of G-quadruplexes. The native contacts T5-T17, G3-A18, G3-A6, and G9-G13 in Form 3 stabilized the G-triplex structure, and the Form 1 had the native contacts A2-T19 and A2-A20 to stabilize this G-triplex structure. The folding free energy barrier of Form 1 G-quadruplex was $\sim$0.83 kbT, but the Form 3 G-quadruplex had the free energy barrier higher than 1.15 kbT. Hence, the Form 3 needs more free energy than the Form 1 for folding into the native state, and the Form 3 is more stable than the Form 1 G-quadruplex. This result is consistent with experimental studies. In experiment, the Form 3 G-quadruplex with basket-type fold had high structural stability, because of the base pairing and the stacking in the loops such as G21 · G9 · G13, T21 · T11, A6 · G3 · A18, and T5 · T17 [17]. Here we found the thrombin aptamer with free energy barrier of $\sim$0.82 kbT as the Form 1 G-quadruplex may have lower stability than the Form 3.

10.3 Conclusions

We have simulated thrombin aptamer, form 1 and form 3 G-quadruplexes with all-atom Gō-model for studying the folding mechanisms for the three G-quadruplexes. The folding processes of the three G-quadruplexes are similar. The compact structures were formed in the initial stage of the folding process. The thrombin aptamer had the compact structure through forming the native contacts in the two ends and the medium loop. The native contacts in the loops of Form 1 had main contribution for the formation of the compact structure. The Form 3 had the native contacts formed between the 5′-end and one strand in order to obtain the compact structure. The G-triplex structures were formed before folding to the native states of the three G-quadruplexes. The G-triplex of thrombin aptamer is consisted of the native contacts G5-G11, G6-G10, G2-G5, G1-G6, G11-G14, and G10-G15. The G-triplex of Form 1 is composed of the native contacts G3-G21, G4-G22, G17-G21, G16-G22, and G15-G23. The G-triplex comprises the native contacts G1-G14, G2-G15, G1-G8 and G2-G7 in Form 3 G-quadruplexes. The Form 3 has higher free energy barrier than the other two G-quadruplex structures, and this structure has more structural stability.

10.4 Materials and Methods

All-Atom Gō-model. The all-atom Gō-model was described previously [34]. This model is available on a web server [38]. In the all-atom Gō-model, all heavy (non-hydrogen) atoms are explicitly included. A single bead of unit mass represents each atom. The harmonic potentials were used for restraining the bond length and angles, and planar dihedrals. The non-bonded atom pairs that are in contact in the native state, are given attractive 6–12 interactions. Nevertheless, all the other non-local interactions are repulsive. Gromacs 4.0.7 software package was used for all simulations [39]. The simulations were started from the unfolded structures. For obtaining the thermodynamic sampling, more than 30 simulation trajectories were performed for the three G-quadruplexes at folding temperature. The Weighted Histogram Analysis Method was used for the calculation of the thermodynamic quantities [40].

Reaction Coordinates. We used the fraction of native residues in contact as the reaction coordinate Q. A native contact is defined as any two atoms in different residues that are within 4 Å of each other and separated by at least 3 bonds [34]. A contact between two atoms is formed if this pair distance is within the 1.2 times their native distance.

References

1. De Cian A, Lacroix L, Douarre C, Temime-Smaali N, Trentesaux C, Riou JF, Mergny JL (2008) Targeting telomeres and telomerase. Biochimie 90:131–155
2. Neidle S, Parkinson GN (2003) The structure of telomeric DNA. Curr Opin Struct Biol 13:275–283
3. Bryan TM, Cech TR (1999) Telomerase and the maintenance of chromosome ends. Curr Opin Cell Biol 11:318–324
4. Lian P, Liu LA, Shi Y, Bu Y, Wei D (2010) Tethered-hopping model for protein-DNA binding and unbinding based on Sox2-Oct1-Hoxb1 ternary complex simulations. Biophys J 98:1285–1293
5. Hanahan D, Weinberg RA (2000) The hallmarks of cancer. Cell 100:57–70
6. Makarov VL, Hirose Y, Langmore JP (1997) Long G tails at both ends of human chromosomes suggest a C strand degradation mechanism for telomere shortening. Cell 88:657–666
7. Gellert MN, Lipsett MN, Davies DR (1962) Helix formation by guanylic acid. Proc Natl Acad Sci USA 48:2013–2018
8. Davis JT (2004) G-quartets 40 years later: from 50-GMP to molecular biology and supramolecular chemistry. Angew Chem Int Ed Engl 43:668–698
9. Wang Y, Patel DJ (1993) Solution structure of the human telomeric repeat d[AG3(T2AG3)3] G-tetraplex. Structure 1:263–282
10. Schaffitzel DL, Berger I, Postberg J, Hanes J, Lipps HJ, Plückthun A (2001) In vitro generated antibodies specific for telomeric guanine-quadruplex DNA react with Stylonychia lemnae macronuclei. Proc Natl Acad Sci USA 98:8572–8577
11. Maizels N (2006) Dynamic roles for G4 DNA in the biology of eukaryotic cells. Nat Struct Mol Biol 13:1055–1059

12. Parkinson GN, Lee MP, Neidle S (2002) Crystal structure of parallel quadruplexes from human telomeric DNA. Nature 417:876–880
13. Phillips K, Dauter Z, Murchie AI, Lilley DM, Luisi B (1997) The crystal structure of a parallel-stranded guanine tetraplex at 0.95 Å resolution. J Mol Biol 273:171–182
14. Schultze P, Macaya RF, Feigon J (1994) Three-dimensional solution structure of the thrombin-binding DNA aptamer d(GGTTGGTGTGGTTGG). J Mol Biol 235:1532–1547
15. Phan AT, Kuryavyi V, Luu KN, Patel DJ (2007) Structure of two intramolecular G-quadruplexes formed by natural human telomere sequences in K+ solution. Nucleic Acids Res 35:6517–6525
16. Luu KN, Phan AT, Kuryavyi V, Lacroix L, Patel DJ (2006) Structure of the human telomere in K+ solution: an intramolecular (3 + 1) G-quadruplex scaffold. J Am Chem Soc 128:9963–9970
17. Lim KW, Amrane S, Bouaziz S, Xu W, Mu Y, Patel DJ, Luu KN, Phan AT (2009) Structure of the human telomere in K+ solution: a stable basket-type G-quadruplex with only two G-tetrad layers. J Am Chem Soc 131:4301–4309
18. Gray RD, Chaires JB (2008) Kinetics and mechanism of K+- and Na+-induced folding of models of human telomeric DNA into G-quadruplex structures. Nucleic Acids Res 36:4191–4203
19. Xu Y, Sato H, Sannohe Y, Shinohara K, Sugiyama H (2008) Stable lariat formation based on a G-quadruplex scaffold. J Am Chem Soc 130:16470–16471
20. Mashimo T, Yagi H, Sannohe Y, Rajendran A, Sugiyama H (2010) Folding pathways of human telomeric type-1 and type-2 G-quadruplex structures. J Am Chem Soc 132:14910–14918
21. Kim E, Yang C, Pak Y (2012) Free-energy landscape of a thrombin-binding DNA aptamer in aqueous environment. J Chem Theory Comput 8:4845–4851
22. Gosavi S, Chavez LL, Jennings PA, Onuchic JN (2006) Topological frustration and the folding of interleukin-1 beta. J Mol Biol 357:986–996
23. Go N (1983) Theoretical studies of protein folding. Annu Rev Biophys Bioeng 12:183–210
24. Nymeyer H, Garcia AE, Onuchic JN (1998) Folding funnels and frustration in off-lattice minimalist protein landscapes. Proc Natl Acad Sci USA 95:5921–5928
25. Koga N, Takada S (2001) Roles of native topology and chain-length scaling in protein folding: a simulation study with a go-like model. J Mol Biol 313:171–180
26. Bryngelson JD, Wolynes PG (1987) Spin glasses and the statistical mechanics of protein folding. Proc Natl Acad Sci USA 84:7524–7528
27. Leopold PE, Montal M, Onuchic JN (1992) Protein folding funnels—a kinetic approach to the sequence structure relationship. Proc Natl Acad Sci USA 89:8721–8725
28. Onuchic JN, Wolynes PG (2004) Theory of protein folding. Curr Opin Struct Biol 14:70–75
29. Li L, Chen Q, Wei DQ (2012) Prediction and functional analysis of single nucleotide polymorphisms. Curr Drug Metab 13:1012–1023
30. Wei DQ (2012) New drug design based on multi-targets and system biology approach in light of real time DNA sequencing technologies. Curr Top Med Chem 12:1309
31. Xiong Y, Liu J, Wei DQ (2011) An accurate feature-based method for identifying DNA-binding residues on protein surfaces. Proteins 79:509–517
32. Sorin EJ, Nakatani BJ, Rhee YM, Jayachandran G, Vishal V, Pande VS (2004) Does native state topology determine the RNA folding mechanism? J Mol Biol 337:789–797
33. Hyeon C, Dima RI, Thirumalai D (2006) Pathways and kinetic barriers in mechanical unfolding and refolding of RNA and proteins. Structure 14:1633–1645
34. Whitford PC, Schug A, Saunders J, Hennelly SP, Onuchic JN, Sanbonmatsu KY (2009) Nonlocal helix formation is key to understanding S-adenosylmethionine-1 riboswitch function. Biophysical J 96:L7–L9
35. Kouza M, Hansmann UH (2012) Folding simulations of the A and B domains of protein G. J Phys Chem B 116:6645–6653
36. Plaxco KW, Simon KT, Baker D (1998) Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 277:985–994

37. Mao XA, Gmeiner WH (2005) NMR study of the folding-unfolding mechanism for the thrombin-binding DNA aptamer d(GGTTGGTGTGGTTGG). Biophys Chem 113:155–160
38. Noel JK, Whitford PC, Sanbonmatsu KY, Onuchic JN (2010) SMOG@ctbp: simplified deployment of structure-based models in GROMACS. Nucleic Acids Res doi:10.1093/nar/gkq498
39. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. J Chem Theory Comput 4:435–447
40. Kumar S, Bouzida D, Swendsen RH, Kollman PA, Rosenberg JM (1992) The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. J Comput Chem 13:1011–1021

Chapter 11
RNA Folding: Structure Prediction, Folding Kinetics and Ion Electrostatics

Zhijie Tan, Wenbing Zhang, Yazhou Shi and Fenghua Wang

Abstract Beyond the "traditional" functions such as gene storage, transport and protein synthesis, recent discoveries reveal that RNAs have important "new" biological functions including the RNA silence and gene regulation of riboswitch. Such functions of noncoding RNAs are strongly coupled to the RNA structures and proper structure change, which naturally leads to the RNA folding problem including structure prediction and folding kinetics. Due to the polyanionic nature of RNAs, RNA folding structure, stability and kinetics are strongly coupled to the ion condition of solution. The main focus of this chapter is to review the recent progress in the three major aspects in RNA folding problem: structure prediction, folding kinetics and ion electrostatics. This chapter will introduce both the recent experimental and theoretical progress, while emphasize the theoretical modelling on the three aspects in RNA folding.

Keywords RNA folding · Structure · Kinetics · Metal ions · Thermodynamics

11.1 Introduction

Recent discoveries reveal the new biological functions of non-coding RNAs such as RNA silence and riboswitch. The functions of non-coding RNAs are intrinsic to RNA structures and stability, and can provide potential RNA-based therapeutic strategies. This demands the quantitative understanding and prediction on RNA

Z. Tan (✉) · W. Zhang (✉) · Y. Shi · F. Wang
Department of Physics and Key Laboratory of Artificial Micro- and Nano-Structures of Ministry of Education, School of Physics and Technology, Wuhan University, Wuhan, China
e-mail: zjtan@whu.edu.cn

W. Zhang
e-mail: wbzhang@whu.edu.cn

D. Wei et al. (eds.), *Advance in Structural Bioinformatics*, Advances in Experimental Medicine and Biology 827, DOI 10.1007/978-94-017-9245-5_11

structure and its stability, i.e., RNA folding problem. RNA structural folding is driven by the intra-molecular forces, such as base pairing/stacking interactions, ion-mediated interactions and the conformational entropies [1, 2]. This chapter is focusing on the three aspects in RNA folding: structure prediction, folding kinetics and ion-mediated electrostatic interactions, respectively.

First, RNA structure prediction is one of central issues in RNA folding problem, since RNA structures involve not only their biological functions such as gene regulation, but also the interactions with other molecules which can provide the potential therapeutic strategies. Generally, RNA folding is hierarchical, i.e., the secondary structure can be formed firstly driven by strong base pairing/stacking interactions, and afterwards, the tertiary structure can be folded by the aggregation of secondary segments and the formation of tertiary contacts [3]. Since the base pairing/stacking interactions are very strong, the secondary structure of RNAs can be relatively stable. Accordingly, the RNA structure prediction can be classified into the secondary structure prediction and tertiary structure prediction. From 1980s, many efforts have been made on the secondary structure prediction based on the experimentally derived parameters and great progress has been made in accurate predictions on RNA secondary structure [4–8]. However, RNAs are often biological functional in their 3-dimensional structures. The lack of experimentally derived structures and the high cost for experimentally determining structures have enabled the computational modelling for predicting RNA structures [9–38]. The tertiary (including 3-dimensional) structure predictions have attracted much attention and important progress has been achieved in recent years, which will be focused on in the first part of this chapter.

Second, in addition to static structures, RNA folding kinetics is also directly tied to RNA biological functions. Experiments suggest that alternative conformations of the same RNA sequence perform different functions [39–41]. The capability of RNA molecules to form multiple (metastable) conformations for different functions is probably used by nature to regulate versatile functions of RNA. Furthermore, it was found that the folding of the functional structures is controlled by folding kinetics rather than by equilibrium thermodynamics. The mechanisms of ribozyme [42, 43], anti-HIV RNA aptamers [44–46], gene expression regulators such as miRNA, siRNA and riboswitches [47–53] and other RNAs are often kinetically controlled. For instance, self-induced riboswitches regulate RNA functions by limiting biologically functional properties of RNA structures to certain time windows. The hok/sok system of plasmid R1 [54, 55] regulates the plasmid maintenance through mRNA conformation rearrangements into different functional forms. For riboswitches, it has been proposed that the transient intermediate structure of RNA can regulate transcription and translation by creating a time window that is necessary for regulatory reactions to occur [56]. After the early work of Porschke [57] and Crothers [58, 59], extensive kinetic experiments, such as temperature-jump, single molecule and time resolved NMR spectroscopy experiments, have been employed to study the RNA [60–64] and DNA [65–67] folding kinetics. The recent progresses in understanding RNA/DNA folding kinetics will become the second part of the chapter.

Third, due to the polyanionic nature of RNAs, RNA folding causes massive build-up of the negative charges [68–73] and a strong intra-chain Coulombic repulsion. However, the folding would attract metal ions in solution and cause significant ion binding to RNA, to effectively reduce the electrostatic energy barrier and stabilize a folded RNA structure. Therefore, RNA folding including folded structure, folding kinetics, and stability are strongly coupled to ion electrostatic interactions [1, 68–73]. The third part of the chapter will focus on the recent progress in qualitative/quantitative understanding on ion roles in RNA folding.

In the following, we will review the recent progress in RNA folding, in the three aspects including structure prediction, folding kinetics, and ion electrostatics.

11.2 RNA Structure Prediction

As described above, the RNA structure prediction can be divided into two levels: secondary structure prediction and tertiary structure prediction. On predicting RNA secondary structure, many computational models have been proposed, based on the experimental thermodynamic parameters [4–8], such as Mfold through free energy minimization method [4], Vienna RNA package with dynamic programming algorithm [5], Sfold through sampling structures with Boltzmann statistics [6], etc.; see Ref. [8] for a recent review on RNA secondary structure prediction. In the following, we will focus on the RNA tertiary structure prediction methods which can be classified into three types: knowledge-based structure modelling, physics-based structure modelling, and knowledge/physics-hybridized structure modelling; see Table 11.1 for a summary on the algorithms for RNA tertiary structure prediction [9–38]. Other reviews are also available [74–79].

11.2.1 Knowledge-Based Structure Prediction

With the rapid increase in the RNA structure data deposited in protein data bank (PDB) and nucleic acid data bank (NDB), the knowledge-based modelling is becoming an important method for predicting RNA structure based on available sequences. The knowledge-based modelling relies on the database of experimentally solved structures and empirically observed structural similarities between the same (similar) sequences.

11.2.1.1 Graphics-Based Method

One kind of knowledge-based modelling is Graphics-based method, which usually involves interactive (user-guided) manipulation of RNA structures based on the assembly of fragments derived from various experimental structures (motifs),

Table 11.1 The algorithms for RNA three-dimensional/tertiary structure prediction

Algorithms and references	Input	Simulation method	Classification
MANIP [9]	Alignment of sequences, 2D structure	Graphical user interface	Graphics-based
S2S/Assemble [11, 12]	Database of known fragment, 2D structure	Graphical user interface	Graphics-based
RNA2D3D [13]	2D structure	Graphical user interface	Graphics-based
RNABuilder [15]	2D structure, tertiary contacts	Scripting interface, MD	Homology-based
ModeRNA [14]	Sequence, 3D structure of template	Scripting interface	Homology-based
YUP [17]	2D structure, tertiary contacts	MD	Physics-based
NAST [18, 19]	2D structure, tertiary contacts	MD	Physics-based
Vfold [21–23]	Sequence	Statistical mechanical calculation, MD	Physics-based
iFold/DMD [24, 25]	Sequence	Replica exchange DMD	Physics-based
FARNA/FARFAR [31, 32]	Sequence, 2D structure constrains	Fragment assembly, MC	Hybrid method
MC-Fold/MC-Sym [34]	Sequence, 2D structure	Fragment assembly, Las Vegas algorithm	Hybrid method

2D secondary; *3D* three-dimensional; *MD* molecular dynamics simulation; *MC* Monte Carlo simulation; *DMD* discrete molecular dynamics simulation

including the algorithms of MANIP [9], S2S/Assemble [11, 12], and RNA2D3D [13].

MANIP. Massire and Westhof have developed a program MANIP [9], which allows the rapid assembly of isolated motifs (each with a specified sequence) into a complex three-dimensional architecture by users. As an interactive tool, MANIP has a toolbox, where the user can find a variety of tools that help to design a three-dimensional structure model. MANIP constitutes a quick and easy way to model small- to large-size structured RNAs, and the use of multiple connections and pairing tables opens the further development perspectives and allows, for instance, the precise modelling of RNA-protein interactions.

S2S/Assemble. Jossinet et al. [11] developed a program S2S (sequence to structure), in which a user can conveniently display, manipulate and interconnect heterogeneous RNA data. Assemble [12], an algorithm complementary to S2S, is an intuitive graphical interface to analyze, manipulate and build complex 3D RNA architectures. S2S/Assemble is a system that combines various tools and web services into a powerful package to edit sequences and structures of RNA. It contains explicit annotation of base pairing and stacking interactions, multiple sequence alignments, a motif library and an automatic procedure to generate 3D models from the annotation. But all interactions have to be annotation manually, and thus it is difficult to perform a high-throughput analysis.

RNA2D3D. With the use of the primary sequence and secondary structure information of an RNA, the program RNA2D3D [13] automatically and rapidly produces a 3-dimensional conformation (the initial) consistent with the available information. At the next step, the overlaps in the initial 3D structure model are removed and conformational changes are made aiming to the targeted features. Subsequently, the refinement needs to be performed by the user through its interactive graphical editing and the special tools such as the compacting, stem-stacking and segment-positioning energy-refinement. The most important advantage of RNA2D3D is that it is applicable to structures of arbitrary branches and pseudoknots. The algorithm has been verified in the modelling of ribozymes, viral kissing loops, and viral internal ribosome entry sites.

Obviously, all these methods are not an automatic algorithm. The graphics-based modelling requires users to set up and refine the RNA structures according to the specific principles, thus requires users' expert knowledge [9–13].

11.2.1.2 Homology-Based Modelling

Another kind of knowledge-based modelling is the homology-based modelling, i.e., comparative modelling, based on the empirical observation that evolutionarily related macromolecules usually retain similar 3D structure despite the divergence on the sequence level [81]. Several algorithms have been developed based on the homology-based modelling, such as ModeRNA [14] and RNABuilder [15].

ModeRNA. As a minimal input, ModeRNA [14] requires the 3D coordinates of template structures and a pairwise sequence alignment between the sequences of

the template and the RNA to be modeled. The ModeRNA provides a flexible scripting framework that can build RNA structures with various strategies, including the fast automated modelling based on template structure and target–template alignment without additional data. The ModeRNA was tested by 99 tRNAs with known structures (experimentally solved and each of them as a target to be modelled on each of the other 98 structures as templates) with RMSD values around 5.0 Å.

RNABuilder. Recently, Flores et al. have developed RNABuilder [15] now known as MMB (a contraction of MacroMolecule Builder), for comparative modelling of RNA structures. It generates RNA structures by treating the kinematics and forces at separate. The coarse-graining force field for an alignment used in this approach consists of forces and torques which act to bring the interacting bases into the base pairing geometry specified by the user. RNABuilder has been used to predict the structure of the 200-nucleotide Azoarcus group I intron in the absence of any information of the solved Azoarcus intron crystal structure. The model accurately depicts the global topology, secondary and tertiary connections, and gives an overall RMSD value of 4.6 Å relative to the crystal structure.

Homology-based modelling can be used to predict any RNA molecules no matter how large or small, as long as the user can find a template and an effective alignment between the template and the target [14, 15, 79]. So this method is also called template-based modelling. Although the PDB/NDB database covers many important families, it may be difficult to find a proper template RNA for a particular target. In addition, creating an accurate and biologically relevant target–template sequence alignment is also a critical issue [79, 80].

11.2.2 Physics-Based Structure Prediction

Physics-based (ab initio) approach is based on the thermodynamics hypothesis [82]: the conformation with the lowest free energy corresponds to the native structure. Since a full-atomic structural model of RNA has a large number of degrees of freedom, which results in the huge computational complexity. For physical simplification, several prediction models with coarse-graining have been proposed at different resolution levels [16].

11.2.2.1 One-Bead Coarse-Grained Model

One-bead model uses one bead to represent a nucleotide, thus significantly reduce the spatial freedoms of an RNA structure. Several algorithms have been developed for predicting RNA 3D structures, such as YUP [17] and NAST [18].

YUP. Yammp Under Python (YUP) [17] is a general-purpose molecular mechanics program for multi-scaled coarse-grained modelling, in which Python is used as a programming/scripting language. It can be used to model RNA structures

as well as DNA and protein structures by extending the Python language through adding three new data types (atom maps, atom vectors and numerous energy types). In general, YUP is an extendable and useful tool for multi-scale modelling, but its potentials are required to be changed by the user according to the problem at hand. In addition, a fragment-based approach is used to add full-atomic details to the coarse-grained structure in YUP.

NAST. The Nucleic Acid Simulation Tool (NAST) [18] is a molecular dynamics simulation tool for predicting 3D structure for large RNA molecules based on secondary structures. Three types of data are also used to rank the conformational clusters produced form molecular dynamics simulations: (1) ideal small-angle X-ray scattering (SAXS) data; (2) experimental and ideal solvent accessibility (SAS) data; and (3) NAST energy (statistical information). NAST has been tested by building the structural models for two RNA molecules–the yeast $tRNA^{Phe}$ (76-nt) and the P4-P6 domain of the *Tetrahymena thermophila* group I intron (158-nt), with the averaged RMSD 8.0 ± 0.3 and 16.3 ± 1.0 Å, respectively. Recently, the authors also developed a fully automated frament- and knowledge-based method, called C2A (Coarse to Atomic) [19], to add full atomic details to coarse-grained models.

Both YUP and NAST are successful for large RNA molecules at nucleotide level, but they are limited by their prior need for secondary structure and the information of some tertiary contacts derived from both experimental and computational methods.

11.2.2.2 Three-Bead Coarse-Grained Model

Beyond the one-bead models [17–20], a number of coarse-grained models with higher resolution have been developed, such as three-bead [21, 24], five-bead [27] and six to seven-bead model [28].

Vfold. Cao and Chen have developed a three-vector virtual bond-based RNA folding model (Vfold) [21] for predicting RNA 3D tertiary folds from the sequence without using the experimental constraints. In Vfold, the loop conformations are produced by the self-avoiding random walks of the virtual bonds on a diamond lattice [22, 23] and the conformational entropy of RNA structures can be calculated. The Vfold model has been tested by a systematic benchmark including a wide range of RNA motifs (such as hairpin, duplex), pseudoknot, and a large RNA (a 122-nt 5S rRNA domain) with rmsd of about 3.5, 6.0 and 7.4 Å, respectively. Due to the rigid lattice constraints, Vfold is inadequate to study the folding dynamics of RNA.

iFoldRNA. iFoldRNA [25] is a web-based methodology for RNA 3D structure prediction and analysis of RNA folding thermodynamics. It is based on discrete molecular dynamics (DMD) and a force field (including base-pairing, base-stacking and loop entropy) [24]. The ifoldRNA has been tested by simulating a set of 153 RNA molecules within an average 4 Å deviation from experimental structures. Despite its rapid conformational sampling efficiency, the CPU time for DMD simulations also depends on RNA length. Ding et al. recently reported the

development of a qualitatively structure refinement approach using hydroxyl radical probing (HRP) measurements to drive DMD simulations for large RNA molecules (80 $\sim$ 230 nt) with complex topologies [26].

The physical-based approaches not only emphasize the necessity of an accurate understanding of RNA tertiary structure but also illustrate the importance of native state dynamics. Although there are many models have been applied [16–28], how to build and choose proper force-fields is still a challenge.

11.2.3 Knowledge/Physics-Hybridized Structure Prediction

In protein structure prediction, the most successful approach is hybrid (*de nove*) modelling which combines the features of physics-based folding with the use of previously solved structures [83–85]. This hybrid (*de nove*) modelling strongly relies on the structural information from databases [79], and based on the principle, there are some existed programs for RNA structure prediction [29–34].

FARNA/FARFAR. Fragment assembly of RNA (FARNA) [31] is developed to predict RNA 3D structure from a sequence, while fragment assembly of RNA with full-atom refinement (FARFAR) [32] adds a refinement with atomic-level interactions to optimize RNA structures generated by FARNA. Based on knowledge-based energy function, FARNA can assemble three-nucleotide all-atom fragments with Monte Carlo algorithm. In a benchmark test of 20 RNA molecules ($\leq$46 nt), FARNA reproduces better than 90 % of Watson-Crick base pairs. Smaller RNAs in the test are accurately reproduced with a resolution of better than 4.00 Å, but the probability of a FARNA prediction within a backbone rmsd of 4.00 Å decreases sharply as a function of RNA length. Nevertheless, combined with secondary structure and multiplexed hydroxyl radical cleavage analysis (MOHCA), FARNA can predict the structure for an RNA as large as 158 nt with the rmsd of 13 Å [32].

MC-Fold/MC-Sym. MC-Fold/MC-Sym pipeline [34] is another full-atomic RNA 3D structure prediction algorithm, which assembles RNA structures from a library of the nucleotide cyclic motif (NCM) [35]. MC-Fold predicts RNA secondary structure using a free energy minimization function, and MC-Sym builds full-atom 3D models of RNA structures based on the scripts generated by MC-Fold and 3D version of the NCM fragments. The predictive power of the pipeline has been confirmed by building 3D structures of precursor microRNA (pre-miRNA), and proposing a new 3D structure of the human immunodeficiency virus (HIV-1) cis-acting 21 frame-shifting element.

Knowledge/physics-hybridized modelling including FARNA/FARFAR and MC-Fold/MC-Sym pipeline is powerful for modelling small RNA molecules [29–32], but larger structures remain a challenge because of the computational requirements for full-atomic modelling. A coarse-grained approach would decrease computational requirements for modelling large structures.

11.3 RNA Folding Kinetics

11.3.1 Kinetic Model

Most approaches to kinetic RNA folding are based on the description of folding in terms of a stochastic process. Each model consists of three key ingredients: (1) The state space, i.e., the set of structures or conformations, (2) a move-set, i.e., the elementary transitions that can occur between such conformations, and (3) transition rates for each of these allowed transitions. The folding process can now be described as a continuous time Markov process, governed by a master equation for the state probabilities.

Consider an ensemble of conformational states. The population $p_i(t)$ for each state i at time t can be described by the following equation (master equation):

$$dp_i/dt = \sum\nolimits_{\Omega}[k_{j\to i}p_j(t) - k_{i\to j}p_i(t)],$$

where Ω is the total number of conformations. $k_{j\to i}$ and $k_{i\to j}$ are the rate of the respective transitions, and they should satisfy the detailed balance condition: $p_i k_{i\to j} = p_j k_{j\to i}$, where p_i and p_j are the Boltzmann distribution of state i and j, $p_i = \frac{1}{Z}\exp(-\frac{\Delta G_i}{k_B T})$, and Z is the partition function $Z = \sum_i \exp(-\frac{\Delta G_i}{k_B T})$.

When Ω is not very large, the above rate equation can be written as the matrix form [86]: $d\mathbf{p}/dt = \mathbf{M}\cdot\mathbf{p}$, where $\mathbf{p}$ is the vector for the population distribution, $\mathbf{M}$ is the $\Omega \times \Omega$ rate matrix with the matrix elements defined by $M_{ij} = k_{j\to i}$ $(i \neq j)$ and $M_{ii} = -\sum_{j\neq i} k_{i\to j}$. The equation can be solved with analytical form and the population kinetics is given by the eigenvalue spectrum for long times:

$$p(t) = \sum_{m=1}^{\Omega} C_m n_m e^{-\lambda_m t}, \tag{11.1}$$

where $-\lambda_m$ and n_m are the m-th eigenvalue and eigenvector of the rate matrix and C_m is the coefficient as determined by the initial condition.

Because the passage of a rate-limiting step is intrinsically related to the folding speed, it is possible to probe and to identify the rate-limiting steps through the folding from different unfolded initial conformations. In a master equation approach, slow and fast folding speeds are directly correlated to the large and small contributions of the rate-limiting slow kinetic modes. Because the contributions from the slow modes can be computed from the corresponding eigenvectors, Zhang and Chen proposed a general transition state searching method to identify the rate-limiting steps from the eigenvectors of the slow modes [87].

When Ω is large, if there exist discrete rate-limiting steps for the kinetic process, it would be possible to "renormalize" the conformational space into a number of conformational clusters. The large ensemble of chain conformations can thus be drastically reduced into a much smaller number of conformational

clusters [88, 89]. Different clusters are separated by the rate-limiting steps. If the rate-limiting steps involve sufficiently high kinetic barrier, the microstates within each cluster would have sufficient time to equilibrate and form a macrostate (in local equilibrium) before crossing the intercluster barriers to enter other kinetically neighboring clusters. The transitions between different clusters (macrostates) determine the overall folding kinetics of the molecule. Otherwise, it needs to simulate the process with Monte Carlo methods [90–93]. Due to the drawbacks such as limited sampling and slow calculation, a few methods have been applied to amend these. For examples, rejectionless Monte-Carlo approach was used to conserve the detailed balance condition [94], the simulated annealing techniques was used in order to accelerate folding [95], optimization techniques such as genetic algorithms rather than Monte-Carlo simulation was also used [96].

11.3.2 Conformation Space

The base pair is the basic subunit for RNA secondary structure, so a base-pair forming/melting corresponds to the smallest possible steps in conformation space [90]. Considering that RNA secondary structure is stabilized mainly by the base stacking interactions, and a single (unstacked) base pair is not stable and can quickly unfold, Zhang and Chen [86] defined an elementary kinetic step for RNA secondary structural change to be the formation/disruption of a stack or a stacked base pair. While this allows the most detailed description of folding pathways, the conformation space is so large that it leads to extremely long simulation runs or restricted to short sequences. To reduce the conformation space, many approaches therefore define the formation or destruction of an entire helix as the basic step [93, 97–100]. Another approach is that several uncorrelated base pairs are changed in a single time step [101]. Folding simulations in this scenario are similar to that of single base pair moves. But the relationship of the helices is different from that of base pairs [102]: the two helices can be compatible, partial compatible, and incompatible. A conformation state should consist of compatible or partial compatible helices [97–100]. Recently, thermodynamics-based RNA folding in a kinetic folding context. Coarse-grained landscapes in conjunction with stochastic sampling algorithms have been used to study the RNA folding kinetics [103]. By using the barrier trees and assuming that the basins of individual local minima are in quasi-equlibrium, the folding kinetics under transcription was studied [104]. Another approach for cotranscription folding combined the thermodynamic computations with coarse-grained local kinetics [105]. Flammn et al. developed a flooding algorithm that decomposes the landscape into basins surrounding local minima connected by saddle points [106]. Wolfinger et al. [107] use a partitioning of the landscape into macrostates, where a macrostate is defined as the set of all starting conformations for which a gradient walk ends in the same local minimum. The effective transition rates between any two macrostates were calculated from the barrier tree. Tang et al. [103, 108] adopt the probabilistic roadmaps to build an

approximated representation of the RNA folding landscape. In the roadmap graph, the vertex set represents valid sampled conformations of the folding landscape and edges are the possible transition path, and the time evolution of the population of different conformations can be calculated through the probabilistic roadmap.

11.3.3 Move Set and Kinetic Rate Models

11.3.3.1 Kinetic Rate Models

The kinetic rate for an elementary kinetic step is usually defined as:

$$k_{i\to j} = k_0 \exp(-\frac{\Delta G^+ - \Delta G_i}{k_B T}),$$

where ΔG^+ is the free energy of the transition state, ΔG_i is the free energy of the state i, k_0 is a constant. The actual models for the base pair kinetic move use:

$$k = k_0 \exp(-\frac{\Delta G}{2k_B T}),$$

where ΔG is the free energy difference between the two states. Schmitz and Steger [95] treat the stacking energy as the barrier when opening a base pair, the loop energy change as the barrier when closing a base pair. Zhang and Chen [86] define the transition rate for the formation (k_+) and the disruption (k_-) of a base stack as the following:

$$k_+ = k_0 \exp(-\frac{\Delta G_+}{k_B T}), \quad k_- = k_0 \exp(-\frac{\Delta G_-}{k_B T}),$$

where the prefactor k_0 is fitted from the experimental data and is equal is $6.6 \times 10^{12}\,s^{-1}$ and $6.6 \times 10^{13}\,s^{-1}$ for an AU and GC stack [109], k_B is the Boltzmann constant, T is the temperature, and $\Delta G_\pm$ is the free energy barrier for the respective transition. Assuming that the barrier for the formation of a stack is caused by the reduction in entropy, $\Delta G_+ = T\Delta S$. If the stack closes a loop, the formation of the stack is accompanied by concomitant entropic decrease for loop closure, thus, the kinetic barrier for loop closure is $\Delta G_+ = T\Delta S = T(\Delta S_{loop} + \Delta S_{stack})$ where ΔS_{loop} is the entropy of the loop and ΔS_{stack} is the entropy of the stack. Assuming that the barrier for the disruption of a base pair is caused by the energetic (enthalpic) cost ΔH to break the hydrogen bonding and the base stacking interactions: $G_- = \Delta H_{stack}$, where ΔH_{stack} is the enthalpy of the stack. Then the rates for the formation and disruption of a stack are:

$$k_+ = k_0 e^{-\Delta S_{stack}/k_B T};$$
$$k_- = k_0 e^{-\Delta H/k_B T}$$

respectively, and the rates for formation and disruption of a loop-closing stack (and the loop) are:

$$k_+^{loop} = k_0 \exp(-(\Delta S_{loop} + \Delta S_{stack})/k_B T), \quad k_-^{loop} = k_0 \exp(-\Delta H/k_B T).$$

11.3.3.2 Model to Calculate the Transition Rate for Helix Based Moves

Tacker et al. [110] propose a rate model similar to that described for single base pairs: the activation energy is the change in loop energies when forming a helix, while opening a helix it is the stacking free energies. The same approach was adopted by in Refs. [111, 112]. Isambert [113] proposed that the free energy barrier for helix formation is the entropic penalty incurred by inserting the nucleus and the rate of is then given by an Arrhenius law using for nucleation. Zhao et al. [102] calculate the rate of a helix move set from that of the stack move set. If two conformations differ only in one helix, the transition between them would be the formation and disruption of the helix. Assuming that after the first stack is closed (with the concurrent formation of a loop), the helix will form along the zipping pathway. The rate of helix formation can be estimated along this zipping pathway. From the empirical thermodynamic parameters [114, 115], it can be found that for most RNA helices, the free energy landscape for a zipping pathway shows a downhill profile after the formation of the third base stack. Therefore, the rate k_f of the helix formation (along a specific pathway) is equal to the rate for the formation of the three-stack state. Considering the (slow) breaking of the stacks, for zipping along the 1 → 2→3 pathway in Fig. 11.1:

$$k_f = k_{12}K_1(1 - K_2'K_1' \sum_0^\infty K_2'K_1) = k_{12}K_1(1 - K_2'K_1' \frac{1}{1 - K_2'K_1}), \tag{11.2}$$

where k_{ij} denotes the rate for the transition from state i to state j, K_1 and K'_1 are the forward (state 2 → 3) and reverse (state 2 → 1) probability of state 2,K_2 and K'_2 are the forward (state 3 → 5 and 3 → 6) and reverse (state 3 → 2) probability of state 3,

$$K_1 = \frac{k_{23}}{k_{23} + k_{21} + k_{24}}, K_1' = \frac{k_{21}}{k_{23} + k_{21} + k_{24}}, K_2 = \frac{k_{35} + k_{36}}{k_{32} + k_{35} + k_{36}}, K_2' = \frac{k_{32}}{k_{35} + k_{36} + k_{32}}. \tag{11.3}$$

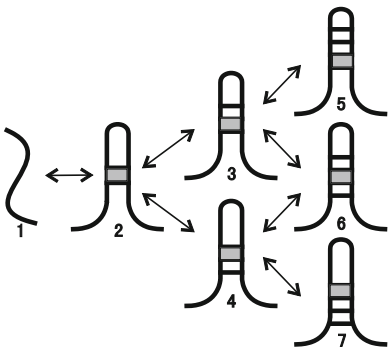


Fig. 11.1 Multiple pathways for the formation of a helix after the first (nucleation) stack formed

For a given RNA molecule, the first base stack can be formed anywhere inside the helix. Therefore, the net rate k_F for the formation of a helix is the sum of the rates (Eq. 11.2) along the two pathways (Fig. 11.1) with the different first (nucleation) base stacks. The rate for the disruption of the helix can be estimated from the equilibrium constant of the helix: $k_U = k_F e^{\Delta G/k_B T}$, where ΔG is the folding free energy of the helix.

If two helices A and B overlap with each other, they cannot coexist in the same structure. The conversion of helix A to helix B through complete unfolding of helix A followed by refolding to B is extremely slow due to the high energy barrier to disrupt all the base stacks in helix A. Zhao et al. [102] proposed that there is a much faster tunneling pathway, which is classified as three process: (1) at first helix A partially disrupted, (2) exchanging, disruption of a base stack in A is accompanied by a concurrent formation of a base stack in B, (3) zipping, helix B grows through a zipping process. The pathway is fast because the formation of the base stacks in B tends to cause an overall downhill shape of the free energy landscape. This is similar to the Morgan-Higgs saddle point approach [116], in which the saddle point height is estimated as the highest point along the path. However, the free energy landscape suggests that there exist multiple high free energy points along the path. This (tunneling) pathway involves a much lower energy barrier to unwinding the helix than the complete unfolding pathway (Fig. 11.2). Based on the tunneling pathway, the rate for helix exchange is estimated as:

$$k_{A\to B} = \frac{\prod_i^n k_i}{\sum_{j=0}^{n-1}\left(\prod_{i=1}^{j} k_i' \prod_{m=j+2}^{n} k_m\right)}, \quad k_{B\to A} = k_{A\to B} e^{-\frac{\Delta G_{AB}}{k_B T}}. \tag{11.4}$$

In the above formula, k_n and k_n' are the rate constants for the process to formation (disruption) and disruption (formation) of a base stack in A (B), respectively.

When the conformation space is consisted of local minima, the transition rate is often calculated by searching the saddle point or from the free energy barrier tree [103–105].

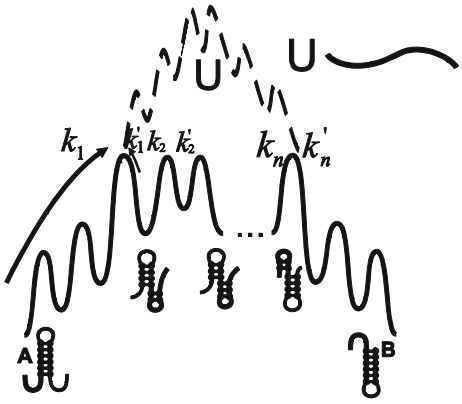


Fig. 11.2 The free energy landscape of the tunneling pathway that connects two overlapping helices *A* and *B*. *U* is the open state. The unfolding of *A* is accompanied by the folding of *B*. k_1 denotes the transition rate for the unfolding of helix *A* to form the first stack of helix *B*. $k_1', k_2, k_2', \ldots, k_n$ denote the transition rates between the neighboring intermediates along the tunneling pathways

11.3.3.3 Folding Kinetics During Transcription

The folding of functional RNA structures are often coupled with the transcription process [117–119]: since transcription is slow compared to local folding processes, the partially synthesized RNA will start folding while the molecule is still being synthesized. For instance, in the auto-catalyzed splicing reaction of tetrahymena group| intron, the functional native structure may form within the timescale of transcription, which is much faster than the refolding of the complete chain in vitro [120–125]. Investigations about the RNA component of *Bacillus subtilis* RNase P folding indicate site-specific pausing could greatly influence the folding result of RNA molecule [126]. Addition of NusA which causes pausing in the process of transcription provides longer duration of temporary RNA chain to undergo the conformational search. Recently, several RNA folding kinetics algorithms were developed in connection with the thermodynamic energetics of the folding system. For instance, by using the barrier trees and assuming that the basins of individual local minima are in quasi-equilibrium, the folding kinetics under transcription was studied [104]. Combining the thermodynamic properties with coarse-grained local folding kinetics, a heuristic approach was also developed to successfully predict cotranscriptional folding for large RNAs [105]. Zhao et al. [127] treat the transcription of a single nucleotide as an elementary time step. The real time for each step is a constant or variable if the nucleotides are synthesized at a constant or variable speed, respectively. The transcriptional pausing at a specific site can be simulated by assigning a large number of effective time steps for the corresponding (paused) step. If the transcription speed of an RNA sequence is v nucleotides per second, the (real) time window for each step would be $1/v$ seconds, i.e., the polymerase spends

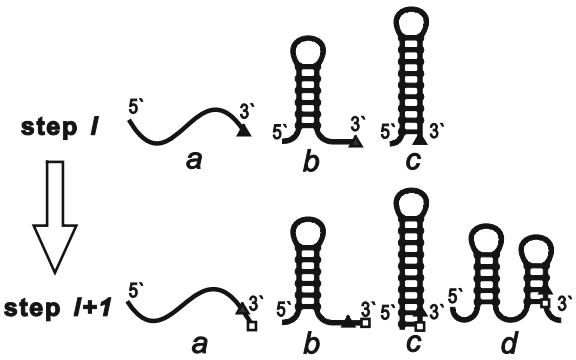


Fig. 11.3 The relationships between l-nt and $(l + 1)$-nt structures: elongation of an open chain (*a*), a dangling tail (*b*), a helix (*c*), and the formation of a new structure (*d*). The filled triangle denotes the last transcribed nucleotide in step l, and the square denotes the last transcribed nucleotide in step $l + 1$

$1/v$ seconds to synthesis a nucleotide. Assuming that at time t the l-nt chain is (newly) transcribed, the population distribution of the l-nt chain conformational space is relaxed from $[P_1(l)_{begin}, P_2(l)_{begin}, \ldots, P_\Omega(l)_{begin}]$ to $[P_1(l)_{end}, P_2(l)_{end}, \ldots, P_\Omega(l)_{end}]$ at time t to time $t + 1/v$, when the $(l + 1)$-th nucleotide is transcribed, here Ω is the number of conformations for an l-nt chain (Fig. 11.3). The beginning population of the $(l + 1)$-th step is inherited from the ending population of the l-th step. However, the RNA chain in the $(l + 1)$-th step is one nucleotide longer than in the l-th step. According to the possible changes of the structures upon the elongation of the chain by one nucleotide, the structures are classified as four types. The population distribution at the beginning of step $l + 1$ can be derived from that of the step at the end: $P(l+1)_{begin} = P(l)_{end}$ for *a*, *b*, & *c*; $P(l+1)_{begin} = 0$ for *d*. Applying this method from the first step to the end of transcription, we compute the folding kinetics for the RNA chain during transcription.

11.4 Metal Ions in RNA Folding

11.4.1 Ions Stabilize RNA/DNA Folded Structure

11.4.1.1 Ion Binding to RNAs/DNAs

Metal ions would like to bind to negatively charged nucleic acids to neutralize the negatively charged RNAs/DNAs [128–139]. The number of binding ions is important to DNA/RNA structure and stability, and can be measured via several experimental methods such as the small angle X-ray scattering (SAXS) [128–139], the ion-counting method [132], and the thermodynamic method [133–138].

The experimental methods have been applied to various RNAs/DNAs, including yeast 58-nt ribosomal RNA fragment [136], tRNA [137, 138], poly(A.U) [133], beet western yellow virus pseudoknot fragment [135], polymeric calf thymus DNA [134], oligomeric DNA/RNA duplexes [132], and DNA triplex [132]; see Ref [72] for a collection on the experimental data for ion binding to RNAs/DNAs. The extensive experimental data have yielded the following major conclusions:

(1) The detailed distributions of binding ions near molecular surface are sensitive to the specific atomistic structure of the RNAs/DNAs [131], and the ion-binding's of different (monovalent and divalent) species of ions appear anti-cooperative [132];
(2) Metal ions can give more efficient charge neutralization for RNA than for DNA. Such difference possibly comes from the higher charge density on backbone of A-form helix has than B-form helix [131].
(3) Multivalent ions (e.g., Mg^{2+}) are much more efficient in charge neutralization than monovalent ions (e.g., Na^{+}). Such unusually higher efficiency of multivalent ions is beyond the mean-field description such as ionic strength, and is more pronounced for larger RNAs with more compact structures.

In addition to the diffusive ion binding, the specific ion binding may make significant contribution to stabilizing specific RNA folded structure and the function of RNA. The specific ion binding may be related to the RNA sequence, the local geometry, and the property of ion and water, and is a challenge in both experiments and modelling.

11.4.1.2 Ion Contribution to Flexibility of Single-Stranded RNA

Single-strand (ss) RNA is a fundamental segment in RNA structure and the flexibility of ssRNA is important to the global stability of RNA. The ion contribution to the flexibility (e.g., persistence length l_p) of single-strand RNAs have been quantified by a variety of experimental approaches, such as force-extension, single molecule fluorescence resonance energy transfer (smFRET), small angle X-ray scattering (SAXS), and fluorescence recovery after photobleaching, over different kinds of ssRNAs/DNAs [140–147]. The major conclusions are in the following:

(1) Mg^{2+} is approximately 60-120 times more efficient than Na^{+} in neutralizing ss RNAs/DNAs, which is beyond the mean-field concept (e.g., ionic strength) [145];
(2) The persistence length of ss RNA/DNA decreases with the increase of $[Na^{+}]$ or $[Mg^{2+}]$, and the ion-concentration dependence is stronger for $[Na^{+}]$ than for $[Mg^{2+}]$ [145, 146];
(3) The dependence of persistence length of ss RNA/DNA is stronger for longer sequence. For long ss generic sequence, there is a crude empirical formula for ion-dependent persistence length: $l_p = 5 + 1.5/\sqrt{I}$, where I is the ionic strength [147].

(4) Poly(A)/poly(U) are more stiff than poly(T) at high salt while the ion-dependence of l_p are similar. Such stronger intrinsic stiffness may result from the stronger self-stacking of Poly(A)/poly(U) [144, 145].

However, there are also questions remained: (1) How do the flexibility of a ss RNA/DNA and its Na^+/Mg^{2+} dependence rely on the surrounding space? (2) How is the Na^+/Mg^{2+}-dependent l_p on the sequence length quantified? To answer the questions remains a challenge due to the high conformational fluctuation of ss chain and possible stronger correlations between Mg^{2+}.

11.4.1.3 Ions Stabilize Helices and Hairpins

Helix is the most fundamental segment of RNA structure (ranging from several to about ten base pairs), and hairpin is the simplest secondary structural motif. The thermodynamic experiments have revealed that the stability of helices and hairpins is sensitive to ionic environments. Most of the experiments were performed in a Na^+ solution or a mixed Na^+/Mg^{2+} solution; see Ref. [72] for a brief summary on the experimental data [148–159]. These thermodynamic data lead to the following major features for ion effects in helix and hairpin stability:

(1) In Na^+ or K^+ solution, the stabilities of DNA/RNA helices/hairpins depend on ion concentration with the approximately linear dependence on the logarithm of Na^+ or K^+ concentration and such dependence is strong at low salt ($<$0.1 M Na^+ or K^+), and relatively weak at a high Na^+ or K^+ concentration ($\geq$0.1 M Na^+);
(2) Compared with Na^+ or K^+, divalent ions (e.g., Mg^{2+}) are more efficient in stabilizing helices/hairpins. For an example, the stability for short DNA/RNA duplexes/hairpins in a 10 mM Mg^{2+} solution is approximately equivalent to the stability in a 1 M Na^+ solution [139, 147–149].

The thermodynamic parameters for the formation of helix and loop have been measured extensively at 1 M Na^+ (i.e., the standard ion condition). These parameters have enabled the accurate predictions on RNA (DNA) secondary structure, stability and kinetics [160–164]. For ion condition other than 1 M NaCl, RNA/DNA thermodynamic data and theoretical modelling for various ionic conditions yields a set of fitted formulas for the thermodynamic parameters of RNA/DNA helices versus Na^+/Mg^{2+} concentrations. In contrast to Na^+ solutions, experimental data on Mg^{2+}-dependent helix/hairpin stability has been relatively limited, and the [Mg^{2+}]-dependent thermodynamic parameters [154, 155] may need to be validated through more extensive experimental data; see the Sect. 11.4.3. For ion-dependent loop formation thermodynamics, the hairpin loop stability has been derived as functions of [Na^+] and [Mg^{2+}], based on the statistical mechanical modelling [156]; see the Sect. 11.4.3.

11.4.1.4 Ions Stabilize Tertiary Structures

Generally, RNA tertiary structure is folded by the aggregation of secondary segments, the minor rearrangements of secondary segments and the formation of tertiary contact. Since RNA tertiary folding generally involves massive charge build-up, ion-RNA interaction is stronger for tertiary structures and consequently the quantitative understanding on ion effects in RNA tertiary folding becomes more challenging. Extensive experiments have investigated how metal ions assist RNA tertiary folding and stabilize tertiary structures for various RNAs, such as tRNA, 58-nt ribosomal RNA fragment, beet western yellow virus pseudoknot fragment, Tetrahymena ribozyme, kissing complex etc.; see Ref. [72] for a summary for the experimental data of ion effects in RNA tertiary folding [165–179]. These experiments have revealed the following important major features on the effects of metal ions, especially Mg^{2+}:

(1) Metal ions of higher charge density (i.e. higher valence and smaller size) are more efficient in stabilizing RNA tertiary folds [166, 169]. For the Tar–tar RNA complex, smaller ions can enhance the folding stability [170].
(2) Mg^{2+} can make a significant contribution to RNA tertiary structure stability even at high Na^+ (or other monovalent ions) concentration, and Mg^{2+} can induce more compact folded structures than Na^+.
(3) For HIV-1 dimerization initiation signal (DIS) type kissing loop-loop complexes, the melting temperature shows much stronger ion-dependence than for the corresponding duplex of the same sequence at the kissing interface [174, 175].
(4) The higher efficiency of Mg^{2+} over Na^+ is much more pronounced for the kissing loop complex than for the duplex [174]. Such phenomena may result from the significantly higher massive built-up when loop-loop kissing.

In addition to the non-specific effects of metal ions shown above, some experiments also suggest that, depending on the sequence and geometry, specific interactions of binding ions with the RNA could make critical contribution to RNA tertiary structure [72, 135, 136]. The unclear understanding on roles of specific binding ions suggest the demand for the further more careful and extensive investigations, especially theoretical investigations, on the role of specific binding ions.

11.4.1.5 Ion-Mediated Structural Collapse

RNA structural collapse during tertiary folding often involves the helix-helix packing, and therefore, the helix-helix interaction is important for RNA tertiary folding. Rau and Parsegian have performed osmotic pressure measurements to quantify the ion-mediated interactions between *long* DNA helices [159, 160], leading to the following general conclusions:

(1) Multivalent ions, such as Co^{3+}, can induce effective attraction between DNA helices, while monovalent ions (e.g., Na^{+}) can only screen the helix–helix electrostatic repulsion [180];
(2) Certain types of divalent ions such Mn^{2+} can induce effective helix-helix attractive force [181], while other divalent ions such as Ca^{2+} cannot. Mg^{2+} in the presence of methanol could induce the effective helix-helix attraction [181]. The different roles of divalent ions might be attributed to the different ion binding affinities to the different groups [1]. For example, Mn^{2+} likes to binding into grooves, while Ca^{2+} likes to binding to phosphate groups [1].

However, in realistic RNA structures, helices are generally very short (ranging from several to around ten base pairs). Ions may have different effects in the effective interactions between short helices from long helices due to the greater rotational freedom and stronger end-effects of short helices. The recent experiments for short helices [68, 182–185] indicate the following conclusions:

(1) For a system of dispersed short DNA helices, the SAXS experiments suggest that Mg^{2+} of high concentration can induce effective helix-helix attraction through end-end stacking [183].
(2) For a system of loop-tethered short helices, the SAXS experiments suggested a possible weak side-side helix-helix attraction in a Mg^{2+} solution of high concentration ($\sim$0.6 M) [68].
(3) The experiments showed that the PB theory underestimates the efficiency of Mg^{2+} in RNA structural collapse by over 10 times [184].
(4) In trivalent ion solution, short DNA duplexes can become condensed while RNA duplexes keep soluble [185].
(5) Mg^{2+} cannot condense long DNA duplexes while could condense short DNA triplex in aqueous solution [181, 184].

However, for the system of short helices, further investigations are still required to make clear: (1) How do the different ions (with different valences and sizes) cause the different effective helix-helix interactions? (2) Is the relaxation state a randomly disordered state or a state with certain order or ion-specific?

11.4.2 Theoretical Modelling for Ion Electrostatics

To quantitatively explore the ion effects in RNA folding, some theoretical approaches have been developed. The application of these theories on the ion-RNA (DNA) system has significantly enhanced the quantitative understanding on the ion role in RNA folding, which will be introduced in the following; see Ref. [186] for a review on the theoretical models.

11.4.2.1 Counterion Condensation Theory

The counterion condensation (CC) theory was developed for describing the interaction between ion and long DNA [187, 188]. In the theory, a DNA helix is approximated as a line-charge, and metal ions around DNA are either in the condensed state (near the DNA surface) or in the free state (away from the vicinity of DNA). The binding of an ion would decrease the electrostatic energy and simultaneously increase the ion entropic free energy. The competition between the two components (electrostatic energy and ion entropy) determines the thermodynamically equilibrium state.

The application of the CC theory on the effect of monovalent ions (e.g., Na^+, K^+) in DNA helix thermodynamics [160, 187] shows a linear dependence of melting temperature T_m of DNA helix on the logarithm of monovalent ion concentration, which is in accordance with the experimental measurements. For the system of multi-body helices, the CC theory predicts that two DNA helices attract each other in both monovalent and multivalent salts. For lower salt concentration, the predicted attractive force becomes stronger while two helices are equilibrated at a larger separation [189]. The predictions are somewhat inconsistent with the experiment measurements [180, 181, 183] and computer simulations [190, 191] on nucleic acid helix-helix interactions.

Although the CC theory has gained the great success in the analysis of DNA thermodynamics, the theory still has the serious shortcomings: (1) The CC theory cannot be strictly employed to the RNA with complex structure in finite salt solutions since the theory is derived based on the assumptions of infinite-length DNA line-charge structural model and a infinite-dilute ion concentration; (2) The CC theory ignores the fluctuation and correlation of condensed ions, by assuming a uniform distribution of condensed ions along DNA. Consequently, the theory may become invalid for the multivalent ion solution where the correlations can be strong.

11.4.2.2 Poisson-Boltzmann Theory

The Poisson-Boltzmann (PB) theory has its early and simplified versions known as Gouy-Chapmann theory and Debye-Huchel theory. These two theories are the simplified versions of the PB theory for different specific systems [186]. The PB equation can be derived based on the Poisson equation for mean electric potential ψ and a Boltzmann distribution for diffusive ions in solutions

$$\nabla \cdot [\varepsilon(\mathbf{r})\varepsilon_0 \nabla \psi(\mathbf{r})] = -4\pi \left[\rho_f + \sum_{\alpha} e c_{\alpha}^0 N_{AV} z_{\alpha} \mathrm{e}^{-z_{\alpha} e \psi(\mathbf{r})/k_B T} \right], \tag{11.5}$$

where $z_\alpha e\psi$ is approximated to be the electrostatic energy of a diffusive ion with ionic charge $z_\alpha e$. ε is the dielectric constant; ρ_f is the charge density of fixed

charges in biomolecules; and c_α^0 is the bulk concentration of ion species α. In the recent two decades, some algorithms have been developed to numerically solve the PB equation, and the PB theory has been widely used in the electrostatics of biomolecules in solutions [192–196]. For electrostatics of biomolecules in aqeous/monovalent ion solutions, the experimental comparisons show that the PB theory makes rather accurate predictions [e.g., 197].

However, a mean-field approximation is employed in deriving the PB equation, i.e., diffusive charges (ions) obey a mean Boltzmann distribution based on the mean electric potential in stead of the potential of mean force. As the result, (1) the PB theory ignores the fluctuation of ions in solution by assuming a mean ion distribution; (2) the PB theory ignores the ion-ion correlations by assuming the mean electric potential for diffusive ions rather than the potential of mean force, and (3) the PB theory ignores the ion finite size by the point-charge approximation. Therefore, the PB could not make reliable predictions on the electrostatic interactions for nucleic acid in multivalent ion solution where ion-ion interactions can be strong [186, 198]. An important example is the (multivalent) ion-mediated like-charge interaction, the PB always predicts repulsive force between two like-charged polyelectrolytes (DNA helices), while the experiments have shown the attractive force in multivalent salts [180, 181]. For ion-mediated RNA structural collapse, the PB theory underestimates the efficiency of Mg^{2+} by over 10 times [179, 182].

11.4.2.3 Modified Models Beyond Mean-Field Approximation

Aiming to improve the prediction for polyelectrolyte-multivalent ions, many efforts have been made in the recent years to overcome the shortcomings of the PB theory. Here, we will introduce several major modified models beyond the mean-field approximation [198–207].

Size-modified Poisson-Boltzmann model. The simplest modification for the PB model is to incorporating discrete ion size into the model. Recently, a size-modified PB model was proposed based on the lattice gas formulism, where the ion solution is discretized into a lattice with grid cells which can be occupied by ions with finite size [199, 200]. The application of the model for 3-dimensional complex nucleic acids shows that the modification can improve the prediction on monovalent ion-binding profiles at high salt concentration by capturing the binding saturation effect at high ion concentration. But for RNA solution with multivalent ions, the model still cannot give reliable predictions since it ignores the ion-ion electrostatic correlations [199, 200].

Modified Poisson-Boltzmann theory based on Kirkwood/BBGY hierarchy. A modified PB model has been developed based on Kirkwood/BBGY hierarchy through taking into account the fluctuation potential and ion-exclusion term in the potential of mean force for diffusive ions [e.g., 201]. The comparisons with the computer simulations show that, the theory gives the improved predictions for

multivalent ion distributions near polyelectrolytes with ideal 3D shapes such as cylinder, sphere and plane. But for realistic nucleic acids (or proteins) with arbitary 3D shape, the numerical solution requires huge computational cost and is computationally inapplicable for realistic nucleic acids/proteins with complex 3D shape because the equation for the fluctuation potential is coupled to the equation for the mean electrostatic potential [201, 202].

Correlation-corrected Poisson-Boltzmann model. Recently, a so-called correlation-corrected Poisson-Boltzmann model was developed to account for the ion-ion correlations, by introducing an effective potential between like-charge ions [203]. Such effective potential is the same as the Coulomb potential at large ion-ion separation, while becomes a reduced repulsive Coulomb potential for a close ion-ion separation. For the electric double layers, the comparisons with the computer simulations showed that the model makes improved predictions on multivalent ion distributions and predicts an attractive force between the two planes in the presence of multivalent ions [203]. However, the model is computationally expensive for RNAs with complex structures. Moreover, such effective potential is somewhat ad hoc and the mode is still lack of the validation on thermodynamics of nucleic acids.

Other theories beyond the mean-field approximation. Other theories beyond the mean-field approximation such as the integration theory [204], the density function theory [205] and the local molecular field theory [206] have been developed to account for the ion-ion correlation effects around nucleic acids/polyelectrolytes. But due to the huge computation cost for realistic nucleic acid system, these theories are also practically inapplicable. Recently, a tightly bound ion (TBI) theory is developed by explicitly treating the strongly correlated ions which reside in the vicinity of nucleic acid surface [186, 198, 207–211]. The extensive experimental comparisons showed that this theory has been shown to make improved predictions on the ion effects for various nucleic acid structures in the presence of Mg^{2+}. In the following, we will focus on the TBI theory and its applications on modelling ion effects in DNA/RNA structure stabilities, including helices, hairpins, tertiary folds, and assembly.

11.4.3 Tightly Bound Ion Theory

As described above, extensive experiments have shown that Mg^{2+} plays a special role in RNA folding: Mg^{2+} is much more efficient than Na^+ in RNA folding and Mg^{2+} can induce more compact structure than Na^+. For example, Mg^{2+} is generally $\sim$1,000 times more efficient than Na^+ in RNA tertiary folding [169]. Aiming to quantitatively understand the role of multivalent ions in RNA folding, Tan and Chen have developed a TBI theory, by accounting for ion-correlations and fluctuations for realistic RNAs in ion solutions [186, 198, 207–211]. In the following, we will introduce the TBI theory with theoretical framework and applications on modelling ion effects in RNA/DNA structure stabilities.

11.4.3.1 Framework of the Tightly Bound Ion Theory

Since RNAs/DNAs are highly charged polyanionic molecules, the positively charged metal ions in solutions would aggregate on nucleic acid surface, causing high ion concentration in the vicinity of RNA/DNA surface. These condensed ions of high concentration would interact (correlate) strongly with each other. The correlation strength between ions can be characterized by the coupling parameter Γ [186, 198]

$$\Gamma(\mathbf{r}) = \frac{(z_\alpha e)^2}{\varepsilon a_{wz}(\mathbf{r}) k_B T} \geq \Gamma_c. \tag{11.6}$$

Previous studies have shown that for ionic system, the change of coupling parameter Γ can induce the gas-liquid transition, and the critical value Γ_c was shown to reside in the range of [2.3, 2.9] [186, 198]. In the TBI theory, according to the critical inter-ion correlation strength Γ_c (2.6, a mean value over [2.3–2.9]), the ions around RNAs/DNAs are divided into two types: (strongly correlated) tightly bound ions in the vicinity of RNA and (weakly correlated) diffusive ions in the outer space. Correspondingly, the space around RNAs is also divided into the tightly bound region and diffusive region. Due to the weak inter-ion correlations, the diffusive ions can be treated by the mean-field PB approach. While for the (strongly correlated) tightly bound ion, the tightly bound region is discretized into different tightly bound cells, each of them around a (negatively) phosphate group. Every tightly bound cell can keep empty or be occupied by an ion, and all possible states of tightly bound cells (either empty or occupied by an ion) give the ensemble of ion-binding modes. The different ion-binding modes (M) in different cells are explicitly considered to account for the effects of ion correlations and fluctuations.

For a nucleic acid-ion system, the partition function Z is given by the summation of the partition function Z_M for all possible modes M

$$Z = \sum_M Z_M, \tag{11.7}$$

where Z_M is given by [177, 186, 198, 208]

$$Z_M = Z^{id}(c_z)^{N_b} \left(\int \prod_{i=1}^{N_b} d\mathbf{R}_i \right) e^{-\left(\Delta G_b + \Delta G_d + \Delta G_b^{pol}\right)/k_B T}. \tag{11.8}$$

Here, Z^{id} is the partition function for the uniform ion solution (without RNAs). N_b is the number of the tightly bound ions for model M. c_z is the bulk concentration of the z-valent ions, and $\mathbf{R}_i$ denotes the position of tightly bound ion i. $\int \prod_{i=1}^{N_b} d\mathbf{R}_i$ is the volume integral over the tightly bound region for the N_b tightly bound ions. ΔG_b is the free energy for the discrete charges in the tightly bound region

(including the tightly bound ions and phosphate charges); ΔG_d is the free energy for the diffusive ions, including the electrostatic interactions between the diffusive ions, and between the diffusive ions and the charges in the tightly bound region as well as the entropic free energy of the diffusive ions; ΔG_b^{pol} is the (Born) self-polarization energy for the discrete charges within the tightly bound region [177, 186, 198, 208].

The TBI theory has been widely applied to quantitatively understanding the ion contributions to RNA secondary and tertiary structure stability, which will be described in the following.

11.4.3.2 Modelling Ion Effects Stability of DNA/RNA Helices

The stability of helices is essential to the global stability and the functions of RNAs (DNAs). Due to the polyanionic nature, metal ions can be important to the stability of DNA/RNA helices. Based on a polyelectrolyte theory, the melting of a helix can be modelled as a two-state model. The free energy change due to the melting can be decoupled into a non-electrostatic contribution ΔG^{NE} and an electrostatic contribution ΔG^{E}, and ΔG^{NE} can be evaluated by combining the experimental data at a reference state with a polyelectrolyte theory (e.g. the TBI theory) [154, 155]

$$\begin{aligned}\Delta G &= \Delta G^{E} + \Delta G^{NE} \\ &= \Delta G^{E} + \left(\Delta G_{1\mathrm{MNa}^+} - \Delta G^{E}_{1\mathrm{MNa}^+}\right).\end{aligned} \tag{11.9}$$

With the use of the TBI theory for treating ion-DNA(RNA) interactions, the Na^+/Mg^{2+} dependence of helix stability can be quantitatively evaluated.

The comparisons with the extensive experimental data showed that the TBI theory makes reliable predictions on the stability of DNA and RNA helices in Na^+/Mg^{2+} solutions [154, 155]. Furthermore, the comprehensive calculations with the TBI theory give a series of empirical formulas for describing the thermodynamics of DNA (RNA) helices in Na^+/Mg^{2+} solutions.

Thermodynamic parameters for DNA helix in Na^+/Mg^{2+}solution.

For a DNA helix in Na^+ solution, the following formulas of Na^+-dependent thermodynamics can be obtained from the TBI theory [154]

$$\begin{aligned}\Delta G[\mathrm{Na}^+] &= \Delta G[1\ \mathrm{M\ Na}^+] + (N-1)\Delta g_1^{DNA}; \\ \Delta S[\mathrm{Na}^+] &= \Delta S[1\ \mathrm{M\ Na}^+] - 3.22(N-1)\Delta g_1^{DNA}; \\ 1/T_m[\mathrm{Na}^+] &= 1/T_m[1\ \mathrm{M\ Na}^+] - 0.00322(N-1)\Delta g_1^{DNA}/\Delta H[1\ \mathrm{M\ Na}^+],\end{aligned} \tag{11.10}$$

where ΔG, ΔS, T_m, ΔH are the free energy change, entropy change, melting temperature, enthalpy change for helix formation at $[Na^+]$ (in molar), or 1 M $[Na^+]$ (standard ion condition). Δg_1^{DNA} is a function associated with electrostatic folding

free energy per base stack of DNA, and is a function of helix length and $[Na^+]$ [154]

$$\begin{aligned} \Delta g_1^{DNA} &= a_1^{DNA} + b_1^{DNA}/N; \\ a_1^{DNA} &= -0.07\ln[Na^+] + 0.012\ln^2[Na^+]; \\ b_1^{DNA} &= 0.013\ln^2[Na^+]. \end{aligned} \tag{11.11}$$

The thermodynamics for DNA helix at any given $[Na^+]$ can be calculated through the above empirical formulas, since those at 1 M $[Na^+]$ (standard ion condition) can be obtained from the nearest neighbor model with the experimental parameters of SantaLucia et al. [160]. The quantitative comparisons with extensive experimental data show that the empirical formulas can give rather accurate estimates for thermodynamics of DNA helices in Na^+ solutions [154].

For the thermodynamics of a DNA helix in Mg^{2+} solution, the TBI model gives the following similar empirical formulas [154]

$$\begin{aligned} \Delta G[Mg^{2+}] &= \Delta G[1\text{ M }Mg^{2+}] + (N-1)\Delta g_2^{DNA}; \\ \Delta S[Mg^{2+}] &= \Delta S[1\text{ M }Mg^{2+}] - 3.22(N-1)\Delta g_2^{DNA}; \\ 1/T_m[Mg^{2+}] &= 1/T_m[1\text{ M }Mg^{2+}] - 0.00322(N-1)\Delta g_2^{DNA}/\Delta H[1\text{ M }Mg^{+}], \end{aligned} \tag{11.12}$$

where Δg_2^{DNA} is given by

$$\begin{aligned} \Delta g_2^{DNA} &= a_2^{DNA} + b_2^{DNA}/N^2; \\ a_2^{DNA} &= 0.02\ln[Mg^{2+}] + 0.0068\ln^2[Mg^{2+}]; \\ b_2^{DNA} &= 1.18\ln[Mg^{2+}] + 0.344\ln^2[Mg^{2+}]. \end{aligned} \tag{11.13}$$

Through the above empirical formulas, the thermodynamics of a DNA helix at a given $[Mg^{2+}]$ can be calculated easily. The experimental comparisons show that the empirical formulas can make reliable estimates for the stability of a short DNA helix ranging from 6-bp to 30-bp at an arbitrary $[Mg^{2+}]$ [154].

Generally, a buffer contains both Na^+ (or K^+) and Mg^{2+} ions. The TBI theory also gives the empirical formulas for DNA helix thermodynamics in a mixed Na^+/Mg^{2+} solution [155]

$$\begin{aligned} \Delta G &= \Delta G[1\text{ M }Na^+] + (N-1)\left(x_{duplex}\Delta g_1^{DNA} + (1-x_{duplex})\Delta g_2^{DNA}\right) + \Delta g_{12}; \\ \Delta S &= \Delta S[1\text{ M }Na^+] - 3.22\left((N-1)\left(x_{duplex}\Delta g_1^{DNA} + (1-x_{duplex})\Delta g_2^{DNA}\right) + \Delta g_{12}\right); \\ 1/T_m &= 1/T_m[1\text{ M }Na^+] - 0.00322\left((N-1)\left(x_{duplex}\Delta g_1^{DNA} + (1-x_{duplex})\Delta g_2^{DNA}\right) + \Delta g_{12}\right)/\Delta H[1\text{ M }Na^+], \end{aligned} \tag{11.14}$$

where x_{duplex} stands for the contribution fraction from Na^+, and Δg_{12} is a crossing term. x_{duplex} and Δg_{12} are given by

$$x_{duplex} = \frac{[Na^+]}{\left([Na^+] + (8.1 - 32.4/N)(5.2 - \ln[Na^+])[Mg^{2+}]\right)};$$
$$\Delta g_{12} = -0.6 x_{duplex}(1 - x_{duplex}) \ln[Na^+] \ln\left((1/x_{duplex} - 1)[Na^+]\right)/N, \quad (11.15)$$

where $[Na^+]$ and $[Mg^{2+}]$ are both in molar. The comparisons with experimental data show that the above formulas give good estimate for the thermodynamics of a DNA helix in mixed Na^+/Mg^{2+} solutions [155].

Thermodynamic parameters for RNA helix in Na^+/Mg^{2+}solution.

For an RNA helix in a Na^+ solution, the thermodynamics can also be formulated by Eq. (11.10), except that Δg_1^{DNA} needs to be replaced by Δg_1^{RNA}. Δg_1^{RNA} can be given by [155]

$$\begin{aligned} \Delta g_1^{RNA} &= a_1^{RNA} + b_1^{RNA}/N; \\ a_1^{RNA} &= -0.075 \ln[Na^+] + 0.012 \ln^2[Na^+]; \\ b_1^{RNA} &= 0.018 \ln^2[Na^+]. \end{aligned} \quad (11.16)$$

The combination of Eq. (11.10) and Eq. (11.16) can give good estimate for an RNA helix in a Na^+ solution, as shown in Ref [155].

Similarly, for an RNA helix in a Mg^{2+} solution, the thermodynamics can be described by Eq. (11.12), except that Δg_2^{DNA} needs to be changed into Δg_2^{RNA}

$$\begin{aligned} \Delta g_2^{RNA} &= a_2^{RNA} + b_2^{RNA}/N^2; \\ a_2^{RNA} &= -0.6/N + 0.025 \ln[Mg^{2+}] + 0.0068 \ln^2[Mg^{2+}]; \\ b_2^{RNA} &= \ln[Mg^{2+}] + 0.38 \ln^2[Mg^{2+}]. \end{aligned} \quad (11.17)$$

For an RNA helix in a mixed Na^+/Mg^{2+} solution, the thermodynamics can be calculated by the following empirical formulas

$$\begin{aligned} \Delta G &= \Delta G[1\ M\ Na^+] + (N-1)\left(x_{duplex}\Delta g_1^{RNA} + (1 - x_{duplex})\Delta g_2^{RNA}\right) + \Delta g_{12}; \\ \Delta S &= \Delta S[1\ M\ Na^+] - 3.22\left((N-1)\left(x_{duplex}\Delta g_1^{RNA} + (1 - x_{duplex})\Delta g_2^{RNA}\right) + \Delta g_{12}\right); \\ 1/T_m &= 1/T_m[1\ M\ Na^+] - 0.00322\left((N-1)\left(x_{duplex}\Delta g_1^{RNA} + (1 - x_{duplex})\Delta g_2^{RNA}\right) + \Delta g_{12}\right)/\Delta H[1\ M\ Na^+], \end{aligned} \quad (11.18)$$

where x_{duplex} and Δg_{12} are given by Eq. (11.15). Δg_1^{RNA} and Δg_2^{RNA} are given by Eqs. (11.16 and 11.17), respectively. As shown in Ref. [155]. The predictions from Eq. (11.18) are quite reliable for the thermodynamics of an RNA helix in a mixed Na^+/Mg^{2+} solution (Fig. 11.4).

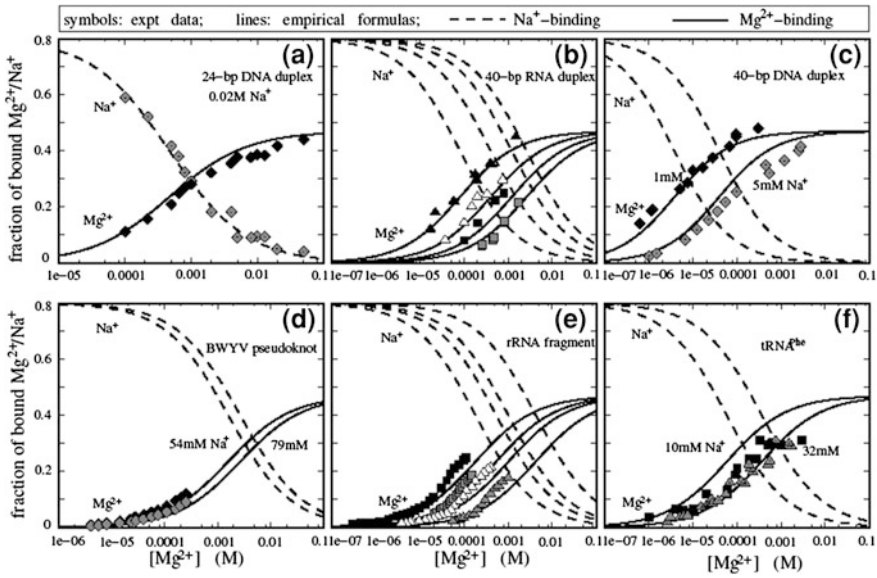


Fig. 11.4 The Mg^{2+} and Na^+ binding fractions per nucleotide for various RNA/DNA molecules. The *solid lines* are from the empirical formulas (Eqs. 11.26 and 11.28); and the symbols are experimental data. **a** 24-bp DNA duplex in $[Mg^{2+}]$ with fixed $[Na^+] = 20$ mM [132]; **b** 40-bp RNA duplex. The experimental data are for poly(A.U) [133]: From the *left* to *right*, $[Na^+] = 10$, 29, 60, and 100 mM, respectively; **c** 40-bp DNA duplex. The experimental data are for the calf thymus DNA [134]; **d** BWYV pseudoknot RNA [135]; **e** 58-nt rRNA fragment [136]. Please note that the experimental data are for mixed Mg^{2+}/K^+ (not Mg^{2+}/Na^+) solution: from *left* to *right*, $[K^+] = 20$, 40, 60, and 150 mM, respectively. Here we show the experimental data for semi-quantitative comparisons. **f** Yeast $tRNA^{Phe}$ [137, 138]

11.4.3.3 Modelling Ion Effects in Stability of DNA/RNA Hairpins

An RNA/DNA hairpin consists of a helix stem and a hairpin loop. The Na^+/Mg^{2+} dependence of a helix can be quantified by the empirical formulas described above. The TBI model can also gives the empirically analytical Na^+/Mg^{2+}-dependent thermodynamics for a single-stranded loop, with the combination with the virtual bond model for the single-stranded loop conformation [156].

For a loop formation in Na^+ solutions, the systematic calculations of the TBI model give the following empirical relation for the folding free energy of an N-nt loop with end-to-end distance l [156]:

$$\Delta G[Na^+] = -k_B T\Big(a_1 \ln(N - l/d + 1) + b_1(N - l/d + 1)^2 - b_1 - (c_1 N - d_1)\Big), \tag{11.19}$$

where $d = 6.4$ Å. The coefficients a_1, b_1, c_1, and d_1 are given by

$$\begin{aligned} a_1 &= (0.02N - 0.026)\ln[\mathrm{Na}^+] + 0.54N + 0.78; \\ b_1 &= (-0.01/(N+1) + 0.006)\ln[\mathrm{Na}^+] - 7/(N+1)^2 - 0.01; \\ c_1 &= 0.07\ln[\mathrm{Na}^+] + 1.8; \\ d_1 &= 0.21\ln[\mathrm{Na}^+] + 1.5. \end{aligned} \tag{11.20}$$

For a loop in Mg^{2+} solutions, the empirical formulas from the TBI theory for the folding free energy can be written as

$$\Delta G[\mathrm{Mg}^{2+}] = -k_B T\Big(a_2\ln(N - l/d + 1) + b_2(N - l/d + 1)^2 - b_2 - (c_2 N - d_2)\Big), \tag{11.21}$$

where a_2, b_2, c_2, and d_2 are given by

$$\begin{aligned} a_2 &= (-1/(N+1) + 0.32)\ln[\mathrm{Mg}^{2+}] + 0.7N + 0.43; \\ b_2 &= 0.0002(N+1)\ln[\mathrm{Mg}^{2+}] - 5.9/(N+1)^2 - 0.003; \\ c_2 &= 0.067\ln[\mathrm{Mg}^{2+}] + 2.2; \\ d_2 &= 0.163\ln[\mathrm{Mg}^{2+}] + 2.53. \end{aligned} \tag{11.22}$$

For a loop in mixed Na^+/Mg^{2+} solutions, the folding free energy is represented by

$$\Delta G[\mathrm{Na}^+/\mathrm{Mg}^{2+}] = x_{loop}\Delta G[\mathrm{Na}^+] + (1 - x_{loop})\Delta G[\mathrm{Mg}^{2+}], \tag{11.23}$$

where x_{loop} stands for the contribution fraction from Na^+ and is given by

$$x_{loop} = \frac{[\mathrm{Na}^+]}{[\mathrm{Na}^+] + (7.2 - 20/N)(40 - \ln[\mathrm{Na}^+])[\mathrm{Mg}^{2+}]}. \tag{11.24}$$

With the use of the above formulas for loop formation, the folding thermodynamics of hairpin loop, bulge loop, internal loop in an arbitrary Na^+/Mg^{2+} solution can be easily calculated [156].

For a hairpin loop, the Na^+/Mg^{2+}-dependent thermodynamics can be obtain by fixing the loop end-to-end distance at ~ 17 Å. Then the thermodynamics of an RNA (or DNA) hairpin can be calculated by the following formula [156]

$$\begin{aligned} \Delta G_{\text{hairpin}} &= \Delta H_{\text{stem}} - T\Delta S_{\text{stem}} + \Delta G_{\text{hairpin loop}}[\mathrm{Na}^+/\mathrm{Mg}^{2+}]; \\ \Delta H_{\text{stem}} &= \Delta H_{\text{stem}}[1\ \mathrm{M}\ \mathrm{Na}^+] + \Delta H_{\text{terminal mismatch}}; \\ \Delta S_{\text{stem}} &= \Delta S[\mathrm{Na}^+/\mathrm{Mg}^{2+}] + \Delta S_{\text{terminal mismatch}}, \end{aligned} \tag{11.25}$$

where ΔH_{stem}, $\Delta H_{\text{terminal mismatch}}$, and $\Delta S_{\text{terminal mismatch}}$ can be obtained from the nearest neighbour model with the measured thermodynamic parameters. $\Delta H_{\text{stem}}[Na^+/Mg^{2+}]$ can be given by the previously introduced empirical formulas (Eq. 11.14 for DNA and Eq. 11.18 for RNA). The extensive experimental comparisons show that the empirical formulas can make rather reliable predictions on the hairpin stability in a Na^+/Mg^{2+} solution; see Ref. [156].

In a very recent single-molecule experiment, the quantitative comparisons with the measurements on a 20-bp RNA hairpin show that, the above empirical formulas are rather accurate in describing the Na^+/Mg^{2+}-dependent thermodynamics for short RNA hairpin [146] (Fig. 11.4).

11.4.3.4 Modelling Ion Binding to RNA Tertiary Structures

As described above (Sect. 11.4.1.1), ion binding is critical for stabilizing RNA folded structure. The TBI theory has been developed for a static atomistic structure, and can quantify the ion atmosphere around an RNA (or DNA) with complex 3D structure [177]. In a mixed Na^+/Mg^{2+} solution, the binding's of Na^+ and Mg^{2+} are competitive and anti-cooperative. The TBI model has given an empirical equivalence relation between Mg^{2+} and Na^+ as [177]

$$\log[Na^+]_{Mg} = A \log[Mg^{2+}] + B, \tag{11.26}$$

where $[Na^+]$ and $[Mg^{2+}]$ are both in millimolar (mM). A and B are two parameters depending on the (low-resolution) RNA (or DNA) structure

$$A = 0.65 + \frac{4.2}{N}\left(\frac{R_g}{R_g^0}\right)^2; \quad B = 1.8 - \frac{9.8}{N}\left(\frac{R_g}{R_g^0}\right)^2, \tag{11.27}$$

where N is the number of nucleotides of an RNA, and R_g is the radius of gyration of the RNA (or DNA) backbone, and R_g^0 is the radius of gyration of an N-nt RNA duplex.

Based on Eq. (11.23), the binding fractions of Na^+ and Mg^{2+} can be calculated through [177]

$$f_{Na^+} = \frac{[Na^+]}{[Na^+] + [Na^+]_{Mg}} f^0_{Na^+}; \quad f_{Mg^+} = \frac{[Na^+]_{Mg}}{[Na^+] + [Na^+]_{Mg}} f^0_{Mg^{2+}}, \tag{11.28}$$

where $[Na^+]_{Mg}$ is given by Eq. (11.20). $f^0_{Na^+}$ and $f^0_{Mg^{2+}}$ are the binding fractions for pure Na^+ and pure Mg^{2+} solutions, respectively. Generally, $f^0_{Na^+} \approx 0.8$, and $f^0_{Mg^{2+}} \approx 0.47$ [177]. As shown in Fig. 11.4, the above empirical formulas could make reliable predictions for Na^+/Mg^{2+} binding to RNAs/DNAs with complex 3D structures.

11.4.3.5 Modelling Salt Contribution to RNA Tertiary Structure Stability

Since RNA folding is hierarchical, the tertiary structure folding can be crudely modelled as a two-state transition from an intermediate (I) to the native (N) state. Similarly to the helix stability, the RNA tertiary folding free energy can be decoupled into two contributions: an electrostatic part and a nonelectrostatic part [178]

$$\begin{aligned}\Delta G &= \Delta G^E[\mathrm{Na}^+/\mathrm{Mg}^{2+}] + \Delta G^{NE};\\ &= \Delta G^E[\mathrm{Na}^+/\mathrm{Mg}^{2+}] + \left(\Delta G[\text{expt } \mathrm{Na}^+] - \Delta G^E[\text{expt } \mathrm{Na}^+]\right);\\ &= \Delta G[\text{expt } \mathrm{Na}^+] + \left(\Delta G^E[\mathrm{Na}^+/\mathrm{Mg}^{2+}] - \Delta G^E[\text{expt } \mathrm{Na}^+]\right),\end{aligned} \tag{11.29}$$

where $\Delta G[\text{expt } \mathrm{Na}^+]$ is the experimental folding free energy at a reference ion condition. ΔG^E can be given by the empirical formulas derived from the TBI model [178].

For an RNA folding in Na^+ solutions, ΔG^E can be calculated by the following empirical formula

$$\Delta G^E[\mathrm{Na}^+] = \Delta G^E[1\text{ M } \mathrm{Na}^+] + a_1 N \ln[\mathrm{Na}^+] + b_1 N \ln^2[\mathrm{Na}^+], \tag{11.30}$$

where a_1 and b_1 are the parameters related to the RNA folded structure. a_1 and b_1 can be formulated by [178]

$$\begin{aligned}a_1 \times \varepsilon^*(T)T^* &= -0.086 + 7/(Nr_g^3 + 65);\\ b_1 \times \varepsilon^*(T)T^* &= 0.008 - 3.6/(N-5)^2,\end{aligned} \tag{11.31}$$

where $r_g = R_g^0/R_g$. $\varepsilon^*(T) = \varepsilon\ (T)/\varepsilon$ (298.15 K) is the relative dielectric constant, and $T^* = T/298.15$ is the relative temperature.

For an RNA folding in Mg^{2+} solutions, the TBI model gives the following empirical formula for ΔG^E [178]

$$\Delta G^E[\mathrm{Mg}^{2+}] = \Delta G^E[1\text{ M } \mathrm{Mg}^{2+}] + a_2 N \ln[\mathrm{Mg}^{2+}] + b_2 N \ln^2[\mathrm{Mg}^{2+}] + c_2 N T^*, \tag{11.32}$$

a_2, b_2, c_2 are given by

$$\begin{aligned}a_2 \times \varepsilon^*(T)T^* &= 0.012 - 1.4/(Nr_g^3 + 75);\\ b_2 \times \varepsilon^*(T)T^* &= 0.0048 - 57/(Nr_g^3 + N + 75)(N + 75);\\ c_2 \times \varepsilon^*(T)T^* &= -0.27 + 0.16/r_g^3 + 1.4/N.\end{aligned} \tag{11.33}$$

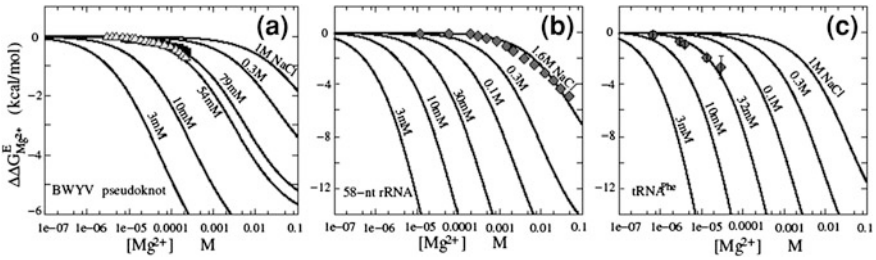


Fig. 11.5 The Mg^{2+}-contribution $\Delta\Delta G^E_{Mg^{2+}}$ to RNA tertiary structure folding free energy as a function of $[Mg^{2+}]$ for three RNA molecules: BWYV pseudoknot (**a**), 58-nt ribosomal RNA fragment (**b**), and yeast $tRNA^{Phe}$ (**c**) at room temperature. *Solid lines*, empirical formulas derived from the TBI model; symbols, experimental data: **a** BWYV pseudoknot in 54 and 79 mM Na^+ solution [135]; **b** 58-nt rRNA fragment in solution with 1.6 M monovalent ions [178, 197]; **c** yeast $tRNA^{Phe}$ in solution with 32 mM Na^+ [138, 179, 197]

For RNA folding in a mixed Na^+/Mg^{2+} solution, ΔG^E is given by the empirical relation

$$\Delta G^E[Na^+/Mg^{2+}] = x_{3^o}\Delta G^E[Na^+] + (1 - x_{3^o})\Delta G^E[Mg^{2+}] + N\Delta g_{12}, \quad (11.34)$$

where x_{3^o} denotes the contribution fraction from Na^+, and Δg_{12} is a crossing term. x and Δg_{12} are given by

$$x_{3^o} = \frac{[Na^+]}{[Na^+] + \left(3.8 - 34/(N-20)r_g^3\right)(1 + 0.2[Na^+])[Mg^{2+}]^{0.64}}; \quad (11.35)$$
$$\Delta g_{12} = -x_{3^o}(1 - x_{3^o})(0.26 - 1.2/(N-20)).$$

With the use of the above empirical formulas for ΔG^E and Eq. (11.29), the Na^+/Mg^{2+}-dependent RNA tertiary folding thermodynamics can be conveniently calculated. Figure 11.5 shows that Eq. (11.34) gives good estimates for the Mg^{2+}-contribution to the total folding stability $\Delta\Delta G^E_{Mg^{2+}} = \Delta G^E_{Na^+,Mg^{2+}} - \Delta G^E_{Na^+,Mg^{2+}=0}$, as compared with the experimental data. It is also shown that Eq. (11.29) with the empirical formulas for ΔG^E (Eqs. 11.30–11.35) can make good evaluation for the Na^+/Mg^{2+}-dependent folding thermodynamics of small RNAs [178].

11.5 Perspectives

Although many RNA 3D structure modelling methods have proposed, further developments and refinements of the existing models are still required. Current algorithms have shown how the use of available experimental data can dramatically improve the structure prediction, e.g. the discrete molecular dynamics simulations

with the use of HRP measurements can predict structure of RNAs ranging in size from 80 to 230 nucleotides [26]. In addition, several algorithms have exploited the hierarchical properties of RNA folding [36–38], and consequently the prediction accuracy can be improved by adding the knowledge of secondary structure and tertiary contacts from experiments to the existed programs. However, there are still some essential problems remaining challenging, including: (1) Could the structures for larger RNA molecules be predicted reliably and efficiently? (2) Could RNA structures be predicted versus different environments (temperature, ion conditions, etc.)?

Despite the significant progress, modelling of RNA folding dynamics remains a challenging problem. The current form of the theories involves several limitations. First, the theories do not treat folding/unfolding of tertiary folds such as pseudoknots. Second, the theories cannot treat, at the explicitly atomistic level, the effects of cofactors such as magnesium ions, ligands and proteins. In the future, we expect that the RNA folding kinetic theories can overcome these limitations and will be applicable to design RNA and DNA molecules with particular dynamic properties, which is of great importance in the emerging fields of synthetic biology and nucleic acid-based nanotechnology.

The extensive investigations have significantly enhanced the qualitative/quantitative understanding on ion effects in RNA folding. However, the quantitative understanding on ion roles is still challenging at least in the following issues: (1) How are the specific properties of ions correlated to their specific roles in RNA folding? (2) Is the efficient role of multivalent ions come from the inter-ion Coulomic correlation? (3) What are the roles of the specific-site binding of Mg^{2+} in RNA tertiary binding [212]? More issues related to RNA ion electrostatics includes: (1) What are the role of ions in RNA-ligand interaction? (2) What is the role of ions in RNA-protein interaction? To answer the questions requires the further development of theoretical modelling [209–211], combined with the progress in experiments.

Most above introduced progress in RNA folding problem were obtained for in vitro systems, while in cells, RNAs are surrounded by many other macromolecules. Therefore, in reality, RNA folds in a possibly interactive and dynamic confined space [213–216]. Limited existed investigations have revealed that the spatial confinement may significantly influence the folded structure and the ion role in folding [179, 214–216]. Further investigations on RNA folding should also involve the complex effects from the other macromolecules in vivo.

References

1. Bloomfield VA, Crothers DM, Tinoco IJ (2000) Nucleic acids: structure, properties and functions. University Science Books, Sausalito
2. Walter NG, Woodson SA, Batey RT (eds) (2009) Non-coding RNA. Non-protein coding RNAs. Springer, Berlin

3. Cruz JA, Westhof E (2009) The dynamic landscapes of RNA architecture. Cell 136:604–609
4. Zuker M, Stiegler P (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 9:133–148
5. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, Schuster P (1994) Fast folding and comparison of RNA secondary structures. Monatsh Chem 125:167–188
6. Ding Y, Lawrence CE (2003) A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res 31:7280–7301
7. Do CB, Woods DA, Batzoglou S (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics 23:90–98
8. Mathews DH, Turner DH (2006) Prediction of RNA secondary structure by free energy minimization. Curr Opin Struct Biol 16:270–278
9. Massire C, Westhof E (1998) MANIP: an interactive tool for modelling RNA. J Mol Graph Model 16(197–205):255–257
10. Zwieb C, Mueller F (1997) Three-dimensional comparative modeling of RNA. Nucleic Acids Symp Ser 36:69–71
11. Jossinet F, Westhof E (2005) Sequence to Structure (S2S): display, manipulate and interconnect RNA data from sequence to structure. Bioinformatics 21:3320–3321
12. Jossinet F, Ludwig TE, Westhof E (2010) Assemble: an interactive graphical tool to analyze and build RNA architectures at the 2D and 3D levels. Bioinformatics 26:2057–2059
13. Martinez HM, Maizel JV, Shapiro BA (2008) RNA2D3D: a program for generating, viewing, and comparing 3-dimensional models of RNA. J Biomol Struct Dyn 25:669–683
14. Rother M, Rother K, Puton T, Bujnicki JM (2011) ModeRNA: a tool for comparative modeling of RNA 3D structure. Nucleic Acids Res 1–16
15. Flores SC, Wan YQ, Russell R, Altman RB (2010) Predicting RNA structure by multiple template homology modeling. Pac Symp Biocomput 15:216–227
16. Paliy M, Melnik R, Shapiro BA (2010) Coarse-graining RNA nanostructures for molecular dynamics simulations. Phys Biol 7:036001
17. Tan RKZ, Petrov AS, Harvey SC (2006) YUP: molecular simulation program for coarse-grained and multiscaled models. J Chem Theory Comput 2:529–540
18. Jonikas MA, Radmer RJ, Laederach A, Das R, Pearlman S, Herschlag D, Altman RB (2009) Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. RNA 15:189–199
19. Jonikas MA, Radmer RJ, Altman RB (2009) Knowledge-based instantiation of full atomic detail into coarse-grain RNA 3D structural models. Bioinformatics 25:3259–3266
20. Taxilaga-Zetina O, Pliego-Pastrana P, Carbajal-Tinoco MD (2010) Three-dimensional structures of RNA obtained by means of knowledge-based interaction potentials. Phys Rev E 81:041914
21. Cao S, Chen SJ (2011) Physics-based de novo prediction of RNA 3D structures. J Phys Chem B 115:4216–4226
22. Cao S, Chen SJ (2006) Predicting RNA pseudoknot folding thermodynamics. Nucleic Acids Res 34:2634–2652
23. Cao S, Chen SJ (2005) Predicting RNA folding thermodynamics with a reduced chain representation model. RNA 11:1184–1897
24. Ding F, Sharma S, Chalasani P, Demidov VV, Broude NE, Dokholyan NV (2008) Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. RNA 14:1164–1173
25. Sharma S, Ding F, Dokholyan NV (2008) iFoldRNA: three-dimensional RNA structure prediction and folding. Bioinformatics 24:1951–1952
26. Ding F, Lavender CA, Weeks KM, Dokholyan NV (2012) Three-dimensional RNA structure refinement by hydroxyl radical probing. Nat Methods 9:603–608
27. Xia Z, Gardner DP, Gutell RR, Ren P (2010) Coarse-grained model for simulation of RNA three-dimensional structures. J Phys Chem B 114:13497–13506

28. Pasquali S, Derreumaux P (2010) HiRE-RNA: a high resolution coarse-grained model for RNA. J Phys Chem B 114:11957–11966
29. Zhang J, Dundas J, Lin M, Chen M, Wang W, Liang J (2009) Prediction of geometrically feasible three-dimensional structures of pseudoknotted RNA through free energy estimation. RNA 15:2248–2263
30. Zhang J, BianYQ Wang W (2012) RNA fragment modeling with a nucleobase discrete-state model. Phys Rev E 85:021909
31. Das R, Baker D (2007) Automated de novo prediction of native-like RNA tertiary structures. Proc Natl Acad Sci 104:14664–14669
32. Das R, Karanicolas J, Baker D (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. Nat Methods 7:291–294
33. Bida JP, Maher LJ III (2012) Improved prediction of RNA tertiary structure with insights into native state dynamics. RNA 18:385–393
34. Parisien M, Major F (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. Nature 452:51–55
35. Lemieux S, Major F (2006) Automated extraction and classification of RNA tertiary structure cyclic motifs. Nucleic Acids Res 34:2340–2346
36. Zhao YJ, Gong Z, Xiao Y (2011) Improvement of the hierarchical approach for predicting RNA tertiary structure. J Biomol Struct Dyn 28:815–826
37. Gong Z, Zhao Y, Xiao Y (2010) RNA stability under different combinations of amber force fields and solvation models. J Biomol Struct Dyn 28(3):431–441
38. Seetin MJ, Mathews DH (2011) Automated RNA tertiary structure prediction from secondary structure and low-resolution restraints. J Comput Chem 32:2232–2244
39. Baumstark T, Schroder AR, Riesner D (1997) Viroid processing: switch from cleavage to ligation is driven by a change from a tetraloop to a loop E conformation. EMBO J 16:599–610
40. Perrotta AT, Been MD (1998) A toggle duplex in hepatitis delta virus self-cleaving RNA that stabilizes an inactive and a salt-dependent pro-active ribozyme conformation. J Mol Biol 279:361–373
41. Schultes EA, Bartel DP (2000) One sequence, two ribozymes: implications for the emergence of new ribozyme folds. Science 289:448–452
42. Kruger K, Grabowski P, Zaug AJ, Sands J, Gottschling DE, Cech TR (1982) Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*. Cell 31:147–157
43. Bartel DP, Szostak JW (1993) Isolation of new ribozymes from a large pool of random sequences. Science 261:1411–1418
44. Joyce GF (1989) Amplication, mutation and selection of catalytic RNA. Gene 82:83–87
45. Ellington AE, Szostak JW (1990) In vitro selection of RNA molecules that bind specific ligands. Nature 346:818–822
46. Tuerk C, Gold L (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science 249:505–510
47. Mironov AS, Gusarov I, Rafikov R, Lopez LE, Shatalin K, Kreneva RA, Perumov DA, Nudler E (2002) Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. Cell 111:747–756
48. Nahvi A, Sudarsan N, Ebert MS, Zou X, Brown KL, Breaker RR (2002) Genetic control by metabolite binding mRNA. Chem Biol 9:1043–1049
49. Winkler WC, Breaker RR (2003) Genetic control by metabolite-binding riboswitches. Chem Bio Chem 4:1024–1032
50. Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism and function. Cell 116:281–297
51. Nudler E, Mironov AS (2004) The riboswitch control of bacterial metabolism. Trends Biochem Sci 29:11–17
52. He L, Hannon GJ (2004) MicroRNAs: small RNAs with a big role in gene regulation. Nature Rev Genet 5:522–531

53. Winkler WC, Nahvi A, Roth A, Collins JA, Breaker RR (2004) Control of gene expression by a natural metabolite-responsive ribozyme. Nature 428:281–286
54. Gerdes K, Wagner EGH (2007) RNA antitoxins. Curr Opin Microbiol 10:117
55. Nagel JHA, Gultyaev AP, Gerdes K, Pleij CWA (1999) Metastable structures and refolding kinetics in hok mRNA of plasmid R1. RNA 5:1408–1419
56. Groeneveld H, Thimon K, Duin J (1995) Translational control of maturation-protein synthesis in phage MS2: a role for the kinetics of RNA folding? RNA 1:79–88
57. Porschke D (1977) Elementary steps of base recognition and helix-coil transitions in nucleic acids. Mol Biol Biochem Biophys 24:191–218
58. Craig ME, Crothers DM, Doty P (1971) Relaxation kinetics of dimer formation by self complementary oligonucleotides. J Mol Biol 62:383–401
59. Crothers DM, Cole PE, Hilbers CW, Shulman RG (1974) The molecular mechanism of thermal unfolding of *Escherichia coli* formylmethionine transfer RNA. J Mol Biol 87:63–88
60. Micura R, Hobartner C (2003) On secondary structure rearrangements and equilibria of small RNAs. Chem Biochem 4:984–990
61. Furtig B, Buck J, Manoharan V, Bermel W, Jaschke A, Wenter P, Pitsch S, Schwalbe H (2007) Time-resolved NMR studies of RNA folding. Biopolymers 86:360–383
62. Harlepp S, Marchal T, Robert J, Leger J, Xayaphoummine A, Isambert H, Chatenay D (2003) Probing complex RNA structures by mechanical force. Eur Phys J E-Soft Matter 12:605–615
63. Jean JM, Hall KB (2001) 2-Aminopurine fluorescence quenching and lifetimes: role of base stacking. Proc Natl Acad Sci USA 98:37–41
64. Liphardt J, Onoa B, Smith SB, Tinoco IJ, Bustamante C (2001) Reversible unfolding of single RNA molecules by mechanical force. Science 292:733–737
65. Bonnet G, Krichevsky O, Libchaber A (1998) Kinetics of conformational fluctuations in DNA hairpin-loops. Proc Natl Acad Sci USA 95:8602–8606
66. Ansari A, Kunznetsov SV, Shen Y (2001) Configurational diffusion down a folding funnel describes the dynamics of DNA hairpins. Proc Natl Acad Sci USA 98:7771–7776
67. Wallace MI, Ying L, Balasubramanian S, Klenerman D (2001) Non-arrhenius kinetics for the loop closure of a DNA hairpin. Proc Natl Acad Sci USA 98:5584–5589
68. Bai Y, Das R, Millett IS, Herschlag D, Doniach S (2005) Probing counterion modulated repulsion and attraction between nucleic acid duplexes in solution. Proc Natl Acad Sci USA 102:1035–1040
69. Chu VB, Herschlag D (2008) Unwinding RNA's secrets: advances in the biology, physics, and modeling of complex RNAs. Curr Opin Struct Biol 18:305–314
70. Draper DE (2008) RNA folding: thermodynamic and molecular descriptions of the roles of ions. Biophys J 95:5489–5495
71. Chen SJ (2008) RNA folding: conformational statistics, folding kinetics, and ion electrostatics. Annu Rev Biophys 37:197–214
72. Tan ZJ, Chen SJ (2011) Importance of diffuse ions binding to RNA. Met Ions Life Sci 9:101–124
73. Bowman JC, Lenz TK, Hud NV, Williams LD (2012) Cations in charge: magnesium ions in RNA folding and catalysis. Curr Opin Struct Biol 22:262–272
74. Cruz JA et al (2012) RNA-puzzles: A CASP-like evaluation of RNA three-dimensional structure prediction. RNA 18:610–625
75. Shapiro BA, Yingling YG, Kasprzak W, Bindewald E (2007) Bridging the gap in RNA structure prediction. Curr Opn Struct Biol 17:157–165
76. Hajdin CE, Ding F, Dokholyan NV, Weeks KM (2010) On the significance of an RNA tertiary structure prediction. RNA 16:1340–1349
77. Laing C, Schlick T (2011) Computational approaches to RNA structure prediction, analysis, and design. Curr Opin Struct Biol 21:1–13
78. Laing C, Schlick T (2010) Computational approaches to 3D modeling of RNA. J Phys Condens Matter 22:283101

79. Rother K, Rother M, Boniecki M, Puton T, Bujnicki JM (2011) RNA and protein 3D structure modeling: similarities and differences. J Mol Model 17:2325–2336
80. Levitt M (1969) Detailed molecular model for transfer ribonucleic acid. Nature 224:759–763
81. Chothia C, Gerstein M (1997) How far can sequences diverge? Nature 385:579–581
82. Anfinsen CB (1973) Principles that govern the folding of protein chains. Science 181:223–230
83. Baker D, Sali A (2001) Protein structure prediction and structural genomics. Science 294:93–96
84. Zhang Y, Skolnick J (2004) Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci USA 101:7594–7599
85. Simons KT, Kooperberg C, Huang E (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. J Mol Biol 268:209–225
86. Zhang WB, Chen SJ (2002) RNA hairpin-folding kinetics. Proc Natl Acad Sci USA 99:1931–1936
87. Zhang WB, Chen SJ (2003) Master equation approach to finding the rate-limiting steps in biopolymer folding. J Chem Phys 118:3413
88. Konishi Y, Ooi T, Scheraga HA (1982) Regeneration of ribonuclease a from the reduced protein rate-limiting steps. Biochemistry 21:4734–4740
89. Zhang WB, Chen SJ (2003) Analyzing the biopolymer folding rates and pathways using kinetic cluster method. J Chem Phys 119:8716–8729
90. Flamm C, Fontana W, Hofacker IL, Schuster P (2000) RNA folding at elementary step resolution. RNA 6:325–338
91. Isambert H, Siggia ED (2000) Modeling RNA folding paths with pseudoknots: application to hepatitis d-virus ribozyme. Proc Natl Acad Sci USA 97:6515–6520
92. Danilova LV, Pervouchine DD, Favorov AV, Mironov AA (2006) RNA kinetics: a web server that models secondary structure kinetics of an elongating RNA. J Bioinf Comp Biol 4:589–596
93. Martinez HM (1984) An RNA folding rule. Nucleic Acids Res 12:323–335
94. Gillespie DT (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J Comput Phys 22:403–434
95. Schmitz M, Steger G (1996) Discription of RNA folding by simulated annealing. J Mol Biol 255:254–266
96. Gultyaev AP, Batenburg FH, Pleij CW (1995) The computer-simulation of RNA folding pathways using a genetic algorithm. J Mol Biol 250:37–51
97. Mironov A, Kister A (1985) A kinetic approach to the prediction of RNA secondary structures. J Biomol Struct Dyn 2:953–962
98. Mironov AA, Lebedev VF (1993) A kinetic model of RNA folding. Biosystems 30:49–56
99. Isambert H, Siggia ED (2000) Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme. Proc Natl Acad Sci USA 97:6515–6520
100. Danilova LV, Pervoud DD, Favorov AA, Mironov AA (2006) RNAKinetics: a web server that models secondary structure kinetics of an elongating RNA. J Bioinform Comput Biol 4:589–596
101. Ndifon W (2005) A complex adaptive systems approach to the kinetic folding of RNA. Biosystems 82:257–265
102. Zhao PN, Zhang WB, Chen SJ (2010) Predicting secondary structural folding kinetics for nucleic acids. Biophys J 98:1617–1625
103. Tang X, Thomas S, Tapia L, Giedroc DP, Amato NM (2008) Simulating RNA folding kinetics on approximated energy landscapes. J Mol Biol 381:1055–1067
104. Hofacker IL, Flamm C, Heine C, Wolfinger MT, Scheuermann G, Stadler PF (2010) BarMap: RNA folding on dynamic energy landscapes. RNA 16:1308–1316

105. Geis M, Flamm C, Wolfinger MT, Tanzer A, Hofacker IL, Middendorf M, Mandl C, Stadler PF, Thurner C (2008) Enhancement of transactivation activity of Rta of Epstein-Barr virus by RanBPM. J Mol Biol 379:242–261
106. Flamm C, Hofacker IL, Stadler PF, Wolfinger MT (2002) Barrier trees of degenerate landscapes. Z Phys Chem 216:155–173
107. Wolfinger MT, Svrcek-Seiler WA, Flamm C, Hofacker IL, Stadler PF (2004) Efficient computation of RNA folding dynamics. J Phys A Math Gen 37:4731–4741
108. Tang X, Kirkpatrick B, Thomas S, Song G, Amato NM (2005) Using motion planning to study RNA folding kinetics. J Comp Biol 12:862–881
109. Zhang WB, Chen SJ (2006) Exploring the complex folding kinetics of RNA hairpins: I. general folding kinetics analysis. Biophys J 90:765–777
110. Tacker M, Fontana W, Stadler PF, Schuster P (1994) Statistics of RNA melting kinetics. Eur Biophys J 23:29
111. Suvernev AA, Frantsuzov PA (1995) Statistical description of nucleic acid secondary structure folding. J Biomol Struct Dyn 13:135–144
112. Jacob C, Breton N, Daegelen P, Peccoud J (1997) Probability distribution of the chemical states of a closed system and thermodynamic law of mass action from kinetics: the RNA example. J Chem Phys 107:2913
113. Isambert H, Siggia ED (2000) Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme. Proc Natl Acad Sci USA 97:6515–6520
114. Xia TB, SantaLucia J, Turner DH (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson–Crick base pairs. Biochemistry 37:14719–14735
115. Mathews DH, Sabina J, Zuker M, Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 288:911–940
116. Morgan SR, Higgs PG (1998) Barrier heights between ground states in a model of RNA secondary structure. J Phys A Math Gen 31:3153–3170
117. Henkin TM, Yanofsky C (2002) Regulation by transcription attenuation in bacteria: how RNA provides instructions for transcription termination/antitermination decisions. BioEssays 24:700–707
118. Merino E, Yanofsky C (2005) Transcription attenuation: a highly conserved regulatory strategy used by bacteria. Trends Genet 21:260–264
119. Franch T, Gultyaev AP, Gerder K (1997) Programmed cell death by hok/sok of plasmid R1: processing at the hok mRNA 3′-end triggers structural rearrangements that allow translation and antisense RNA binding. J Mol Biol 273:38–51
120. Heilman-Miller SL, Woodson SA (2003) Effect of transcription on folding of the *Tetrahymena* ribozyme. RNA 9:722–733
121. Brehm SL, Cech TR (1983) The fate of an intervening sequence RNA: excision and cyclization of the *Tetrahymena* ribosomal RNA intervening sequence in vivo. Biochemistry 22:2390–2397
122. Zhang F, Ramsay ES, Woodson SA (1995) In vivo facilitation of *Tetrahymena* group I intron splicing in *Escherichia coli* pre-ribosomal RNA. RNA 1:284–292
123. Treiber DK, Williamson JR (2001) Beyond kinetic traps in RNA folding. Curr Opin Struct Biol 11:309–314
124. Woodson SA (2002) Folding mechanisms of group I ribozymes: role of stability and contact order. Biochem Soc Trans 30:1166–1169
125. Zhang LB, Bao P, Michael JL, Zhang Y (2009) Slow formation of a pseudoknot structure is rate limiting in the productive co-transcriptional folding of the self-splicing Candida intron. RNA 15:1986–1992
126. Pan T, Artsimovitch I, Fang X, Landick R, Sosnick TR (1999) Folding of a large ribozyme during transcription and the effect of the elongation factor NusA. Proc Natl Acad Sci USA 96:9545–9550

127. Zhao PN, Zhang WB, Chen SJ (2011) Cotranscriptional folding kinetics of ribonucleic acid secondary structures. J Chem Phys 135:245101
128. Das R, Mills TT, Kwok LW, Maskel GS, Millett IS, Doniach S, Finkelstein KD, Herschlag D, Pollack L (2003) Counterion distribution around DNA probed by solution X-ray scattering. Phys Rev Lett 90:188103
129. Andresen K, Qiu X, Pabit SA, Lamb JS, Park HY, Kwok LW, Pollack L (2008) Mono- and trivalent ions around DNA: a small-angle scattering study of competition and interactions. Biophys J 95:287–295
130. Pabit SA, Qiu X, Lamb JS, Li L, Meisburger SP, Pollack L (2009) Both helix topology and counterion distribution contribute to the more effective charge screening in dsRNA compared with dsDNA. Nucleic Acids Res 37:3887–3896
131. Kirmizialtin S, Pabit SA, Meisburger SP, Pollack L, Elber R (2012) RNA and its ionic cloud: solution scattering experiments and atomically detailed simulations. Biophys J 102:819–828
132. Bai Y, Greenfeld M, Travers KJ, Chu VB, Lipfert J, Doniach S, Herschlag D (2007) Quantitative and comprehensive decomposition of the ion atmosphere around nucleic acids. J Am Chem Soc 129:14981–14988
133. Krakauer H (1971) The binding of Mg^{++} ions to polyadenylate, polyuridylate, and their complexes. Biopolymers 10:2459–2490
134. Clement RM, Sturm J, Daune MP (1973) Interaction of metallic cations with DNA VI. Specific binding of Mg^{2+} and Mn^{2+}. Biopolymers 12:405–421
135. Soto M, Misra V, Draper DE (2007) Tertiary structure of an RNA pseudoknot is stabilized by "diffuse" Mg^{2+} ions. Biochemistry 46:2973–2983
136. Grilley D, Misra V, Caliskan G, Draper DE (2007) Importance of partially unfolded conformations for Mg^{2+}-induced folding of RNA tertiary structure: structural models and free energies of Mg^{2+} interactions. Biochemistry 46:10266–10278
137. Rialdi G, Levy J, Biltonen R (1972) Thermodynamic studies of transfer ribonucleic acids. I. Magnesium binding to yeast phenylalanine transfer ribonucleic acid. Biochemistry 11:2472–2479
138. Romer R, Hach R (1975) tRNA conformation and magnesium binding. A study of a yeast phenylalanine-specific tRNA by a fluorescent indicator and differential melting curves. Eur J Biochem 55:271–284
139. Stellwagen E, Dong Q, Stellwagen NC (2007) Quantitative analysis of monovalent counterion binding to random-sequence, double-stranded DNA using the replacement ion method. Biochemistry 46:2050–2058
140. Smith SB, Cui YJ, Bustamante C (1996) Overstretching B-DNA: The elastic response of individual double-stranded and single-stranded DNA molecules. Science 271:795–799
141. Murphy MC, Rasnik I, Cheng W, Lohman TM, Ha T (2004) Probing single-stranded DNA conformational flexibility using fluorescence spectroscopy. Biophys J 86:2530–2537
142. Tinland B, Pluen A, Sturm J, Weill G (1997) Persistende length of single-stranded DNA. Macromolecules 30:5763–5765
143. McIntosh DB, Saleh O (2011) Slat species-dependent electrostatic effects on ssDNA elasticity. Macromolecules 44:2328–2333
144. Sim AYL, Lipfert J, Herschlag D, Doniach S (2012) Salt dependence of the radius of gyration and flexibility of single-stranded DNA in solution probed by small-angle x-ray scattering. Phys Rev E 86:021901
145. Chen H, Meisburger SP, Pabit SA, Sutton JL, Webb WW, Pollack L (2012) Ionic strength-dependent persistence lengths of single-stranded RNA and DNA. Proc Natl Acad Sci USA 109:799–804
146. Bizarro CV, Alemany A, Ritort F (2012) Non-specific binding of Na^{+} and Mg^{2+} to RNA determined by force spectroscopy methods. Nucleic Acids Res 40:6922–6935
147. Tan ZJ, Chen SJ (2006) Electrostatic free energy landscape for nucleic acid helix assembly. Nucleic Acids Res 34:6629–6639

148. Williams P, Longfellow CE, Freier SM, Kierzek R, Turner DH (1989) Laser temperature-jump, spectroscopic, and thermodynamic study of salt effects on duplex formation by dGCATGC. Biochemistry 28:4283–4291
149. Nakano S, Fujimoto M, Hara H, Sugimoto N (1999) Nucleic acid duplex stability: influence of base composition on cation effects. Nucleic Acids Res 27:2957–2965
150. Serra MJ, Baird JD, Dale T, Fey BL, Retatagos K, Westhof E (2002) Effects of magnesium ions on the stabilization of RNA oligomers of defined structures. RNA 8:307–323
151. Owczarzy R, Moreira BG, You Y, Behlke MA, Walder JA (2008) Predicting stability of DNA duplexes in solutions containing magnesium and monovalent cations. Biochemistry 47:5336–5353
152. Kuznetsov SV, Ren CC, Woodson SA, Ansari A (2008) Loop dependence of the stability and dynamics of nucleic acid hairpins. Nucleic Acids Res 36:1098–1112
153. Vieregg J, Cheng W, Bustamante C, Tinoco I Jr (2007) Measurement of the effect of monovalent cations on RNA hairpin stability. J Am Chem Soc 129:14966–14973
154. Tan ZJ, Chen SJ (2006) Nucleic acid helix stability: effects of salt concentration, cation valency and size, and chain length. Biophys J 90:1175–1190
155. Tan ZJ, Chen SJ (2007) RNA helix stability in mixed Na^+/Mg^{2+} solution. Biophys J 92:3615–3632
156. Tan ZJ, Chen SJ (2008) Salt dependence of nucleic acid hairpin stability. Biophys J 95:738–752
157. Nixon PL, Giedroc DP (1998) Equilibrium unfolding (folding) pathway of a model H-type pseudoknotted RNA: the role of magnesium ions in stability. Biochemistry 37:16116–16129
158. Stellwagen E, Muse JM, Stellwagen NC (2011) Monovalent cation size and DNA conformational stability. Biochemistry 50:3084–3094
159. Anthony PC, Sim AY, Chu VB, Doniach S, Block SM, Herschlag D (2012) Electrostatics of nucleic acid folding under conformational constraint. J Am Chem Soc 134:4607–4614
160. SantaLucia JJ (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proc Natl Acad Sci USA 95:1460–1465
161. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31:3406–3415
162. Chen SJ, Dill KA (2000) RNA folding energy landscapes. Proc Natl Acad Sci USA 97:646–651
163. SantaLucia J, Hicks D (2004) The thermodynamics of DNA structural motifs. Annu Rev Biophys Biomol Struct 33:415–440
164. Zhang WB, Chen SJ (2006) Exploring the complex folding kinetics of RNA hairpins: II. Effect of sequence, length, and misfolded states. Biophys J 90:778–787
165. Theimer A, Giedroc DP (2000) Contribution of the intercalated adenosine at the helical junction to the stability of the gag-pro frameshifting pseudoknot from mouse mammary tumor virus. RNA 6:409–421
166. Koculi E, Hyeon C, Thirumalai D, Woodson SA (2007) Charge density of divalent metal cations determines RNA stability. J Am Chem Soc 129:2676–2682
167. Takamoto K, He Q, Morris S, Chance MR, Brenowitz M (2002) Monovalent cations mediate formation of native tertiary structure of the *Tetrahymena* thermophila ribozyme. Nature Struct Biol 9:928–933
168. Moghaddam S, Caliskan G, Chauhan S, Hyeon C, Briber RM, Thirumalai D, Woodson SA (2009) Metal ion dependence of cooperative collapse transitions in RNA. J Mol Biol 393:753–764
169. Heilman-Miller SL, Thirumalai D, Woodson SA (2001) Role of counterion condensation in folding of the *Tetrahymena* ribozyme. I. Equilibrium stabilization by cations. J Mol Biol 306:1157–1166
170. Lambert D, Leipply D, Shiman R, Draper DE (2009) The influence of monovalent cation size on the stability of RNA tertiary structures. J Mol Biol 390:791–804

171. Walter NG, Burke JM, Millar DP (1999) Stability of hairpin ribozyme tertiary structure is governed by the interdomain junction. Nature Struct Biol 6:544–549
172. Pljevaljcic G, Millar DP, Deniz AA (2004) Freely diffusing single hairpin ribozymes provide insights into the role of secondary structure and partially folded states in RNA folding. Biophys J 87:457–467
173. Leipply D, Draper DE (2011) Evidence for a thermodynamically distinct Mg^{2+} ion associated with formation of an RNA tertiary structure. J Am Chem Soc 133:13397–13405
174. Weixlbaumer A, Werner A, Flamm C, Westhof E, Schroeder R (2004) Determination of thermodynamic parameters for HIV DIS type loop-loop kissing complexes. Nucleic Acids Res 32:5126–5133
175. Lorenz C, Piganeau N, Schroeder R (2006) Stabilities of HIV-1 DIS type RNA loop-loop interactions in vitro and in vivo. Nucleic Acids Res 34:334–342
176. Vander Meulen KA, Butcher SE (2012) Characterization of the kinetic and thermodynamic landscape of RNA folding using a novel application of isothermal titration calorimetry. Nucleic Acids Res 40:2140–2151
177. Tan ZJ, Chen SJ (2010) Predicting ion binding properties for RNA tertiary structures. Biophys J 99:1565–1576
178. Tan ZJ, Chen SJ (2011) Salt contribution to RNA tertiary structure folding stability. Biophys J 101:176–187
179. Tan ZJ, Chen SJ (2012) Ion-mediated RNA structural collapse: effect of spatial confinement. Biophys J 103:827–836
180. Rau DC, Parsegian VA (1992) Direct measurement of the intermolecular forces between counterion-condensed DNA double helices. Evidence for long range attractive hydration forces. Biophys J 61:246–259
181. Rau DC, Parsegian VA (1992) Direct measurement of temperature-dependent solvation forces between DNA double helices. Biophys J 61:260–271
182. Bai Y, Chu VB, Lipfert J, Pande VS, Herschlag D, Doniach S (2008) Critical assessment of nucleic acid electrostatics via experimental and computational investigation of an unfolded state ensemble. J Am Chem Soc 130:12334–12341
183. Qiu X, Andresen K, Kwok LW, Lamb JS, Park HY, Pollack L (2007) Inter-DNA attraction mediated by divalent counterions. Phys Rev Lett 99:038104
184. Qiu X, Parsegian VA, Rau DC (2010) Divalent counterion-induced condensation of triple-strand DNA. Proc Natl Acad Sci USA 107:21482–21486
185. Li L, Pabit SA, Meisburger SP, Pollack L (2011) Double-stranded RNA resists condensation. Phys Rev Lett 106:108101
186. Tan ZJ, Chen SJ (2009) Predicting electrostatic force in RNA folding. Methods Enzymol 469:465–487
187. Manning GS (1978) The molecular theory of polyelectrolyte solutions with applications to the electrostatic properties of polynucleotides. Q Rev Biophys 11:179–246
188. Schurr MJ (2009) Polyanion models of nucleic acid-metal ion interactions. In: Hud NV (ed) Nucleic acid-metal ion interactions. Royal Society of Chemistry, London, pp 307–344
189. Ray J, Manning GS (2000) Formation of loose clusters in polyelectrolyte solutions. Macromolecules 33:2901–2908
190. Lyubartsev P, Nordenskiold L (1995) Monte Carlo simulation study of ion distribution and osmotic pressure in hexagonally oriented DNA. J Phys Chem 99:10373–10382
191. Dai L, Mu Y, Nordenskiöld L, van der Maarel JR (2008) Molecular dynamics simulation of multivalent-ion mediated attraction between DNA molecules. Phys Rev Lett 100:118301
192. Gilson MK, Sharp KA, Honig B (1987) Calculating the electrostatic potential of molecules in solution: method and error assessment. J Comput Chem 9:327–335
193. Boschitsch H, Fenley MO (2007) A new outer boundary formulation and energy corrections for the nonlinear Poisson-Boltzmann equation. J Comput Chem 28:909–921
194. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA (2000) Electrostatics of nanosystems: application to microtubules and the ribosome. Proc Natl Acad Sci USA 98:10037–10041

195. Zhou YC, Feig M, Wei GW (2008) Highly accurate biomolecular electrostatics in continuum dielectric environments. J Comput Chem 29:87–97
196. Lu B, Cheng X, Huang J, McCammon JA (2010) AFMPB: an adaptive fast multipole Poisson-Boltzmann solver for calculating electrostatics in biomolecular systems. Comput Phys Commun 181:1150–1160
197. Misra VK, Shiman R, Draper DE (2003) A thermodynamic framework for the magnesium-dependent folding of RNA. Biopolymers 69:118–136
198. Tan ZJ, Chen SJ (2005) Electrostatic correlations and fluctuations for ion binding to a finite length polyelectrolyte. J Chem Phys 122:044903
199. Chu VB, Bai Y, Lipfert J, Herschlag D, Doniach S (2007) Evaluation of ion binding to DNA duplexes using a size-modified Poisson-Boltzmann theory. Biophys J 93:3202–3209
200. Kirmizialtin S, Silalahi AR, Elber R, Fenley MO (2012) The ionic atmosphere around A-RNA: Poisson-Boltzmann and molecular dynamics simulations. Biophys J 102:829–838
201. Gavryushov S (2008) Electrostatics of B-DNA in NaCl and CaCl2 solutions: ion size, interionic correlation, and solvent dielectric saturation effects. J Phys Chem B 112:8955–8965
202. Grochowski P, Trylska J (2008) Continuum molecular electrostatics, salt effects and counterion binding. A review of the Poisson-Boltzmann theory and its modifications. Biopolymers 89:93–113
203. Forsman J (2004) A simple correlation-corrected Poisson-Boltzmann theory. J Phys Chem B 108:9236–9245
204. Vlachy V (1999) Ionic effect beyond Poisson-Boltzmann theory. Annu Rev Phys Chem 50:145–165
205. Wang K, Yu YX, Gao GH (2008) Density functional study on the structural and thermodynamic properties of aqueous DNA-electrolyte solution in the framework of cell model. J Chem Phys 128:185101
206. Chen YG, Weeks JD (2006) Local molecular field theory for effective attractions between like charged objects in systems with strong Coulomb interactions. Proc Natl Acad Sci USA 103:7560–7565
207. Tan ZJ, Chen SJ (2006) Ion-mediated nucleic acid helix-helix interactions. Biophys J 91:518–536
208. Tan ZJ, Chen SJ (2008) Electrostatic free energy landscapes for DNA helix bending. Biophys J 94:3137–3149
209. Chen G, Tan ZJ, Chen SJ (2010) Salt-dependent folding energy landscape of RNA three-way junction. Biophys J 98:111–120
210. Chen G, Chen SJ (2011) Quantitative analysis of the ion-dependent folding stability of DNA triplexes. Phys Biol 8:066006
211. He Z, Chen SJ (2012) Predicting ion-nucleic acid interactions by energy landscape-guided sampling. J Chem Theo Compt 8:2095–2102
212. Cate JH, Gooding AR, Podell E, Zhou K, Golden BL, Kundrot CE, Cech TR, Doudna JA (1996) Crystal structure of a group I ribozyme domain: principles of RNA packing. Science 273:1678–1685
213. Zhou HX, Rivas G, Minton AP (2008) Macromolecular crowding and confinement: biochemical, biophysical, and potential physiological consequences. Annu Rev Biophys 37:375–397
214. Lambert D, Leipply D, Draper DE (2010) The osmolyte TMAO stabilizes native RNA tertiary structures in the absence of Mg^{2+}: evidence for a large barrier to folding from phosphate dehydration. J Mol Biol 404:138–157
215. Pincus DL, Hyeon C, Thirumalai D (2008) Effects of trimethylamine N-oxide (TMAO) and crowding agents on the stability of RNA hairpins. J Am Chem Soc 130:7364–7372
216. Kilburn D, Roh JH, Guo L, Briber RM, Woodson SA (2010) Molecular crowding stabilizes folded RNA structure by the excluded volume effect. J Am Chem Soc 132:8690–8696

Part III
The Interactions Between Biological Macromolecules and Ligands

Chapter 12
Binding Modes and Interaction Mechanism Between Different Base Pairs and Methylene Blue Trihydrate: A Quantum Mechanics Study

Peijun Xu, Jinguang Wang, Yong Xu, Huiying Chu, Hujun Shen, Depeng Zhang, Meixia Zhao, Jiahui Liu and Guohui Li

Abstract Different quantum mechanic methods have been evaluated for the calculation of binding modes and interactions between intercalators with different DNA base pairs by comparing with the results of MP2, which is very expensive, indicating that WB97XD method under 6-311+G* basis set is able to efficiently reproduce MP2 results. We discovered that the methylene blue trihydrate intercalated into the DNA base pairs, and DNA intercalation increased the distance between DNA base pairs, depending on the types of DNA bases. According to the binding energy results, it was found that the intercalation of methylene blue trihydrate into AA-TT base pair was more favorable in the orientation of nitrogen than other directions and intercalation, and the electric charge was transferred from

Huiying Chu, Jinguang Wang and Peijun Xu have been contributed equally to this paper.

P. Xu · D. Zhang · M. Zhao · J. Liu
School of Physics and Electronic Technology, Liaoning Normal University, Dalian, Liaoning, China

J. Wang
The First Affiliated Hospital, Dalian Medical University, Dalian, China

Y. Xu
Guangzhou Institute of Biomedicine and Health, Guangzhou, China

H. Chu · H. Shen · G. Li (✉)
Laboratory of Molecular Modeling and Design, State Key Laboratory of Molecular Reaction Dynamics, Dalian Institute of Chemical Physics, Chinese Academy of Science, Dalian, Liaoning, China
e-mail: ghli@dicp.ac.cn

D. Wei et al. (eds.), *Advance in Structural Bioinformatics*, Advances in Experimental Medicine and Biology 827, DOI 10.1007/978-94-017-9245-5_12

methylene blue trihydrate to the AA-TT base pair. The analysis of change in the charge density shows that changes often take place in the heavy atom in the middle of the system which the charge density changes most remarkable.

Keywords DNA base pairs · Intercalators methylene blue trihydrate · Quantum mechanic methods · Charge density

12.1 Introduction

DNA is an informational molecule encoding the genetic instructions used in the development and functioning of all known living organisms and many viruses, therefore there are many researches on it [1–3]. Small molecules interact with DNA or RNA through the several modes: minor groove binding, major groove binding, intercalation and other types of binding [4]. When a small molecule inserts into the two adjacent base pairs of a DNA strand, sandwich-like structure is formed, which is called DNA intercalation [5, 6].

A small molecule is called DNA intercalator if it can insert into the base pairs and contains planar, polycyclic and aromatic conformations [4]. The interaction between DNA and its intercalator is usually considered as stacking interaction [6–8]. Up to now, intercalators of DNA are used in chemotherapy such as ethidium, proflavine and anthracycline etc., understanding their interactions with nucleic acids is of benefit to many fields—especially for medicinal chemistry, where it may aid the rational design of novel drugs and make it possible to govern their behavior, for instance, by triggering intercalative capabilities [5, 9]. The derivatives and their specificities for certain nucleic acid sequences make intercalating agents especially useful as nucleic acid dyes [10]. There have been numerous experimental and theoretical studies carried out on intercalation systems, especially the antitumor drugs and antiseptics molecules [11].

Methylene blue trihydrate (Fig. 12.1) has a variety of biomedical and biologically therapeutic applications, and used widely as a dye and therapeutic agent [12, 13]. Methylene blue trihydrate has been also used in the detection of the environmental pollutant in the experimental research as a dye. So far, the calculations

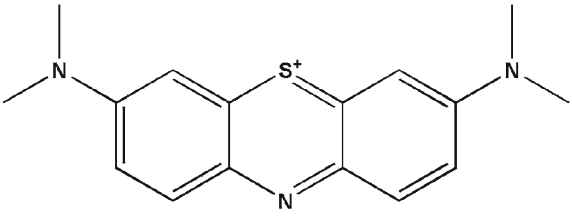


Fig. 12.1 Structure of methylene blue trihydrate

of the interaction between intercalators and base pairs are mostly based on empirical or semiempirical methods and there are only a few low-level ab initio studies [4, 5, 14, 15]. The reason is that the size of the intercalation system is usually large and high-level methods such as second-order Møller-Plesset perturbation theory (MP2) and coupled cluster (CCSD (T)) become impractical, and thus, the interaction mechanism between the intercalators and DNA is still under clear. Previous researches have elucidated that the study of intercalation systems with large size, one should consider two aspects [16]. First, intercalator molecules usually have side chains with various sizes. To capture the binding properties, at least one intercalator and one DNA base pair have to be included in the calculation [4]. Second, the intercalating process can also be affected by some factors such as its surrounding environment, such as twisting of the DNA backbone, and entropic effects, etc. [17].

Density functional theory (DFT) has been widely accepted as a useful tool for understanding and predicting the electronic properties of materials [18]. Compared to quantum chemical methods such as MP2 or CCSD (T), standard DFT methods are often better choices for large size systems because of their computational efficiency. They are efficient for both molecular complexes and single molecules, while remain sufficient accuracy. Nevertheless, for van der Waals (vdW) complexes and sparse matter, the dispersion energy becomes more important, such that standard density functionals are often failed. To remedy the failure, some functionals with long range correction have been developed, and can treat long-range dispersion interactions reasonably well. Therefore, in this paper, we adapted special DFT method to study the binding modes and interaction mechanism between intercalators and different DNA base pairs.

Due to lacking the crystal structure information of the methylene blue trihydrate interacting with DNA, the most favorable conformation between the methylene blue trihydrate and DNA base pairs are studied through the theoretical methods. In this paper, several calculation results of semiempirical and DFT methods are compared with those derived from MP2 method. The suitable method is selected to study the conformations between the methylene blue trihydrate and DNA base pairs, and the electric charge change and electrostatic density are calculated as well. Our results may help the understanding of interaction between methylene blue trihydrate and DNA, and provide useful protocols for other related studies.

12.2 Computational Details

The structures and properties of isolated intercalators and base pairs are determined by the calculations using QM methods at the ab initio level and DFT level. Then their complexes were evaluated using the selected ab initio and DFT calculations. All the calculations were performed using Gaussian 09 package [19]. A variety of methods were applied on the two adenine-thymine (A:T) base pairs at different distances in order to estimate their results and compare with MP2

calculation. The different methods, including (1) semiempirical method, such as AM1 [20], PM6 [21] and PDDG [22], (2) DFT methods that contain Becke 3-Parameter (Exchange), Lee, Yang and Parr methods (B3LYP) [23], the long range corrected version of B3LYP using the Coulomb-attenuating method (CAM-B3LYP) [24], Long range-corrected version of wPBE (LC-wPBE) [25], the latest functional from Head-Gordon and coworkers, which includes empirical dispersion (WB97XD) [26], and the hybrid functional of Zhao and Truhlar (M06) [27], and (3) ab initio methods, such as Hatree-Fork method (HF) [28], MP2 [29].

The geometries of two different DNA base pairs and their complex with methylene blue trihydrate were partially optimized using WB97XD functional under 6-311+G* basis set, which is accorded with MP2 results well in all the test methods. The geometries of adenine-thymine (A:T) and guanine-cytosine (G:C) base pairs, were fully optimized at the MP2/6-311+G* level. Molecular geometries of the DNA bases and base pairs (with the exception of the thymine methyl group) still remain planar. Atomic point charges for each molecule were obtained using the method of natural bond orbital analysis at the WB97XD/6-311+G (d, p) level.

12.3 Results and Discussion

12.3.1 The Performance of the QM Methods

The two adenine-thymine (A:T) base pairs are applied into the evaluation of QM methods and functional. Because the large size of systems are involved and computational speed of DFT is comparable with MP2 method, DFT methods, which contain the long range correction, are consider as good and useful tools for understanding and predicting the electronic properties of materials. The model used in the QM calculations and MP2 results are shown in Fig. 12.2.

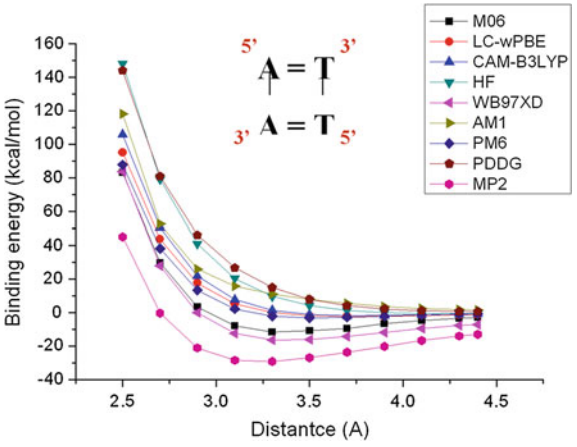


Fig. 12.2 The calculated model and the results of the MP2 calculation (take the two separated base pair of the energy calculation, which is in 100 Å, as the zero point)

Table 12.1 The calculated results of energies under different functional in MP2 method and WB97XD method

Distance (Å)	Binding energy (kcal/mol)	Binding energy (kcal/mol)	Binding energy (kcal/mol)	Binding energy (kcal/mol)	Binding energy (kcal/mol)	Binding energy (kcal/mol)
	mp2/6-31+G*	mp2/6-311+G*	mp2/6-311++G**	Wb97xd/6-31+G*	Wb97xd/6-311+G*	Wb97xd/6-311++G**
2.5	48.55	45.13	45.72	85.14	83.66	84.58
2.7	1.91	−0.41	0.38	29.19	27.91	28.69
2.9	−19.42	−20.96	−20.36	1.17	−0.02	0.54
3.1	−27.26	−28.45	−27.68	−11.25	−12.42	−12.00
3.3	−28.39	−29.15	−28.45	−15.20	−16.30	−16.00
3.5	−26.52	−26.90	−26.28	−15.06	−16.04	−15.85
3.7	−23.56	−23.60	−23.05	−13.27	−14.14	−14.00
3.9	−20.41	−20.16	−19.66	−11.09	−11.83	−11.72
4.1	−17.15	−16.68	−16.36	−8.90	−9.50	−9.43
4.3	−14.56	−13.98	−13.70	−7.24	−7.74	−7.68
4.4	−13.65	−13.05	−12.78	−6.69	−7.15	−7.09

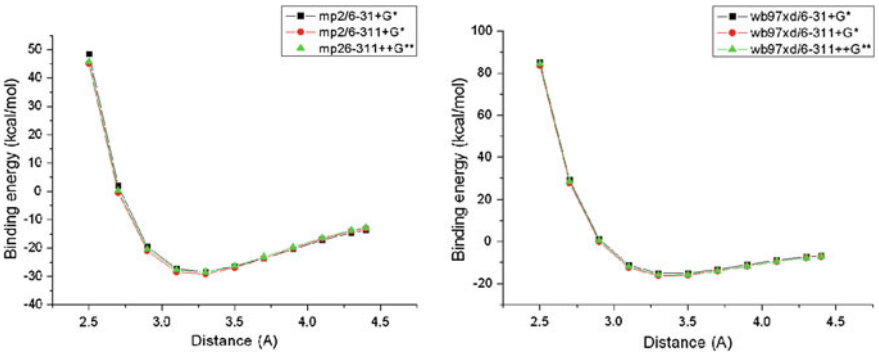


Fig. 12.3 The binding energy of the WB97XD and MP2 calculations at different functional

Among these QM methods, WB97XD is the best one reproducing the result of MP2 method. Therefore, all the calculations for the following topics are performed using WB97XD method. Furthermore, we also evaluate the performance of WB97XD and MP2 with different basis sets including 6-31+G*, 6-311+G*, 6-311++G**, and their final results, shown in Table 12.1 and Fig. 12.3, demonstrate that the basis set would not affect the results a lot, and the difference among the results with various basis set is less than 1 kcal/mol. Thus, the middle level basis set (6-311+G*) was used in the following calculations.

12.3.2 The Conformation of Methylene Blue Trihydrate Inserting into Two Base Pairs

The conformations of two base pairs, which used in this paper, contain 8 conformations which are completely different, and these base pairs are all studied by the QM calculations. The experimental results of distance between proflavine and the guanine-cytosine (G:C) base pair is 3.4–3.5 Å. Therefore, the distance between the two base pairs, which are intercalated by the proflavine, is about 6.8–7.0 Å. The conformation of methylene blue trihydrate is similar to the structure of the proflavine, and distance of two base pairs is set to be 5.3 Å. The methylene blue trihydrate is intercalated into the base pairs between which the distance is set to be 5.3 Å. The new conformation of methylene blue trihydrate and two base pairs are performed with partial optimization under WB97XD/6-311+G* basis set, with freezed two base pairs. Then the conformation of methylene blue trihydrate was kept as the same as the optimized structure, and two base pairs were allowed to move at the same time to enlarge or shorten the distance between them. The moving conformations are evaluated by single point energy calculation to obtain the most stable geometry of the complex conformation of methylene blue trihydrate with base pairs. It should be noticed in the Fig. 12.1 that the methylene blue trihydrate has two sides, one is S side, another is N side, so methylene blue trihydrate can intercalate into DNA base pairs at both sides. Therefore, intercalations from these two sides are considered for the calculations. The most favorable distance between the two base pairs was determined through the above calculations. The most stable conformations, which obtained from above, were further undergoing partial optimization with the two freezed base pairs.

The conformation and the results of the binding energies are shown in Fig. 12.4 and Table 12.2. From the Table 12.2, it can be seen that binding energy of the methylene blue trihydrate complexed with AA-TT base pair is the lowest. And when the methylene blue trihydrate intercalates into the AA-TT base pair with the N side, the distance of AA-TT is 6.6 Å, and it needs less energy to push the base pair open. Therefore, the methylene blue trihydrate intercalates into AA-TT base pair in the orientation of N is the most favorable conformation based on the above calculation of binding energies.

12.3.3 The Analysis of the Charge Transfer and Charge Density

The analysis of charge transfer between the methylene blue trihydrate before and after intercalation is shown in Table 12.3. From Table 12.3, we can see that the conformation with most favorable binding energy, which is the methylene blue trihydrate intercalates into AA-TT base pair in the orientation of N, transfers 0.223 charges from methylene blue trihydrate to the AA-TT base pairs. Another

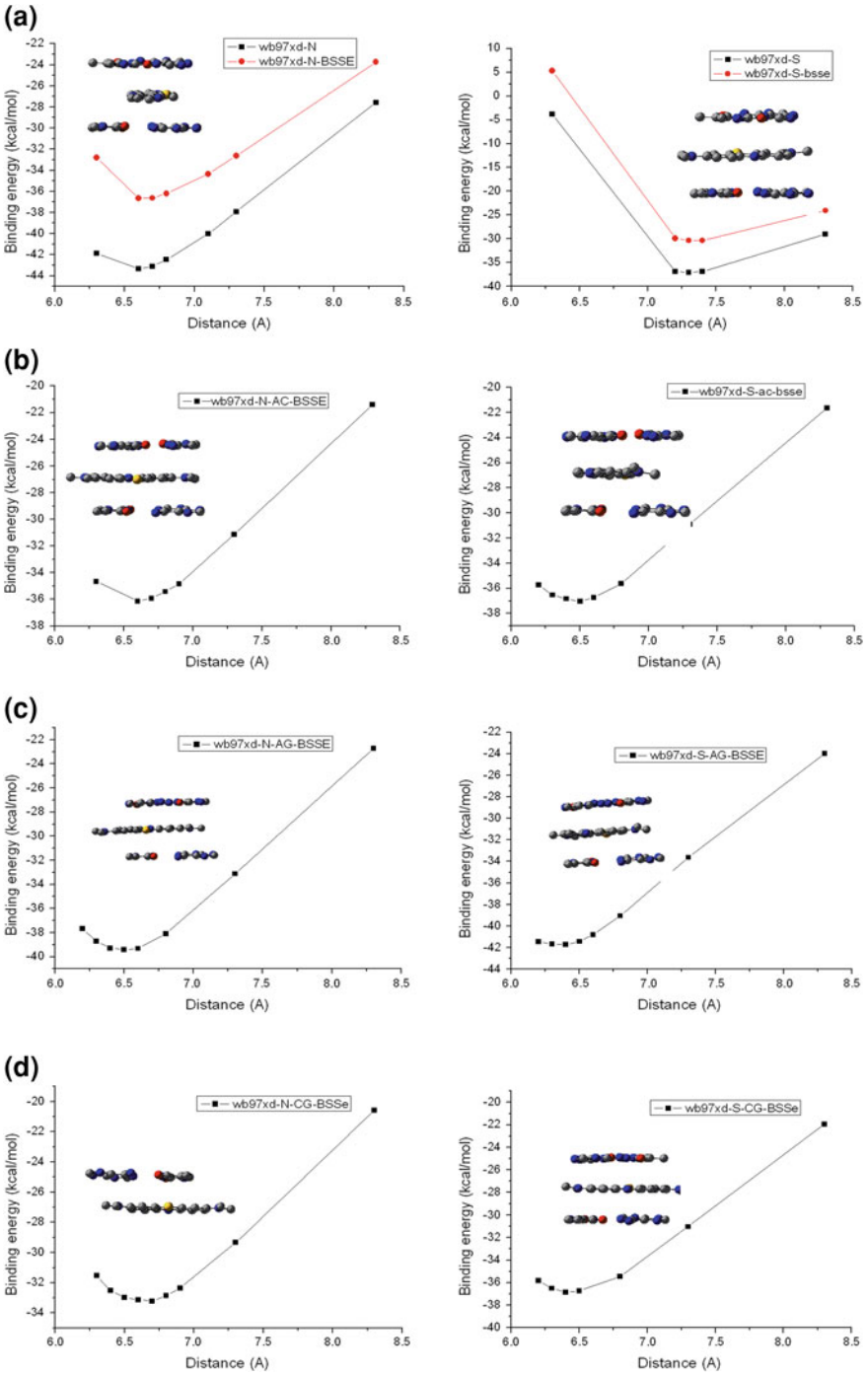


Fig. 12.4 The conformation of the methylene blue trihydrate and two base pairs. **a** AA-TT; **b** AC-GT; **c** AG-CT; **d** CG-AT; **e** GC-GC; **f** AT-AT; **g** TA-TA; **h** TG-AC

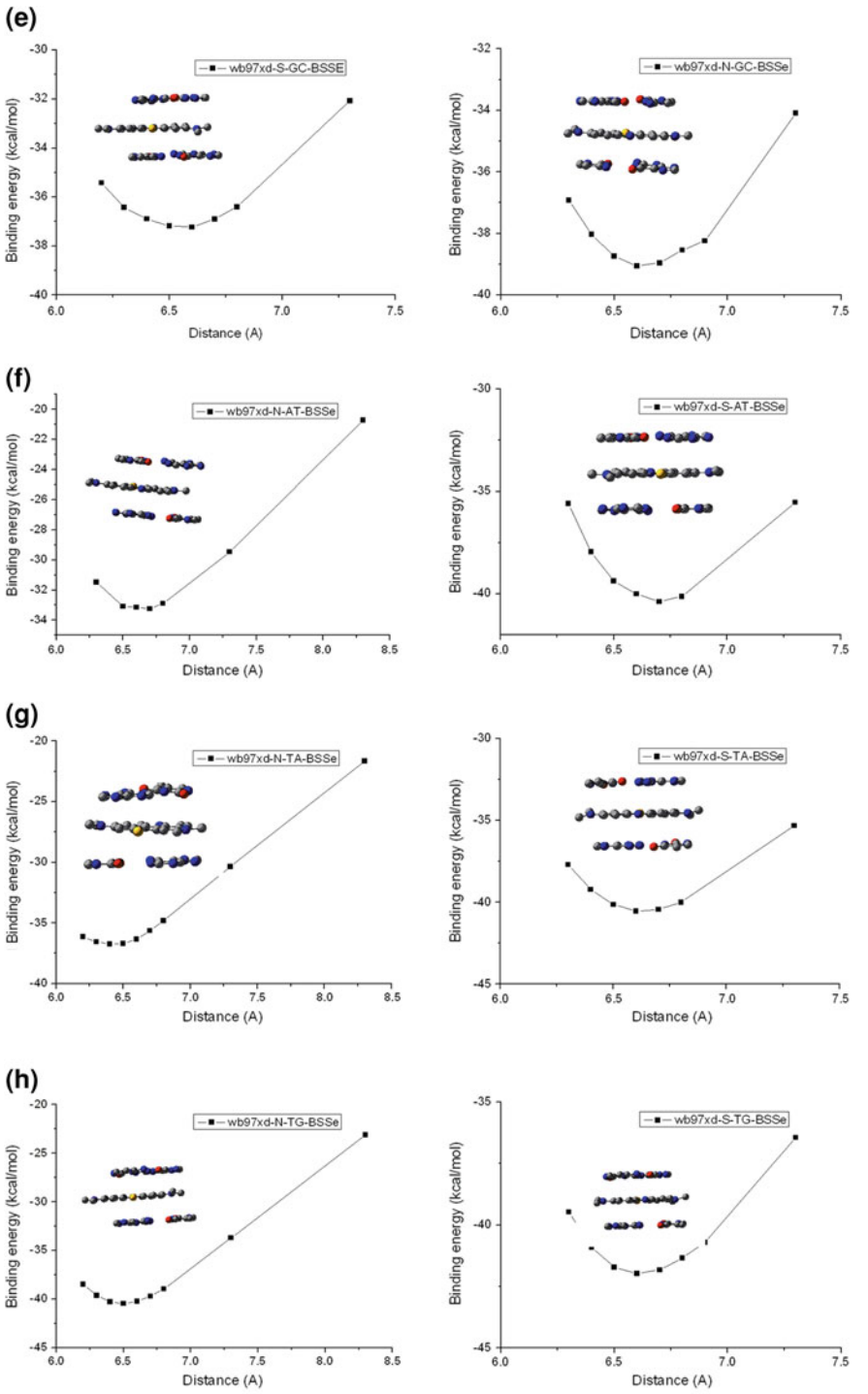


Fig. 12.4 (continued)

Table 12.2 The binding energy of the different base pairs

Base pair	Orientation	Distance (Å)	Energy (kcal/mol)
AA-TT	N	6.6	−43.34
AA-TT	S	7.3	−37.07
AC-GT	N	6.6	−35.94
AC-GT	S	6.5	−37.02
AG-CT	N	6.5	−39.43
AG-CT	S	6.4	−41.73
AT-AT	N	6.6	−33.15
AT-AT	S	6.7	−40.39
CG-CG	N	6.6	−33.23
CG-CG	S	6.4	−36.84
GC-GC	N	6.6	−37.23
GC-GC	S	6.7	−39.91
TA-TA	N	6.4	−36.75
TA-TA	S	6.6	−40.54
TG-CA	N	6.5	−40.47
TG-CA	S	6.6	−41.97

Table 12.3 The change of the charge between the methylene blue trihydrate before and after intercalation

Base pair	Before intercalation	After intercalation	The change of the charge
aa-N-6.6	1	0.777	−0.223
aa-S-7.3	1	0.944	−0.056
ac-N-6.6	1	0.941	−0.059
ac-S-6.5	1	0.939	−0.061
ag-N-6.5	1	0.947	−0.053
ag-S-6.4	1	0.95	−0.05
at-N-6.6	1	0.957	−0.043
at-S-6.7	1	0.946	−0.054
cg-N-6.6	1	0.958	−0.042
cg-S-6.4	1	0.917	−0.083
gc-N-6.6	1	1.007	0.007
gc-S-6.7	1	0.987	−0.013
ta-N-6.4	1	0.883	−0.117
ta-S-6.6	1	0.933	−0.067
tg-N-6.5	1	0.937	−0.063
tg-S-6.6	1	0.934	−0.066

conformation, which is the methylene blue trihydrate intercalates into TA-TA base pair in the orientation of N, is the second significant charge-transfer situation, the charge transfer quantity from methylene blue trihydrate to base pair is 0.117, but it is not favorable in the binding energy. Therefore, the conformation, which the binding energy is lowest, transfers the most charges, but the conformations of favor binding energy do not all transfer the more charges.

The analysis of changes of charge after the methylene blue trihydrate intercalation is shown in Fig. 12.5. From Fig. 12.5, we can see that the change of the charge of heavy atoms in the middle of system is most remarkable, except for the AC-S-6.5, AG-N-6.5, CG-N-6.6, CG-S-6.4, TG-N-6.5. It means that the charge transfers from methylene blue trihydrate to different base pairs, is most probably taken place in these positions. The amount of charge change of heavy atoms is large in some base pairs, such as AC-N-6.6, AG-S-6.4, CG-N-6.6, and GC-N-6.6, but the total change of charges is relatively small, which is −0.059, −0.050, −0.042, 0.007, respectively, because of the charge is re-distribute within the inter-molecule. Therefore, the changes of charge density often take place at the heavy atoms in the middle part of system which the charge density changes most remarkable.

The change of charge density of the system after the methylene blue trihydrate intercalation of all the base pair are analyzed through Multiwfn software [30] and shown in Fig. 12.6. In the model of AA-N-6.6, the decrease of charge density is concentrated at the side chain of methyl, the adjacent ring and the other side ring of molecule. And the decreased region of AA-N-6.6 in the center of three rings in methylene blue trihydrate is weaker and smaller than that in the model of AA-S-7.3. In the model of AA-S-7.3, the decreased charge density is happened in side chain near the N atom. In the model of AC-N-6.6, the decreased part of charge density is smaller than that in the model of AC-S-6.5, and the domain is concentrated in the side chain of the methyl, and between N atom of the side chain and the other side ring of the molecule. The decreased domain of AC-N-6.6 can be found in the center of these three rings in methylene blue trihydrate, but in the model of AC-S-6.5 only one domain in the centre of middle ring can be identified. The decreased domain of charge density is concentrated at both the side chain and methyl group in the side of the N side, and the other side ring of molecule in the model of AG-N-6.5. The decreased domain of the AG-N-6.5 in the center of the three rings in methylene blue trihydrate, but in the model of AG-S-6.4 two domains in the centre of ring can be found. In the model of AG-S-6.4, the decreased portion of charge density is located at the opposite side of that in the model of AG-N-6.5. For the models of AT-N-6.6 and AT-S-6.7, the similar decreased region of charge density is concentrated in C of the ring which is adjacent to the side chain N. The other decreased domains of two models are found at different position of the other side ring in methylene blue trihydrate, but in the model of AT-N-6.6 this area is large than that in the AT-S-6.7. The decreased range of model CG-S-6.4 is remarkable smaller than in the CG-N-6.6. The most decreased range can be found between the side chain and adjacent ring. In the model of GC-S-6.7, the decreased domain of charge density is distributed symmetrically and concentrated at the two sides of side chain N–C bond. But the decreased ranges of GC-N-6.6 model is found in the side chain, center and outside of

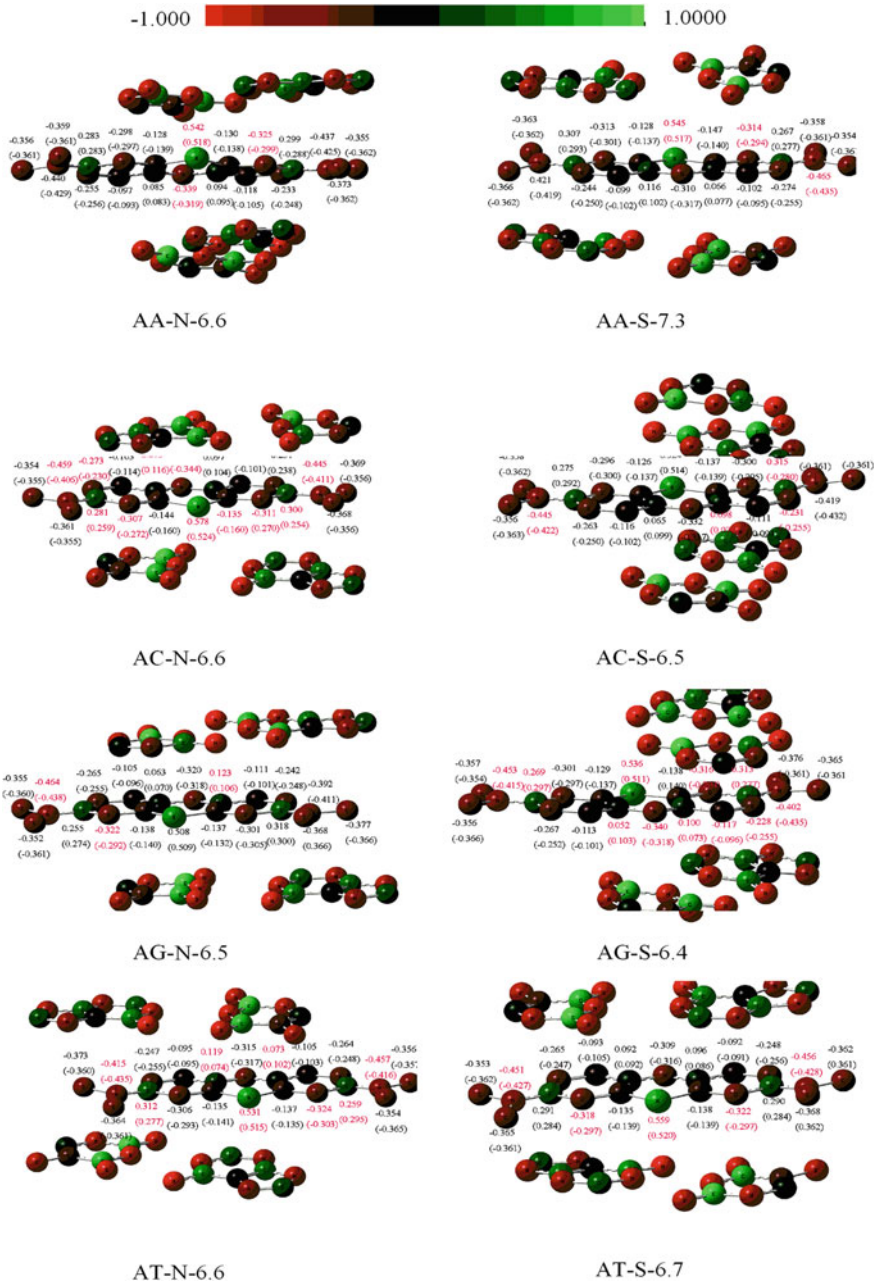
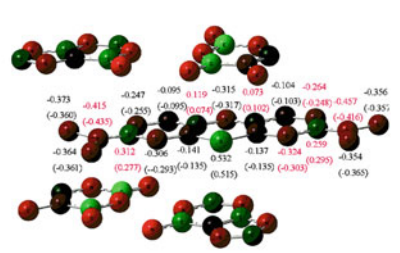
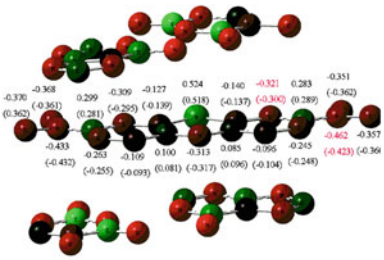


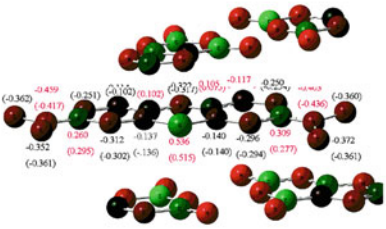
Fig. 12.5 The change of the charge of the system after the methylene blue trihydrate intercalation. The charges of methylene blue trihydrate, which is before the intercalation, are shown in parentheses, and the value of the change of charge, which is larger than 0.02 is colored in *red*



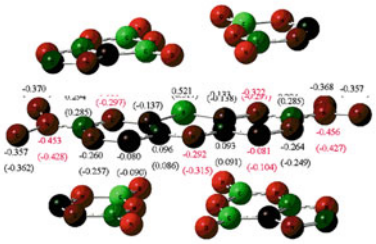
CG-N-6.6



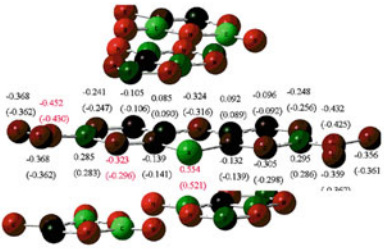
CG-S-6.4



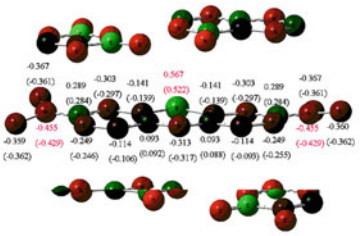
GC-N-6.6



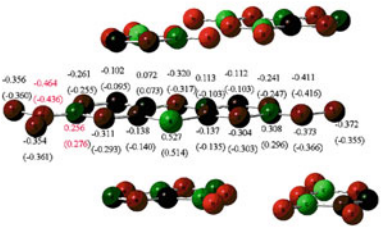
GC-S-6.7



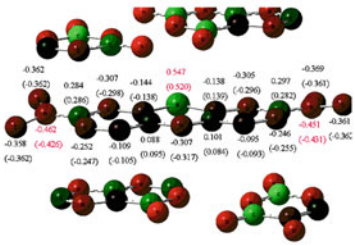
TA-N-6.4



TA-S-6.5



TG-N-6.5



TG-S-6.6

Fig. 12.5 (Continued)

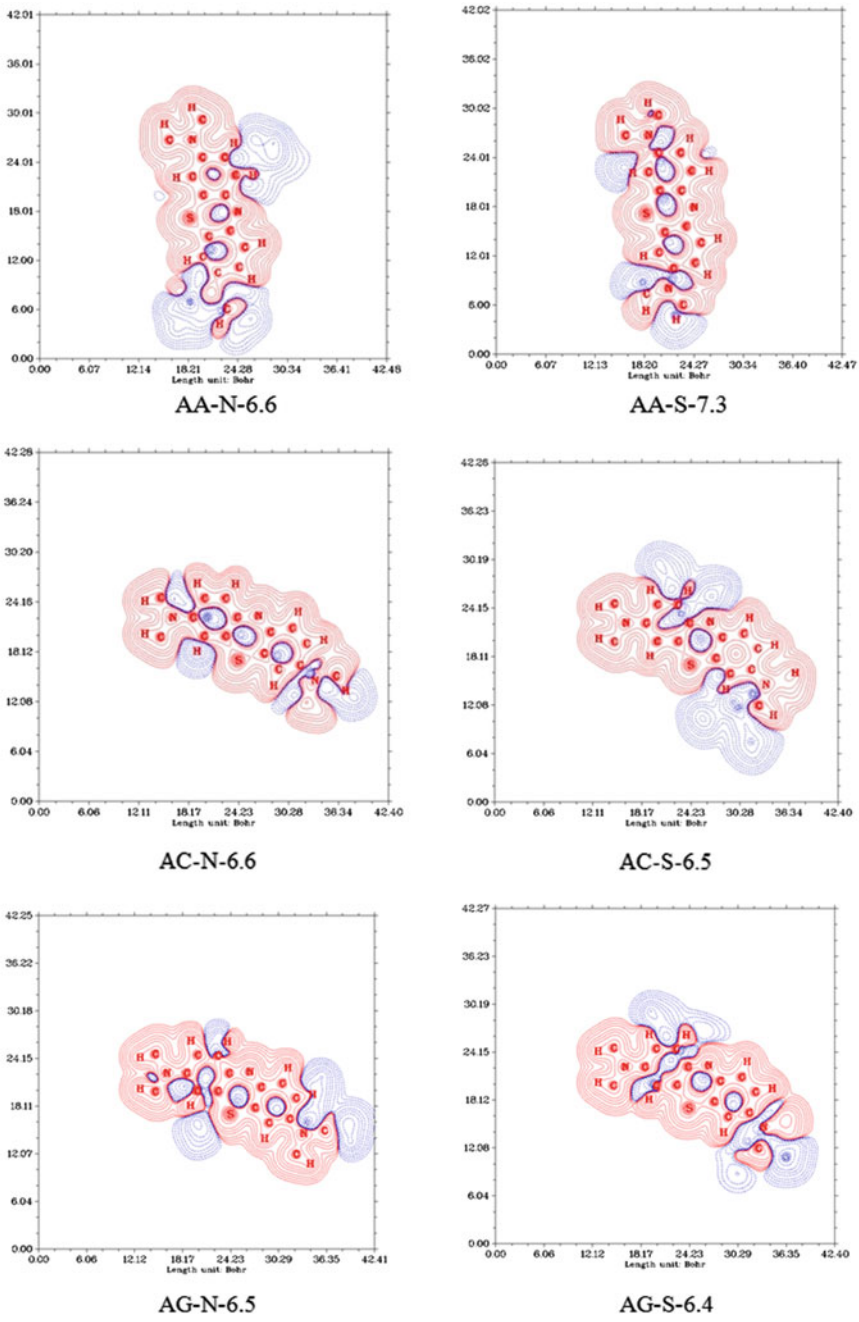


Fig. 12.6 The change of the charge density of the system after the methylene blue trihydrate intercalation. The change of the charge density is shown in the plane of methylene blue trihydrate, in which the increase of the charge density is shown in *red*, and the decrease is shown in *blue*

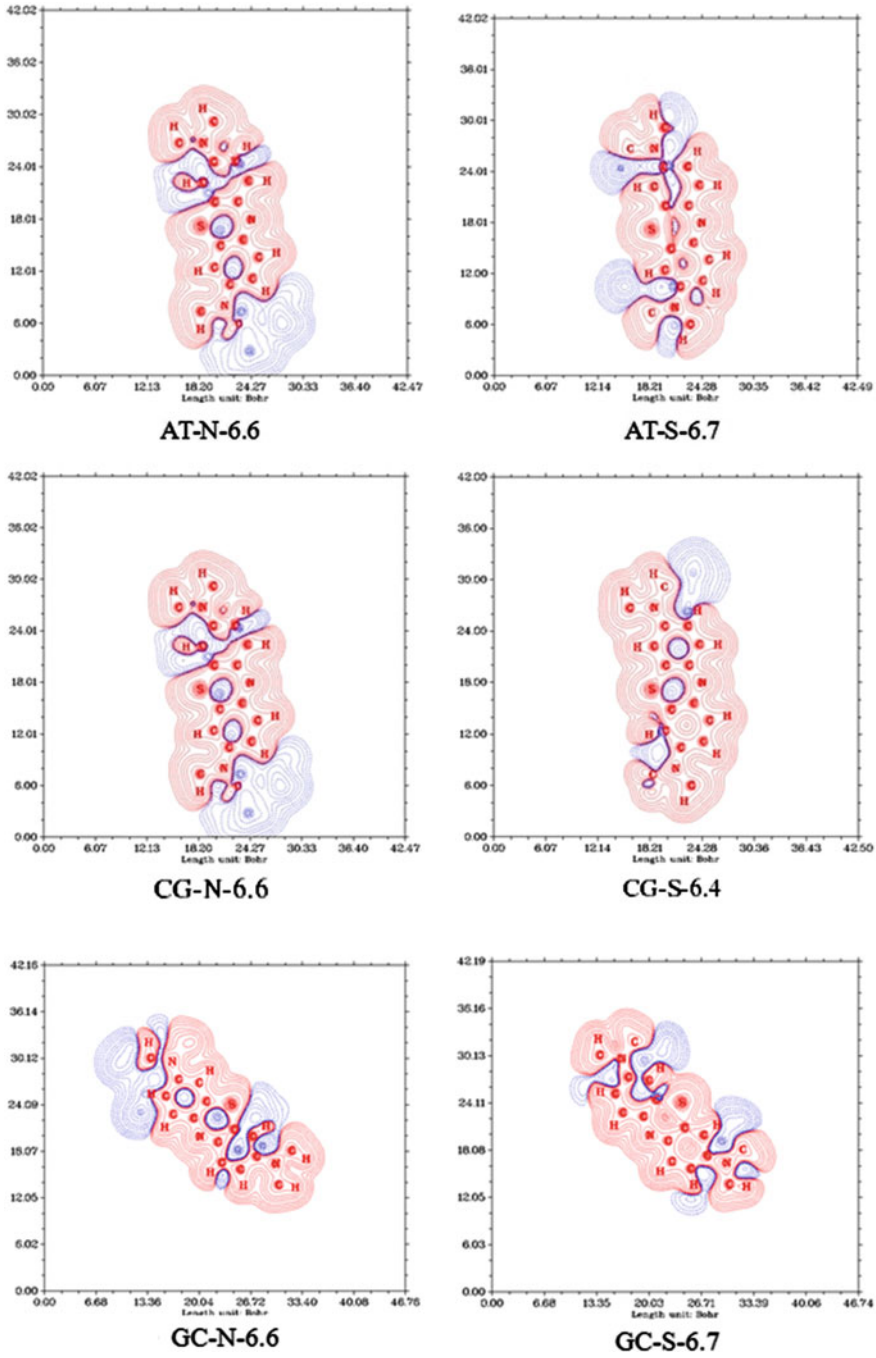


Fig. 12.6 (Continued)

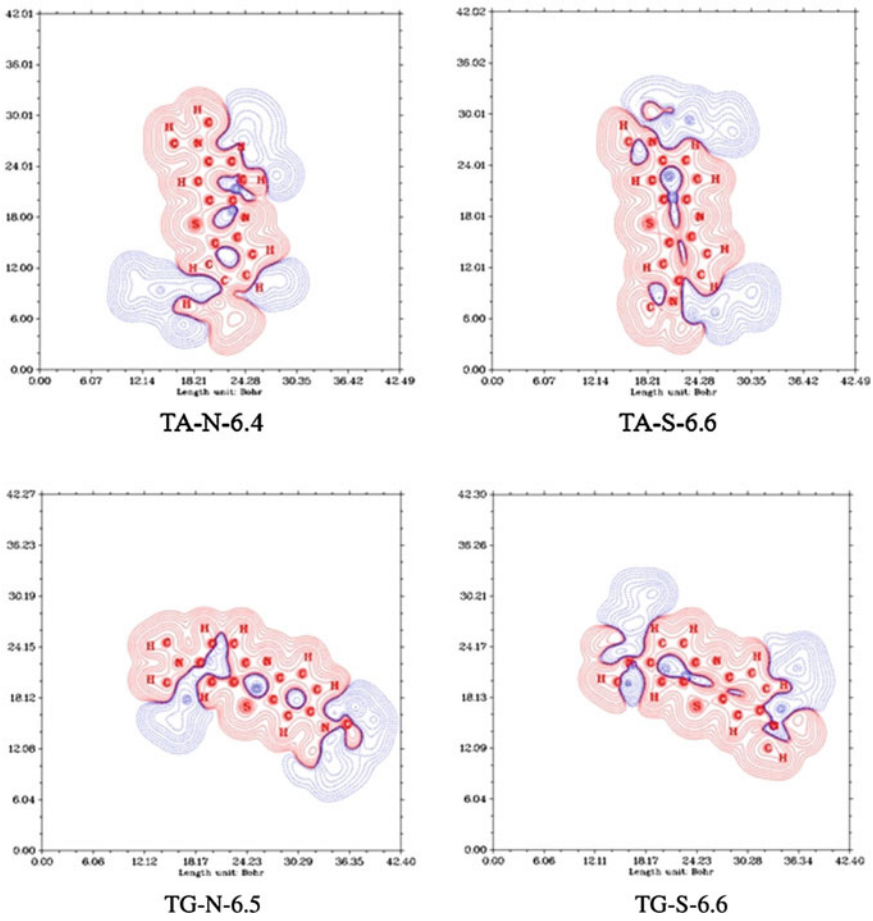


Fig. 12.6 (Continued)

the other side ring in methylene blue trihydrate, and in the model of GC-N-6.6 this range is larger than that in the GC-S-6.7. In the model of TA-S-6.6, the decreased domains are distributed in the N side of middle ring, and concentrated around the side chain. Although the decreased area in the model TA-N-6.4 is close to the model TA-S-6.6, the ranges are distributed in the two sides of the molecule. In the model of TG-N-6.5, the decreased domain is concentrated in the side of the S atom of the middle ring, but it is distributed in the N side in the model TG-S-6.6.

From the above analysis, it can be drawn that if $\lll$ the decreased domain located in the centre of the ring of methylene blue trihydrate in S or N orientation is more than the other orientation, and the binding energy of this orientation is higher than the other, which means that the binding between the methylene blue trihydrate and base pair is less stable.

12.4 Conclusion

In conclusion, we have shown that in the absence of crystal structure of intercalators with DNA base pairs, the calculated method can be performed to predict the binding modes between the intercalators and DNA base pair. The different methods including semiempirical method, such as AM1, PM6 and PDDG, DFT methods, B3LYP, CAM-B3LYP, LC-wPBE, WB97XD, and the M06, and ab initio level methods, such as Hatree-Fork method (HF), were verified for the appropriate calculation of the binding energy between the methylene blue trihydrate and DNA base pairs, and their complexes. Among these methods, WB97XD with 6-311+G* basis set is selected for the detail analysis of changes of charge density. Our results show that the methylene blue trihydrate intercalated into the DNA base pair can enlarge the distance between the base pair, and the different base pair was enlarged to the different distance. From the analysis of binding energies, the methylene blue trihydrate intercalated into AA-TT base pair in the orientation of N side is the most favorable conformation and it transfers the charges of 0.223 from methylene blue trihydrate to the AA-TT base pair. The analysis of change of the charge shows that the changes of charge of the methylene blue trihydrate often take place on the heavy atoms in the middle of system, where the charge changes most remarkable. And with the analysis of changes of the charge density, we can see that if the decreased domain of charge density located in the centre of the ring of methylene blue trihydrate in one orientation is more than the other orientation, the binding between methylene blue trihydrate and base pair is less stable.

References

1. Li L, Chen Q, Wei DQ (2012) Prediction and functional analysis of single nucleotide polymorphisms. Curr Drug Metab 13:1012–1023
2. Wei DQ (2012) New drug design based on multi-targets and system biology approach in light of real time DNA sequencing technologies. Curr Top Med Chem 12:1309
3. Xiong Y, Liu JA, Wei DQ (2011) An accurate feature-based method for identifying DNA-binding residues on protein surfaces. Proteins 79(2):509–517
4. Řeha D, Kabeláč M, Ryjáček F, Šponer J, Šponer JE, Elstner M et al (2002) Intercalators. 1. Nature of stacking interactions between intercalators (ethidium, daunomycin, ellipticine, and 4', 6-diaminide-2-phenylindole) and DNA base pairs. Ab initio quantum chemical, density functional theory, and empirical potential study. J Am Chem Soc 124:3366–3376
5. Langner KM, Kedzierski P, Sokalski WA, Leszczynski J (2006) Physical nature of ethidium and proflavine interactions with nucleic acid physical nature of ethidium and proflavine interactions with nucleic acid. J Phys Chem B 110:9720–9727
6. Graves DE, Velea LM (2000) Intercalative binding of small molecules to nucleic acids. Curr Org Chem 9:915–929
7. Chaires JB (1997) Energetics of drug-DNA interactions. Biopolymers 44(3):201–215
8. Kubar T, Hanus M, Ryjacek F, Hobza P (2005) Binding of cationic and neutral phenanthridine intercalators to a DNA oligomer is controlled by dispersion energy:

quantum chemical calculations and molecular mechanics simulations. Chem-A Eur J 12(1):280–290
9. Waring MJ (1981) DNA modification and cancer. Annu Rev Biochem 50:159–192
10. Starcevic K, Karminski-Zamola G, Piantanida I, Zinic M, Suman L, Kralji M (2005) Photoinduced switch of a DNA/RNA inactive molecule into a classical intercalator. J Am Chem Soc 127(4):1074–1075
11. Fantacci S, De Angelis F, Sgamellotti A, Marrone A, Re N (2005) Photophysical properties of $[Ru(phen)_2(dppz)]^{2+}$ intercalated into DNA: An integrated Car-Parrinello and TDDFT Study. J Am Chem Soc 127(41):14144–14145
12. Auerbach SS, Bristol DW, Peckham JC, Travlos GS, Hébert CD, Chhabra RS (2010) Toxicity and carcinogenicity studies of methylene blue trihydrate in F344 N rats and B6C3F1 mice. Food Chem Toxicol l48(1):169–177
13. Hejtmancik MR, Ryan MJ, Toft JD, Persing RL, Kurtz PJ, Chhabra RS (2002) Hematological effects in F344 rats and B6C3F1 mice during the 13-week gavage toxicity study of methylene blue trihydrate. Toxicol Sci 65(1):126–134
14. Bondarev DA, Skawinski WJ, Venanzi CA (2000) Nature of intercalator amiloride-nucelobase stacking. An empirical potential and ab initio electron correlation study. J Phys Chem B 104:815–822
15. Šponer J, Gabb HA, Leszczynski J, Hobza P (1997) Base-base and deoxyribose-base stacking interactions in B-DNA and Z-DNA: a quantum-chemical study. Biophys J 73(1):76–87
16. Li S, Cooper VR, Thonhauser T, Lundqvist BI, Langreth DC (2009) Stacking interactions and DNA intercalation. J Phys Chem B 113:11166–11172
17. Elcock AH, Rodger A, Richards WG (1996) Theoretical studies of the intercalation of 9-hydroxyellipticine in DNA. Biopolymers 39:309–326
18. Cooper VR, Thonhauser T, Puzder A, Schröder E, Lundqvist BI, Langreth DC (2008) Stacking interactions and the twist of DNA. J Am Chem Soc 130:1304–1308
19. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Rob MA, Cheeseman JR et al (2003) Gaussian 03 Inc., Wallingford CT
20. Dewar MJS, Holder AJ (1990) AM1 parameters for aluminum. Organometallics 9:508–511
21. Stewart JJP (2007) Optimization of parameters for semiempirical methods. V. Modification of NDDO approximations and application to 70 elements. J Mol Model 13:1173–1213
22. Tirado-Rives J, Jorgensen WL (2008) Performance of B3LYP density functional methods for a large set of organic molecules. J Chem Theor Comput 4:297–306
23. Becke AD (1993) Density-functional thermochemistry. III. The role of exact exchange. J Chem Phys 98:5648–5652
24. Yanai T, Tew D, Handy N (2004) A new hybrid exchange-correlation functional using the Coulomb-attenuating method (CAM-B3LYP). Chem Phys Lett 393:51–57
25. Tawada Y, Tsuneda T, Yanagisawa S, Yanai T, Hirao K (2004) A long-range-corrected time-dependent density functional theory. J Chem Phys 120:8425–8433
26. Chai JD, Head-Gordon M (2008) Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections. Phys Chem Chem Phys 10:6615–6620
27. Zhao Y, Truhlar DG (2006) A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions. J Chem Phys 125:1–18
28. McWeeny R, Dierksen G (1968) Self-consistent perturbation theory. 2. Extension to open shells. J Chem Phys 49:4852–4856
29. Head-Gordon M, Maurice D, Oumi M (1995) A perturbative correction to restricted open-shell configuration-interaction with single substitutions for excited-states of radicals. Chem Phys Lett 246:114–121
30. Lu T, Chen F (2012) Multiwfn: a multifunctional wavefunction analyzer. J Comp Chem 33:580–592

Chapter 13
Drug Inhibition and Proton Conduction Mechanisms of the Influenza A M2 Proton Channel

Ruoxu Gu, Limin Angela Liu and Dongqing Wei

Abstract The influenza A virus matrix protein 2 (M2 protein) is a pH-regulated proton channel embedded in the viral membrane. Inhibition of the M2 proton channel has been used to treat influenza infections for decades due to the crucial role of this protein in viral infection and replication. However, the widely-used M2 inhibitors, amantadine and rimantadine, have gradually lost their efficiencies because of naturally-occurring drug resistant mutations. Therefore, investigation of the structure and function of the M2 proton channel will not only increase our understanding of this important biological system, but also lead to the design of novel and effective anti-influenza drugs. Despite the simplicity of the M2 molecular structure, the M2 channel is highly flexible and there have been controversies and arguments regarding the channel inhibition mechanism and the proton conduction mechanism. In this book chapter, we will first carefully review the experimental and computational studies of the two possible drug binding sites on the M2 protein and explain the mechanisms regarding how inhibitors prevent proton conduction. Then, we will summarize our recent molecular dynamics simulations of the drug-resistant mutant channels and propose mechanisms for drug resistance. Finally, we will discuss two existing proton conduction mechanisms and talk about the remaining questions regarding the proton-relay process through the channel. The studies reviewed here demonstrate how molecular modeling and simulations have complemented experimental work and helped us understand the M2 channel structure and function.

Keywords M2 protein · Inhibition · Binding sites · Molecular dynamics simulations

R. Gu · D. Wei (✉)
State Key Laboratory of Microbial Metabolism, College of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China
e-mail: dqwei@sjtu.edu.cn

L.A. Liu
Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

D. Wei et al. (eds.), *Advance in Structural Bioinformatics*, Advances in Experimental Medicine and Biology 827, DOI 10.1007/978-94-017-9245-5_13

13.1 Introduction

The influenza A virus matrix protein M2 (A/M2) is a pH-regulated proton selective ion channel in the viral envelop [1–3]. In the life cycle of the virus, the M2 channel plays two critical roles. First, the acidification of the viral interior through the M2 channel facilitates the release of influenza ribonucleoproteins into the infected cell. Second, the M2 channel regulates the pH value of the Golgi lumen of the host cell and prevents premature conformational changes of other viral proteins to ensure proper viral assembly [2, 3].

Due to these crucial physiological roles of the M2 protein in the influenza viral life cycle, the M2 channel has been the target of anti-influenza drugs for several decades. Amantadine and rimantadine, for instance, are adamantane-based anti-influenza drugs that function as M2 inhibitors and they have been used for treating influenza for more than thirty years [4]. However, these two drugs have gradually lost their efficacy during the past decades because of naturally-occurring drug resistant mutations [5, 6]. Understanding the mechanisms of drug resistance in these mutant channels is of paramount importance for designing novel anti-influenza drugs. Here, we will review recent experimental and theoretical studies of the wild-type and drug-resistant mutant channels. These studies not only give us insight into the proton conduction mechanisms and drug inhibition mechanisms of the M2 protein, but also guide the design of novel anti-influenza drugs.

13.2 Structure of the M2 Proton Channel

The M2 channel is a homo-tetramer constituted by four transmembrane (TM) peptides [7] arranged in a left-handed way with their N- and C-terminal domains residing at the extracellular side and the cytoplasmic side, respectively [8]. Each of the four monomer subunits contains 97 amino acids, including a short extracellular domain (residues 1–24), a transmembrane (TM) domain (residues 25–46), and a cytoplasmic domain (residues 47–97) [6, 8]. The extracellular domain is a signal peptide that facilitates channel incorporation into the membrane bilayer. The transmembrane domain consists of four α helices and is responsible for proton conduction and could be inhibited by adamantane-based drugs (Fig. 13.1) [9]. The cytoplasmic domain contains a short amphipathic helix and a disordered tail. The amphipathic helices of four subunits stabilize the channel by forming a base that is nearly perpendicular to the transmembrane bundle on the membrane surface, whereas the disordered tails interact with matrix protein 1 (M1) that packs around the ribonucleoproteins [10].

The transmembrane domain itself is capable of proton conduction and could be inhibited by adamantane-based inhibitors [9]. Four single transmembrane helices are packed to construct a hydrophilic channel pore that contains structured water molecules across the membrane bilayer (Fig. 13.1). Several pore-facing residues

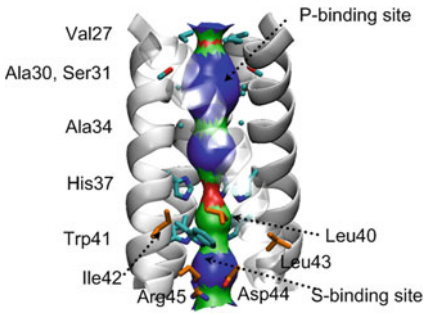


Fig. 13.1 Structure of the transmembrane domain of the M2 protein. The M2 protein is shown in cartoon model in *white*. The pore facing residues (Val27, Ala30, Ser31, Gly34, His37, and Trp41), which constitute the drug binding site in the channel pore (the P-binding site), are shown in stick model in *cyan*. The residues constituting the drug binding site on the protein surface (the S-binding site) are shown in stick model in *orange*. The channel pore radius profile is shown in *blue* (pore radii larger than 2.8 Å), *green* (pore radii larger than 1.4 Å and smaller than 2.8 Å) and *red* (pore radii smaller than 1.4 Å), respectively

are lined along the channel pore. These residues directly affect the pore radii and the hydrophilicity of the channel pore, due to their physical shapes and different electrostatic properties. These residues are Val27, Ser31 or Ala30, Gly34, His37 and Trp41, from the N-terminal side to the C-terminal side of the channel [11] (Fig. 13.1).

The hydrophobic side chains of Val27 residues form a gate at the N-terminal entrance of the channel (Fig. 13.1). The Val27 residues are also crucial for channel inhibition by forming hydrophobic interactions with the M2 inhibitors, which will be discussed in detail in the following sections.

The Ala30/Ser31 and Gly34 residues are located in the middle portion of the proton channel and form a hydrophilic cavity occupied by water molecules in the absence of M2 inhibitors (Figs. 13.1 and 13.2a). Highly-ordered water molecules are stabilized by hydrogen bond interactions with the carbonyl groups of these residues and the hydroxyl groups of Ser31.

The His37–Trp41 quartet is the functional core for the proton conduction process through the channel. This quartet is highly conserved in the M2 proteins of type A and type B influenza viruses. The Trp41 residues occlude the channel at the C-terminal end and act as a channel gate to prevent outward ion flux from the viral interior to the extracellular side (Fig. 13.1) [12]. The His37 and Trp41 residues change their conformations under different pH_{out} values (pH_{out} refers to the pH value of the extracellular side of the influenza virus) to open the channel and allow proton transfer across the bilayer.

The Asp44 and Arg45 residues (Fig. 13.1) are not pore-facing residues, but these two residues form hydrogen bonds and salt bridges among themselves to stabilize the tetramer protein. These two residues also bind and stabilize water molecules at the C-terminal end of the M2 channel. These water molecules are believed to facilitate the proton conduction process.

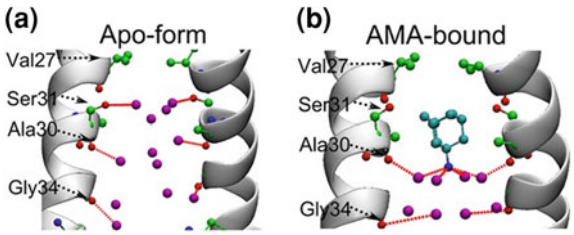


Fig. 13.2 The three-layer water structure in the M2 channel pore. The figure is drawn based on our molecular dynamics simulations [43]. Both the water structures in the apo form (panel **a**) and the amantadine bound form (panel **b**) are shown. The M2 protein is shown in cartoon model in *white*. The pore facing residues (Val27, Ala30, Ser31, and Gly34) and the amantadine molecule are shown in *ball* and stick model with the carbon atoms of the residues, the carbon atoms of the inhibitor, and the oxygen and nitrogen atoms in *green*, *cyan*, *red* and *blue*, respectively. The stable water molecules in the channel pore are shown as magenta spheres, whereas the hydrogen bonds between these water molecules and the Ala30, Ser31, and Gly34 residues and the amantadine molecule are labeled by *red lines*

13.2.1 M2 Protein Structure Is Sensitive to Its Environments

Although the M2 protein appears to have a very simple structure compared to ion channels from higher organisms (such as K^+ and Ca^{2+} channels), how the M2 channel conducts protons under low environmental pH and how drugs bind to the channel and inhibit M2 channel function have not been fully understood. This is mainly attributed to the flexibility of the M2 protein, which exhibits a wide range of structural and functional properties due to the varying experimental conditions and peptide chain lengths used in the studies. The M2 channel transmembrane domain is directly contacted by the lipid molecules, leading to significant sensitivity of the structure and function of the channel to the bilayer environment. The channel pore of many higher-organism ion channels (e.g., the nicotinic acetylcholine receptors (nAChR) and the K^+ channels) [13, 14], in contrast, is constructed with the help of other transmembrane helices. In these channels, the pore-constituting helices are surrounded and protected by these additional transmembrane helices, so that direct contacts between the lipids and the pore helices are reduced significantly and the sensitivity of the channel pore to the bilayer environment is decreased. Such a shielding mechanism is absent in the M2 channel, whose structure has been found to show large differences in different lipid environments [15]. In addition, the C-terminal intracellular amphipathic helices that are crucial for the structural stability and the proton conduction of the channel are missing in most of the experimental studies, which may also lead to biased results.

Based on close examination and comparison of existing experimental and computational studies, we postulated that the experimental conditions, including drug binding, environmental pH value, as well as the bilayer composition, strongly affect the structure of the M2 protein. We will briefly review these factors below.

Interested readers may also refer to two recent reviews by Hong and DeGrado [16] and Zhou and Cross [15] for more discussion on this topic.

Drug binding has significant influence on the conformation of the M2 protein. The M2 channel exists in a variety of conformational states in the apo-form, whereas inhibitor-binding significantly reduces the conformational flexibility of the channel. Thomason et al. [17] found that the drug-bound M2 channel tetramer packed more tightly than the apo-form, while Ma et al. [9] observed that inhibitors facilitated the tetramerization of the M2 peptides.

The M2 channel structure is also regulated by environmental pH value. The M2 tetramer reaches maximal stability around pH_{out} value of 6.5. The structural stability of the M2 channel decreases in lower or higher pH_{out} values [9, 18]. In low pH environments where the channel opens and allows proton conduction, the channel pore is more hydrated and the distances between the N-terminal ends of the transmembrane helices are larger, implying a loosely-packed tetramer structure [17]. The TM helical kink around Gly34 is eliminated under low pH_{out} values according to comparisons of the M2 structures solved under different pH values (PDB ID: 2RLF, 3LBW, 3C9 J, see Fig. 2 in Ref. [19]). This conformational change could result in more loosely-packed tetramer structure and wider channel pore radii, which would facilitate the conduction of protons.

A variety of studies have revealed the influence of the molecular composition of bilayers on the conformational equilibrium and the function of the M2 membrane proteins. Such influence is achieved by binding of lipids and lipid-soluble molecules (e.g., cholesterol) to specific positions of membrane proteins or by physical properties of bilayers such as the dimension of the hydrophobic region [20]. These elements also have significant effects on the conformational states of the M2 transmembrane domain, as reviewed by Cross et al. [21, 22] and Zhou and McCammon [15]. The length of the transmembrane helix of the M2 channel (res. 25–46, $\sim$33 Å) is larger than the width of the hydrophobic region of bilayers ($\sim$25 Å), therefore the transmembrane helices have to tilt at an angle to reconcile this dimension mismatch. Duong-LY et al. [23] investigated the M2 conformations in different bilayers and found that the helical tilt angle was highly correlated with the bilayer width. The wider the bilayer was, the smaller the helix tilt angle (with respect to the bilayer normal) and the longer the M2 channel would become. A tilt angle ranging from 30° to 38° have been reported [11].

However, both Kovacs et al. [24] and Duong-LY et al. [23] found that the changes of helical tilt angles of the M2 channel cannot be explained solely by the bilayer width. For example, the helical tilt angles in DMPC ($\sim$23 Å) and DOPC ($\sim$27 Å) bilayers were $\sim$37° and $\sim$33°, respectively, where significant changes of the bilayer width only resulted in small changes of the tilt angle [23, 24]. This result indicates that the conformation of the M2 channel is restrained by other factors besides the bilayer width. For instance, the N-terminal end of the transmembrane domain does not have an anchoring residue and its exact position in the bilayer interfacial region is flexible. This variation in the N-terminal position may also affect the channel conformation. Different molecular compositions of bilayers introduce

different lateral chemical groups that affect the equilibrium of the different conformational states of the M2 protein, as discussed by Duong-LY et al. [23].

The different M2 conformations in different bilayer environments affect its functions, including inhibitor binding and proton conduction. For instance, the proton conduction rate of the M2 channel varies in different bilayers [25–27]. The conformation of the M2 transmembrane domain was found to shift toward the apo-form rather than the drug-binding state in cholesterol-containing bilayers [20]. Therefore, investigating the M2 conformational states under different environments is crucial for understanding its proton conduction mechanism and for designing novel channel inhibitors.

13.2.2 Differences Among the M2 Structures in the Protein Data Bank

There are 12 structures of the wild-type and drug-resistant mutant M2 channels in the Protein Data Bank solved by different techniques under different conditions (crystal and solution NMR structures in micelle environments and solid state NMR structures in bilayer environments, see Table 13.1 in Ref. [28]). The pore radius profiles and the tetramer assembly of these structures were compared in detail in our previous publications [11, 28]. We will only briefly summarize the differences of these structures here.

In some of the early structures such as 1NYJ [29] and 2H95 [30], the channel pore radii are very large in either the N- or the C-terminal ends of the channel, implying that the tetramer was not assembled tightly. The most recent M2 structures showed more tightly-packed tetramers with two minima of the pore radii at the two gates around Val27 and Trp41, respectively (Fig. 13.1). Such conformation is believed to closely resemble the M2 conformation in psychological conditions. The most representative M2 structures in the Protein Data Bank are 2KQT (ssNMR, bilayer), 3LBW (X-ray, micelle), 2RLF (sNMR, micelle), and 2L0 J (ssNMR, bilayer). Among them, 2KQT and 3LBW are transmembrane-only constructs, whereas 2RLF and 2L0 J contain the short C-terminal intracellular amphipathic helices. We would like to note that, the differences between these M2 structures are not only attributed to the different experimental conditions but also due to technical issues such as the precision with which these structures were resolved. We direct interested readers to several recent reviews of the M2 channel structure for a brief overview of the existing structural work of the transmembrane domain of the M2 channel [10, 21] as well as comparisons of the precision and reliability of these structures [16].

We would like to note here that, the structures of the intracellular amphipathic helices are different in 2RLF and 2L0 J, mainly because of the different lipid environments in which these structures were solved. 2RLF was solved in micelle and the C-terminal helices were connected to the transmembrane helices via a

Table 13.1 Proton conduction rate of several M2 mutants where each monomer chain contains a single amino acid mutation

Mutation	Proton conduction*	Rationale for observed conduction results	Refs.
V27A	↑	Smaller side chains facilitate pore hydration	[5, 27]
V27T, G34E, A30P, A30T	↓	Larger side chains disrupt pore hydration	[5]
S31N	–	–	[5]
V27S	–	–	[5]
S31A	↓	Hydrophobic side chains disrupt pore hydration	[27]
D44A, W41A	↓	Mutation disrupt proton release at the exit	[5, 27]

* ↑ indicates increase of the proton conduction rate; ↓ indicates decrease of proton conduction rate; – indicates no obvious change.

short loop (residue 47–51) and were located in the solution. However, in the bilayer environment (2L0 J), the C-terminal helices are connected to the transmembrane helices through a rigid turn (residue 47) and are positioned on the bilayer surface [31]. In 2L0 J, the conformation of the four amphipathic helices is stabilized by extensive hydrophobic interactions among these helices. The positively-charged residues (Lys49, Arg53, His57, Lys60, and Arg61) are exposed to the negatively-charged lipid head groups to anchor these helices on the membrane surface.

13.3 Drug Inhibition Mechanism of the M2 Proton Channel

13.3.1 Two Different Drug Binding Sites

As mentioned above, residues Ala30, Ser31 and Gly34 constitute a hydrophilic cavity that is capable of accommodating water molecules and small drug molecules (Figs. 13.1 and 13.2). Drug binding in the channel pore may inhibit proton conduction across the bilayer by physically occluding the channel pore. According to the NMR experiments (PDB ID: 2KQT) [32], inhibitor-binding in this pocket was stabilized by hydrophobic interactions between the adamantane group and Val27 side chains as well as an extensive hydrogen bonding network between the positively-charged ammonium group and the pore-facing residues and water molecules (Fig. 13.2b).

This pore-binding pocket (P-binding site) has been considered the only drug binding site for several decades until in 2008 another possible drug binding site

was reported by Schnell and Chou on the protein surface (the surface-binding site or the S-binding site, PDB ID: 2RLF) [33]. This new drug-binding site is constituted by Leu40, Leu43 and Asp44 from one subunit and Ile42 and Arg45 from the adjacent subunit (Fig. 13.1). The adamantane group of the inhibitor interacted favorably with the hydrophobic side chains of Leu40, Leu43 and Ile42, whereas the positively-charged ammonium group interacted with the polar patch formed by Asp44 and Arg45. The surface-bound inhibitors could stabilize helical packing and may prevent proton conduction allosterically. This drug-protein binding model was found in micelle environments (PDB ID: 2RLF). In contrast, in bilayer environment (PDB ID: 2L0 J), the polar residues Asp44 and Arg45 were buried by hydrophobic residues of the intracellular amphipathic helices and the S-binding site was more hydrophobic than that in 2RLF. Therefore, the S-binding site was absent in 2L0 J [31].

Both drug-binding sites have been observed in experimental studies but the P-binding site has been believed to be the pharmacologically-relevant binding site [11, 32, 34]. The S-binding site is believed to only exist under specific conditions such as high drug concentration. The type B influenza virus M2 channel (B/M2) is amantadine/rimantadine insensitive [3]. The chimeric ion channel constructed by substituting the N-terminal half of the B/M2 TM domain (res. 6 to 18) with the corresponding segment of the A/M2 (residue 24–36, in the P-binding site) was partially ($\sim$50 %) sensitive to amantadine [35]. Transferring of residues 37–45 of A/M2 (outside the P-binding site) to the corresponding position of B/M2 resulted in a chimeric channel that was not sensitive to amantadine, indicating that only the N-terminal part of the channel is associated with drug inhibition [35, 36]. Pielak et al. [37] also solved an sNMR structure of a chimeric channel with a pore-bound inhibitor in micelle environments, which proved the importance of the P-binding site.

13.3.2 Free Energy Properties of the Two Binding Site

As two binding sites co-exist on the protein, extensive studies especially theoretical studies were conducted in order to explore the structural and energetic properties of these two sites. In 2008, Chuang et al. defined binding hot spots on both the X-ray crystal structure (PDB entry: 3BKD and 3C9 J) and the sNMR structure (PDB entry: 2RLF) by computational solvent mapping method [38]. The fact that binding hot spots were found at both sites on both structures implies co-existence of the two drug binding sites. However, the hot spots at the P-binding site were more preferred energetically than those at the S-binding site [39].

Our group performed molecular dynamics simulations of the M2 proton channel with inhibitors binding at different sites to investigate the stability and dynamic properties of the M2-inhibitor complexes [11]. We found that, in short molecular dynamics simulations, drugs could bind at both sites stably. In the P-binding site, the drug was bound with its ammonium group pointing to the His37

residues and spaned the N-terminal portion of the channel pore. In the S-binding site, extensive hydrophobic interactions were formed between the adamantane group of the inhibitor and the Leu40, Leu43 and Ile42 residues, whereas the positively charged ammonium group hydrogen-bonded with the protein, the lipids, as well as water molecules simultaneously as the S-binding site was close to the lipid-water interface. All of these interactions were consistent with experimental results [32, 33, 40].

We then conducted free energy calculations for both sites. Umbrella sampling methods were employed and three different reaction coordinates were designed to calculate the free energy changes of rimantadine binding in the channel pore from the N-terminal solution, rimantadine dissociating from the surface binding site and entering into the lipid-water interfacial region, as well as rimantadine penetrating the membrane bilayer. These calculations revealed a binding free energy difference of $\sim$7 kcal/mol between the P-binding site and the S-binding site with the P-binding site being the more stable site (Fig. 13.3a, b). Although the P-binding site binds inhibitor more stably, an energy barrier of $\sim$10 kcal/mol was found in the vicinity of Val27 for rimantadine to enter the channel pore from the N-terminal side (Fig. 13.3a, b). This energy barrier may be due to the small pore radius of the M2 channel at this position, as well as the need for dehydration of the positively charged ammonium group when the inhibitor passes through the hydrophobic gate. In conclusion, our calculations showed that the P-binding site was more stable for drug binding but a higher energy barrier needs to be overcome for binding to occur. The S-binding site was less stable for drug binding but it was nearly barrierless and was easily accessed. The dissociation of the drug molecule from the P-binding site needs to overcome a large energy barrier ($\sim$28 kcal/mol), which nicely explains the stability of drug-binding at this site. Dissociation of the drug molecule from the S-binding site needs to overcome a small energy barrier ($\sim$2 kcal/mol). Therefore, the P-binding site is the thermodynamic binding site where the drug molecule binds slowly and stably and dissociates even more slowly, whereas the S-binding site is a kinetic binding site where the drug molecule binds readily but less stably and dissociates easily.

We first conducted the above calculations by using the GROMOS united-atom force field for the protein and the rimantadine molecule and then we repeated them by using the OPLS all-atom force field. Both force fields revealed the same free energy differences of the two binding sites with the OPLS force file resulted in lower absolute binding free energies and slightly lower energy barrier. This result may imply that the OPLS force field is more suitable for describing the protein and the induced-fit process of drug-protein binding.

Our calculations also found an energy well of $\sim$9 kcal/mol for the rimantadine at the lipid-water interface (lipid binding site or L-binding site). We postulated that in bilayer environment rimantadine is primarily bound in the L-binding site. In order to enter into the channel pore, it needs to dissociate from the L-binding site first and then penetrates the hydrophobic gate at the channel entrance (Fig. 13.4). Both processes needed to overcome very high energy barrier ($\sim$9 kcal/mol and $\sim$10 kcal/mol), which explains the slow inhibition of the M2 proton channel

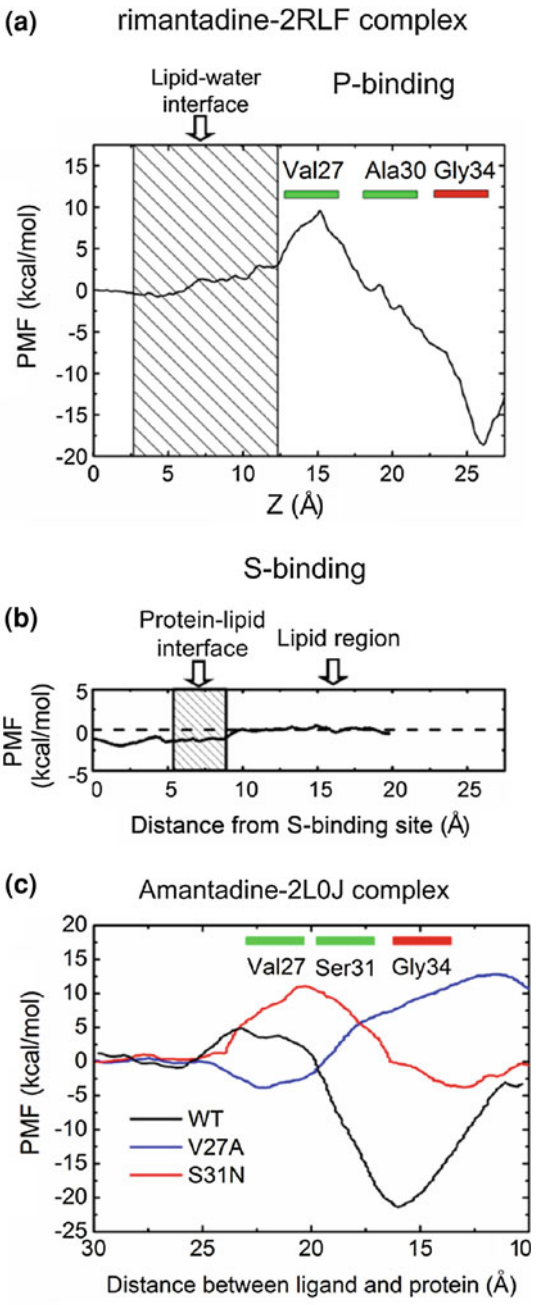


Fig. 13.3 Free energy profiles of inhibitor binding with the M2 protein. Panel **a** and Panel **b** show the free energy profiles of rimantadine binding to the P-binding site from the N-terminal solution and dissociating from the S-binding site on the protein surface and entering into the water-lipid interface (2RLF was used as the initial structure in the simulations, where the GROMOS united-atom force filed was used for the protein), respectively. Panel **c** shows the free energy profiles of amantadine binding to the channel pore of the wild-type and the drug resistant mutant M2 proteins (2L0J was used as the initial structure in the simulations, where the OPLS all-atom force filed was used for the protein) from the N-terminal entrance

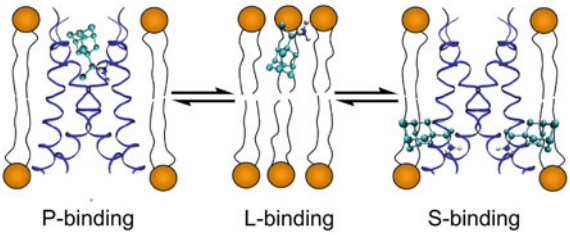


Fig. 13.4 Different drug binding states in the M2-bilayer system. In the membrane environments, the rimantadine could bind in the channel pore of the M2 protein (the P-binding site) or on the protein surface of the M2 protein (the S-binding site). There is also another binding site at the lipid-water interfacial region (the L-binding site). The L-binding site exists on both sides of the bilayer and the S-binding site is close to the L-binding site on the intracellular side

[6, 41, 42]. However, the S-binding site is close to the L-binding site at the intracellular side of the bilayer and the drug molecule bound on the protein surface could easily dissociate from the S-binding site (Fig. 13.4).

13.3.3 Structures of the Drug Resistant Channels

As mentioned earlier, the adamantane-based molecules have lost their drug efficacy gradually because of naturally-occurring drug resistant mutations [6]. The currently found drug resistant mutants could be classified into three categories according to the positions of the mutated residues on the protein. They include pore-facing mutations where the residues pointing toward the channel pore are mutated (V27A, S31 N, G34E, A30T) and the N- and C-terminal interhelical-facing mutations where the residues at the helical interfaces at the N- (L26F) and C-terminal (L38F, D44A) ends of the M2 channel are mutated, respectively (see Fig. 1 in Ref. [43]). The pore-facing mutations affect the pore radii and the hydrophilicity of the channel pore directly by varying the size and electrostatic properties of the mutated residues, whereas the interhelical facing mutations affect drug-binding allosterically by modifying the interactions between the transmembrane helices and changing the channel tetramer packing and overall structure

In order to investigate the structural changes of the mutant channels and their interactions with the M2 inhibitors, we conducted molecular dynamics simulations of the apo-form, the amantadine-bound form as well as the rimantadine-bound form of the wild-type and several drug-resistant mutant channels [43]. We simulated four different mutants and they are V27A, S31 N, L26F, and L38F, respectively. The S31 N mutation is the dominant naturally-occurring mutation among all current drug resistant influenza viruses [44, 45]. There have been some debates regarding the molecular mechanism of drug resistance for this mutant [33, 40]. The other three mutations were simulated to study the effects of mutations at different positions in the channel.

Our MD simulations showed that, the pore radii of the V27A, L26F, and L38F mutant channels were enlarged at the N-terminal side of the channel, which could be explained by the different volumes of the new residues. For V27A, a larger pore facing residue (Val27) was mutated to a much smaller alanine and the pore radii were increased significantly by $\sim$2 Å. For L26F and L38F, the leucine residues lying at the helical interfaces were mutated to larger phenylalanine residues that destabilized the helical packing and hence increased the channel pore radii by $\sim$0.5 Å. Our simulations are consistent with the results of Wang et al.'s work [46].

However, there are some debates regarding the structure of the S31 N mutant. In our simulations using 2L0 J as the initial structure, the Ser31 residue was located in the channel pore and mutation of it to larger asparagine decreased the channel pore radius significantly by $\sim$1.5 Å. However, in other structures such as 2RLF, the Ser31 residue lies at the helical interfaces and simulations using this structure showed destabilized tetramer and increased channel pore radii. Both conformations of residue 31 were found in NMR structures 2L0 J and 2LYO and each conformation account for $\sim$50 % of the snapshots of these two structures. Pore-facing conformation in these experimental structures also resulted in smaller pore radii than the helical facing conformation, as we showed in Fig. 3 in Ref. [28]. However, the positions of residue 31 in these structures were not reliable due to the precision of these structures according to Wang et al. [47] and at present, most of the existing molecular modeling and simulations support the pore-facing conformation and hence the decreased channel pore radius in the S31 N mutant.

The above-mentioned mutations also affected the water structures in the channel pore. In the wild-type M2 protein, a three-layer model of stable water structure was found around Gly34 in the absence of inhibitors (Fig. 13.2a). The Ser31 side chains, the backbone oxygen atoms of Ala30, as well as the backbone oxygen atoms of Gly34 constitute three sub-sites at the binding cavity, where well-ordered water molecules are stabilized via hydrogen bond interactions with these residues, as shown in Fig. 13.2a. Our water density maps based on MD simulations confirmed these findings [43]. The water molecules in these three layers were dynamic in our simulations and exchanged with the bulk water frequently.

In the V27A, L26F, and L38F mutant channels, this water structure was destroyed or destabilized due to the increased channel pore radii. In the wild-type channel, the Val27 residues formed a hydrophobic gate at the N-terminal entrance and the water molecules were trapped in the drug binding cavity. However, in the drug resistant mutants, the channel pore radii were increased, and the water molecules in the binding site became more dynamic and the water structure was destroyed or destabilized. In the S31 N mutant, the first layer of water molecules disappeared because of the large asparagine side chains that occupy this site, but the other two water layers became more stable.

13.3.4 Interactions Between Inhibitors and Drug Resistant Channels

MD simulations showed that, drug binding in the pore-binding site of both the wild-type and mutant channels decreased the channel pore radii around Val27 and enhanced this hydrophobic gate, consistent with Yi et al.'s work [48].

In order to analyze the energetic properties of drug-binding to the channel pore of the drug resistant channels, we conducted free energy calculations using the umbrella sampling method [28]. Figure 13.3c showed the free energy profiles of amantadine binding to the P-binding site of the wild-type channel and the S31 N and V27A mutants.

2L0 J were used as the initial structure in these simulations and the shape of the free energy profile of the wild-type channel is highly similar to the results using 2RLF (compare Fig. 13.3a and c). An energy well of $\sim$20 kcal/mol was found in the drug binding site, whereas an energy barrier of $\sim$5 kcal/mol was found in the vicinity of the hydrophobic gate. However, for the V27A and S31 N mutants, the binding free energies of amantadine were only $\sim$3 kcal/mol, implying unstable binding compared to the wild-type. There is no energy barrier for drug binding in the channel pore of V27A, which means that the pore-binding site in the V27A mutant is not stable and the drug molecule binds and dissociates easily. The energy barrier for drug binding with the S31 N mutant was increased to $\sim$12 kcal/mol due to the significantly decreased channel pore radius, implying that drug binding in this mutant was much more difficult than in the wild-type channel (Fig. 13.3c).

We would like to note that, the positions of the energy well in these two mutant channels were different from that of the wild-type channel, indicating different inhibitor binding positions (Fig. 13.3c). In the V27A mutant, the inhibitor binding site lied around Ala27, where a new hydrophobic pocket was formed because of the mutation. The strong hydrophobic interactions between the drug molecule and Val27 were significantly weakened. For the S31 N mutant, the inhibitor binding site was located in a deeper position in the channel pore because the binding cavity of the wild-type channel was partially occupied by the large asparagine side chains. The hydrophobic interactions between the adamantane group and the Val27 residues were shielded by the large, polar asparagine side chains. We proposed that the abolished hydrophobic interactions between the inhibitor and the hydrophobic gate may account for the unstable binding of drug molecules in these mutants and the subsequent drug resistance.

13.4 Proton Conduction Mechanism of the M2 Proton Channel

The M2 channel undergoes extensive conformational changes to facilitate the proton conduction process when the His37 residues are protonated. By comparing different M2 structures in the protein data bank, Acharya et al. [19] proposed that

the helical kink around Gly34 was eliminated under low environmental pH values and the helices took on a straighter conformation in the open channel. The ssNMR experiments by Hu et al. [49] found more ideal helical conformation of the His37 backbone structure and broadened conformational distribution of the M2 helices at low pH values. Their observations suggested that the M2 helices adjusted their helical kink and tilt angle in order to open the channel for proton transfer at low pH values. The C-terminal amphipathic base also changed conformations at low pH_{out}. Nguyen et al. [50] found significant structural differences of the C-terminal helices under different pH values. The amphipathic helices rotated and were positioned in a deeper position in the bilayer and were farther away from each other under low environmental pH values than under neutral pH values. Their result suggests that the C-terminal base may participate in channel opening by changing its conformation. The M2 channel opening mechanism may be somewhat similar to other ion channels, such as the KcsA K^+ channel [13] and the nicotinic acetylcholine receptors (Na^+/Ca^{2+} channels) [14], where the pore-constituting TM helices rotate and tilt to yield a wider channel pore for ion conduction.

13.4.1 Two Proposed Proton Conduction Models

There are two popular proton conduction models that have been proposed to explain the proton transfer process through the M2 channel. In the first model, called the "water wire" model, when the channel changes into an open conformation at low pH_{out}, the His37 and Trp41 adjust their side chains so that a continuous water wire is formed through which protons are transferred [27, 51, 52]. In the second model, called the "proton relay" model, the His37 residues go through transitions among several protonation and conformational states at low pH_{out} values and relay protons from one side of the channel to the other [7, 26, 27, 53, 54]. His37 residues may acquire protons from the N-terminal side of the channel at low pH_{out} and become protonated. They subsequently donate the protons to the cytoplasmic side and return to their initial protonation states and conformations by tautomerization or ring flip. One of the advantages of the "proton relay" model is that it is easy to explain the ion selectivity: other monovalent ions, such as Na^+ and K^+, cannot bind with the histidine residues and hence will not be relayed to the viral interior [53].

The major difference between these two models is whether a stable continuous water wire through which protons are transferred is formed. When pH_{out} is low, if the channel pore of the protonated M2 is large enough to accommodate a continuous water wire, the "water wire" model would take precedence. If the channel pore radius is small and does not allow the formation of a continuous water wire, then the "proton relay" model becomes the only plausible proton conduction mechanism.

As discussed earlier, the M2 channel structure is highly flexible and dynamic. As a result, both proton conduction models may accurately describe the proton conduction mechanisms in M2 channels under different conditions. For instance, the proton conduction has two saturation steps, one pseudo-saturation step occurs at $pH_{out} \sim 5.5$ and another full saturation step exists at a pH_{out} value of 4 [27]. Since the pKa values of the third and forth His37 residues are 6.3 and <5, respectively [55], three His37 residues may be protonated in the first saturation step whereas in the second full saturation step, all four histidine residues become protonated. When the four His37 residues are protonated, the electrostatic repulsion may cause the channel pore to widen, hence forming the continuous water wire. Therefore, it is possible that the "water wire" model may take effect in the full saturation step. The "proton relay" model probably takes effect in the pseudo-saturation step where the channel pore size is still relatively small and cannot accommodate a continuous water wire.

Several experimental studies have found that the full-length M2 channel transfers protons more slowly than the truncated transmembrane domain [9, 27]. For instance, Ma et al. [9] found that the transmembrane domain of the M2 channel plus a few C-terminal residues (residues 22–50) conducted protons at a rate of 7.6 s^{-1} per tetramer, whereas the full length channel conducted protons at a rate of 4.8 s^{-1} per tetramer. As discussed previously, the full-length channel is a more stable and compact structure due to the C-terminal intracellular base. Therefore, in the channel construct containing only the transmembrane domain, the channel pore radius became larger in the "water-wire" model, and the His37 residue was more flexible in the "proton relay" model. Both mechanisms would result in larger proton conduction rate in the truncated channel. We would also like to note that, the different proton conduction rates are also probably due to the different mechanisms that are at play under different conditions. It is possible that the proton conduction in full-length channel occurs primarily in the fashion as described in the "proton relay" model, whereas in the transmembrane only constructs, the channel is more open and the proton conduction might occur in a fashion as described in the "water wire" model, which may transfer protons much faster. More extensive experiments would be needed to delineate the proton conduction mechanism in in vivo systems.

13.4.2 Proton Conduction Modeled by Molecular Simulations

Molecular modeling of the M2 channel has been used to explore the proton conduction mechanisms. However, the results are highly sensitive to the choice of the starting M2 channel structures. For instance, molecular dynamics simulations based on early M2 structures (loosely-packed conformation) (e.g., PDB ID: 1NYJ) supported the "water wire" model by finding that three protonated His37 residues

resulted in a channel pore wide enough to allow the formation of a water wire [51, 56]. The lowest energy barrier for protons diffusing through the water wire was found to be ~7 kcal/mol when three His37 residues were protonated [51].

In comparison, molecular simulations based on M2 structures in closed and tightly-packed conformations solved in recent years mostly support the proton relay model. For instance, Carnevale et al. studied the "proton relay" model by using both MD simulations and quantum mechanics calculations starting from the recent crystal structure (PDB: 3LBW) [57]. They found that the protons were stored in water molecules in the vicinity of Ser31. The protons hopped among the water molecules at the N-terminal half of the channel in a similar way as in bulk water and were transferred to His37 in a nearly barrierless manner [57]. The energy barrier of proton conductance mainly came from the His37 side chain flipping.

In order for molecular modeling methods to capture in vivo proton conduction mechanisms, there are two important considerations to keep in mind. First, the intracellular amphipathic domain that is important in maintaining proper channel conformational flexibility and dynamics needs to be included. Second, the protonation states and conformational states of the His37 and Trp41 residues under different pH_{out} need to be studied thoroughly and extensively, to delineate the events (including transitions among different protonation states and conformational changes such as ring flips in His37) in the "proton relay" model. Quantum mechanical calculations where His37 residues may change protonation states as proton conduction occurs would be a valuable method for this study [58].

13.4.3 Details of the Proton Relay Process

The HxxxW quartet at the C-terminal part of the transmembrane helices is the functional core of the pH-gated M2 proton channel. Since proton conduction through the M2 channel requires the conformational changes of both His37 and Trp41, it is therefore important to understand their side chain conformations under low pH values.

Two recent studies by Hu et al. and Sharma et al. [31, 59] applied NMR experiments and molecular dynamics simulations to explore the proton conduction mechanisms of the M2 channel at low environmental pH values. Although both of their studies support the proton relay model, they found different side chain conformations for the His37 residues, leading to two different proton relay processes.

Hu et al.'s study [59] found that the His37 residues packed tightly in an edge to face conformation (Fig. 13.5a) [31, 59]. The His37 residues underwent multiple reorientations and very large side chain rotations (change of χ_2 angle at about 180°) to relay protons (Fig. 13.5b). They estimated a proton conduction energy barrier of ~25 kcal/mol. The energy barrier for histidine side chain reorientation is > 14 kcal/mol [59]. The proton hopping energy barrier through a continuous

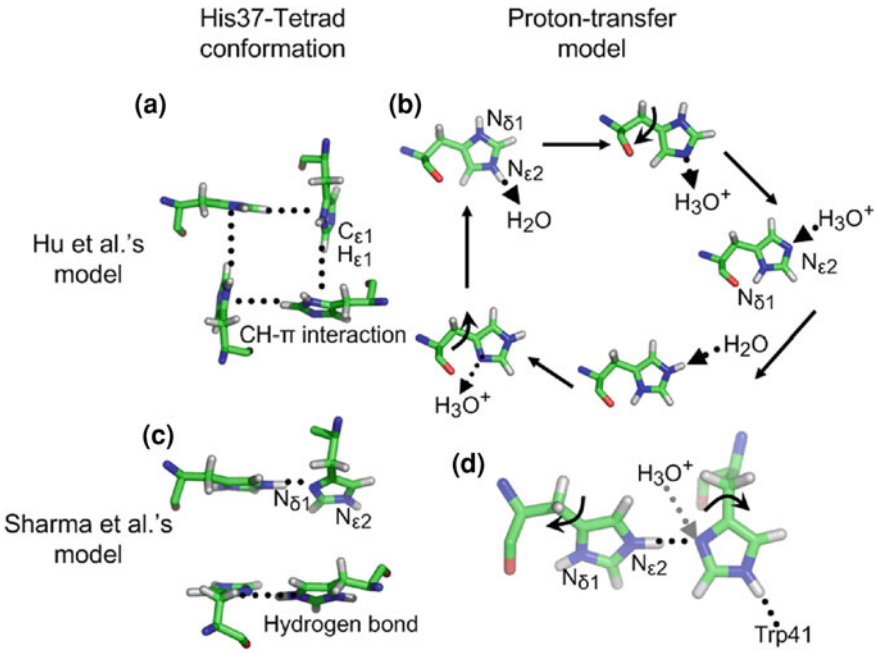


Fig. 13.5 Details of the "proton relay" model. Panel **a** shows the conformations of the His37 tetrad under neutral pH in Hu et al.'s model, whereas panel **c** is for Sharma et al.' model. In Hu et al.'s model, the four histidine residues pack tightly in an edge to face conformation in which the $C_{\varepsilon 1}$–$H_{\varepsilon 1}$ of one histidine interacts with the electron rich imidazole ring of its adjacent histidine (CH-π interaction). In Sharma et al.'s model, two His-HisH$^+$ dimers are formed by sharing a hydrogen atom between the $N_{\delta 1}$ atom of one histidine residue and the $N_{\varepsilon 2}$ atom of the adjacent His37. Panels **b** and **d** show the proton relay process in Hu et al.'s model and Sharma et al.'s model, respectively. In Hu et al.'s model, much larger side chain conformational changes of His37 residues are found. In Sharma et al.'s model, one proton is transferred to the $N_{\delta 1}$ atom of one histidine residue and the His-HisH$^+$ dimer dissociates into two monomers in low pH environment. The His37 and Trp41 residues then change their side chain conformations to release protons to the cytoplasmic side. Panels **a** and **b** are drawn based on Cady et al.'s structure (PDB ID: 2KQT) [32] and panels **c** and **d** are drawn based on Sharma et al.'s structure (PDB ID: 2L0 J) [31]

water wire is in the range of 7–10 kcal/mol [51]. These energy values suggest that the proton relay conduction mechanism was likely at play in the Hu et al.'s study.

Sharma et al.'s study [31] suggested that the His-Trp quartet changed its conformations between three states, namely the "histidine-locked state", the "activated state", and the "conducting state". Through the transitions among these three states, protons were relayed from the extracellular side to the cytoplasmic side of the channel (Fig. 13.5d). At physiological pH, one proton was believed to be shared between two adjacent histidine residues and the channel was locked by two His-HisH$^+$ dimers (Fig. 13.5c) [31, 60]. In the Sharma et al.'s proton conduction model, at low pH_{out}, one hydronium ion approaches the $N_{\delta 1}$ atom of one of

the deprotonated histidine residues from the N-terminal side to break the His-$HisH^+$ dimer (Fig. 13.5d). The "histidine locked state" is then transitioned to the "activated state". The cation-π interaction between the His37 and Trp41 side chains stabilizes this "activated state". The Trp41 residues form a gate at the C-terminal end of the channel. When the Trp41 residues change their conformations, the "activated state" is transitioned to the "conducting state" where the cation-π interaction is perturbed and the protons are donated to the water molecules at the C-terminal side of the HxxxW quartet. Unlike Hu et al.'s model where large conformational changes of the His37 residues are required during proton conduction, the conversions among the three states in Sharma et al.'s model are accomplished by small changes of the χ_2 angles (<45°) of the His37 and Trp41 side chains. The differences of these two models may be attributed to the different protein chain lengths used in the studies. The C-terminal intracellular helices were included in Sharma et al.'s experiments, which resulted in a more stable channel conformation and reduced flexibility of the His37 residues.

The most important differences between these two proton relay models are the conformations of His37 residues and how the extra protons are stored in the vicinity of His37 under different pH values. In Hu et al.'s model, the protons are stored in the water molecules interacting with the His37 residues under natural pH values and His37's undergo extensive conformational changes to relay them from one side of the channel to another. In Sharma et al.'s model, the protons are stored in the channel by the His-$HisH^+$ dimers and relatively smaller conformational changes of the His37 residues are necessary in proton conduction. Both models have extensive experimental and theoretical supports, which have been reviewed carefully in two recent reviews [15, 16].

13.4.4 Effects of Mutations on Proton Conduction

Table 13.1 summarizes several single amino acid mutations of the M2 channel and their effects on the proton conduction properties. It is generally believed that mutations that increase the pore size would allow easier conduction of protons (either through a continuous water wire or proton transfer via His37 protonation state and conformational change) through the channel, and vice versa. For example, substitution of the Val27 residue with the smaller alanine largely abolishes the N-terminal gate, allowing easier entry of water molecules and pore hydration, hence accelerates proton conduction [5, 61]. On the other hand, mutations that decrease the pore size such as V27T, G34E, A30T, and A30P interfere with pore hydration and slow down proton conduction (Table 13.1). The Ser31 residue was often observed to be facing the pore, and its mutation into the hydrophobic alanine was believed to cause poor pore hydration and decreased proton conduction [5, 61]. However, the Ser31 to asparagine mutation has shown some conflicting proton conduction rates. Based on the sNMR structure (PDB ID: 2RLF), the Ser31 residue lies in between two α helices, and its mutation into larger

asparagine would increase pore size and consequently increase the proton conduction rate. In contrast, the Ser31 residue faces the pore in the crystal structure (PDB ID: 3C9 J) and mutation into asparagine would decrease the pore size and the proton conduction rate. Holsinger et al. [5] found that the proton conduction rate of the S31 N mutant channel was nearly the same as that of the wild-type channel at an environmental pH of 6.2. Therefore, the effects of the S31 N mutation on the proton conduction in the M2 channel remain to be fully understood. The Asp44 residues form hydrogen bonds with several water molecules that may accept protons at the C-terminus of the channel. Replacing Asp44 with the hydrophobic alanine causes an increase of the energy barrier for proton exit and hence decreases the proton conduction rate [27] (Table 13.1). Other mutations may also exist that could affect the conformations of the His37 and Trp41 residues, which would lead to changes in the proton conduction rate.

13.5 Conclusions

The M2 proton channel in the influenza viral envelop has a relatively simple molecular structure, but exhibits a wide range of structural and functional properties due to the flexibility of its transmembrane domain. Due to its crucial physiological role in the viral infection and replication, it has been used as a target for anti-influenza drugs. However, its structural flexibility led to difficulties and controversies in understanding the drug inhibition mechanism and the proton conduction mechanism.

At present, two drug binding sites are believed to co-exist in the channel pore and on the protein surface, respectively. Extensive computational and experimental studies showed that, drug binding in the channel pore is more stable than binding on the protein surface, and therefore, the pore binding site is the pharmacologically relevant site.

The adamantane-based M2 inhibitors have lost their channel inhibition efficacy due to naturally occurring drug resistant mutations, among which the S31 N mutant is the most dominant one. Our MD simulations found significant changes of the channel pore radii and the water structures in the channel pore of the mutant proteins. We also calculated the binding free energies of amantadine with the V27A and S31 N mutants, and found that, the hydrophobic interactions between the inhibitor and the Val27 side chains are either abolished or shielded. The weakened hydrophobic interactions between the drug and the protein may be responsible for the observed drug resistance.

Two proton conduction models have been proposed and they are the “water wire” model and the “proton relay” model. The “proton relay” model is gaining more support in recent studies. However, there are still extensive debates regarding the details of the proton relay process, mainly because the conformations of the His37 and Trp41 residues under different pH values are yet to be fully investigated and understood.

In summary, environmental conditions and the intrinsic protein flexibility affect the structures and functions of the M2 protein. Such factors may also play important roles in other membrane proteins.

References

1. Cady SD, Luo W, Hu F et al (2009) Structure and function of the influenza A M2 proton channel. Biochemistry 48(31):7356–7364
2. Pinto LH, Lamb RA (2006) Influenza virus proton channels. Photochem Photobiol Sci 5(6):629–632
3. Pinto LH, Lamb RA (2006) The M2 proton channels of influenza A and B viruses. J Biol Chem 281(14):8997–9000
4. Chizhmakov IV, Geraghty FM, Ogden DC et al (1996) Selective proton permeability and pH regulation of the influenza virus M2 channel expressed in mouse erythroleukaemia cells. J Physiol 494(Pt 2):329–336
5. Holsinger LJ, Nichani D, Pinto LH et al (1994) Influenza A virus M2 ion channel protein: a structure-function analysis. J Virol 68(3):1551–1563
6. Wang C, Takeuchi K, Pinto LH et al (1993) Ion channel activity of influenza A virus M2 protein: characterization of the amantadine block. J Virol 67(9):5585–5594
7. Sakaguchi T, Tu Q, Pinto LH et al (1997) The active oligomeric state of the minimalistic influenza virus M2 ion channel is a tetramer. Proc Natl Acad Sci USA 94(10):5000–5005
8. Lamb RA, Zebedee SL, Richardson CD (1985) Influenza virus M2 protein is an integral membrane protein expressed on the infected-cell surface. Cell 40(3):627–633
9. Ma C, Polishchuk AL, Ohigashi Y et al (2009) Identification of the functional core of the influenza A virus A/M2 proton-selective ion channel. Proc Natl Acad Sci 106(30):12283–12288
10. Wang J, Qiu JX, Soto C et al (2011) Structural and dynamic mechanisms for the function and inhibition of the M2 proton channel from influenza A virus. Curr Opin Struct Biol 21(1):68–80
11. Gu RX, Liu LA, Wei DQ et al (2011) Free energy calculations on the two drug binding sites in the M2 proton channel. J Am Chem Soc 133(28):10817–10825
12. Tang Y, Zaitseva F, Lamb RA et al (2002) The gate of the influenza virus M2 proton channel is formed by a single tryptophan residue. J Biol Chem 277(42):39880–39886
13. Cuello LG, Jogini V, Cortes DM et al (2010) Structural mechanism of C-type inactivation in K+ channels. Nature 466(7303):203–208
14. Miller PS, Smart TG (2010) Binding, activation and modulation of Cys-loop receptors. Trends Pharmacol Sci 31(4):161–174
15. Zhou HX, Cross TA (2013) Modeling the membrane environment has implications for membrane protein structure and function: influenza A M2 protein. Protein Sci 22(4):381–394
16. Hong M, DeGrado WF (2012) Structural basis for proton conduction and inhibition by the influenza M2 protein. Protein Sci 21(11):1620–1633
17. Thomaston JL, Nguyen PA, Brown EC et al (2013) Detection of drug-induced conformational change of a transmembrane protein in lipid bilayers using site-directed spin labeling. Protein Sci 22(1):65–73
18. Salom D, Hill BR, Lear JD et al (2000) pH-dependent tetramerization and amantadine binding of the transmembrane helix of M2 from the influenza A virus. Biochemistry 39(46):14160–14170
19. Acharya R, Carnevale V, Fiorin G et al (2010) Structure and mechanism of proton transport through the transmembrane tetrameric M2 protein bundle of the influenza A virus. Proc Natl Acad Sci 107(34):15075–15080

20. Cady S, Wang T, Hong M (2011) Membrane-dependent effects of a cytoplasmic helix on the structure and drug binding of the influenza virus M2 protein. J Am Chem Soc 133(30):11572–11579
21. Cross TA, Dong H, Sharma M et al (2012) M2 protein from influenza A: from multiple structures to biophysical and functional insights. Curr Opin Virol 2(2):128–133
22. Cross TA, Sharma M, Yi M et al (2011) Influence of solubilizing environments on membrane protein structures. Trends Biochem Sci 36(2):117–125
23. Duong-Ly KC, Nanda V, DeGrado WF et al (2005) The conformation of the pore region of the M2 proton channel depends on lipid bilayer environment. Protein Sci 14(4):856–861
24. Kovacs F, Denny JK, Song Z et al (2000) Helix tilt of the M2 transmembrane peptide from influenza A virus: an intrinsic property. J Mol Biol 295(1):117–125
25. Lear JD (2003) Proton conduction through the M2 protein of the influenza A virus; a quantitative, mechanistic analysis of experimental data. FEBS Lett 552(1):17–22
26. Mould JA, Li HC, Dudlak CS et al (2000) Mechanism for proton conduction of the M2 ion channel of influenza A virus. J Biol Chem 275(12):8592–8599
27. Pielak RM, Chou JJ (2010) Kinetic analysis of the M2 proton conduction of the influenza virus. J Am Chem Soc 132:17695–17697
28. Gu R-X, Liu LA, Wei D-Q (2013) Structural and energetic analysis of drug inhibition of the influenza A M2 proton channel. Trends Pharmacol Sci 34(10):571–580
29. Nishimura K, Kim S, Zhang L et al (2002) The closed state of a H+ channel helical bundle combining precise orientational and distance restraints from solid state NMR. Biochemistry 41(44):13170–13177
30. Hu J, Asbury T, Achuthan S et al (2007) Backbone structure of the amantadine-blocked trans-membrane domain M2 proton channel from influenza A virus. Biophys J 92(12):4335–4343
31. Sharma M, Yi M, Dong H et al (2010) Insight into the mechanism of the influenza A proton channel from a structure in a lipid bilayer. Science 330(6003):509–511
32. Cady SD, Schmidt-Rohr K, Wang J et al (2010) Structure of the amantadine binding site of influenza M2 proton channels in lipid bilayers. Nature 463(7281):689–692
33. Schnell JR, Chou JJ (2008) Structure and mechanism of the M2 proton channel of influenza A virus. Nature 451(7178):591–595
34. Rosenberg MR, Casarotto MG (2010) Coexistence of two adamantane binding sites in the influenza A M2 ion channel. Proc Natl Acad Sci 107(31):13866–13871
35. Jing X, Ma C, Ohigashi Y et al (2008) Functional studies indicate amantadine binds to the pore of the influenza A virus M2 proton-selective ion channel. Proc Natl Acad Sci 105(31):10967–10972
36. Ohigashi Y, Ma C, Jing X et al (2009) An amantadine-sensitive chimeric BM2 ion channel of influenza B virus has implications for the mechanism of drug inhibition. Proc Natl Acad Sci 106(44):18775–18779
37. Pielak RM, Oxenoid K, Chou JJ (2011) Structural investigation of rimantadine inhibition of the AM2-BM2 chimera channel of influenza viruses. Structure 19(11):1655–1663
38. Brenke R, Kozakov D, Chuang GY et al (2009) Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. Bioinformatics 25(5):621–627
39. Chuang GY, Kozakov D, Brenke R et al (2009) Binding hot spots and amantadine orientation in the influenza A virus M2 proton channel. Biophys J 97(10):2846–2853
40. Stouffer AL, Acharya R, Salom D et al (2008) Structural basis for the function and inhibition of an influenza virus proton channel. Nature 451(7178):596–599
41. Balannik V, Wang J, Ohigashi Y et al (2009) Design and pharmacological characterization of inhibitors of amantadine-resistant mutants of the M2 ion channel of influenza A virus. Biochemistry 48(50):11872–11882
42. Wang J, Cady SD, Balannik V et al (2009) Discovery of spiro-piperidine inhibitors and their modulation of the dynamics of the M2 proton channel from influenza A virus. J Am Chem Soc 131(23):8066–8076

43. Gu R-X, Liu LA, Wang Y-H et al (2013) Structural comparison of the wild-type and drug-resistant mutants of the influenza A M2 proton channel by molecular dynamics simulations. J Phys Chem B 117(20):6042–6051
44. Deyde VM, Xu X, Bright RA et al (2007) Surveillance of resistance to adamantanes among influenza A (H3N2) and A (H1N1) viruses isolated worldwide. J Infect Dis 196(2):249–257
45. Saito R, Sakai T, Sato I et al (2003) Frequency of amantadine-resistant influenza A viruses during two seasons featuring cocirculation of H1N1 and H3N2. J Clin Microbiol 41(5):2164–2165
46. Wang J, Ma C, Fiorin G, et al (2011) Molecular dynamics (MD) simulation directed rational design of inhibitors targeting drug-resistant mutants of influenza A Virus M2. J Am Chem Soc 133(32):12834–12841
47. Wang J, Wu Y, Ma C et al (2013) Structure and inhibition of the drug-resistant S31 N mutant of the M2 ion channel of influenza A virus. Proc Natl Acad Sci 110(4):1315–1320
48. Yi M, Cross TA, Zhou HX (2008) A secondary gate as a mechanism for inhibition of the M2 proton channel by amantadine. J Phys Chem B 112(27):7977–7979
49. Hu F, Luo W, Cady SD et al (2011) Conformational plasticity of the influenza A M2 transmembrane helix in lipid bilayers under varying pH, drug binding, and membrane thickness. Biochimica et Biophysica Acta (BBA)-Biomembranes 1808(1):415–423
50. Nguyen PA, Soto CS, Polishchuk A et al (2008) pH-induced conformational change of the influenza M2 protein C-terminal domain. Biochemistry 47(38):9934–9936
51. Chen H, Wu Y, Voth GA (2007) Proton transport behavior through the influenza A M2 channel: insights from molecular simulation. Biophys J 93(10):3470–3479
52. Okada A, Miura T, Takeuchi H (2001) Protonation of histidine and histidine-tryptophan interaction in the activation of the M2 ion channel from influenza A virus. Biochemistry 40(20):6053–6060
53. Pinto LH, Dieckmann GR, Gandhi CS et al (1997) A functionally defined model for the M2 proton channel of influenza A virus suggests a mechanism for its ion selectivity. Proc Natl Acad Sci USA 94(21):11301–11306
54. Sansom MSP, Kerr ID, Smith GR et al (1997) The influenza A virus M2 channel: a molecular modeling and simulation study. Virology 233(1):163–173
55. Hu J, Fu R, Nishimura K et al (2006) Histidines, heart of the hydrogen ion channel from influenza A virus: toward an understanding of conductance and proton selectivity. Proc Natl Acad Sci 103(18):6865–6870
56. Kass I, Arkin IT (2005) How pH opens a H+ channel: the gating mechanism of influenza A M2. Structure 13(12):1789–1798
57. Carnevale V, Fiorin G, Levine BG et al (2010) Multiple proton confinement in the M2 channel from the influenza A virus. J Phys Chem C 114(48):20856–20863
58. Dong H, Yi M, Cross TA et al (2013) Ab initio calculations and validation of the pH-dependent structures of the His37-Trp41 quartet, the heart of acid activation and proton conductance in the M2 protein of Influenza A virus. Chem Sci 4(7):2776–2787
59. Hu F, Luo W, Hong M (2010) Mechanisms of proton conduction and gating in influenza M2 proton channels from solid-state NMR. Science 330(6003):505–508
60. Hu J, Fu R, Cross TA (2007) The chemical and dynamical influence of the anti-viral drug amantadine on the M2 proton channel transmembrane domain. Biophys J 93(1):276–283
61. Pielak RM, Chou JJ (2010) Solution NMR structure of the V27A drug resistant mutant of influenza A M2 channel. Biochem Biophys Res Commun 401:58–63

Chapter 14
Exploring the Ligand-Protein Networks in Traditional Chinese Medicine: Current Databases, Methods and Applications

Mingzhu Zhao and Dongqing Wei

Abstract While the concept of "single component–single target" in drug discovery seems to have come to an end, "Multi-component–multi-target" is considered to be another promising way out in this field. The Traditional Chinese Medicine (TCM), which has thousands of years' clinical application among China and other Asian countries, is the pioneer of the "Multi-component–multi-target" and network pharmacology. Hundreds of different components in a TCM prescription can cure the diseases or relieve the patients by modulating the network of potential therapeutic targets. Although there is no doubt of the efficacy, it is difficult to elucidate convincing underlying mechanism of TCM due to its complex composition and unclear pharmacology. Without thorough investigation of its potential targets and side effects, TCM is not able to generate large-scale medicinal benefits, especially in the days when scientific reductionism and quantification are dominant. The use of ligand-protein networks has been gaining significant value in the history of drug discovery while its application in TCM is still in its early stage. This article firstly surveys TCM databases for virtual screening that have been greatly expanded in size and data diversity in recent years. On that basis, different screening methods and strategies for identifying active ingredients and targets of TCM are outlined based on the amount of network information available, both on sides of ligand bioactivity and the protein structures. Furthermore, applications of successful *in silico* target identification attempts are discussed in details along with experiments in exploring the ligand-protein networks of TCM. Finally, it will be concluded that the prospective application of ligand-protein networks can be used not only to predict protein targets of a small molecule, but also to explore the mode of action of TCM.

Keywords Traditional Chinese Medicine · Multiple components · Multiple targets · Ligand-Protein networks · TCM databases

M. Zhao · D. Wei (✉)
State Key Laboratory of Microbial Metabolism, College of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China
e-mail: dqwei@sjtu.edu.cn

D. Wei et al. (eds.), *Advance in Structural Bioinformatics*, Advances in Experimental Medicine and Biology 827, DOI 10.1007/978-94-017-9245-5_14

14.1 Introduction

Drug discovery was once an empirical process when the effect of the medicine was purely based on phenotype readout, while the mode of action of drug molecules remained unknown. Later, reductionists began to research on the molecular mechanism of the drug-target interactions, believing that the drug is like a magic bullet towards the functioning targets [1]. This means a drug takes action on the disease by interacting with one specific therapeutic target. The idea of each drug being like a key (or ligand) matching each 'lock' (or protein) has guided the modern drug discovery practice for the last several decades. However, in the recent years, more and more evidence has shown that many drugs exert their activities by modulating multi-targets [2–4]. Besides, some drugs interact with anti-targets and induce strong side effects [5, 6]. Therefore, it is inappropriate to stick to the paradigm that drug interact with only one target. How to modulate a set of targets to achieve efficacy, while avoiding others to reduce the risk of side effects remains a central challenging task for pharmaceutical industry.

The Traditional Chinese Medicine (TCM), which has been widely used in China as well as other Asian countries for a long history, is considered to be the pioneer of the "Multi-component—multi-target" pharmacology [7, 8]. Thousands of years' clinic practices in TCM have accumulated a considerable number of formulae that exhibit reliable in vivo efficacy and safety. Based on the methodology of holism, hundreds of different components in a TCM prescription can cure the diseases or relieve the patients by modulating a serial of potential therapeutic targets [9].

In recent years, great efforts have been made on modernization of TCM, most on identification of effective ingredients and ligands in TCM formulae and functioning targets [10, 11]. Several databases of TCM formulae, ingredients and compounds with chemical structures have been established such as Traditional Chinese Medicine Database (TCMD) [12]. However, the molecular mechanisms responsible for their therapeutic effectiveness are still unclear. On one hand, experimental validation of new drug-target interactions still remains very limiting and expensive, and very few new drugs and targets are identified as clinical applications every year [13, 14]. On the other hand, the complex composition and polypharmacology of TCM make it even harder to conduct a full set of experiments between compounds and targets and elucidate the multi-target mode of action from the holistic view on the biological network level.

On the contrary, *in silico* methods can predict a large number of new drug-target interactions, construct the drug-target networks, and explore the functional mechanism underlying the multi-component drug combinations at the molecular level. In the present stage, there have already been successfully applications in interpreting the action mechanism of TCM from the perspective of drug-target networks, although the quantity is limited. Compared with the huge amounts of TCM formulae and components, only a small portion of drug-target pairs have been validated by the laborious and costly biochemical experiments. This

motivates the needs for constructing models that could predict genuine interacting pairs between ligands and targets, based on the existing small number of known ligand-target bindings.

In this article, we firstly investigate TCM databases for *in silico* methods that have been greatly expanded in size and data diversity in recent years. On that basis, different screening methods and strategies for identifying active ingredients and targets of TCM are outlined based on the amount of information available, both on sides of ligand bioactivity and the protein structures. Finally, successful applications in this area have been summarized and reviewed, including experimental and computational examples. Learning from the methods in modern western medicine (WM), different computing models and strategies can be used to confirm the effective components and related targets in TCM in order to build the ligand-target networks. One of the research directions of the modernization of TCM is to clarify the mode of action of TCM based on ligand-protein networks.

14.2 Databases for TCM

Data availability is the first consideration before any virtual screening or data mining task could be undertaken. The TCM databases can be classified in accordance to several categories, namely formulae, herbs, and compounds. The formula of TCM is a combination of herbs for treating a disease, while compounds are the bioactive molecules within herbs. In this section, we have summarized a list of databases for TCM herbs, formulations and compounds, as shown in Table 14.1.

The elementary units of TCM databases are compounds, the bioactive components that exert efficacy through binding to therapeutic targets. Most of the compounds in TCM databases have two-dimensional structure, while some of them have three-dimensional structures deduced by force filed. In most TCM databases, the information of both herbs and compounds are collected while some even have formulae information as well.

The Traditional Chinese Medicine Database (TCMD) contains 23,033 chemical constituents and over 6,760 herbs that mainly come from Yan et al. [12]. The query keywords for the database include molecular formula, substructure, botanical identity, CAS number, pharmacological activity and traditional indications. Only a small proportion of herbs in TCMD have full coverage of compounds while most have partial coverage. Chinese Herb Constituents Database (CHCD) contains information on 8,264 compounds derived from 240 commonly used herbs with both botanical and Chinese pinyin names, the part of the herbs that contain the compounds, pharmacological and toxicological information, and other useful information [15]. Qiao et al. [16] have developed 3D structural database of biochemical components which covers 10,564 constituents from 2,073 herbs with 3D structures built and optimized using the MMFF94 force field [17]. This database uses MySQL as the data engine and contains detailed information such as basic

Table 14.1 Basic information for main TCM databases

Database	Description	ULR or References
Traditional Chinese Medicine Database (TCMD)	6,760 herbs, 23,033 compounds	[12]
Chinese Herb Constituents Database (CHCD)	240 herbs, 8,264 compounds	[15]
3D structural database of biochemical components	2,073 herbs, 10,564 compounds	[16]
TCM Database@Taiwan	453 herbs, 20,000 compounds	[18]
Traditional Chinese Medicine Information Database (TCM-ID)	1,197 formulae, 1,313 herbs, ~9,000 compounds	[19]
TCM Drugs Information System	1,712 formulae, 2,738 herbs, 16,500 compounds, 868 dietotherapy prescription	[20]
Comprehensive Herbal Medicine Information System for Cancer (CHMIS)-C	203 formulae, 900 herbs, 8,500 compounds	[21]
China Natural Products Database (CNPD)	45,055 compounds	[22]
Marine Natural Products Database (MNPD)	8,078 compounds, 3,200 with bioactivity data	[23]
Bioactive Plant Compounds Database (BPCD)	2,794 compounds	[15]
Acupuncture.com.au	TCM formulations	http://www.acupuncture.com.au/education/herbs/herbs.html
Dictionary of Chinese Herbs	TCM formulae, toxicity and side effects	http://alternativehealing.org/Chinese herbs dictionary.htm
Plants For a Future	Herb medical usage and potential side effects	http://www.pfaf.org

molecular properties, optimized 3D structures, herb origin and clinical effects. The TCM Database@Taiwan was reported to be the world's largest traditional Chinese medicine database. The web-based database contains more than 20,000 pure compounds isolated from 453 TCM herbs [18]. Both simple and advanced query methods are acceptable in terms of molecular properties, substructures, TCM ingredients and TCM classifications.

In addition to herbs and compounds, Traditional Chinese Medicine Information Database (TCM-ID) [19], TCM Drugs Information System [20], and Comprehensive Herbal Medicine Information System for Cancer (CHMIS-C) [21] also collect the information of TCM formulae. TCM-ID is developed by Zhejiang University together with National University of Singapore on all aspects of TCM

herbs. TCM-ID currently takes in 1197 TCM formulae, 1,313 herbs and around 9,000 compounds. It covers ~4,000 disease conditions and more than half of the compounds have valid 3D structures. The data are collected from creditable TCM books as well as Journals and the records can be retried by different sets of query keywords. TCM Drugs Information System based on networks of five large databases has also been developed [20]. It includes information of 1,712 formulae, 2,738 herbs, 16,500 compounds, 868 dietotherapy prescriptions from the integration of Chinese herb database, Chinese patent medicine database, effective components database of Chinese herbs, Chinese medical dietotherapy prescription database, and Chinese medical recipe database. Herbal Medicine Information System for Cancer (CHMIS-C) integrates the information of 203 formulae that are commonly used to treat cancer clinically as well as 900 herbs and 8,500 compounds. The compounds in this database are linked to the entries in National Cancer Institute's database and drugs approved by the U.S. Food and Drug Administration.

The China Natural Products Database (CNPD) [22], Marine Natural Products Database (MNPD) [23], and Bioactive Plant Compounds Database (BPCD) [15] only focus on the structures of the compounds in TCM and do not contain pertinent information on formulae and herbs. CNPD is built to meet the needs for drug discovery using natural products including TCM and collects the 2D and 3D structures of more than 45,055 compounds. MNPD has a collection of 8,078 compounds from 10,000 marine natural products, of which 3,200 have bioactivity data. BPCD contains information on 2,794 active compounds against 78 molecular targets, as well as the subunits of the target structures to which the compounds bind.

There are other databases from the internet focusing only on the clinical efficacy or side effects of formulae and herbs, without details of compounds. Acupuncture.com.au collects the TCM formulae according to their clinical action and efficacy. Both the English and Chinese names of TCM herbs are recorded to facilitate studies using both traditional and modern methods. The Dictionary of Chinese Herbs contains information on both clinical usage and side effects of the TCM herbs. It also includes the samples of TCM formulae for treating diseases such as cancer, dengue fever, diabetes, and hepatitis B. Besides, the compatibility of TCM herbs and certain drugs are listed to provide biochemical explanation for drug designers. The Plants for a Future database allows querying of herbs with special medicinal usage, and also lists the potential side effects, medical usage, and physical characteristics.

14.3 In Silico Methods for Ligand-Protein Interactions

The computational methods for drug discovery based on ligand-protein networks have been increasingly developed and applied in the area of TCM and other drugs in recent years [7, 8]. These methods mainly fall into the territories of ligand-based

approach, target-based approach, and machine learning. Of course, these methods of predicting ligand-protein interactions are not isolated, and researchers often use them jointly to achieve better computational results, which can be easily shown in the following case studies.

14.3.1 Ligand-Based Approach

The ligand-based approach, also known as the chemical approach, is to reorganize pharmacological characteristics and protein associations, by means of ligand similarities rather than genomic space such as sequence, structural or pathway information. The basic assumption for ligand-based approach is that regardless that similar chemical structures may interact with proteins in different ways, similar ligands tend to bind to similar targets more than not [24]. The core of ligand-based approach is the calculation of chemical similarities, with the help of chemical descriptors. Before this approach can work, one need to answer how to describe molecular structures in a way that computers can recognize. Currently there are plenty of molecular descriptors to indicate the similarity of two different ligands.

14.3.1.1 Chemical Descriptors

In order to predict the ligand-target interactions, prior knowledge should be acquired in terms of the ligand information for the target [25]. By comparing the chemical structures of the new ligands against the know ligand set of a targeting protein, a threshold is usually set to decide whether the new ligand and the targeting protein can interact.

The most commonly used structure representation is the topological fingerprints that encode the sub-structural information [26]. In these fingerprints, the atom-centred feature pairs have also been proven to be very successful in many applications of virtual screening. The widely used samples of fingerprints are 2D *Daylight* [25] and *Scitegic* extended connectivity fingerprints [27] with atom types and the bond connectivity among them.

Although 2D fingerprints have been proven to be extremely robust and reliable in many chemoinformatics approaches, they seem hard to credit to be informative, therefore, consistent efforts have been made to develop more comprehensive three-dimensional fingerprints [28, 29]. 3D fingerprints encode the 3D geometry or scaffolds of molecular structures. One method to encode a compound is based on geometrical configuration of molecular structures. A common methodology in descriptors *Flexible* is the superposition of molecules onto one or multiple conformations of a reference bioactive ligand [30–32].

There are other topological descriptors based on molecular features that have been developed to compare ligand profiles. The *SHED* (SHannon Entropy Descriptors) is derived from distributions of atom-centered pairs and calculates the

variability in a feature-pair distribution [33]. Gregori-Puigjane then used the *SHED* descriptor to *in silico* profiling of 767 drugs against 684 related targets and revealed the promiscuity of the drugs targeting aminergic G protein-coupled receptors (GPCRs) [34]. On the other hand, *RED* (Renyi entropy descriptors) is another topological descriptor that measures the molecular features generalized Renyi Entropy. The scaffolds can also be used to predict the bioactivity of the compounds on target sets. In this particular research, 24,000 unique scaffolds were extracted from 458 target sets and the external test shows that to the high-priority virtual scaffolds have the predictive activities [35].

Also in the area of chemicogeneric-based predictive methods to screen ligand-target interactions Weil proposed a novel fingerprint encoding both ligand and target properties. The ligand properties are represented by common descriptors, while the cavity information of the target is incorporated by a fixed length bit string. This fingerprint shows preference to support vector machine (SVM) classifiers and the resulting precision is as high as 90 % in separating true and false pairs [36].

14.3.1.2 Similarity Coefficient

The most common way to compare molecular fingerprints for similarity analysis is by means of Tanimoto Coefficient (T_C, also known as Jaccard index) [37–39], which compares the number of bits shared between the two fingerprints to all possible matched bits between them,

$$T_C = N_{AB}/(N_A + N_B - N_{AB})$$

where N_A is number of features (ON bits) in compound A, N_B is the number of features (ON bits) in compound B, and N_{AB} is the number of features (ON bits) common to both A and B. If the Tanimoto Coefficient of two molecules is larger than 0.85, then they are considered to have a higher structurally similarity [40].

14.3.1.3 Ligand-Based Predictions

One advantage of the ligand-based similarity searching approach is that it does not need alignment between multiple molecules. The ligand-based approach describes a protein by the chemicgenomic space of its ligands. With the ligand-based descriptions of a protein, one can predict which targets are likely to be hit by a ligand, given its known structure.

In the area of ligand-based virtual screening, researchers have tried to evaluate whether novel ligand-target pairs could be identified, based on the chemical knowledge of ligands and ligand-target interactions. G protein-coupled receptors (GPCRs) are a family of effective drug targets with significant therapeutic value. Many researchers have built SVM models as well as substructural analysis to describe GPCRs from the perspective of ligand chemicogenerics [41]. Especially,

the de-orphanization of receptors without known ligands was employed using the ligands of the related receptors. For 93 % of the orphan receptors, the prediction results are better than random, while for 35 % the performance was good.

A powerful ligand-based prediction method based on features of protein ligands is the Similarity Ensemble Approach (SEA), which was originally used to investigate protein similarity based on chemical similarity between their ligand sets with the main idea that similar ligands might tend to share same targets [3]. SEA calculates Z-score and E-value by summing up the T_C over a threshold between two ligand sets as indicators to evaluate the possible interaction between two ligand sets in a way similar to BLAST. The similarity threshold for T_C is chosen in a way that the Z-score best observes the extreme value distribution (EVD). This method was then applied to predict new molecular targets for known drugs [42]. The Author investigated 3000 FDA-approved drugs against hundreds of targets and found 23 new cases of drug-target interactions. By *in vitro* experiments, five of them were validated to be positive with affinities less than 100 nM. Besides Keiser's research, SEA was also used to investigate the off-target effect of the some commercial available drugs against the target protein farnesyltransferase (PFTase) [43] and two drug loratadine and miconazole were found to be able to bind to PFTase.

The pharmacophore model is perhaps the most widely used methods that make use of the 3D structure representations of molecules [44]. A pharmacophore is defined to be the molecular features pertinent to bioactivity aligned in three dimensional spaces, including hydrogen bonding, charge transfer, electrostatic and hydrophobic interactions [45]. The underlying methodology of pharmacophore model was defined by different researchers [46]. Recently, this model was successfully applied in mesangial cell proliferation inhibitor discovery and virtual screening of potential ligands for many targets such as HIV integrase and CCR5 antagonist [47–50]. In 3D pharmacophore model, the molecular spatial features and volume constraints represent the intrinsic interactions of small bioactive ligands with protein receptors. Wolber tried to extract ligand pharmacophores from protein cavities based on a define set of six types of chemical structures [51], and develop the algorithms for ligand extraction and interpretation as well as pharmacophore creation for multiple targets.

Pharmacophore screening only considers those compounds who are direct mimics of the ligand from which the pharmacophore has been generated and may neglect the other positive binding modes as well. In fact, the pharmacophore model limits to only one mode of action for small molecules [52]. However, this limitation can be conquered by combining multiple pharmacophore models with different modes of action. This method is called Virtual Parallel Screening and has been successfully applied to the identification of Natural Products' activity [52, 53]. In such work, The PDB-based pharmacophores was firstly used for target fishing for TCM constituents. Results shown 16 constituents of *Ruta graveolens* were screened against a database of pharmacophores and good congruity was found between the potential predictions and their corresponding IC50 values.

Quantitative structure-activity relationships (QSAR) was first established in early 1960s when computational means were used to quantitatively describe

pharmacodynamics and pharmacokinetic effects in biology systems and the chemical structures of compounds [54]. Generally speaking, any mathematical model or statistical method that builds relationship between molecular structures and biological properties may be considered as QSAR. The idea of QSAR is easy while training and application of QSAR is much difficult since similar structures may interact with totally different targets due to the diversity and complexity of biology [55]. Furthermore, the intrinsic noise in data to describe both the chemical space and biological effects brings much trouble in accurate modeling [56]. Despite these difficulties, in case robust biological data is available and few outliers coexist, thousands of QSAR models have been generated and stored in related database in the past 40 years [57, 58].

14.3.2 Target-Based Approach

The target-based approach, predicts ligand-target interactions by the structural information of protein targets as well as ligands. The target-based approach depends highly on the availability of the structural information of targets, either from wet experiments or numerical simulations [59, 60]. On one hand, these methods aim to predict the conformation and orientation of the ligand within the protein cavity. On the other hand, the binding affinity of the ligand and protein is simulated with scoring functions. The main target-based approach is docking, which predicts the preferred orientation of one molecule to another when they bound to each other to form a stable complex [61]. Usually, docking is implemented to search appropriate ligands for known targets with the lowest fitting energy. On the contrary, inverse docking seeks to fish targets from known ligands 'from scratch' and also plays an important role in virtual screening.

Despite more than 20 years' research, docking and scoring ligands with proteins are still challenging processes and the performance is highly dependent on targets [62–64]. Docking cannot be applied to proteins whose 3D structures are not identified [65]. The high-resolution structure of the protein target is preferably obtained from X-ray crystallography and NMR spectroscopy. However, approximately half of the currently approved drugs bind to the membrane proteins, whose structures are extremely difficult to be acquired experimentally. Alternatively, homology modeling is usually adopted to build a putative geometry and docking cavity [66]. Besides, threading and ab initio structure prediction together with molecular dynamics (MD) and Monte Carlo simulations are utilized to predict the target structures. However, the fidelity of homology modeling, threading and ab initio structures is still questioned by many researchers. Other important challenges of docking are the dynamic behavior, the large number of degrees of freedom and the complexity of the potential energy surface. This confines docking to be a low throughput method on a very small scale, which fails to predict interactions on the level of millions of ligands and targets.

To alleviate the situation that docking depends on the nature of targets, multiple active site has been used to compensate the ligand-dependent biases and the Consensus scoring has been also suggested to reduce the false positives in virtual screening [67]. The accuracy of scoring functions still remain the main weakness of docking approach [68]. Also, docking is starting to adopt the conformation information derived from protein-bound ligands as a strategy to overcome the limitations of current scoring functions and can predict the orientation of the ligands into the protein cavity [69]. Besides, molecular dynamics-assisted docking method has been applied in virtual screening against the individual targets in HIV to search for multi-target drug-like agents and KNI-765 was identified to be potential inhibitors [70].

Regardless of the all the limitations, virtual screening based on docking and inverse docking has been successfully utilized to identify and predict novel bioactive compounds in the past 10 years. Using the combinatorial small molecule growth algorithm, Grzybowski applied the docking to the design of picomolar ligands for the human carbonic anhydrase II [71, 72]. Inverse docking was firstly developed to identify multiple proteins to which a small molecule can bind or weakly bind. In some cases, the bioactivity of the TCM compounds is well recognized, while the underlying mode of action is not very clear. In 2001, INVDOCK [73] has been developed to search for the targets for TCM constitutes, and employed a database of protein cavities derived from PDB entries. The results of inverse docking involving multiple-conformer shape-matching alignment showed that 50 % of the computer-predicted potential protein targets were implicated or experimentally validated. The same approach was used to determine potential drug toxicity and side effects in early stages of drug development and results showed that 83 % of the experimentally known toxicity and side effects were predicted [74]. Zahler tried the inverse docking method to find potential kinase targets for three indirubin derivatives and examined 84 unique protein kinases in total [75]. Recently, one indirubin compound was found to possess therapeutic effects against myelogenous leukemia [76].

Docking is usually used as the second step to further validate the ligand-target binding features after the first round of virtual screening by other ligand-based approaches [77–80]. Wei applied the docking together with similarity search and molecular simulation to search for Anti-SAS drugs [81], find the binding mechanism of H5N1 Influenza Virus with ligands [82], detect possible drug leads for Alzheimer's Disease [83, 84] and identify the binding sites for several novel amide derivatives in the nicotinic acetylcholine receptors (AChRs) [85].

14.3.3 Machine Learning

The ligand-based approach and target-based approach predict potential ligand-target bindings by means of chemical similarity and structural information. Machine learning is a high-throughput method of artificial intelligence that enables

computers to learn from data of knowns, including ligand chemistry, structural information and ligand-protein networks and to predict unknowns, such as new drugs, targets and drug-target pairs. This method gains stability and credibility, and has strong ability for classifications among large numbers of ligand-protein pairs that otherwise would be impossible to be connected based on chemical similarity alone.

Machine learning is to exact features from data automatically by computers [86]. Basically, machine learning can be categorized into unsupervised learning and supervised learning. In unsupervised learning, the objective is to extract and conjecture patterns and interactions among a series of input variables and there is no outcome to train the input variables. The common approaches in unsupervised learning are clustering, data compression and outlier detection, such as principal component based methods [87]. In supervised learning, the objective is to predict the value of an outcome variable based on the input variables [88]. The data is commonly divided into training and validation datasets, which are used in turn to finalize a robust model. The variable the supervised model predicts is typically the binding probability of ligands and targets.

Nidhi trained a multiple-category Laplacian-modified naive Bayesian model from 964 target classes in WOMBAT and predict the top three potential targets for compounds in MDDR with or without known targets information [89]. On average, the prediction accuracy with compounds with known targets is 77 %. Bayesian classifier was usually used in early prediction, while the Winnow algorithm was reported more recently [90]. With the same training datasets, the prediction result is slightly different with the Multiple-category laplacian. This indicates that it is necessary to apply different prediction methods and make comparisons even on the same training dataset.

The Gaussian interaction profile kernels, which represented the drug-target interactions, were used in Regularized Least Squares in combined with chemical and genomic space to achieve the prediction with precision-recall curve (AUPR) up to 92.7 [91]. Based on simple physicochemical properties extracted from protein sequences, the potential drug targets were related to the existing ones by several models [92]. The supervised bipartite graph inference is used to represent the drug interaction networks and can be solely be able to predict new interactions, or together with chemical and genomic space [93, 94]. Besides, semi-supervised learning method (Laplacian regularized least square FLapRLS) was also explored to effectively predict the results by integration of genomic and chemical space [95].

The Support Vector Machine (SVM) is a powerful classification tool in which appropriate kernel functions are selected to map the data space into higher dimensional space without increasing the computational difficulties. The performance of SVM is usually stronger than other probability based models. Wale and Karypis [96] made comparisons between a Bayes Classifier together with binary SVM, cascaded SVM, a ranking-based SVM, Ranking Perception and the combination of SVM and Ranking Perception in terms of the ability to predict the targets for small compound, and found that the cascaded SVM has better

performance than the Bayes models and the combination of SVM and Ranking Perceptron has the best performance of all. Zhao et al. developed a SVM model based on the chemical-protein interactions from STITCH [97] using new features from ligand chemical space and interaction networks. Four new D-amino acid oxidase inhibitors were successfully predicted by this model and validate by wet experiments, and one may have a new application in therapy of psychiatric disorders other than being an antineoplastic agent [98].

Random forest, a form of multiple decision trees, recently has been applied to screen TCM database for potential inhibitors against several therapeutically important targets [99]. With the use of binding information from another database, random forest was performed to find multiple hits out of 8,264 compounds in 240 Chinese Herbs on an unbalanced dataset. Among all the predictions, 83 herb-target predictions were proved by literature search. Three Potential inhibitors of the human, aromatase enzyme (CYP19) myricetin, liquiritigenin and gossypetin, were screened by Random Forest as well as molecular docking studies. The virtual screening results were subsequently confirmed experimentally by *in vitro* assay [100].

Linear regression models have also been applied to predict ligand-target pairs. Zhao developed a computational framework, drugCIPHER to infer drug-target interactions based on pharmacology and genomic space [101]. In this framework, three linear regress models were created to relate drug therapeutic similarity, chemical similarity and target similarity on the basis of a protein-protein interaction network. The drugCIPHER achieved the performance with AUC of 0.988 in the training set and 0.935 in the test set and 501 new drug-target interactions were found, implying potential novel applications or side effects.

Although machine learning has strong performance in classification of protein-ligand interactions, its shortcoming is obvious. The process of some machine learning methods is implicit, like a black box, from which we cannot have an intuitive biological or physical relevance between proteins and ligands. SVM maps the classification problem into higher space, and acquires excellent performance with high computational efficiency. The tradeoff is that it can hardly explicitly create relationship between a protein and a ligand. Therefore, even with a very strong prediction tool, we can hardly move forward with innovations in theory of protein-ligand interactions.

14.3.4 Case Studies

14.3.4.1 Inhibiting Biological Transmethylation Reaction

Wei et al. focused on the discovery of potential inhibitors against S-adenosyl-homocysteine hydrolase (SAH), a key reactant in duplication of virus life cycle. A similarity search in Traditional Chinese Medicine Database was performed and 17 hits with high similarity were retrieved. Followed by docking, they proposed the

potential inhibitors by comparing best docked solutions and possible modification for the best inhibitors [79].

14.3.4.2 New D-Amino Acid Oxidase Inhibitor Discovery

Zhao et al. have developed a support vector machine (SVM) model based on the chemical-protein interactions from STITCH using new features from ligand chemical space and interaction networks. The model is used to search for the potential D-amino acid oxidase inhibitors from STITCH database and the predicted results are finally validated by wet experiments. Out of the ten candidates obtained, seven D-amino acid oxidase inhibitors have been verified, in which four are newly found, and one may have a new application in therapy of psychiatric disorders other than being an antineoplastic agent [98].

14.3.4.3 Drug Discovery for AIDs

From docking experiments for more than 9,000 compounds extracted from various Chinese medicines, Gao found that the compound agaritine distinguished itself from all the others in binding to the HIV protease with the most favorable free energy. It has been observed thru an extensive docking study that some of agaritine derivatives had markedly stronger binding interaction with the HIV protease than agaritine, suggesting that these derivatives might be good candidates for developing drugs for AIDS therapy [77].

14.3.4.4 Treating Alzheimer's Disease

To find new drug candidates for treating Alzheimer's disease, Zheng used the similarity search technique and GTS-21 as a template to search the Traditional Chinese Medicines Database. Then those molecules with higher score were selected for docking studies against the alpha7 nicotinic acetylcholine receptor. Though an in-depth structural analysis, it was found Mol 7,235 might be a promising candidate and need further experimental validation before it becomes an effective drug for treating Alzheimer's disease [78].

14.4 Applications of Ligand-Protein Networks in TCM Pharmacology

Network-based pharmacology explores the possibility to develop a systematic and holistic understanding of the mode of actions of multi-drugs by considering their multi-targets in the context of molecular networks. It has also been suggested that

relatively weak patterns of inhibition of many targets may prove more satisfactory than the highly potent single target inhibitors routinely developed in the course of a drug discovery program [102]. In drug discovery, the use of networks incorporating multiple components and the corresponding multiple targets, is one of the driving force to propel the current development in TCM pharmacology. Several successful examples have been accumulated both in experiments and *in silico* analysis, as shown in Table 14.2.

14.4.1 Experimental Study

Many bioactive compounds in TCM herbs may have synergetic effort with many non-TCM drugs in markets. Tannin, a component derived from a TCM, can be combined with HIV triple cocktail therapy to yield everlasting efforts in preventing HIV virus propagation. The underlying mechanism is that Tannin suppresses the activity of HIV-1 reverse transcriptase, protease and intergrase and cut off virus fusion and virus entry into the host cells [103]. Recently, Li proposed a new idea to induce immunetolerance in T cells by using matrine, a chemical derived from the root of Sophora flavescens AIT, targeting both the PKCy pathway and the NFAT pathway in cocktail preparations for treating AIDS [104].

Lam et al. recently showed in murine colon 38 allograft model that a formula containing 4 herbs (PHY906) has synergetic effect on reducing side effects and enhancing efficacy induced by CPT-11, a power anticancer agent with strong toxicity. The reason is that PHY906 can repair the intestinal epithelium by facilitating the intestinal progenitor or stem cells and several Wnt signaling components and suppress a batch of inflammatory responses like factor kB, cyclooxygenase-2, and inducible nitric oxide synthase [105].

Multi-component and multi-target interactions are the main mode of action for TCM formula, which exerts synergetic effects as a whole preparation rather than the primary active compound in TCM alone. Xie et al. demonstrated that other components in "Qingfu Guanjieshu" (QFGJS) could effectively influence the pharmacokinetic behavior and metabolic profile of paeonol in rats, indicating the synergy of herbal components. This synergy may be the result of enhanced adsorption of paeonol in the gastrointestinal tract induced by P-glycoprotein-mediate efflux change [106]. Another similar study, showed that paeoniflorin from the root of *Paeonia lactiflora* were markedly enhanced when co-administrated with sinomenine, the stem of *Sinomenium acutum*. Sinomenine promotes intestinal transportation via inhibition of P-glycoprotei, and affect the hydrolysis of paeoniflorin via interaction with b-glycosidase [107].

Huang-Lian-Jie-Du-Tang (HLJDT) is a TCM formula with anti-inflammatory efficacy, but the action mechanism is still not very clear. Zeng et al. investigated the effects of its component herbs and pure components on eicosanoid generation and found out the active components and their precise targets on arachidonic acid (AA) cascade. Results showed that Rhizoma coptidis and Radix scutellariae were

Table 14.2 Summary of multi-target drugs/preparations with TCM pharmacology based on ligand-protein networks

Disease	Methods and experiments	Formula, herbs and components	TCM pharmacology	References
AIDS	Experiments	Tannin	Tannin suppresses the activity of HIV-1 reverse transcriptase, protease and intergrase and cut off virus fusion and virus entry into the host cells	[103]
AIDS	Experiments	Matrine from the root of *Sophora flavescens*	Matrine is effective in inducing T cell anergy by targeting both the MAPKs pathway and the NFAT pathway	[104]
Anti-tumor	Experiments	PHY906: *Glycyrrhiza uralensis* Fisch (G), *Paeonia lactiflora* Pall (P), *Scutellaria baicalensis* Georgi (S), and *Ziziphus jujuba* Mill (Z).	PHY906 reduces CPT-11-induced gastrointestinal toxicity in the treatment of colon or rectal cancer by several mechanisms. It both repairs the intestinal epithelium by facilitating the generation of intestinal progenitor or stem cells and several Wnt signaling components and suppresses inflammatory responses like factor kB, cyclooxygenase-2, and inducible nitric oxide. synthase	[105]
Anti-inflammatory and analgesic effects	Experiments	Qingfu Guanjieshu (QFGJS): Paeonol and other components	The pharmacokinetic behavior and metabolites of paeonol are greatly promoted by other components in	[106]

(continued)

Table 14.2 (continued)

Disease	Methods and experiments	Formula, herbs and components	TCM pharmacology	References
			QFGJS. This may be the result of enhanced adsorption of paeonol in the gastrointestinal tract by P-glycoprotein-mediate efflux change	
Inflammatory and arthritic diseases	Experiments	Paeoniflorin from the root of *Paeonia lactiflora* and sinomenine from the stem of Sinomenium acutum	Paeoniflorin is markedly enhanced when co-administrated with Sinomenine, which promotes of intestinal transportation via the inhibition of P-glycoprotein, and affects the hydrolysis of Paeoniflorin via interaction with b-glycosidase	[107]
Anti-inflammatory	Experiments	Huang-Lian-Jie-Du-Tang (HLJDT): Rhizoma coptidis and Radix scutellariae	Baicalein derived from Radix scutellariae showed significant inhibitory effect on 5-LO and 15-LO while coptisine from Rhizoma coptidis showed medium inhibitory effects on LTA(4)H	[108]
Acute promyelocytic leukemia (APL)	Experiments	Realgar-Indigo naturalis: tetraarsenic tetrasulfide (A), indirubin (I), and tanshinone IIA (T)	ATI leads to ubiquitination/ degradation of promyelocytic leukemia (PML)-retinoic acid receptor oncoprotein, reprogramming of myeloid differentiation	[109]

(continued)

Table 14.2 (continued)

Disease	Methods and experiments	Formula, herbs and components	TCM pharmacology	References
			regulators, and G1/G0 arrest in APL cells by mediating multiple targets. A acts as the principal component of the formula, whereas T and I serve as adjuvant ingredients	
Chronic myeloid leukemia (CML)	Experiments	Imatinib (IM) and arsenic sulfide [As(4)S(4) (AS)]	AS targets BCR/ABL through the ubiquitination of key lysine residues, leading to its proteasomal degradation, whereas IM inhibits the PI3 K/AKT/mTOR pathway	[110]
Inflammation	Pharmacophore-assisted docking	Twelve examples of compounds from CHCD	The screened compounds target cyclo-oxygenases 1 & 2 (COX), p38 MAP kinase (p38), c-Jun terminal-NH(2) kinase (JNK) and type 4 cAMP-specific phosphodiesterase (PDE4)	[111]
Type II diabetes mellitus (T2DM)	Molecular docking (LigandFit), clustering and drug-target network analysis	676 compounds in eleven herbs from Tangminling Pills	Multiple active components in Tangminling Pills interact with multiple targets. The 37 targets were classified into 3 clusters, and proteins in each cluster were highly relevant to each other. 10 known compounds were selected according to their network attribute ranking in drug-target and drug-drug network	[112]

(continued)

Table 14.2 (continued)

Disease	Methods and experiments	Formula, herbs and components	TCM pharmacology	References
Cardiovascular disease	Similarity search and alignment, docking (LigandFit)	Xuefu Zhuyu decoction (XFZYD): 501 compounds, 489 drug/drug like compounds	Active components in XFZYD mainly target rennin, ACE and ACE2 in Renin-Angiotensin. System (RAS), which modulates the cardiovascular physiological function	[113]
9 types of cancer, 5 diseases with dysfunction, and 2 cardiovascular disorders	Distance-based Mutual Information Model (DMIM)	Liu-wei-di-huang formula (LWDH) Shan-zhu-yu (Fructus Corni), Ze-xie (Rhizoma Alismatis), Dan-pi (Cortex Moutan), Di-huang (Radix Rehmaniae), Fu-ling (Poria Cocos) and Shan-yao (Rhizoma Dioscoreae)	The interactions between TCM drugs and disease genes in cancer pathways and neuro-endocrine-immune pathways were inferred to contribute to the action of LWDH formula	[114]
Cardiovascular diseases	Quantitative composition-activity relationship model (QCAR) (SVM and linear regression)	*Shenmai*, Qi-Xue-Bing-Zhi-Fang (QXBZF)	The proportion of active components of *Shenmai* and QXBZF were optimized based on clinical outcome (collateral and infarct rate of heart) using QCAR. The interactions of multiple weak bindings among different compounds and targets may contribute to the synergetic effect of multi-component drugs	[115, 116]
Anticoagulant	Network-based computational scheme utilizing multi-target	Six argatroban intermediates and a series of components from	A ligand can have impact on multiple targets based on the docking scores, and	[117]

(continued)

Table 14.2 (continued)

Disease	Methods and experiments	Formula, herbs and components	TCM pharmacology	References
	docking score (Ligandfit and AutoDock)	24 TCMs widely used for cardiac system diseases	those with highest target network efficiency are regarded as potential anticoagulant agents. Factor Xa and thrombin are two critical targets for anticoagulant compounds and the catalytic reactions they mediate were recognized as the most fragile biological matters in the human clotting cascade system	
Alzheimer disease	Systematical target network analysis framework	Ginkgo biloba, Huperzia serrata, Melissa officinalis and Salvia officinalis	AD symptoms-associated pathways, inflammation-associated pathways, cancer-associated pathways, diabetes mellitus associated pathways, Ca2þ-associated pathways and cell proliferation pathways are densely targeted by herbal ingredients	[118]
Depression	Literature search and network analysis	Hyperforin (HP), hypericin (HY), pseudohypericin (PH), amentoflavone (AF) and several flavonoids (FL) from St. John's Wort (SJW)	Active components in SJW mainly intervene with neuroactive ligand-receptor interaction, the calcium signaling pathway, and the gap junction related pathway	[119]
			Pertinent targets include NMDA-receptor, CRF1	

(continued)

Table 14.2 (continued)

Disease	Methods and experiments	Formula, herbs and components	TCM pharmacology	References
			receptor, 5-hydroxytryptamine receptor 1D, dopamine receptor D1, etc	
Rheumatoid arthritis (RA)	Integrative Platform of TCM Network Pharmacology including drugCIPHER	Qing-Luo-Yin (QLY), including four herbs: Ku-Shen (*Sophora flavescens*), Qing-Feng-Teng (*Sinomenium acutum*), Huang-Bai (*Phellodendron chinensis*) and Bi-Xie (*Dioscorea collettii*), which contain several groups of ingredients such as Saponins and Alkaloids	The target network of QLY is involved in RA-related key processes including angiogenesis, inflammatory response, and immune response. The four herbs in QLY work in concert to promote efficiency and reduce toxicity. Specifically, the synergetic effect of Ku-Shen (*jun* herb) and Qing-Feng-Teng (*chen* herb) may come from the feedback loop and compensatory mechanisms	[120]

the key herbs responsible for the suppressive effect of HLJDT on eicosanoid generation. Further experiments on the pure components of HLJDT revealed that baicalein derived from *Radix scutellariae* has significant inhibitory effect on 5-LO and 15-LO while coptisine from *Rhizoma coptidis* show medium inhibitory effects on LTA(4)H. Besides, baicalein and coptisine were proved to have synergetic inhibition on LTB(4) by the rat peritoneal macrophages [108].

A TCM formula, Realgar-Indigo naturalis formula (RIF), was applied to treat Acute promyelocytic leukemia (APL) and showed a high complete remission (CR rate) [109]. In RIF, multiple agents within one formula were found to work synergistically. A small-scale combinational study using Chou and Talalay combination index method was performed and three main active components of RIF and six core proteins they targets in mediating the auti-tumor effect were identified. The main active ingredients of RIF are tetraarsenic tetrasulfide (A), indirubin (I), and tanshinone IIA (T), from *Realgar*, *Indigo naturalis*, and *Salvia miltiorrhiza*, respectively. A acts as the principal component of the formula, whereas T and I serve as adjuvant ingredients. ATI leads to ubiquitination/degradation of

promyelocytic leukemia (PML)-retinoic acid receptor oncoprotein, reprogramming of myeloid differentiation regulators, and G1/G0 arrest in APL cells by mediating multiple targets. Using multi-omics technologies, Zhang later proved that combined use of imatinib and arsenic sulfide from toxic herbal remedy exerted better therapeutic effects in a BCR/ABL-positive mouse model of chronic myeloid leukemia (CML) than either drug as a single agent. AS targets BCR/ABL through the ubiquitination of key lysine residues, leading to its proteasomal degradation, whereas IM inhibits the PI3 K/AKT/mTOR pathway [110].

14.4.2 Computational Framework

To target the complex, multi-factorial diseases more effectively, the network biology incorporating ligand-protein networks has been applied in multi-target drug development as well as modernization of traditional Chinese medicine in the systematic and holistic way. Zhao reviewed the available disease-associated networks, drug-associated networks that can be used to assist the drug discovery and elaborate the network-based TCM pharmacology [119]. Klipp discussed the possibility to use networks to aid the drug discovery process and focused on networks and pathways in which the components are related by physical interactions or biochemical process [121]. Leung investigated the possibility of network-based intervention for curing system diseases by means of network-based computational models and using medicinal herbs to develop into new wave of network-based multi-target drugs. It was concluded that further integration across various 'omics' platform and computational tools would accelerate the drug discovery based on network [122].

Barlow et al. screened among Chinese herbs for compounds that may be active against 4 targets in inflammation, by means of pharmacophore-assisted docking. The results showed that the twelve examples of compounds from CHCD inhibit multiple targets including cyclo-oxygenases 1 & 2 (COX), p38 MAP kinase (p38), c-Jun terminal-NH(2) kinase (JNK) and type 4 cAMP-specific phosphodiesterase (PDE4).The distribution of herbs containing the predicted active inhibitors was studied in regards to 192 Chinese Formulae and it was found that these herbs were in the formulae that were traditionally used to treat fever, headache and so on [111].

Many Traditional Chinese Medicines (TCMs) are effective to relieve complicated diseases such as type II diabetes mellitus (T2DM). Gu et al. employed the molecular docking and network analysis to elucidate the action mechanism of a medical composition-Tangminling Pills which had clinical efficacy for T2DM. It was found that multiple active components in Tangminling Pills interact with multiple targets in the biological network of T2DM. The 37 targets were classified into 3 clusters, and proteins in each cluster were highly relevant to each other. 10 known compounds were selected according to their network attribute ranking in drug-target and drug-drug network [112].

XFZYD, a recipe derives from Wang Q. R. in Qing dynasty, was widely used in cardiac system disease. From similarity search and alignment, the chemical space of compounds in XFZYD was found to share a lot of similarities with that of drug/drug-like ligands set collected from cardiovascular pharmacology while the chemical pattern in XFZYD are more diverse than drug/drug-like ligands for cardiovascular pharmacology. Docking protocol between compounds in XFZYD and targets related to cardiac system disease using LigandFit show that many molecules have good binding affinity with the targeting enzymes and most have interactions with more than one single target. The active components in XFZYD mainly target rennin, ACE and ACE2 in Renin-Angiotensin System (RAS), which modulates the cardiovascular physiological function. It was proved that promiscuous drugs in TCM might be more effective for treating cardio system diseases, which tends to result from multi-target abnormalities, but not from a single defect [113].

A lot of integrative computational tools and models have been developed and widely used to optimize the combination regimen of multi-components drugs and elucidating the interactive mechanism among ligand-target networks.

Li et al. built a method called Distance-based Mutual Information Model (DMIM) to identify useful relationships among herbs in numerous herbal formulae. DMIM combines mutual information entropy and distance between herbs to score herb interactions and construct herb network. Novel anti-angiogenic herbs, Vitexicarpin and Timosaponin A-III were discovered to have synergistic effects. Based on herb network constructed by DMIM from 3,865 collateral-related herbs, the interactions between TCM drugs and disease genes in cancer pathways and neuro-endocrine-immune pathways were inferred to contribute to the action of Liu-wei-di-huang formula, one of the most well-known TCM formula as potential treatment for a variety of diseases including cancer, dysfunction of the neuro-endocrine-immune-metabolism system and cardiovascular [114].

Wang et al. adopted a new method based upon lattice experimental design and multivariate regression to model the quantitative composition-activity relationship (QCAR) of *Shenmai*, a Chinese medicinal formula. This new strategy for multi-component drug design was then successfully applied in searching optimal combination of three key components (PD, PT and OP) of *Shenmai*. Experimental outcome of infarct rate of heart in mice with different dosage combination of the three components were finally measured and the fitted relationship equation showed that the optimal values of PD, PT and OP were 21.6, 39.2 and 39.2 %, respectively [115]. The proportion of two active components of Qi-Xue-Bing-Zhi-Fang, PF and FP, was also optimized in similar way using several fitting technique like linear regression, artificial neural network and support vector regression [116]. Although the underlying mechanism of drug synergy for the two formulae was still not very clear, the interactions of multiple weak bindings among different compounds and targets might be the contributory factors.

A network-based multi-target computational scheme for the whole efficacy of a compound in a complex disease was develop for screening the anticoagulant activities of a serial of argatroban intermediates and eight natural products respectively. Aimed at the phenotypic data of drugs, this scheme predicted

bioactive compounds by integrating biological network efficiency analysis with multi-target docking score, which evolves from the traditional virtual screening method that usually predicted binding affinity between single drug molecule and target. A ligand can have impact on multiple targets based on the docking scores, and those with highest target network efficiency are regarded as potential anti-coagulant agents. Factor Xa and thrombin are two critical targets for anticoagulant compounds and the catalytic reactions they mediate were recognized as the most fragile biological matters in the human clotting cascade system [117].

Sun et al. [118] presented a systematic target network analysis framework to explore the mode of action of anti-Alzheimer's disease (AD) herb ingredients based on applicable bioinformatics resources and methodologies on clinical anti-AD herbs and their corresponding target proteins. The results showed that just as many FDA-approved anti-AD drugs do the compounds of these herbs binds to targets in AD symptoms-associated pathway. Besides, they also interact closely with many successful therapeutic targets related to diseases such as inflammation, cancer and diameters. This suggests that the possible cross-talks between these complicated diseases are prevalent in TCM anti-AD herbs [123]. Moreover, pathways of Ca(2+) equilibrium maintaining, upstream of cell proliferation and inflammation were found to be were intensively hit by the anti-AD herbal ingredients.

Based on the available experimental results, Zhao analyzed the molecular mechanism with the aid of pathways and networks and theoretically proved the multi-target effect of St. John's Wort [119]. A comprehensive literature search was conducted and the neurotransmitter receptors, transporter proteins, and ion channels on which the SJW active compounds show effects were collected. Three main pathways that SJW intervenes were found by mapping these proteins onto KEGG pathways. Active components in SJW mainly intervene with neuroactive ligand-receptor interaction, the calcium signaling pathway, and the gap junction related pathway, pertinent to targets including NMDA-receptor, CRF1 receptor, 5-hydroxytryptamine receptor 1D, dopamine receptor D1. The networks show that the effect of SJW is similar to that of combinations of different types of antidepressants. However, the inhibitory effects of the SJW on each of the pathway are lower than other individual agents. Accordingly, the significant antidepressant efficacy and lower side effects are due to the synergetic effect of low-dose multi-target actions.

Zhang et al. established an integrative platform of TCM network pharmacology to discover herbal formulae on basis of systematic network. This platform incorporates a set of state-of-the-art network-based methods to explore the action mechanism, identify activate ingredients, and create new synergetic combinations of components. The Qing-Luo-Yin (QLY), an antirheumatoid arthritis (RA) formula was studied comprehensively using the new platform. It is found the target network of QLY is involved on RA-related key processes including angiogenesis, inflammatory response, and immune response. The four herbs in QLY work in concert to promote efficiency and reduce toxicity, as the *jun*, *chen*, *zuo*, *shi* in Chinese, respectively. Specifically, the synergetic effect of Ku-Shen (*jun* herb) and Qing-Feng-Teng (*chen* herb) may come from the feedback loop and compensatory mechanisms [120].

14.5 Discussion and Conclusion

In recent years, the bottleneck in western medicine has brought unprecedented opportunities in TCM research and development. For decades, the fundamental research has achieved great success, and laid the foundation of modern western medicine and the philosophical idea of "reductionism" was considered to own the credit.

The counterparty of "reductionism" in Chinese medicine is the philosophical idea of holism, which has thousands years' history of practice in China as well as other Asian countries. Using this methodology, the effectiveness of TCM can only be verified from a large number of clinical trials given the unclear composition and unknown relationship among various components. This implicit effect without clear clarification at the molecular level has been hindering the modernization of TCM. How to learn from the accumulative knowledge of western medicine, in order to identify the effective compositions and explore the molecular mechanism of the efficacy is an urgent problem that needs to be solved in TCM.

The hypothesis of "multi-drug, multi-target and multi-gene" in fact bridges the gap between TCM and western medicine and is also a manifestation of unity of opposites on "reductionism" and "holism". TCM uses the holistic method to investigate the effects of multi-component formula across the whole organism, such as the use of a variety of "ZHENG" in TCM theory [124]. However, the only option we have to uncover the underlying mechanism of TCM at the molecular level is to make use of the theory of reductionism. Of course, for complex systems, the reduction method can only reach to a certain depth since it becomes more troublesome as we get deeper. Therefore, some researchers tend to reduce the mechanism of TCM to the level of "multi-drug, multi-target, multi-gene" at present, and for further reduction to the level of "single-drug, single-target, single gene", the problem of emergentism [125] in philosophy needs to be addressed properly. The theory of emergentism believes that some unique features or "ultimate features" of a system can never be reduced to properties at lower levels, nor the former can be predicted or explained by the latter, as shown in Fig. 14.1.

So far, ligand-protein network or "multi-drug, multi-target, multi-gene" is one of the few basic modules that can clearly reveal the pharmacology of TCM and is expected to be the future direction of the modernization of TCM. But just relying on experimental scientists to build ligand-protein interactions non-exhaustively will slow down both the modernization of TCM and the development of its industry. Therefore, the use of cross-platform database (TCM compounds and recipe database, see Sect. 14.2 in this paper) and the improvement on modeling technique (computational method of ligand-protein interactions, see Sect. 14.3 in this paper), will afford the basis of *in silico* research for future modernization and development of TCM. It can be foreseen that, one future direction is to use these TCM databases and predictive models to reveal the pharmacological effect of TCM, through the establishment of ligand-protein networks or, "multi-drug, multi-target, multi-gene" relationships. Nevertheless, the pharmacological mechanism of

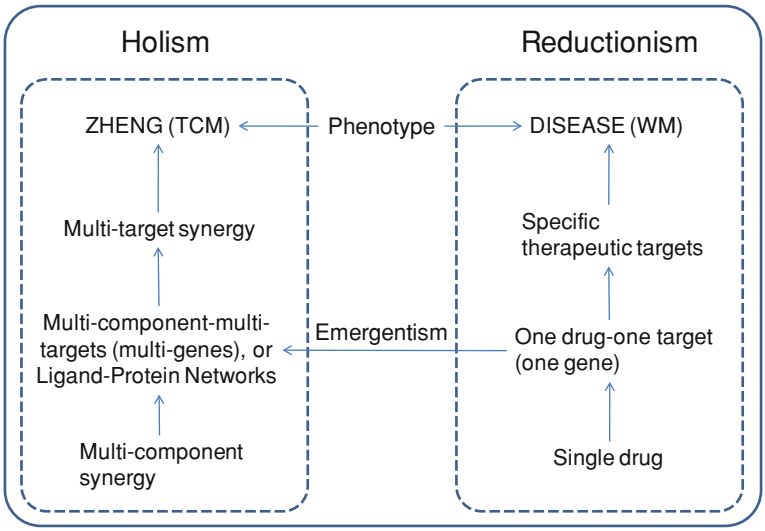


Fig. 14.1 Unity of opposites on holism in traditional Chinese medicine and reductionism in Western medicine. Emergentism constructs the framework of the understanding of holism in TCM via accumulative practice of reductionism in WM

TCM can be very complex and may not be well explained only with the known ligand-protein network. After all, this is a process of reeling silk from cocoons and also one of the best choices we have right now.

The increasing availability of ligand-protein networks is a unique chance to boost success in the modernization of TCM based on the accumulative knowledge of TCM formulae and practices based on the assumption that TCM exerts the pharmacological efficacy in multi-drug-multi-target way. Although preliminary research has been initiated in this area, there is still a long way to go to further leverage these networks and modeling techniques. Virtual screening and informatics in the drug discovery area have already been proven to be quite useful either to predict potential new drug and target candidates for experimentalists or explore the functional mechanism at the molecular level. A large number of drug-target interactions have thus been gained and the resulted drug-target networks will also be quite beneficial to investigate the underlying mechanism of multi-component drugs, such as the TCM. With further applications of these methods in TCM area, we are expecting to reveal the mode of action underlying polypharmacology of TCM. This grants us the possibility to discover novel effective drug leads, understand the synergistic mechanism of drug combinations, and more importantly, develop drug portfolios against epidemic, chronic disease, cancer and other complex disease that are almost incurable by western medicine.

References

1. Kaufmann SH (2008) Paul Ehrlich: founder of chemotherapy. Nat Rev Drug Discov 7(5):373
2. Paolini GV et al (2006) Global mapping of pharmacological space. Nat Biotechnol 24(7):805–815
3. Keiser MJ et al (2007) Relating protein pharmacology by ligand chemistry. Nat Biotechnol 25(2):197–206
4. Hopkins AL (2009) Drug discovery: predicting promiscuity. Nature 462(7270):167–168
5. Vedani A, Dobler M, Lill MA (2006) The challenge of predicting drug toxicity in silico. Basic Clin Pharmacol Toxicol 99(3):195–208
6. Klabunde T, Evers A (2005) GPCR antitarget modeling: pharmacophore models for biogenic amine binding GPCRs to avoid GPCR-mediated side effects. ChemBioChem 6(5):876–889
7. Ekins S, Mestres J, Testa B (2007) In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. Br J Pharmacol 152(1):9–20
8. Ekins S, Mestres J, Testa B (2007) In silico pharmacology for drug discovery: applications to targets and beyond. Br J Pharmacol 152(1):21–37
9. Lukman S, He Y, Hui SC (2007) Computational methods for traditional Chinese medicine: a survey. Comput Methods Programs Biomed 88(3):283–294
10. Ehrman TM, Barlow DJ, Hylands PJ (2010) Phytochemical informatics and virtual screening of herbs used in Chinese medicine. Curr Pharm Des 16(15):1785–1798
11. Feng Y et al (2006) Knowledge discovery in traditional Chinese medicine: state of the art and perspectives. Artif Intell Med 38(3):219–236
12. Yan X, Zhou J, Xu Z (1999) Concept design of computer-aided study on traditional Chinese drugs. J Chem Inf Comput Sci 39(1):86–89
13. Haggarty SJ et al (2003) Multidimensional chemical genetic analysis of diversity-oriented synthesis-derived deacetylase inhibitors using cell-based assays. Chem Biol 10(5):383–396
14. Kuruvilla FG et al (2002) Dissecting glucose signalling with diversity-oriented synthesis and small-molecule microarrays. Nature 416(6881):653–657
15. Ehrman TM, Barlow DJ, Hylands PJ (2007) Phytochemical databases of Chinese herbal constituents and bioactive plant compounds with known target specificities. J Chem Inf Model 47(2):254–263
16. Qiao X et al (2002) A 3D structure database of components from Chinese traditional medicinal herbs. J Chem Inf Comput Sci 42(3):481–489
17. Cheng A et al (2000) GB/SA water model for the Merck molecular force field (MMFF). J Mol Graph Model 18(3):273–282
18. Chen CY (2011) TCM Database@Taiwan: the world's largest traditional Chinese medicine database for drug screening in silico. PLoS ONE 6(1):e15939
19. Chen X et al (2006) Database of traditional Chinese medicine and its application to studies of mechanism and to prescription validation. Br J Pharmacol 149(8):1092–1103
20. Qiao X et al (2002) Research and development of traditional Chinese medicine drugs. Acta Phys Chim Sin 18:394–398
21. Fang X et al (2005) CHMIS-C: a comprehensive herbal medicine information system for cancer. J Med Chem 48(5):1481–1488
22. Shen J et al (2003) Virtual screening on natural products for discovering active compounds and target information. Curr Med Chem 10(21):2327–2342
23. Lei J, Zhou J (2002) A marine natural product database. J Chem Inf Comput Sci 42(3):742–748
24. Bender A, Glen RC (2004) Molecular similarity: a key technique in molecular informatics. Org Biomol Chem 2(22):3204–3218
25. Sastry M et al (2010) Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. J Chem Inf Model 50(5):771–784

26. Keiser MJ, Irwin JJ, Shoichet BK (2010) The chemical basis of pharmacology. Biochemistry 49(48):10267–10276
27. Shoichet BK et al (2008) Quantifying the relationships among drug classes. J Chem Inf Model 48(4):755–765
28. Koutsoukas A et al (2011) From in silico target prediction to multi-target drug design: current databases, methods and applications. J Proteomics 74(12):2554–2574
29. Rush TS 3rd et al (2005) A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. J Med Chem 48(5):1489–1495
30. Lemmen C, Lengauer T (2000) Computational methods for the structural alignment of molecules. J Comput Aided Mol Des 14(3):215–232
31. Mestres J, Veeneman GH (2003) Identification of "latent hits" in compound screening collections. J Med Chem 46(16):3441–3444
32. Jain AN (2004) Ligand-based structural hypotheses for virtual screening. J Med Chem 47(4):947–961
33. Gregori-Puigjane E, Mestres J (2006) SHED: Shannon entropy descriptors from topological feature distributions. J Chem Inf Model 46(4):1615–1622
34. Delgado-Soler L et al (2009) RED: a set of molecular descriptors based on Renyi entropy. J Chem Inf Model 49(11):2457–2468
35. Hu Y, Bajorath J (2011) Combining horizontal and vertical substructure relationships in scaffold hierarchies for activity prediction. J Chem Inf Model 51(2):248–257
36. Weill N, Rognan D (2009) Development and validation of a novel protein-ligand fingerprint to mine chemogenomic space: application to G protein-coupled receptors and their ligands. J Chem Inf Model 49(4):1049–1062
37. Willett P (1987) Similarity and clustering in chemical information systems. Chemometrics Series, Letchworth, Hertfordshire, England. Research Studies Press, Wiley, New York, 254 p (xii)
38. Willett P (2005) Searching techniques for databases of two- and three-dimensional chemical structures. J Med Chem 48(13):4183–4199
39. Willett P (2006) Similarity-based virtual screening using 2D fingerprints. Drug Discov Today 11(23–24):1046–1053
40. Brown RD, Martin YC (1996) Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. J Chem Inf Comput Sci 36(3):572–584
41. van der Horst E et al (2010) A novel chemogenomics analysis of G protein-coupled receptors (GPCRs) and their ligands: a potential strategy for receptor de-orphanization. BMC Bioinform 11:316
42. Keiser MJ et al (2009) Predicting new molecular targets for known drugs. Nature 462(7270):175–181
43. DeGraw AJ et al (2010) Prediction and evaluation of protein farnesyltransferase inhibition by commercial drugs. J Med Chem 53(6):2464–2471
44. Mason JS, Good AC, Martin EJ (2001) 3-D pharmacophores in drug discovery. Curr Pharm Des 7(7):567–597
45. Maclean D et al (2000) Glossary of terms used in combinatorial chemistry. J Comb Chem 2(6):562–578
46. Langer T, Hoffmann RD (2006) Pharmacophores and pharmacophore searches. Wiley-VCH, Weinheim (John Wiley, Chichester: distributor)
47. Nicklaus MC et al (1997) HIV-1 integrase pharmacophore: discovery of inhibitors through three-dimensional database searching. J Med Chem 40(6):920–929
48. Koide Y et al (2002) Development of novel EDG3 antagonists using a 3D database search and their structure-activity relationships. J Med Chem 45(21):4629–4638
49. Debnath AK (2003) Generation of predictive pharmacophore models for CCR5 antagonists: study with piperidine- and piperazine-based compounds as a new class of HIV-1 entry inhibitors. J Med Chem 46(21):4501–4515

50. Kurogi Y et al (2001) Discovery of novel mesangial cell proliferation inhibitors using a three-dimensional database searching method. J Med Chem 44(14):2304–2307
51. Wolber G, Langer T (2005) LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. J Chem Inf Model 45(1):160–169
52. Rollinger JM (2009) Accessing target information by virtual parallel screening—The impact on natural product research. Phytochem Lett 2(2):53–58
53. Rollinger JM et al (2009) In silico target fishing for rationalized ligand discovery exemplified on constituents of *Ruta graveolens*. Planta Med 75(3):195–204
54. Sharples D (1976) Factors affecting the binding of tricyclic tranquillizers and antidepressants to human serum albumin. J Pharm Pharmacol 28(2):100–105
55. Verma RP, Hansch C (2005) An approach toward the problem of outliers in QSAR. Bioorg Med Chem 13(15):4597–4621
56. Polanski J et al (2006) Modeling robust QSAR. J Chem Inf Model 46(6):2310–2318
57. Kurup A (2003) C-QSAR: a database of 18,000 QSARs and associated biological and physical data. J Comput Aided Mol Des 17(2–4):187–196
58. Hansch C et al (2002) Chem-bioinformatics: comparative QSAR at the interface between chemistry and biology. Chem Rev Columbus 102(3):783–812
59. Shoichet BK (2004) Virtual screening of chemical libraries. Nature 432(7019):862–865
60. Klebe G (2006) Virtual ligand screening: strategies, perspectives and limitations. Drug Discov Today 11(13–14):580–594
61. Lengauer T, Rarey M (1996) Computational methods for biomolecular docking. Curr Opin Struct Biol 6(3):402–406
62. Kitchen DB et al (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov 3(11):935–949
63. Leach AR, Shoichet BK, Peishoff CE (2006) Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. J Med Chem 49(20):5851–5855
64. Ghoshal N, Manoharan P, Vijayan RSK (2010) Rationalizing fragment based drug discovery for BACE1: insights from FB-QSAR, FB-QSSR, multi objective (MO-QSPR) and MIF studies. J Comput Aided Mol Des 24(10):843–864
65. Cheng AC et al (2007) Structure-based maximal affinity model predicts small-molecule druggability. Nat Biotechnol 25(1):71–75
66. Evers A, Gohlke H, Klebe G (2003) Ligand-supported homology modelling of protein binding-sites using knowledge-based potentials. J Mol Biol 334(2):327–345
67. Vigers GP, Rizzi JP (2004) Multiple active site corrections for docking and virtual screening. J Med Chem 47(1):80–89
68. Gao Z et al (2008) PDTD: a web-accessible protein database for drug target identification. BMC Bioinform 9:104
69. Fradera X, Mestres J (2004) Guided docking approaches to structure-based design and screening. Curr Top Med Chem 4(7):687–700
70. Clemente JC et al (2006) Structure of the aspartic protease plasmepsin 4 from the malarial parasite Plasmodium malariae bound to an allophenylnorstatine-based inhibitor. Acta Crystallogr D Biol Crystallogr 62(Pt 3):246–252
71. Grzybowski BA et al (2002) Combinatorial computational method gives new picomolar ligands for a known enzyme. Proc Natl Acad Sci USA 99(3):1270–1273
72. Bissantz C (2003) Conformational changes of G protein-coupled receptors during their activation by agonist binding. J Recept Signal Transduct Res 23(2–3):123–153
73. Chen YZ, Zhi DG (2001) Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule. Proteins 43(2):217–226
74. Chen YZ, Ung CY (2001) Prediction of potential toxicity and side effect protein targets of a small molecule by a ligand-protein inverse docking approach. J Mol Graph Model 20(3):199–218
75. Zahler S et al (2007) Inverse in silico screening for identification of kinase inhibitor targets. Chem Biol 14(11):1207–1214

76. MacDonald ML et al (2006) Identifying off-target effects and hidden phenotypes of drugs in human cells. Nat Chem Biol 2(6):329–337
77. Gao WN et al (2007) Agaritine and its derivatives are potential inhibitors against HIV proteases. Med Chem 3(3):221–226
78. Zheng H et al (2007) Screening for new agonists against Alzheimer's disease. Med Chem 3(5):488–493
79. Wei H et al (2007) Molecular insights of SAH enzyme catalysis and implication for inhibitor design. J Theor Biol 244(4):692–702
80. Wang SQ et al (2007) Virtual screening for finding natural inhibitor against cathepsin-L for SARS therapy. Amino Acids 33(1):129–135
81. Wei DQ et al (2006) Anti-SARS drug screening by molecular docking. Amino Acids 31(1):73–80
82. Gong K et al (2009) Binding mechanism of H5N1 influenza virus neuraminidase with ligands and its implication for drug design. Med Chem 5(3):242–249
83. Gu RX et al (2009) Possible drug candidates for Alzheimer's disease deduced from studying their binding interactions with alpha7 nicotinic acetylcholine receptor. Med Chem 5(3):250–262
84. Chen SG et al (2013) Virtual screening for alpha7 nicotinic acetylcholine receptor for treatment of Alzheimer's disease. J Mol Graph Model 39:98–107
85. Arias HR et al (2011) Novel positive allosteric modulators of the human alpha7 nicotinic acetylcholine receptor. Biochemistry 50(23):5263–5278
86. Mitchell TM (1997) Machine learning. McGraw-Hill series in computer science. McGraw-Hill, New York, 414 p (xvii)
87. Strombergsson H, Kleywegt GJ (2009) A chemogenomics view on protein-ligand spaces. BMC Bioinform 10(Suppl 6):S13
88. Jensen LJ, Bateman A (2011) The rise and fall of supervised machine learning techniques. Bioinformatics 27(24):3331–3332
89. Nidhi et al (2006) Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. J Chem Inf Model 46(3):1124–1133
90. Nigsch F et al (2009) Computational toxicology: an overview of the sources of data and of modelling methods. Expert Opin Drug Metab Toxicol 5(1):1–14
91. van Laarhoven T, Nabuurs SB, Marchiori E (2011) Gaussian interaction profile kernels for predicting drug-target interaction. Bioinformatics 27(21):3036–3043
92. Li Q, Lai L (2007) Prediction of potential drug targets based on simple sequence properties. BMC Bioinform 8:353
93. Yamanishi Y et al (2008) Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics 24(13):i232–i240
94. Bleakley K, Yamanishi Y (2009) Supervised prediction of drug-target interactions using bipartite local models. Bioinformatics 25(18):2397–2403
95. Yu W et al (2011) Predicting drug-target interactions based on an improved semi-supervised learning approach. Drug Dev Res 72(2):219–224
96. Wale N, Karypis G (2009) Target fishing for chemical compounds using target-ligand activity data and ranking based methods. J Chem Inf Model 49(10):2190–2201
97. Kuhn M et al (2008) STITCH: interaction networks of chemicals and proteins. Nucleic Acids Res 36(Database issue):D684–D688
98. Zhao M et al (2014) Predicting Protein-Ligand interactions based on chemical preference features with its application to new D-Amino acid oxidase inhibitor discovery. Curr Pharm Des [epub ahead of print]
99. Ehrman TM, Barlow DJ, Hylands PJ (2007) Virtual screening of Chinese herbs with random forest. J Chem Inf Model 47(2):264–278
100. Paoletta S et al (2008) Screening of herbal constituents for aromatase inhibitory activity. Bioorg Med Chem 16(18):8466–8470
101. Zhao S, Li S (2010) Network-based relating pharmacological and genomic spaces for drug target identification. PLoS ONE 5(7):e11764

102. Li S, Zhang B (2013) Traditional Chinese medicine network pharmacology: theory, methodology and application. Chin J Nat Med 11(2):110–120
103. Borkow G, Lapidot A (2005) Multi-targeting the entrance door to block HIV-1. Curr Drug Targets Infect Disord 5(1):3–15
104. Li T et al (2009) Matrine induces cell anergy in human Jurkat T cells through modulation of mitogen-activated protein kinases and nuclear factor of activated T-cells signaling with concomitant up-regulation of anergy-associated genes expression. Biol Pharm Bull 33(1):40–46
105. Lam W et al (2010) The four-herb Chinese medicine PHY906 reduces chemotherapy-induced gastrointestinal toxicity. Sci Transl Med 2(45):45ra59
106. Xie Y et al (2008) Study on the pharmacokinetics and metabolism of paeonol in rats treated with pure paeonol and an herbal preparation containing paeonol by using HPLC-DAD-MS method. J Pharm Biomed Anal 46(4):748–756
107. Liu ZQ et al (2005) Pharmacokinetic interaction of paeoniflorin and sinomenine: pharmacokinetic parameters and tissue distribution characteristics in rats and protein binding ability in vitro. J Pharmacol Sci 99(4):381–391
108. Zeng H et al (2011) The inhibitory activities of the components of Huang-Lian-Jie-Du-Tang (HLJDT) on eicosanoid generation via lipoxygenase pathway. J Ethnopharmacol 135(2):561–568
109. Wang L et al (2008) Dissection of mechanisms of Chinese medicinal formula Realgar-Indigo naturalis as an effective treatment for promyelocytic leukemia. Proc Natl Acad Sci USA 105(12):4826–4831
110. Zhang QY et al (2009) A systems biology understanding of the synergistic effects of arsenic sulfide and Imatinib in BCR/ABL-associated leukemia. Proc Natl Acad Sci USA 106(9):3378–3383
111. Ehrman TM, Barlow DJ, Hylands PJ (2010) In silico search for multi-target anti-inflammatories in Chinese herbs and formulas. Bioorg Med Chem 18(6):2204–2218
112. Gu J et al (2011) Drug-target network and polypharmacology studies of a traditional Chinese medicine for type II diabetes mellitus. Comput Biol Chem 35(5):293–297
113. Huang Q, Qiao X, Xu X (2007) Potential synergism and inhibitors to multiple target enzymes of Xuefu Zhuyu Decoction in cardiac disease therapeutics: a computational approach. Bioorg Med Chem Lett 17(6):1779–1783
114. Li S et al (2010) Herb network construction and co-module analysis for uncovering the combination rule of traditional Chinese herbal formulae. BMC Bioinform 11(Suppl 11):S6
115. Wang Y et al (2010) A novel methodology for multicomponent drug design and its application in optimizing the combination of active components from Chinese medicinal formula Shenmai. Chem Biol Drug Des 75(3):318–324
116. Wang Y, Wang X, Cheng Y (2006) A computational approach to botanical drug design by modeling quantitative composition-activity relationship. Chem Biol Drug Des 68(3):166–172
117. Li Q et al (2011) A network-based multi-target computational estimation scheme for anticoagulant activities of compounds. PLoS ONE 6(3):e14774
118. Sun Y et al (2012) Towards a bioinformatics analysis of anti-Alzheimer's herbal medicines from a target network perspective. Brief Bioinform
119. Zhao J, Jiang P, Zhang W (2009) Molecular networks for the study of TCM pharmacology. Brief Bioinform 11(4):417–430
120. Zhang B, Wang X, Li S (2013) An integrative platform of TCM network pharmacology and its application on a herbal formula, Qing-Luo-Yin. Evid Based Complement Altern Med 2013:12
121. Klipp E, Wade RC, Kummer U (2010) Biochemical network-based drug-target prediction. Curr Opin Biotechnol 21(4):511–516
122. Leung EL et al (2012) Network-based drug discovery by integrating systems biology and computational technologies. Brief Bioinform

123. Wang T et al (2013) Inferring pathway crosstalk networks using gene set co-expression signatures. Mol BioSyst
124. Kanawong R et al (2012) Automated tongue feature extraction for ZHENG classification in traditional Chinese medicine. Evid Based Complement Alternat Med 2012:912852
125. Soto AM, Sonnenschein C (2005) Emergentism as a default: cancer as a problem of tissue organization. J Biosci 30(1):103–118

Part IV
Functional Analysis of Biological Macromolecules

Chapter 15
Evolutionary Optimization of Transcription Factor Binding Motif Detection

Zhao Zhang, Ze Wang, Guoqin Mai, Youxi Luo, Miaomiao Zhao and Fengfeng Zhou

Abstract All the cell types are under strict control of how their genes are transcribed into expressed transcripts by the temporally dynamic orchestration of the transcription factor binding activities. Given a set of known binding sites (BSs) of a given transcription factor (TF), computational TFBS screening technique represents a cost efficient and large scale strategy to complement the experimental ones. There are two major classes of computational TFBS prediction algorithms based on the tertiary and primary structures, respectively. A tertiary structure based algorithm tries to calculate the binding affinity between a query DNA fragment and the tertiary structure of the given TF. Due to the limited number of available TF tertiary structures, primary structure based TFBS prediction algorithm is a necessary complementary technique for large scale TFBS screening. This study proposes a novel evolutionary algorithm to randomly mutate the weights of different positions in the binding motif of a TF, so that the overall TFBS prediction accuracy is optimized. The comparison with the most widely used algorithm, Position Weight Matrix (PWM), suggests that our algorithm performs better or the same level in all the performance measurements, including sensitivity, specificity, accuracy and Matthews correlation coefficient. Our data also suggests

Zhao Zhang and Miaomiao Zhao have been contributed equally to this paper.

Z. Zhang · Z. Wang
School of Computer Science and Software Engineering, Tianjin Polytechnic University, Tianjin, China

Z. Zhang · G. Mai · Y. Luo · M. Zhao · F. Zhou (✉)
Shenzhen Institutes of Advanced Technology and Key Laboratory for Health Informatics, Chinese Academy of Sciences, Shenzhen, Guangdong, China
e-mail: FengfengZhou@gmail.com; ff.zhou@siat.ac.cn

Y. Luo
School of Science, Hubei University of Technology, Wuhan, Hubei, China

D. Wei et al. (eds.), *Advance in Structural Bioinformatics*, Advances in Experimental Medicine and Biology 827, DOI 10.1007/978-94-017-9245-5_15

that it is necessary to remove the widely used assumption of independence between motif positions. The supplementary material may be found at: http://www.healthinformaticslab.org/supp/ .

Keywords Binding sites · Transcription factor · Position weight matrix · Motif

15.1 Introduction

Transcription of genic regions into RNA molecules is the first step of the biological central dogma, and is dynamically controlled by various transcription factors (TFs) [1]. A TF regulates a gene's transcription through its dynamic binding to a short (5–20 bps) DNA sequence upstream to the regulated gene. This DNA sequence is the TF's binding site (TFBS), which is usually highly specific to this TF and is called a motif [2]. Mutations within TFBSs will change the host's transcription regulatory network, and lead to species specific phenotypes or genetic diseases [3].

There are two major high-throughput strategies to screen the binding sites of a TF in the host genome. Firstly, various high-throughput experimental techniques were developed to screen the TFBSs under the given cell culture conditions, including DNase I footprinting [4], electrophoretic mobility shift assay [5], ChIP-on-chip [6] and ChIP-Seq [7], etc. The dynamic landscape of the transcription regulatory network may be elucidated through these screening techniques. But they are usually costly and labor-intensive, and can only detect the binding sites of one TF under one cell culture condition at a time. Considering the 2,886 transcription factors curated in the human DNA-binding domain (DBD) database [8], and the dynamic nature of transcription regulation, it can be anticipated that the transcription regulatory landscape is significantly under-estimated.

Computational TFBS screening techniques have been used to infer the comprehensive list of TFBSs. The majority of *in silico* TFBS screening techniques assumes that the binding sites of a given TF have a fixed length, and calculates the similarity score of a query DNA sequence compared with the local oligo-nucleotide frequency patterns in the known TFBSs [9]. The computational techniques include the position weigh matrix (PWM) [10], WebLogo [11], and position specific pairwise score [12], etc. The introduction of TF's structural information will greatly reduce the false positive rates, as demonstrated by Facelli [13], Saito et al. [14]. But there are only 300 unique human TF structures in the PDB database [ref], and the limited availability of the experimentally detected TF structures restricts the extensive application of these methods [15].

This study hypothesizes that positions contribute differently to the motif scoring based on their nucleotide frequency patterns, and formulates the position contribution as a weight for the position. The vector of weights for different motif positions were randomly mutated by an evolutionary algorithm, with the

optimization goal to maximize the overall accuracy. The prediction performance suggests that our algorithm performs similarly or better than the position specific scoring strategies.

15.2 Materials and Methods

15.2.1 Data Resources

The proposed algorithm is applied to the following seven transcription factors (TFs), i.e. Ebox, Myc, P53, Q6MAZ, Q601MAZ, V_SREBP_Q3-SREBP (abbreviated as Q3), and V_SREBP2_Q6-SREBP2 (abbreviated as Q6). The known binding sites of these seven transcription factors were manually collected from the database TRANSFAC in August 2012 [16]. Only those binding sites without an "N" letter were kept for further analysis. The target gene sequences and their promoter regions were extracted from the database ENSEMBL [17].

15.2.2 Motif Screening Problem

The mathematical model of the transcription factor binding site (TFBS) screening problem (sTFBS) is formulated as follows. For a given transcription factor (TF), its known fixed-length binding sites are defined to be the positive dataset $P = \{M_1, M_2, \ldots, M_n\}$, where $|M_i| = L$. A negative dataset $N = \{B_1, B_2, \ldots, B_m\}$ is randomly extracted from the promoter regions of the genes regulated by the given TF, where $|B_j| = L$, B_j has no "N" letters and B_j does not overlap with M_i. Considering the promoter region is much larger than a TFBS, we set $m = 10 \times n$. A TFBS screening model is denoted as the classification function $f(X) \in \{P, N\}$, where $X \in P \cup N$.

Firstly, a similarity score between two fixed-length DNA fragments $V = \{v_1, v_2, \ldots, v_L\}$ and $U = \{u_1, u_2, \ldots, u_L\}$ is defined to be $Score(V, U) = (w_1 \times S(v_1, u_1) + w_2 \times S(v_2, u_2) + \cdots + w_L \times S(v_L, u_L))$, where the weight vector $W = \langle w_1, w_2, \ldots, w_L \rangle$ is the pre-calculated combination pattern, and $w_i \in [0, 1]$. The nucleotide similarity score matrix $S(v_i, u_i)$ is defined to be 2 if $v_i = u_i$, 1 for A versus G or C versus T, and -1 for the other pairs [18]. The combination pattern $W = \langle w_1, w_2, \ldots, w_L \rangle$ will be optimized by an evolutionary algorithm, as described in the next section.

This study chose the simple nearest neighbor algorithm as the classification model $f(X)$.

Algorithm *SNN*
Input: The positive and negative datasets are $P=\{M_1, M_2, \ldots, M_n\}$ and $N=\{B_1, B_2, \ldots, B_m\}$, respectively, where $|M_i|=|B_j|=L$. The query sequence is Q, where $|Q|=L$.
Procedure:

```
1.  MaxScoreP=Score(Q, M1); MaxScoreN=Score(Q, B1);
2.  for(i=1;i<=n;i++)
3.  {
4.      CurrentScore= Score(Q, Mi);
5.      if( CurrentScore>MaxScoreP ) { MaxScoreP=CurrentScore; }
6.  }
7.  for(i=1;i<=m;i++)
8.  {
9.      CurrentScore= Score(Q, Bi);
10.     if( CurrentScore>MaxScoreN ) { MaxScoreN=CurrentScore; }
11. }
12. if( MaxScoreP>MaxScoreN ) return P;
13. else return N;
```

Position Weight Matrix (PWM) algorithm assumes that positions in a fixed-length motif are independent to each other and calculates how a query sequence is similar to the set of known motif occurrences [10, 19]. Firstly, a position conservation factor M_i is calculated as $M_i = \sum_{b\in\{A,T,C,G\}} (f_i(b)/N - P_0(b))^2/P_0(b)$, $i = 1, 2, \ldots, L$, where $f_i(b)$ is the observed frequencies of nucleotide b at position i in the set of known motif occurrences, and $P_0(b)$ is the background frequency of nucleotide b. Then the position probability matrix (PPM) is calculated as:

$$PPM = \begin{pmatrix} P_1(A) & P_2(A) & \cdots & P_n(A) \\ P_1(T) & P_2(T) & \cdots & P_n(T) \\ P_1(C) & P_2(C) & \cdots & P_n(C) \\ P_1(G) & P_2(G) & \cdots & P_n(G) \end{pmatrix},$$

where $P_j(b) = \{f_j(b) + s(b)\}/\left\{N + \sum_{b\in\{A,T,C,G\}} s(b)\right\}$, and $s(b) = P_0(b)\sqrt{N}$ is a smoothing factor.

Then the position weight matrix (PWM) is calculated as

$$PWM = \begin{pmatrix} w_1(A) & w_2(A) & \cdots & w_n(A) \\ w_1(T) & w_2(T) & \cdots & w_n(T) \\ w_1(C) & w_2(C) & \cdots & w_n(C) \\ w_1(G) & w_2(G) & \cdots & w_n(G) \end{pmatrix},$$

where $w_i(b) = \ln\{P_i(b)/P_0(b)\}$.

The standardized similarity score of a query sequence Q is defined to be

$$S(Q) = \frac{\sum_{i=1}^{L} M_i w_i(Q_i) - \sum_{i=1}^{L} M_i \min\{w_i(b)\}}{\sum_{i=1}^{L} M_i \max\{w_i(b)\} - \sum_{i=1}^{L} M_i \min\{w_i(b)\}},$$

where Q_i is the i^{th} nucleotide in Q, and $b \in \{\text{A, T, C, G}\}$. For a cutoff S_0, only if $S(Q) \geq S_0$, Q is defined as a binding motif of the transcription factor.

15.2.3 Prediction Performance Measurements and Evaluation

Given the positive dataset $P = \{M_1, M_2, \ldots, M_n\}$, and the negative dataset $N = \{B_1, B_2, \ldots, B_m\}$, where $|M_i| = |B_j| = L$. M_i is a true positive or false negative if $SNN(M_i) = P$ or N, respectively, whereas B_j is a true negative or false positive if $SNN(B_j) = N$ or P, respectively. For the classification model $SNN(X)$, the numbers of true positives, false negatives, true negatives and false positives are abbreviated as TP, FN, TN and FP, respectively. The classification performance of the model is measured by sensitivity (Sn), specificity (Sp), accuracy (Ac) and Matthews correlation coefficient (MCC) [20, 21], which are defined as follows. $Sn = TP/(TP + FN)$, $Sp = TN/(TN + FP)$, $Ac = (Sn + Sp)/2$, and $MCC = (TP \times TN - FP \times FN)/sqrt((TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN))$, where $sqrt(t)$ is the squared root of t.

A line plot will be generated for the evolutionarily optimized combination pattern $W = \langle w_1, w_2, \ldots, w_L \rangle$ for the comparison with the WebLogo plot. TFBS screening algorithms usually use the visual technique WebLogo to demonstrate the DNA compositions at each position in the TFBS, and a higher plotted position suggests a larger information content [11]. An initial weight vector $W^0 = \langle w_1^0, w_2^0, \ldots, w_L^0 \rangle$ is generated from a transcription factor's WebLogo plot, by scaling the information content at position i to [0, 1] as w_i^0.

Two validation strategies are adopted to evaluate the classification algorithm SNN's prediction performance. Firstly, the algorithm SNN is investigated for its leave-one-out (LOO) cross validation performance, i.e. iteratively choosing one data entry and investigating its prediction by the classification model trained on the rest data sets. The LOO validation strategy has been widely used to measure how a TFBS or other functional element prediction algorithm performs [22, 23]. To further investigate the dataset dependency of the proposed SNN algorithm, this study conducted 3-fold cross validation (3FCV) strategy [24–26]. The basic idea is to randomly split the positive and negative datasets into 3 equal-size subsets $\{P_1, P_2, P_3\}$ and $\{N_1, N_2, N_3\}$, respectively. The prediction results are iteratively investigated for $\{P_i, N_i\}$ using the SNN trained on $P \backslash P_i$ and $N \backslash N_i$, where $i = 1$, 2, and 3. A self validation (denoted as Self) is also used to evaluate the self consistency, which is to evaluate how a classification model performs on the training dataset.

15.2.4 Evolutionary Optimization Algorithm

This study proposed an evolutionary optimization algorithm to screen for the weight vector with the best overall accuracy *Ac* of the algorithm *SNN*, as shown in Fig. 15.1. The basic idea of an evolutionary optimization algorithm (EOA) is to simulate the natural selection process [27, 28]. Each generation of individuals produce children through the operations of crossing and mutation from a pair of parents. A fitness function is defined to describe how each children fit the natural selection pressure. A better fitness leads to a higher chance to survive into the next generation. The population size is usually fixed to a constant value [11, 29–37].

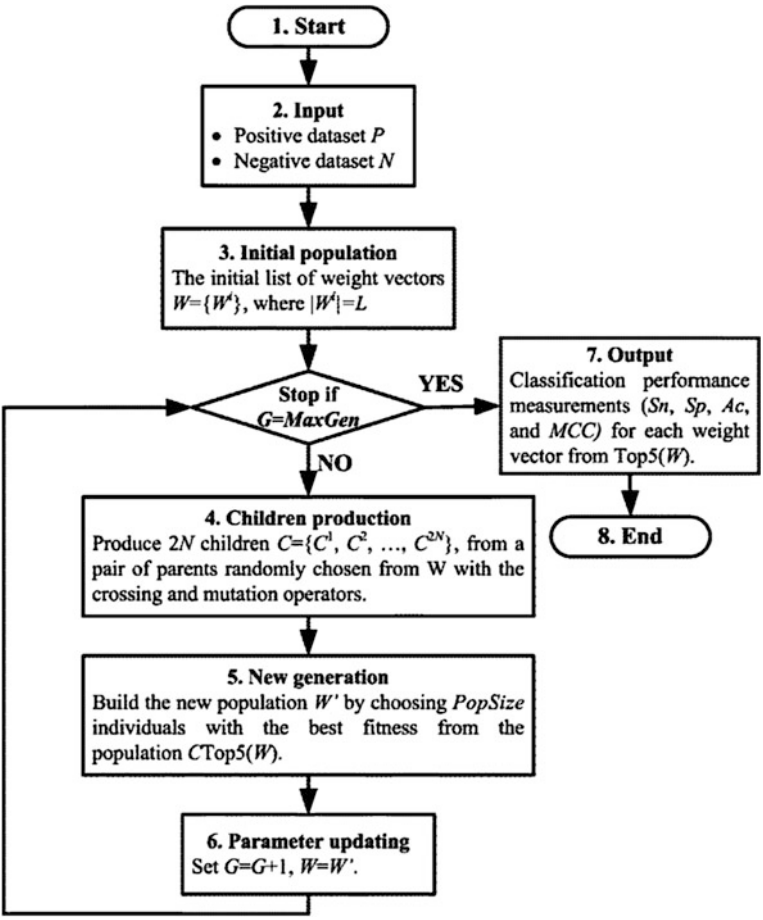


Fig. 15.1 Procedure of the evolutionary optimization algorithm. 5 weight vectors with the best accuracies *Ac* will be output

The initial population W consists of *PopSize* individual weight vectors, *i.e.* W^i, where $i \in \{1, 2, \ldots, PopSize\}$. Each individual W^i is an L-dimension vector $W^i = \langle W_0^i, W_1^i, \ldots, W_L^i \rangle$, where W_j^i is a random value between 0 and 1.

MaxGen generations of natural mutation and selection are conducted to find the fittest weight vectors. For a given weight vector W^i, an SNN classification model is built, and the overall classification accuracy *Ac* with the 4-fold cross validation is defined to be the fitness function $Ac(W^i)$, as used in step 5. For the population of weight vectors W, $Top5(W)$ consists of 5 weight vectors with the best fitness in the population. The final top 5 weight vectors together with the performance measurements of their classification models are output.

15.3 Results and Discussion

15.3.1 Best Parameters for EOA

There are two parameters for the evolutionary algorithm EOA, i.e. the population size *PopSize* and the generation number *MaxGen*. Previous studies suggested that $PopSize = 100$ performs well for the evolutionary optimization problems with individual vector size ~ 10 [38]. So we firstly fix $PopSize = 100$, and investigate how the optimization goal, *Ac*, changes with the increased number of generations, i.e. *MaxGen*. The parameter *MaxGen* is set between 0 and 5,000, and the step size is 100. Q6MAZ and Q3 quickly reach the peak *Ac* value 1.00 after just $MaxGen = 200$ generations of optimizations, as shown in Fig. 15.2a. The TF genes Ebox, Myc and P53 also reach very high *Ac* values (>97 %) at just $MaxGen = 200$. If we choose the Ac value at MaxGen = 5,000 as the final result, all the six investigated TFs reach this peak value at MaxGen = 3,000, as shown in Fig. 15.2a.

We further investigate how the parameter *PopSize* impacts the optimization performance of EOA, as shown in Fig. 15.2 and Supplementary Figure S1. By choosing $PopSize \in \{20, 40, 60, 80, 100, 120, 140, 160, 200\}$, the overall accuracy Ac is calculated for generation $G \in \{0, 100, 200, \ldots, 4{,}900, 5{,}000\}$ of EOA on each of the six TFs. Figure 15.2 shows that the TFBS prediction problem of Q6 is the most difficult to be optimized, and reaches the peak values at generations 3,800, 3,000 and 2,600 for $PopSize = 60$, 100 and 140, respectively. All the other five TFs reach the peak *Ac* values before the optimization generation 3,000. Similar patterns can be observed for other population sizes *PopSize*, as in Supplementary Figure S1.

Considering that the running time of the evolutionary algorithm EOA increases linearly with the product $PopSize \times MaxGen$, and the above data, this study will set $PopSize = 100$ and $MaxGen = 3{,}000$ for the following experiments.

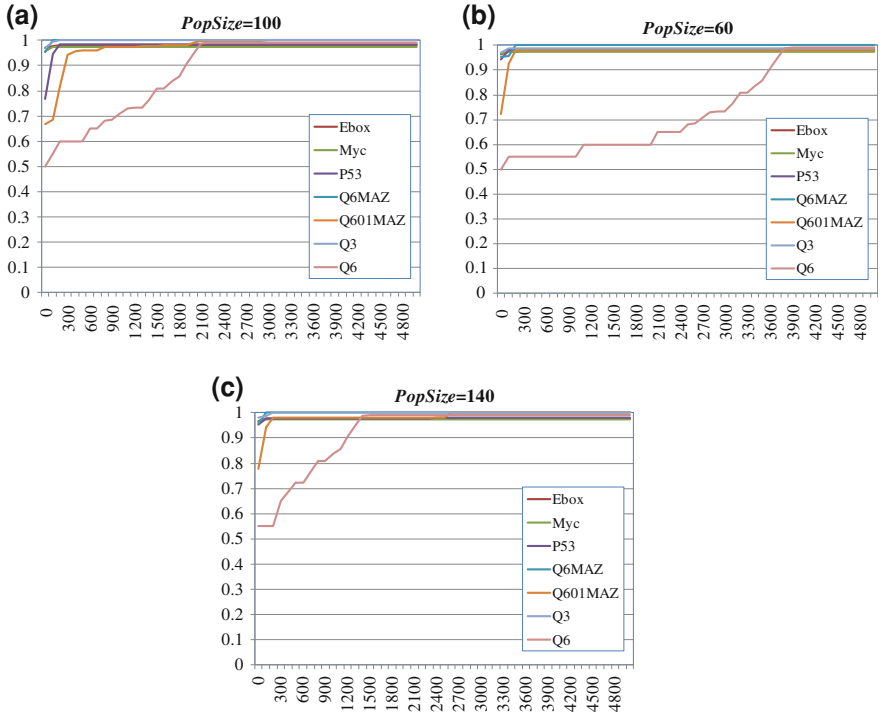


Fig. 15.2 Distributions of overall classification accuracy, *Ac*, for different generation numbers. The population sizes *PopSize* are fixed to **a** 100, **b** 60 and **c** 140, respectively

15.3.2 Comparison of PWM and SNN(W0)

We firstly compare the widely used PWM algorithm with the SNN algorithm. WebLogo is also widely used to demonstrate the information content or conservation at each position of a motif [11]. The higher a position is, the larger information content this position has, as shown in Fig. 15.3. And the binding sites of all the seven TFs do show significant patterns in information content of some motif positions. So we hypothesize that the information content from WebLogo plot may represent well the weight of each motif position for the SNN algorithm, and the weight vector is denoted as W_0.

Both PWM and SNN score the similarity of a query DNA sequence to the known TFBSs, and this study chooses the cutoff score with $Sn \sim Sp$ for the comparison. In general, the SNN(W_0) algorithm performs similarly well or slightly worse compared with the PWM algorithm, as shown in Table 15.1. Both algorithms produce $\sim$90 % or larger overall accuracy *Ac* for the TFBS motif screening problem, and the TF Q3 even receives 100 % accurate separation of the positive and negative data entries from both algorithms under the two validation strategies.

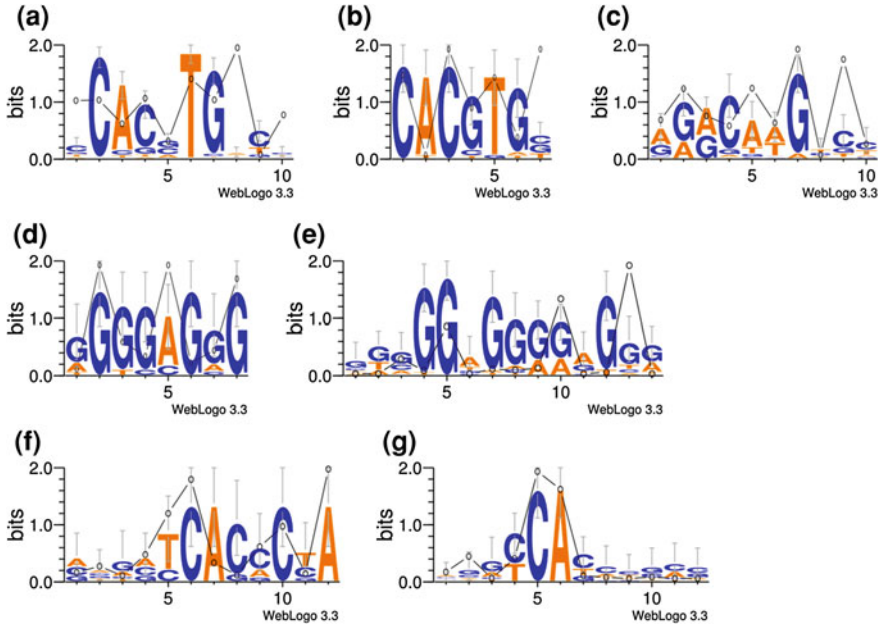


Fig. 15.3 WebLogo plots for the TFs. **a** Ebox, **b** Myc, **c** P53, **d** Q6MAZ, **e** Q601MAZ, **f** Q3 and **g** Q6. *The line plot* is for the evolutionarily optimized weight vector by the SNN + EOA algorithms for each TF

The biggest difference between the two algorithms is for the TFBS motif screening problem of Myc, where SNN(W_0) performs 5.01 and 5.48 % worse in *Ac* than PWM using the LOO and 3FCV validations, respectively. So our first hypothesis about the usage of W_0 is reasonable but may need further optimization.

15.3.3 Comparison of PWM and SNN + EOA

The next hypothesis is that there may exist a weight vector $W = \langle w_1, w_2, \ldots, w_L \rangle$ with increased *Ac* value for the SNN algorithm. Besides the position independent measurements, e.g. PWM or WebLogo, there is no available knowledge about how to optimize the weight vector. So we choose to use the evolutionary optimization algorithm to search for a weight vector with optimal overall accuracy Ac by just random mutations in the weight vectors, as described in Sect. 15.2.4.

After the optimization of $MaxGen = 3{,}000$ generations of $PopSize = 100$ individuals (weight vectors), the motif screening algorithm SNN outperforms the PWM algorithm in any performance measurements for all the seven TFs, as shown in Table 15.2. The PWM algorithm achieves 100 % accuracy for the LOO validation of Q6MAZ and both LOO and 3FCV validations of Q3, and the

Table 15.1 Prediction performances of the algorithms PWM and SNN(W_0)

		LOO					3FCV				
	Method	Cutoff	*Sn*	*Sp*	*Ac*	*MCC*	Cutoff	*Sn*	*Sp*	*Ac*	*MCC*
Ebox	PWM	0.6500	0.9664	0.9454	0.9559	0.7660	0.6400	0.9580	0.9420	0.9500	0.7312
	SNN(W_0)	1.3830	0.9832	0.8521	0.9176	0.5761	1.4720	0.9664	0.9134	0.9399	0.6789
Myc	PWM	0.5900	0.9524	0.9429	0.9477	0.7450	0.5600	0.9524	0.9333	0.9429	0.7187
	SNN(W_0)	1.0310	1.0000	0.7952	0.8976	0.5108	1.0310	0.9524	0.8281	0.8881	0.5176
P53	PWM	0.7500	0.8478	0.9674	0.9076	0.7590	0.6800	0.9130	0.9196	0.9163	0.6594
	SNN(W_0)	1.3640	1.0000	0.8913	0.9457	0.6235	1.3220	1.0000	0.9000	0.9500	0.6708
Q6MAZ	PWM	0.7600	1.0000	1.0000	1.0000	1.0000	0.7400	1.0000	0.9917	0.9959	0.9568
	SNN(W_0)	1.3900	1.0000	0.9667	0.9833	0.8515	1.3900	1.0000	0.9667	0.9833	0.8515
Q601MAZ	PWM	0.7100	0.9630	0.9963	0.9796	0.9593	0.6600	0.9630	0.9778	0.9704	0.8722
	SNN(W_0)	1.3210	0.9259	0.9778	0.9519	0.8497	1.3000	0.9630	0.9741	0.9685	0.8572
Q3	PWM	0.6600	1.0000	1.0000	1.0000	1.0000	0.6600	1.0000	1.0000	1.0000	1.0000
	SNN(W_0)	1.4000	1.0000	1.0000	1.0000	1.0000	1.4000	1.0000	1.0000	1.0000	1.0000
Q6	PWM	0.8800	1.0000	0.9550	0.9775	0.8050	0.8300	1.0000	0.9550	0.9775	0.8050
	SNN(W_0)	1.4400	1.0000	0.9500	0.9750	0.7889	1.4400	1.0000	0.9550	0.9775	0.8050

The leave-one-out (LOO) validation and 3-fold cross validation (3FCV) strategies are used for the seven TFs

Table 15.2 Prediction performances of the algorithms PWM and SNN + EOA

		3FCV					LOO				
	Method	Cutoff	*Sn*	*Sp*	*Ac*	*MCC*	Cutoff	*Sn*	*Sp*	*Ac*	*MCC*
Ebox	PWM	0.6400	0.9580	0.9420	0.9500	0.7312	0.6500	0.9664	0.9454	0.9559	0.7660
	SNN + EOA	1.5000	0.9748	0.9445	0.9597	0.7639	1.6040	0.9832	0.9773	0.9803	0.8825
Myc	PWM	0.5600	0.9524	0.9333	0.9429	0.7187	0.5900	0.9524	0.9429	0.9477	0.7450
	SNN + EOA	1.6320	0.9524	0.9667	0.9596	0.8224	1.8000	0.9524	0.9952	0.9738	0.9476
P53	PWM	0.6800	0.9130	0.9196	0.9163	0.6594	0.7500	0.8478	0.9674	0.9076	0.7590
	SNN + EOA	1.4870	1.0000	0.9413	0.9707	0.7702	1.6000	1.0000	0.9717	0.9859	0.8704
Q6MAZ	PWM	0.7400	1.0000	0.9917	0.9959	0.9568	0.7600	1.0000	1.0000	1.0000	1.0000
	SNN + EOA	1.7500	1.0000	1.0000	1.0000	1.0000	1.7500	1.0000	1.0000	1.0000	1.0000
Q601MAZ	PWM	0.6600	0.9630	0.9778	0.9704	0.8722	0.7100	0.9630	0.9963	0.9796	0.9593
	SNN + EOA	1.9000	1.0000	1.0000	1.0000	1.0000	1.9000	1.0000	1.0000	1.0000	1.0000
Q3	PWM	0.6600	1.0000	1.0000	1.0000	1.0000	0.6600	1.0000	1.0000	1.0000	1.0000
	SNN + EOA	1.5000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Q6	PWM	0.8300	1.0000	0.9550	0.9775	0.8050	0.8800	1.0000	0.9550	0.9775	0.8050
	SNN + EOA	1.9000	1.0000	0.9850	0.9925	0.9223	1.9010	1.0000	0.9850	0.9925	0.9223

The leave-one-out (LOO) validation and 3-fold cross validation (3FCV) strategies are used for the seven TFs

SNN + EOA algorithm achieves such perfect classification. For the other transcription factors, SNN + EOA outperforms PWM by 0.97–7.83 % in overall accuracy Ac. The measurements MCC $\in [-1, 1]$ evaluates how the prediction results match the positive and negative datasets, and a larger MCC means a better prediction. Besides the two TFs Q6MAZ and Q3 that both algorithms perform equally well, SNN + EOA improves the MCC of PWM algorithm by 0.0327–0.2026. The PWM algorithm does not perform well on the dataset of the well-known tumor suppressor P53, as in Table 15.2. It only achieves $Sn = 84.78$ % and $Sp = 96.74$ % for the LOO validation of P53, and the overall accuracy is only 90.76 %. SNN + EOA achieves a slightly better specificity ($Sp = 97.17$ %) and a much better sensitivity ($Sn = 100$ %). A similar improvement is also achieved by SNN + EOA for the 3FCV validation of P53.

It's also interesting to observe that the weight vector achieving the best prediction performance does not match the position independent measurement WebLogo, as shown in Fig. 15.3. For the tumor suppressor P53, the optimized weight vector does not agree with WebLogo at positions 4, 5 and 9, as shown in Fig. 15.3c. The information content at position 4 is larger than that at position 5, but their weights in the optimized vector weighs the two positions reversely. And although the information content at position 9 only ranks 8th, position 9 has the second largest weight. Similar discrepancy exists for all the seven investigated TFs, as in Fig. 15.3, and suggests that a concerted weighing of different positions is necessary for motif screening and other similar problems.

Acknowledgments This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB13040400), Shenzhen Peacock Plan (KQCX20130628112914301), Shenzhen Research Grant (ZDSY20120617113021359), China 973 program (2010CB732606), the MOE Humanities Social Sciences Fund (No.13YJC790105) and Doctoral Research Fund of HBUT (No. BSQD13050). Computing resources were partly provided by the Dawning supercomputing clusters at SIAT CAS.

References

1. Crick F (1970) Central dogma of molecular biology. Nature 227(5258):561–563
2. Ameur A, Rada-Iglesias A, Komorowski J, Wadelius C (2009) Identification of candidate regulatory SNPs by combination of transcription-factor-binding site prediction, SNP genotyping and haploChIP. Nucleic Acids Res 37(12):e85
3. Wray GA (2007) The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 8(3):206–216
4. Galas DJ, Schmitz A (1978) DNAase footprinting a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res 5(9):3157–3170
5. Dent C, Latchman D (1993) The DNA mobility shift assay. In: Transcription factors: a practical approach, pp 1–3
6. Pillai S, Chellappan SP (2009) ChIP on chip assays: genome-wide analysis of transcription factor binding and histone modifications. In: Chromatin protocols. Springer, Berlin, pp 341–366

7. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316(5830):1497–1502
8. Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA (2008) DBD–taxonomically broad transcription factor predictions: new content and functionality. Nucleic Acids Res 36(Database issue):D88–D92
9. Stormo GD (2000) DNA binding sites: representation and discovery. Bioinformatics 16(1):16–23
10. Ben-Gal I, Shani A, Gohr A, Grau J, Arviv S, Shmilovici A, Posch S, Grosse I (2005) Identification of transcription factor binding sites with variable-order Bayesian networks. Bioinformatics 21(11):2657–2666
11. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14(6):1188–1190
12. Quader S, Huang CH (2012) Effect of positional dependence and alignment strategy on modeling transcription factor binding sites. BMC Res Notes 5:340
13. Gorin AA, Zhurkin VB, Wilma K (1995) *B*-DNA twisting correlates with base-pair morphology. J Mol Biol 247(1):34–48
14. Oshchepkov DY, Vityaev EE, Grigorovich DA, Ignatieva EV, Khlebodarova TM (2004) SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition. Nucleic Acids Res 32(suppl 2):W208–W212
15. Rose PW, Bi C, Bluhm WF, Christie CH, Dimitropoulos D, Dutta S, Green RK, Goodsell DS, Prlic A, Quesada M et al (2013) The RCSB Protein Data Bank: new resources for research and education. Nucleic Acids Res 41(Database issue):D475–D482
16. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K et al (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 34(Database issue):D108–D110
17. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S et al (2013) Ensembl 2013. Nucleic Acids Res 41(Database issue):D48–D55
18. String Alignment using Dynamic Programming.(http://www.biorecipes.com/DynProgBasic/code.html)
19. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E (2003) MATCH: a tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res 31(13):3576–3579
20. Xue Y, Zhou F, Zhu M, Ahmed K, Chen G, Yao X (2005) GPS: a comprehensive www server for phosphorylation sites prediction. Nucleic Acids Res 33(Web Server issue):W184–W187
21. Zhou FF, Xue Y, Chen GL, Yao X (2004) GPS: a novel group-based phosphorylation predicting and scoring method. Biochem Biophys Res Commun 325(4):1443–1448
22. Sheffield NC, Thurman RE, Song L, Safi A, Stamatoyannopoulos JA, Lenhard B, Crawford GE, Furey TS (2013) Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. Genome Res 23(5):777–788
23. Zhou Q, Liu JS (2004) Modeling within-motif dependence for transcription factor binding site predictions. Bioinformatics 20(6):909–916
24. Cheng C, Ung M, Grant GD, Whitfield ML (2013) Transcription factor binding profiles reveal cyclic expression of human protein-coding genes and non-coding RNAs. PLoS Comput Biol 9(7):e1003132
25. Zhou F, Xu Y (2010) cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. Bioinformatics 26(16):2051–2052

26. Qian J, Lin J, Luscombe NM, Yu H, Gerstein M (2003) Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. Bioinformatics 19(15):1917–1926
27. Potts JC, Giddens TD, Yadav SB (1994) The development and evaluation of an improved genetic algorithm based on migration and artificial selection. IEEE Trans Syst Man Cybern 24(1):73–86
28. Tam KY (1992) Genetic algorithms, function optimization, and facility layout design. Eur J Oper Res 63(2):322–346
29. Anastassopoulos G, Adamopoulos A, Galiatsatos D, Drosos G (2013) Feature extraction of osteoporosis risk factors using artificial neural networks and genetic algorithms. Stud Health Technol Inform 190:186–188
30. Santiso EE, Musolino N, Trout BL (2013) Design of linear ligands for selective separation using a genetic algorithm applied to molecular architecture. J Chem Inf Model 53(7):1638–1660
31. Chen JB, Chuang LY, Lin YD, Liou CW, Lin TK, Lee WC, Cheng BC, Chang HW, Yang CH (2013) Genetic algorithm-generated SNP barcodes of the mitochondrial D-loop for chronic dialysis susceptibility. Mitochondrial DNA
32. Sale M, Sherer EA (2013) A genetic algorithm based global search strategy for population pharmacokinetic/pharmacodynamic model selection. Brit J Clin Pharmacol
33. Yoon Y, Kim YH (2013) An efficient genetic algorithm for maximum coverage deployment in wireless sensor networks. IEEE Trans Cybern
34. Azadnia AH, Taheri S, Ghadimi P, Mat Saman MZ, Wong KY (2013) Order batching in warehouses by minimizing total tardiness: a hybrid approach of weighted association rule mining and genetic algorithms. Sci World J 2013:246578
35. Chuang LY, Cheng YH, Yang CH, Yang CH (2013) Associate PCR-RFLP assay design with SNPs based on genetic algorithm in appropriate parameters estimation. IEEE Trans Nanobiosci 12(2):119–127
36. Khotanlou H, Afrasiabi M (2012) Feature selection in order to extract multiple sclerosis lesions automatically in 3D brain magnetic resonance images using combination of support vector machine and genetic algorithm. J Med Signals Sens 2(4):211–218
37. Kou J, Xiong S, Fang Z, Zong X, Chen Z (2013) Multiobjective optimization of evacuation routes in stadium using superposed potential field network based ACO. Comput Intell Neurosci 2013:369016
38. Deb K, Pratap A, Agarwal S, Meyarivan T (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans Evol Comput 6(2):182–197

Chapter 16
Prediction of Serine/Threonine Phosphorylation Sites in Bacteria Proteins

Zhengpeng Li, Ping Wu, Yuanyuan Zhao, Zexian Liu and Wei Zhao

Abstract As a critical post-translational modification, phosphorylation plays important roles in regulating various biological processes, while recent studies suggest that phosphorylation in bacteria is also critical for functional signaling transduction. Since identification of phosphorylation substrates and sites is fundamental for understanding the phosphorylation mediated regulatory mechanism, a number of studies have been contributed to this area. Since experimental identification of phosphorylation sites is time-consuming and labor-intensive, computational predictions attract much attention for its convenience to provide helpful information. However, although there are a large number of computational studies in eukaryotes, predictions in bacteria are still rare. In this study, we present a new predictor of cPhosBac to predict phosphorylation serine/threonine in bacteria proteins. The predictor is developed with CKSAAP algorithm, which was combined with motif length selection to optimize the prediction, which achieves promising performance. The online service of cPhosBac is available at: http://netalign.ustc.edu.cn/cphosbac/.

Keywords Phosphorylation · Prediction · Bacteria · CKSAAP

Z. Li · P. Wu
Institute of Applied Microbiology, Anhui Science and Technology University, Fengyang, Anhui, China

Y. Zhao
School of Arts and Media, Hefei Normal University, Hefei, Anhui, China

Z. Liu (✉) · W. Zhao (✉)
Hefei National Laboratory for Physical Sciences at Microscale and School of Life Sciences, University of Science and Technology of China, Hefei, Anhui, China
e-mail: lzx@mail.ustc.edu.cn

W. Zhao
e-mail: zhaowei@ustc.edu.cn

D. Wei et al. (eds.), *Advance in Structural Bioinformatics*, Advances in Experimental Medicine and Biology 827, DOI 10.1007/978-94-017-9245-5_16

16.1 Introduction

In 1992, Edmond H. Fischer and Edwin G. Krebs were award the Nobel Prize in Physiology or Medicine for their discovery that reversible protein phosphorylation is a critical biological regulatory mechanism in biology [1]. So far, numerous studies have been contributed to dissect the biological functions and regulatory relationships of phosphorylation [2–4]. Although most of these researches were carried out in eukaryotes, phosphorylation in bacteria also attracted great attention for its great functional importance in regulation of cellular signaling [5–9]. Previously, studies on phosphorylation in bacteria were focused on phosphorylation of histidine and aspartate, which play critical roles in the two-components systems for signal transduction [8, 9]. However, recent discoveries indicate that serine/threonine phosphorylation system in bacteria also play important roles in the regulation of cellular processes [5, 7] and might be critical for the virulence of pathogens [5, 7]. Since researches in this area are less intensive, more efforts should be contributed.

Identification of phosphorylation substrates and sites is fundamental to understanding the molecular regulatory mechanisms and biological functions of phosphorylation, while recent advancement of large-scale technologies such as high-throughput mass spectrometry greatly promoted the discoveries of phosphorylation events [10, 11]. Although most of the studies are carried out in eukaryotes, pioneering scientists also conducted large-scale identification of serine/threonine phosphorylation in bacteria [12, 13]. Besides experimental efforts, a large number of computational studies have been carried out to predict and analyze the phosphorylation data [10, 14, 15]. Various predictors, databases and analyzing tools were developed in this area [14, 15]. However, most of these tools are developed for the phosphorylation in eukaryotes, while only NetPhosBac was constructed for serine/threonine phosphorylation in bacteria [16]. In this regard, more efforts should be contributed to improve the prediction.

In this study, we developed a novel predictor of cPhosBac to predict serine/threonine phosphorylation sites in bacteria. The well-constructed dataset was retrieved from NetPhosBac. The composition of *k*-spaced amino acid pairs (CKSAAP) method [17–21] was employed to encode the sequence context surrounding the phosphorylation sites, while the support vector machine (SVM) algorithm was used to classify the positive sites from negative sites. The motif length selection algorithm was adopted to optimize the length of sequence surrounding the phosphorylation sites. Through careful evaluation with 4-, 6-, 8- and 10-fold cross validation, it was found that the prediction performance of cPhosBac is promising. Furthermore, the comparison between cPhosBac and NetPhosBac was conducted, while the result indicate that cPhosBac is more accuracy than NetPhosBac. Taken together, it was proposed that cPhosBac could serve as a useful tool to predict serine/threonine phosphorylation in bacteria. The online service of cPhosBac is available at: http://netalign.ustc.edu.cn/cphosbac/.

16.2 Materials and Methods

16.2.1 Data Preparation and Analysis

The dataset was retrieved from NetPhosBac [16], which contains 152 phosphorylation serine/threonine sites in 119 substrates. The phosphorylation sites were identified by previous studies [12, 13], while a homology reduction was conducted to avoid overestimation of prediction performance during the construction of NetPhosBac [16]. The authors set the ratio of negative/positive as 5, so their training dataset contain 152 positive sites and 841 negative sites. However, in our study, all the non-phosphorylated serines/threonines were regarded as negative sites, which resulted in 152 positive sites and 5761 negative sites. To present the differences between positive and negative sites, a two sample logo was created by Two Sample Logo software [22].

16.2.2 The CKSAAP Method

A previously developed sequence encoding method, the composition of *k*-spaced amino acid pairs [18, 19], was employed. The sequence window was presented by the combination of multiple *k*-spaced amino acid pairs. For instance, the space number *k* for "DxxD" and "DxxxxxD" is 2 and 5, respectively. In this study, beside the 20 types of amino acids, "-" was employed to complete the sequence window for N-terminal or C-terminal phosphorylation sites. So, the number of pair types is 441 (21 × 21). Then the composition of each possible *k*-spaced amino acid pair i could be calculated as:

$$\mathrm{CKSAAP}[\mathrm{i} = 1, 2, \ldots, (k+1) \times 441] = \mathrm{N_i}/(\mathrm{W} - k - 1)$$

$\mathrm{N_i}$ represents the count of the *k*-spaced amino acid pair i while W is the window size. In this study, the value of *k* was exhausted tested from 0 to 7, while the optimized *k* was set to be 5 due to its better performance, which resulted in a 2205-dimensional feature vector.

The SVM-light package (http://svmlight.joachims.org/) was employed to build the SVM classifier, while the parameters were adopted from previous study [19].

16.2.3 Performance Improvement by Motif Length Selection

Previously, the sequence window size of the CKSAAP method in a number of studies was manually selected without any optimization [19–21]. Recently, Xue

et al. [23] introduced a motif length selection algorithm to determine the sequence window. We adopted the algorithm as follows:

The *phosphorylation site peptide* PSP(m, n) was defined as the sequence window of a serine/threonine (K) residue flanked by m amino acids upstream and n amino acids downstream. Then the combinations of PSP(m, n) (m = 1, …, 10; n = 1, …, 10) were extensively tested, while the optimized sequence window of PSP(m, n) with the highest AROC value of 10-fold cross validation was determined.

16.2.4 Performance Evaluation

As previously described, four measurements of sensitivity (*Sn*), specificity (*Sp*), accuracy (*Ac*), and Mathew's Correlation Coefficient (*MCC*) were employed to evaluate the prediction performance. The four measurements were defined as below:

$$Sn = \frac{TP}{TP + FN}, Sp = \frac{TN}{TN + FP}, Ac = \frac{TP + TN}{TP + FP + TN + FN},$$

and

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}.$$

In this study, the 4-, 6-, 8- and 10-fold cross-validations were performed, while the Receiver Operating Characteristic (ROC) curves and area under ROCs (AROCs) were analyzed.

16.3 Results

16.3.1 Sequence Analysis of Phosphorylation Sites in Bacteria

Although there are a number of sites identified, the sequence features and motifs of phosphorylation in bacteria are still unknown. In this study, the two sample logo was constructed to present the sequence features in the phosphorylation and non-phosphorylation sites. With the Two Sample Logo software [22], the enriched and depleted amino acid types around the phosphorylation sites were presented in Fig. 16.1. Interestingly, it was observed that the enriched phosphorylated residues were serine (Position 11 in Fig. 16.1), while threonine was depleted. The −1 position of the phosphorylation sites (Position 10 in Fig. 16.1) enrich charged

Fig. 16.1 The two sample logo for the phosphorylation dataset

residues, including positively charged residues lysine (K) and histidine (H) and negatively charged residue aspartate (D), while hydrophobic residue tyrosine (Y) is depleted in this position. Non-charged residues including asparagine (N), glycine (G) and methionine (M) were also enriched in the −1, −2 and −3 positions (Position 10, 9 and 8 in Fig. 16.1). In the position +1 (Position 12 in Fig. 16.1), positively charged residue histidine (H) and negatively charged residue aspartate (D) were enriched and depleted, respectively. Other detailed features were presented in Fig. 16.1.

Although the motif for phosphorylation in bacteria is unknown currently, we conducted the motif analysis for the dataset with the Motif-All software [24]. The identified motifs by the software are presented in Table 16.1. It is obvious that the sequence features were consistent with the motifs. For example, the motif [pS]H represent the sequence feature of histidine preference in position +1.

Table 16.1 Identified motifs by Motif-All for the dataset

Motif	Count in positive	Count in negative	Z-score	*p*-value
----------[pS]H---------	15	118	5.76	4.21E-09
--L------D[pS]----------	8	23	6.27	1.76E-10
------L-G-[pS]----------	8	43	5.07	1.95E-07
-------G--[pS]L---------	8	44	5.02	2.54E-07
--------G-[pS]----A-----	9	47	5.45	2.56E-08
--------G-[pS]-----L----	9	47	5.45	2.56E-08
--------G-[pS]------I---	8	30	5.81	3.18E-09

16.3.2 The Motif Length Selection Algorithm to Optimize the Sequence Window

Previously, a number of studies have been carried out with CKSAAP employed as the prediction method [17–21]. In these studies, the length of peptide considered for prediction was manually determined [17–21]. However, recent study showed that different motif lengths for prediction could generate various performances [23]. Interestingly, it was observed that the self-consistence performances always increase when the motif for prediction elongates, while the leave-one-out validation performances firstly increase to a peak and then decrease when the length of peptide for prediction increase [23]. Since the self-consistence performances represent the accuracy of prediction for the training dataset while the leave-one-out or n-fold validation could represent the prediction ability for the new or unknown sites, it is obvious that leave-one-out or n-fold validation should be better to be employed as the measurement when optimizing the parameters.

In this study, we adopted the motif length selection approach to determine the motif length for prediction in CKSAAP algorithm. The area under receiver operating characteristic curve (AROC) values of 10-fold cross validations were calculated to evaluate the performance. The heatmap of performances for different combination of PSP(m, n) was presented in Fig. 16.2 with the ggplot program (http://had.co.nz/ggplot2/) in the R package (http://www.r-project.org/) [25]. The result showed that the best combination of PSP(m, n) is PSP(3, 6), which achieved a AROC value of 0.7147.

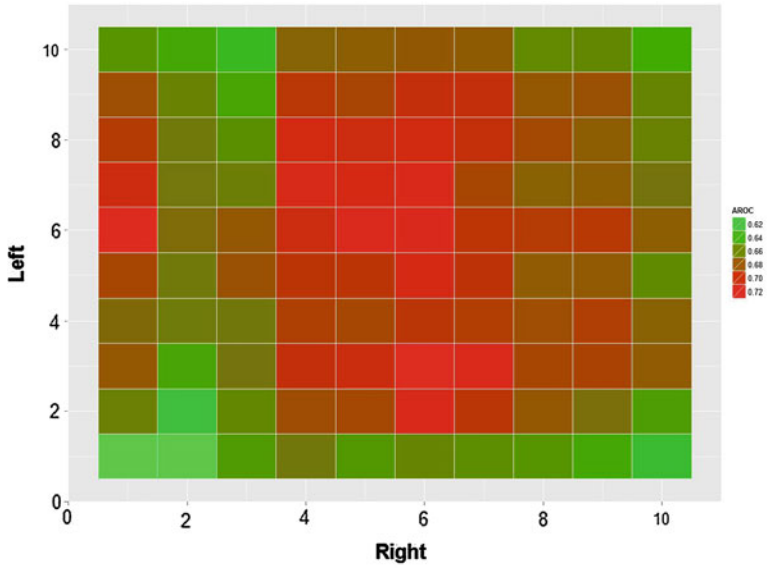


Fig. 16.2 The heatmap of the AROC values for the motif length selection. *Left* and *right* represent the m and n of the PSP(m, n)

16.3.3 Performance Evaluation and Comparison

To evaluate the performance of our prediction, the 4-, 6-, 8-, 10-fold cross validations were carried out. The receiver operating characteristic (ROC) curves for the validations were presented in Fig. 16.3. Since the 4-, 6-, 8-, 10-fold cross validations performances were consistent, it was indicated that the prediction is robust. Since the n-fold cross validations could represent the prediction of new or unknown sites, the results show that our prediction achieves promising performance. For the 10-fold cross validation, the prediction achieved an accuracy of 84.00 %, sensitivity of 44.00 %, specificity of 85.00 %, MCC of 0.1253. For the 8-fold cross validation, the performance is accuracy of 84.18 %, sensitivity of 43.16 %, specificity of 85.27 %, MCC of 0.1244. For the 6-fold cross validation, the prediction achieved an accuracy of 84.03 %, sensitivity of 44.00 %, specificity of 85.07 %, MCC of 0.1258. For the 4-fold cross validation, the performance is accuracy of 84.08 %, sensitivity of 42.37 %, specificity of 85.18 %, MCC of 0.1204.

Since our prediction achieved promising performance, we developed a new predictor of cPhosBac (CKSAAP algorithm to predict Phosphorylation in Bacteria) for prediction of phosphorylation sites in bacteria. Since our dataset was identical with NetPhosBac [16], we compared the performance between them. The result was presented as ROC curves in Fig. 16.4. It is obvious that the performance of cPhosBac is much better than NetPhosBac. Since the authors showed that the performance of NetPhosBac is better than other tools, cPhosBac should be better than other tools.

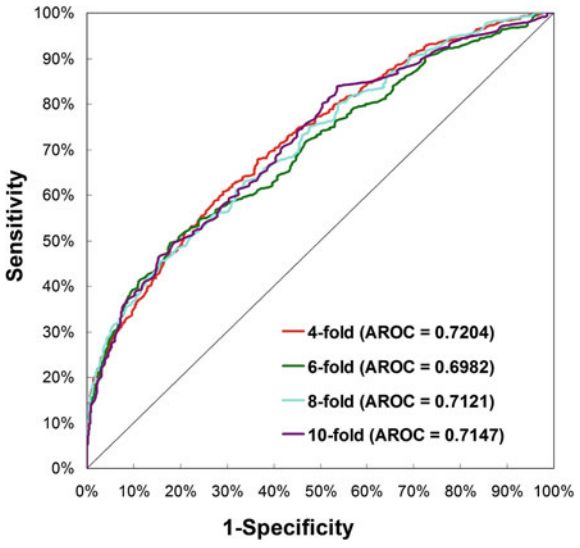


Fig. 16.3 The ROC *curves* of the 4-, 6-, 8-, 10-fold cross validations. The AROC values were calculated and presented

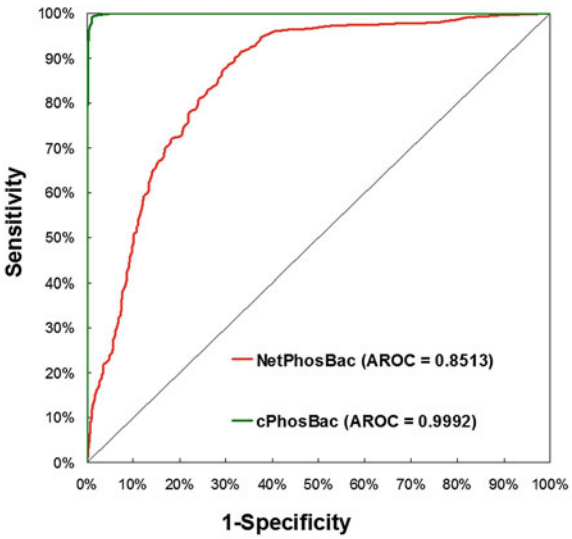


Fig. 16.4 The comparison between cPhosBac and NetPhosBac, the same dataset was submitted to the two predictor for prediction and then the results were compared

16.3.4 The Webserver of CPhosBac

With the promising prediction performance, we developed the webserver of cPhosBac, which is available at http://netalign.ustc.edu.cn/cphosbac/. User could submit one protein sequence or multiple sequences in FASTA format to predict phosphorylation sites as Fig. 16.5a. Then user could click "Submit" button to perform prediction. After the server carried out the prediction, the positive hits will be shown in a tabular format (Fig. 16.5b), which could be downloaded.

16.4 Discussion

As a critical regulatory mechanism, phosphorylation plays important roles in regulation of cellular processes in both eukaryotes and prokaryotes [2–4, 11]. Identification of phosphorylation substrates and their sites is critical to dissecting the molecular mechanisms [7, 11]. Although phosphorylation in eukaryotes is a hot topic, research on serine/threonine phosphorylation in bacteria is much less intensive [11]. Since computational studies could provide helpful information for further experimental investigation, predictor for phosphorylation in bacteria is urgently needed.

In this study, we developed a new predictor of cPhosBac with the CKSAAP method and SVM algorithm. Motif length selection approach is also adopted to optimize the prediction. Although the cPhosBac achieved promising performance, there are a number of improvements could be conducted in the future. Firstly, the prediction performance could be further improved, complex feature selection

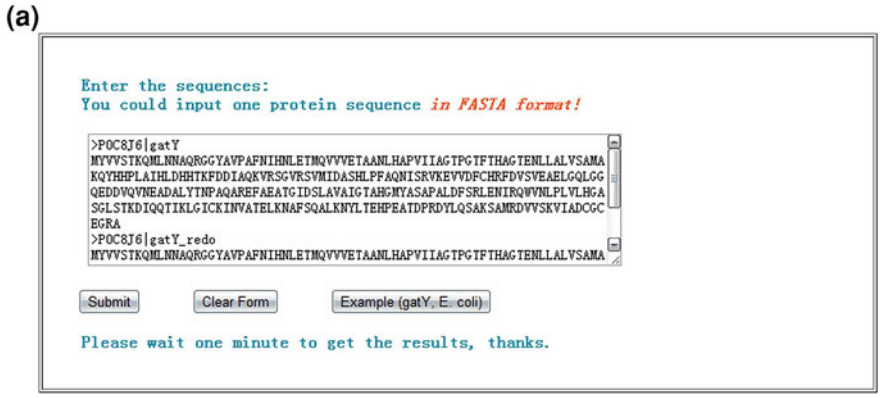


Fig. 16.5 The snapshots of the cPhosBac predictor. **a** One raw sequence and multiple sequences in FASTA format were both allowed for input. **b** The results were presented in tabular format and could be downloaded

could be carried out to provide better prediction, while structural features including secondary structure, solvent-accessible surface areas might provide more selectivity for prediction. Furthermore, since phosphorylation was reversibly regulated by kinase and phosphotases, it should be more valuable to provide kinase-specific prediction.

Taken together, we anticipate that computational prediction, followed by experimental investigation, will help advancing studies of serine/threonine phosphorylation in bacteria.

Acknowledgments This work was supported, in whole or in part, by Provincial Key Research Program of Universities in Anhui (KJ2012A063), Innovation Foundation of USTC for Young Scientists (WK2070000028).

References

1. Raju TN (2000) The nobel chronicles 1992: Fischer EH (b 1920), Krebs EG (b 1918). Lancet, 355:2004
2. Hunter T (2009) Tyrosine phosphorylation: 30 years and counting. Curr Opin Cell Biol 21:140–146
3. Johnson LN (2009) The regulation of protein phosphorylation. Biochem Soc Trans 37:627–641
4. Pawson T, Scott JD (2005) Protein phosphorylation in signaling—50 years and counting. Trends Biochem Sci 30:286–290
5. Cousin C, Derouiche A, Shi L, Pagot Y, Poncet S, Mijakovic I (2013) Protein-serine/threonine/tyrosine kinases in bacterial signaling and regulation. FEMS Microbiol Lett 346(1):11–19
6. Cozzone AJ (1988) Protein-phosphorylation in prokaryotes. Annu Rev Microbiol 42:97–125
7. Ohlsen K, Donat S (2010) The impact of serine/threonine phosphorylation in staphylococcus aureus. Int J Med Microbiol 300:137–141
8. Deutscher J, Francke C, Postma PW (2006) How phosphotransferase system-related protein phosphorylation regulates carbohydrate metabolism in bacteria. Microbiol Mol Biol Rev 70:939–1031
9. Hoch JA (2000) Two-component and phosphorelay signal transduction. Curr Opin Microbiol 3:165–170
10. Song C, Ye M, Liu Z, Cheng H, Jiang X, Han G, Songyang Z, Tan Y, Wang H, Ren J et al (2012) Systematic analysis of protein phosphorylation networks from phosphoproteomic data. Mol cell proteomics: MCP 11:1070–1083
11. Macek B, Mann M, Olsen JV (2009) Global and site-specific quantitative phosphoproteomics: principles and applications. Annu Rev Pharmacol Toxicol 49:199–221
12. Macek B, Gnad F, Soufi B, Kumar C, Olsen JV, Mijakovic I, Mann M (2008) Phosphoproteome analysis of E. coli reveals evolutionary conservation of bacterial Ser/Thr/Tyr phosphorylation. Mol cell proteomics: MCP 7:299–307
13. Macek B, Mijakovic I, Olsen JV, Gnad F, Kumar C, Jensen PR, Mann M (2007) The serine/threonine/tyrosine phosphoproteome of the model bacterium bacillus subtilis. Mol cell proteomics: MCP 6:697–707
14. Ren J, Gao X, Liu Z, Cao J, Ma Q, Xue Y (2011) Computational analysis of phosphoproteomics: progresses and perspectives. Curr Protein Pept Sci 12:591–601
15. Xue Y, Gao X, Cao J, Liu Z, Jin C, Wen L, Yao X, Ren J (2010) A summary of computational resources for protein phosphorylation. Curr Protein Pept Sci 11:485–496
16. Miller ML, Soufi B, Jers C, Blom N, Macek B, Mijakovic I (2009) NetPhosBac—a predictor for Ser/Thr phosphorylation sites in bacterial proteins. Proteomics 9:116–125
17. Chen K, Jiang Y, Du L, Kurgan L (2009) Prediction of integral membrane protein type by collocated hydrophobic amino acid pairs. J Comput Chem 30:163–172
18. Chen K, Kurgan LA, Ruan J (2007) Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs. BMC Struct Biol 7:25
19. Chen YZ, Tang YR, Sheng ZY, Zhang Z (2008) Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs. BMC Bioinformatics 9:101
20. Chen Z, Chen YZ, Wang XF, Wang C, Yan RX, Zhang Z (2011) Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs. PLoS ONE 6:e22930
21. Chen Z, Zhou Y, Song J (1834) Zhang Z (2013) hCKSAAP_UbSite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties. Biochim Biophys Acta 8:1461–1467
22. Vacic V, Iakoucheva LM, Radivojac P (2006) Two sample logo: a graphical representation of the differences between two sets of sequence alignments. Bioinformatics 22:1536–1537

23. Xue Y, Liu Z, Cao J, Ma Q, Gao X, Wang Q, Jin C, Zhou Y, Wen L, Ren J (2011) GPS 2.1: enhanced prediction of kinase-specific phosphorylation sites with an algorithm of motif length selection. Protein Eng Des Sel 24:255–260
24. He Z, Yang C, Guo G, Li N, Yu W (2011) Motif-All: discovering all phosphorylation motifs. BMC Bioinformatics 12(Suppl 1):S22
25. Team RC (2012) R: a Language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria

Chapter 17
Bioinformatics Tools for Discovery and Functional Analysis of Single Nucleotide Polymorphisms

Li Li and Dongqing Wei

Abstract With the high speed DNA sequencing of genome, databases of genome data continue to grow, and the understanding of genetic variation between individuals grows as well. Single nucleotide polymorphisms (SNPs), a main type of genetic variation, are increasingly important resource for understanding the structure and function of the human genome and become a valuable resource for investigating the genetic basis of disease. During the past years, in addition to experimental approaches to characterize specific variants, intense bioinformatics techniques were applied to understand effects of these genetic changes. In the genetics studies, one intends to understand the molecular basis of disease, and computational methods are becoming increasingly important for SNPs selection, prediction and understanding the downstream effects of genetic variation. The review provides systematic information on the available resources and methods for SNPs discovery and analysis. We also report some new results on DNA sequence-based prediction of SNPs in human cytochrome P450, which serves as an example of computational methods to predict and discovery SNPs. Additionally, annotation and prediction of functional SNPs, as well as a comprehensive list of existing tools and online recourses, are reviewed and described.

Keywords Single nucleotide polymorphism · SNPs · Prediction · SNPs discovery · SNPs annotation · Functional analysis

L. Li · D. Wei (✉)
State Key Laboratory of Microbial Metabolism, College of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China
e-mail: dqwei@sjtu.edu.cn

D. Wei et al. (eds.), *Advance in Structural Bioinformatics*, Advances in Experimental Medicine and Biology 827, DOI 10.1007/978-94-017-9245-5_17

17.1 Introduction

With the completion of the Human Genome Project, a large number of subtle variations (polymorphisms) among the population have been found [1–3]. The most abundant (about 90 % of all human genetic variation) type of these variations is single nucleotide polymorphisms (SNPs), with more than 9 million reported in public databases [4, 5]. SNPs, are DNA sequence variations that occur when a single nucleotide (A,T,C, or G) in the genome sequence is altered at least 1 % of the population [6].

SNPs may fall within coding sequences of genes, non-coding regions of genes, or in the intergenic regions between genes. Nonsynonymous SNPs produce either be missense or nonsense change, where a missense change results in a different amino acid, while a nonsense change results in a premature stop codon. Synonymous SNPs (sometimes called a silent mutation) lead to the same polypeptide sequence. SNPs that are not in protein-coding regions may have consequences for gene splicing, transcription factor binding, or the sequence of non-coding RNA (SNPs functional classes are shown in Fig. 17.1).

SNPs have a major impact on how humans respond to and respond to pathogens, chemicals, drugs, vaccines, and other agents. This makes SNPs valuable for biomedical research and for developing pharmaceutical products or medical diagnostics and personalized medicine [7].

Over the past several years a great deal of effort has been devoted to developing accurate, rapid, and cost-effective technologies for SNP analysis, yielding a large number of distinct approaches [8, 9]. Besides experimental methods to characterize specific variants, there has been intense bioinformatics research to understand the molecular effects of these genetic changes. These efforts have focused on two general areas. First, bioinformatics studies have been involved in genomic experimental

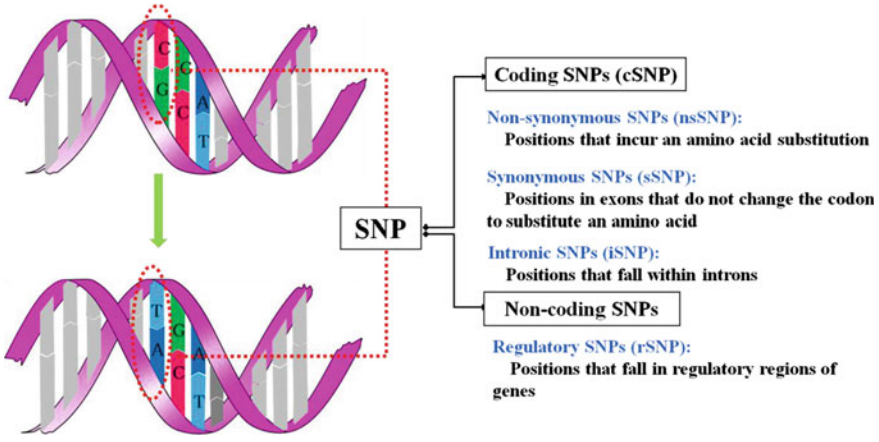


Fig. 17.1 SNP functional classes

assays. Second, researchers annotated genetic variation data and developed statistical methods to predict mutations. These efforts have given rise to many databases, web resources, and tools for prioritizing candidate single nucleotide polymorphisms (SNPs) or hypothesizing the molecular causes of genetic disease. In this paper, bioinformatics methods and tools for SNPs discovery are presented. Further, annotation and prediction of functional SNPs are reviewed and described.

17.2 SNPs Discovery

17.2.1 Bioinformatics Tools and Resources for SNPs Discovery and Analysis

Generally, the discovery and selection of SNPs are carried out by sequencing. SNPs discovery based on the different sites isolating from the sequence, assesses frequency of the error in total numbers of the selected sequences, isolates parologous and then determines genotype.

Bioinformatics techniques play an important role in SNP discovery and analysis. These methods annotate genes that contain SNPs, allow researchers to retrieve data about SNPs based on gene of interest, genetic or physical map location, or expression pattern. PolyPhred, PolyBayes and novoSNP are the widely used bioinformatics tools.

PolyPhred (http://droog.gs.washington.edu/polyphred/) is a program that compares fluorescence-based sequences across traces obtained from different individuals to identify heterozygous sites for single nucleotide substitutions. PolyPhred's functions are integrated with the use of three other programs: Phred, Phrap, and Consed. PolyPhred identifies potential heterozygotes using the base calls and peak information provided by Phred and the sequence alignments provided by Phrap. Potential heterozygotes identified by PolyPhred are marked for rapid inspection using the Consed tool.

PolyBayes (http://genome.wustl.edu/tools/software/polybayes.cgi) is a computer program for the automated analysis of single-nucleotide polymorphism (SNP) discovery in redundant DNA sequences. The software integrates algorithmic solutions to three of the main challenges in sequence-based SNP discovery: Multiple sequence alignment, Paralog identification, and SNP detection. This program produces a list of candidate polymorphic sites, each site with an associated SNP probability score that has been demonstrated to accurately forecast the true positive rate in subsequent validation experiments.

novoSNP (http://www.molgen.ua.ac.be/bioinfo/novosnp/) is a program to find SNPs and small indels in resequencing projects. It takes a reference sequence and a number of sequencing trace files as input, and generates a list of possible variations with a quality score. novoSNP allows easily filter, sort and check the variations found visually and keep track of verifications.

17.3 Predict SNPs by Computational Methods

Because only 1 % of mutations might be expected to confer more than modest individual effects in association studies, the selection of predictive candidate variants for complex disease analyses is formidable. Technologic advances in SNPs discovery have led to massive informational resources that can be difficult to master across disciplines.

Computational methods are successfully employed to predict SNPs function such as whether they are likely to be neutral or deleterious. However, few researches have applied computational approaches to predict SNPs to discover potential SNP sites.

Yan et al. [10] firstly demonstrated SNP prediction and compared machine learning techniques and pattern discovery algorithms for the prediction of SNPs in human. They selected six pattern discovery algorithms (YMF, Projection, Weeder, MotifSampler, AlignACE and ANN-Spec) and two machine learning techniques (Random Forests and K-Nearest Neighbours), then applied them to the DNA sequences. Six methods perform fairly poorly in predicting SNPs, with error rates between 35 and 51 %. Machine learning algorithms perform better than the pattern discovery methods by ~6 %: the average prediction error for Random Forests and KNN is about 44 %, while the pattern discovery methods have around 50 % prediction error.

Li et al. improved the prediction performance using support vector machine (SVM) model based on the CYP450 SNP sites and the physical and chemical properties(polarity, volume, hydropathy, charge, flexibility, isoelectric point, refractivity) of the amino acids on protein flanking sequence. They demonstrate

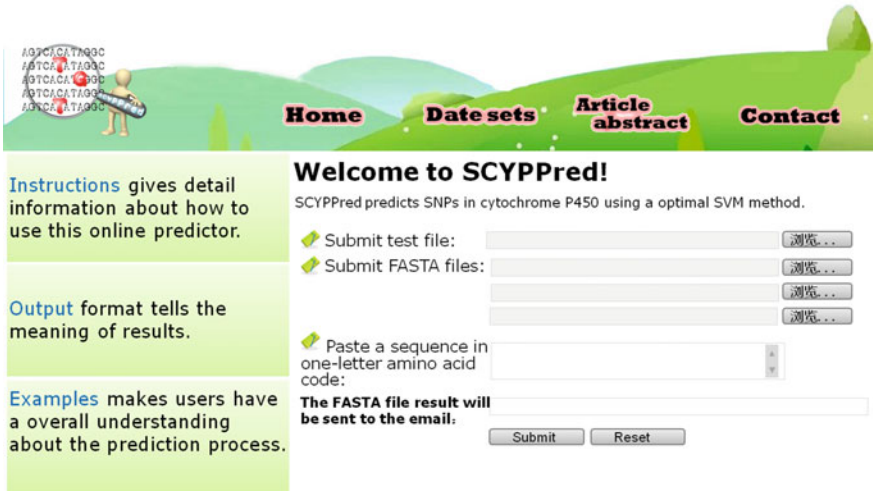


Fig. 17.2 A semi-screenshot of the home page of the web-server SCYPPred

the accuracy of this method achieves 65 % and provide SCYPPred, an online tool freely accessible at http://snppred.sjtu.edu.cn (the homepage is shown in Fig. 17.2) to predict SNPs.

Predicting SNPs leads significant insights into the structural determinants of DNA stability and species evolution. As such, these preliminary results strongly support additional work to improve the ability to predict SNPs.

17.4 An Example of Computational Methods to Predict SNPs–DNA Sequence-Based Prediction of SNPs in Human Cytochrome P450

SNPs prediction can be cast into a binary classification task at the nucleotide level, namely predicting for each nucleotide site in the human cytochrome P450, whether it has the latent to be a SNP.

17.4.1 Datasets

We focused on CYP2 subfamily as they involved in almost 80 % drug metabolism and exhibited a large degree of inter- and intra-species variability in regulation and catalytic activities. Non-synonymous SNPs (nsSNP) change the protein expression, phenotype and directly cause diseases, so, cDNA sequences (only include nsSNPs) were used to construct the datasets. cDNA sequences of human CYP2 subfamily were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/nucleotide/) and 282 SNPs in the human CYP450 SNPs were sourced from Human Cytochrome P450 Allele Nomenclature Committee (http://www.cypalleles.ki.se). Besides, we collected 18,629 remained sites as non-SNPs.

17.4.2 Extracted Features and Encoded the Vector

Our aim was to employ least attributes were employed while achieve higher prediction accuracy. We exploited a set of features only from the DNA primary sequence (summarized in Table 17.1), not require the complex structure information. In the following we give a detailed description of the features.

D1 encode the types of the neighbouring nucleotide: adenine (A), thymine (T), guanine (G) and cytosine (C) were separately converted into 1, 2, 3 and 4.

Table 17.1 Sequence-based features

Feature	Description
D_1	Nucleotide composition
D_2	Neighbouring SNPs
D_3	CpG dinucleotides

D2 encode the SNPs near the candidate site.

For a pair of SNP alleles, linkage disequilibrium (LD) is a measure of deviation from random association (i.e., no recombination). In recent years, LD analysis has become a topic of great interest in the field of SNP association studies and an effective approach to connect structural SNPs to phenotypes [11–14]. Because of the close relation between LD and SNPs, we have taken the existent neighbouring SNPs into account in our research.

D3 encode the occurrence of CpG dinucleotides at the candidate site.

The relation between the CpG dinucleotides and SNPs had been pointed out in papers [15–19], which showed SNPs occurs at a high rate at CpG dinucleotides due to the frequent methylation of CpG and the deamination of methylated cytosine to thymine. This feature was firstly proposed to predict SNPs; and we believed that the special sequence composition is related to SNPs.

For classification, we used window size of 25 (adjacent 25 residues upstream and downstream from the SNP) for forming a 102-dimensional vector to describe the SNP information. In this vector, the 1–51 dimensional characters were the converted numerical vales of nucleotide types including the target site and flanking sequence: adenine (A), thymine (T), guanine (G) and uracil (C) were separately converted into 1, 2, 3 and 4 (the detail was showed in Fig. 17.3a). The 52–101 dimensional characters were associated with the Neighbouring SNPs information. SNPs and non-SNPs were separately encoded into 1 and 0 (showed in Fig. 17.3b). The 102th dimensional character was the occurrence of CpG dinucleotides at the candidate site. If the candidate site is cytosine (C) and the right neighbouring nucleotide is guanine (G), the candidate site is encoded to 1. In another cases, the candidate site is encoded to 0 (showed in Fig. 17.3c).

17.4.3 Group Training

Only 282 out of 18,911 nucleotide sites are SNPs, hence, the dataset is extremely unbalanced with a ratio between positive and negative examples of about 1:66, which usually lead a learning bias to the majority class. In order to deal with this, we have resorted to group training approach [20, 21]. According to this method, the primary training datasets (282 SNPs and 18,629 non-SNPs) was divided into

Fig. 17.3 Features definition

46 subsets at random; each of them has 400 SNPs and 200 non-SNPs (as shown in Fig. 17.4). 46 models are generated after group training (Additional file 1), and the best 23 models (showed in Table 17.2) were selected and integrated to give the final prediction with the following decision criterion: $Y = \sum_{i=1}^{15} Y_i$. If $Y > 0$, it is predicted as SNP; in contrast, if $Y < 0$, it is considered as non-SNP.

Additionally, confusion matrix, a visualization tool typically used in supervised learning, was applied to evaluate the computational models. As shown in Table 17.3, confusion matrix is a table with the true values in rows and the predicted ones in columns. The diagonal elements represent correctly classified cases while the cross-diagonal elements stand for misclassified cases. So, with the confusion matrix, we could compute the sensitivity, specificity, as well as the overall accuracy of our predicting results (Table 17.4).

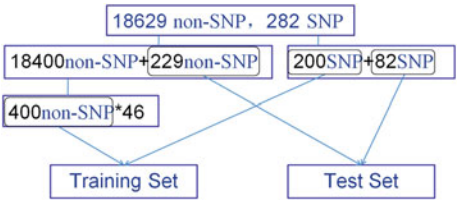


Fig. 17.4 Group prediction

Table 17.2 The performance of selected 23 groups

Group	Performance (%)			
	Specificity	Sensitivity	Accuracy	AUC
1	75.1092	53.6585	69.4534	64.3839
2	79.4760	48.7805	71.3826	64.1282
3	76.4192	45.1220	68.1672	60.7706
4	85.1528	45.1220	74.5981	65.1374
5	79.0393	45.1220	70.0965	62.0806
6	82.9694	43.9024	72.6688	63.4359
7	81.2227	43.9024	71.3826	62.5626
8	83.8428	43.9024	73.3119	63.8726
9	89.9563	42.6829	77.4920	66.3196
10	82.0961	40.2439	71.0611	61.1700
11	88.2096	40.2439	75.5627	64.2268
12	90.8297	39.0244	77.1704	64.9270
13	86.8996	39.0244	74.2765	62.9620
14	82.5328	39.0244	71.0611	60.7786
15	89.5197	37.8049	75.8842	63.6623
16	88.2096	36.5854	74.5981	62.3975
17	86.8996	36.5854	73.6334	61.7425
18	90.8297	36.5854	76.5273	63.7075
19	86.4629	34.1463	72.6688	60.3046
20	88.2096	34.1463	73.9550	61.1780
21	93.0131	31.7073	76.8489	62.3602
22	85.1528	30.4878	70.7395	57.8203
23	89.9563	30.4878	74.2765	60.2221

Table 17.3 Confusion matrix of our datasets

Confusion matrix		Predicted	
		Non-SNP	SNP
Actual	Non-SNP	196	33
	SNP	43	39

$$Specificity = \frac{TN}{TN+FP} = \frac{196}{196+33} = 85.5895\,\%$$
$$Sensitivity = \frac{TP}{TP+FN} = \frac{39}{39+43} = 47.561\,\%$$
$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} = \frac{39+196}{39+36+196+43} = 75.5627\,\%$$
$$ACU = 0.5 * (Specificity + Sensitivity) = 66.5752\,\%$$

Table 17.4 The performance of the integrated model

	Performance (%)			
	Specificity	Sensitivity	Accuracy	AUC
Integrated model	85.5895	47.561	75.5627	66.5752

17.5 Comparison with Other Methods

We compared our results with the most recent methods for both DNA sequence-based and protein sequence-based prediction.

Table 17.5 reports comparisons with other DNA sequence-based methods, including six pattern discovery algorithms(Align ACE, ANN, Motif Sampler, YMF, Weeder and Projection) and two machine learning techniques(Random Forests and KNNs) [10]. All the eight methods only encode information on SNPs with 25 base-pairs of 5′ and 25 base-pairs of 3′ flanking sequence. Our model reached an SNP prediction accuracy of 75.56 %, which increased by 18 % over the reported best accuracy of 57 %. The results indicate that features and SVM employed in our methods allow improving performance on prediction.

Table 17.6 shows comparisons with protein sequence-based method (SCYP-Pred) recently developed in our previous work. Previous results suggested that protein sequence managed to predict SNPs with high accuracy. In this paper, our method successfully improved the accuracy (from 66.3 to 75.6 %), the specificity (from 66.3 to 85.6 %) and the AUC value (from 65.4 to 66.6 %). These improvements we achieved show that DNA sequence can indeed be effectively used in predictions. Nevertheless, further studies are needed to make fully exploit.

Table 17.5 Comparison with state-of-art DNA sequence-based approaches

Other DNA sequence-based approach	Accuracy (%)
Align ACE	50.4
ANN	50.2
Motif sampler	50.1
YMF	51.0
Weeder	50.0
Projection	49.7
Random forests	57.0
KNNs	54.0
Our method	75.6

Table 17.6 Comparison with protein sequence-based approaches

	SCYPPred (protein sequence-based approach) (%)	Our method (protein sequence-based approach) (%)
Accuracy	66.3	75.6
Specificity	66.3	85.6
AUC	65.4	66.6

17.6 SNPs Functional Analysis

17.6.1 Functional Annotation of SNPs (Web Resources Are Reviewed in Table 17.7)

Coding SNPs are located in the exons of genes, in which three nucleotides "code" for the amino acids that are used to build proteins. SNPs may change (nonsynonymous) or not change (synonymous) the amino acid. There are several ways a nonsynonymous SNP (nsSNP) can affect gene product function. The most probable effect is a partial or complete loss of function of the mutated gene product. Less likely, SNPs in an exonic portion of a splice junction will make a noncoding intron be retained or make the exon be skipped, which may result in the loss of some amino acids or an unstable messenger RNA (mRNA) transcript. SNPs occur within an exonic splicing enhancer (ESE), where various components of the splicing machinery localize to splice the pre-mRNA [22], may result in intron retention or exon skipping.

One of the most common annotations of a SNP is identification of its location [23]. GoldenPath, assembled the UCSC Genome Browser and genome, is an excellent resource for visualisation of SNP locations and other genome annotations [24]. The other primary genome resource is Ensembl [25]. which provides visualise variation in and around genes, and their data annotations are of high quality. SNPper, focuses on SNP selection for genetic studies, is another powerful resource for SNP analysis [26, 27].

Functional information is beginning to integrate in many locus databases, such as protein structure, into their annotation sets. The NCBI databases, such as dbSNP and OMIM [28], and Ensembl [29] provide visualisation access and some annotations related to function, based on experiment. For protein structural annotations of variation in dbSNP and Swiss-Prot, MutDB58 was developed to annotate known variation data with information relevant to identifying the molecular effects of a mutation or polymorphism. Effects on protein structure can be very subtle and may not be obvious; hence, visualizing protein structure is useful to biochemistry experts. Several web-based databases annotate protein structure and query services, including Large Scale human SNP annotation (LS-SNP http://modbase.compbio.ucsf.edu/LS-SNP/) [30], SNPs3D (http://snps3d.org/) [31], MutDB (http://www.mutdb.org/) [32], and PolyDoms (http://polydoms.cchmc.org/polydoms/) [33].

Table 17.7 Useful web resources for SNPs functional annotation

Web resources	URL	Comments
dbSNP [96, 97]	http://www.ncbi.nlm.nih.gov/SNP/	Archive for genetic variation, including SNP data
Ensembl [98]	http://www.ensembl.org/	Produce and maintain automatic annotation on SNPs
UCSC [99]	http://genome.ucsc.edu/	Genome database, provides genome browser, gene sorter, blat search function
JSNP [100]	http://snp.ims.u-tokyo.ac.jp/	A repository of Japanese SNP data
HGVBase [101]	http://hgvbase.cgb.ki.se/	Public genotype phenotype database
HGMD [102]	http://www.hgmd.org/	Mutation database with many annotations
Swiss-Prot [103]	http://us.expasy.org/	Protein database with extensive variant annotations
CGAP-GAI [104]	http://cgap.nci.nih.gov/	Cancer Gene Anatomy Project at the National Cancer Institute, a tool for viewing candidate SNPs in the context of EST assemblies
SNPper [105]	http://snpper.chip.org/	Novel software for SNP analysis, a web-based application to automate the tasks of extracting SNPs from public databases
BioPerl [106]	http://www.bioperl.org/	A programming application program interface (API) for bioinformatics analysis, open-source software
Genewindow [107]	http://www.genewindow.nci.nih.gov/	Interactive tool for visualization of variation, represent genomic variation intuitively
LS-SNP [30, 108]	http://modbase.compbio.ucsf.edu/LS-SNP/	Large scale nsSNP annotation software
MutDB [109]	http://mutdb.org/	Annotate genomic variants with data that assists in functional annotation, contains protein structure annotations and comparative genomic annotations
PolyDoms [33]	http://polydoms.cchmc.org/	Genome database for the nsSNPs, a whole genome database for the identification nsSNP
PolyMAPr [73]	http://pharmacogenomics.wustl.edu/	Programs for polymorphism database mining, annotation, and functional analysis
PromoLign [110]	http://polly.wustl.edu/promolign/main.html	A database for upstream region analysis and SNPs
PupaSuite [111]	http://pupasuite.bioinfo.cipf.es/	A tool for the selection of relevant SNPs within a gene

(continued)

Table 17.7 (continued)

Web resources	URL	Comments
SNP function portal [75]	http://brainarray.mbni.med.umich.edu/Brainarray/Database/SearchSNP/snpfunc.asp	A web database for exploring the function implication of SNP alleles
SNP@Promoter [112]	http://variome.kobic.re.kr/SNPatPromoter/	A database of human SNPs within the putative promoter regions
SNP3D [113]	http://www.snps3d.org/	Annotations of structure, systems biology, evolution and alternative splicing

All the web resources in this table are free charged

LS-SNP stands out as a useful resource because it provides annotations of nsSNPs that have been mapped to homology models from the MODBASE (http://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi) [34] dataset.

17.6.2 Prediction of the Functional SNPs

17.6.2.1 Predicting Functional SNPs in Coding Sequence (Web Resources Are Reviewed in Table 17.8)

According to the wide availability of functional data on proteins and the evidence that regulatory and coding SNPs are most likely to affect disease [35–39], much effort has been invested in predicting the function of non-synonymous mutations. The key point is identifying whether a particular mutation will be tolerated [40].

Researchers have taken several approaches to predict the function of nsSNPs. Almost all methods use categories, discrete or continuous valued features to predict a deleterious mutation. Features usually based on sequence, structure, or known function. To classify whether a mutation will be tolerated, a training set is usually constructed of mutations known to be deleterious. Some researches use experimental amino acid substitutions as training sets [41–44], for example, saturation mutagenesis experiments where mutation severity is determined in activity assays [41, 43, 45–47]. Others use substitutions based on disease-associated human alleles [30, 31, 48, 49], such as multiple sequence alignments where tolerance to mutation is derived from evolutionary analyses of sequence positions [50], or known deleterious human mutations [45, 51].

The earliest studies using sequence conservative properties to analysis mutations by a BLOSUM62 matrix [52], which does not take into account the sequence or structural context of the mutation. Further efforts were employed to include position-specific conservation estimates and protein structural information [47, 53, 54]. Sorting Intolerant From Tolerant (SIFT) [55], based on a position-specific scoring matrix (PSSM), find that 25 % of nsSNPs in dbSNP are likely to affect

Table 17.8 Web resources and tools for predicting the function of SNPs

Web resources	URL	Comments
SIFT [55]	http://blocks.fhcrc.org/sift/SIFT.html	Online tool for sequence-based annotation of mutations, nonsynonymous amino acid SNP effect
PolyPhen [50]	http://www.bork.embl-heidelberg.de/PolyPhen/	Server for functional analysis of mutations, nonsynonymous amino acid SNP effect
SNP3D [31]	http://www.snps3d.org/	Annotations of structure, systems biology, evolution and alternative splicing
SNPeffect [114]	http://snpeffect.vib.be/index.php	Annotations based on structure, catalysis and cellular process, nonsynonymous amino acid SNP effect
PupasView [115]	http://pupasview.bioinfo.ochoa.fib.es/	Nonsynonymous amino acid SNP effect, exonic splicing enhancer SNP identification, splice site SNP identification, intronic/upstream, downstream regulatory region, or undiscovered exon SNP identification
SNPselector [116]	http://primer.duhs.duke.edu/	Haplotype-tagging SNPs, splice site SNP identification, intronic/upstream, downstream regulatory region, or undiscovered exon SNP identification
TAMAL [117]	http://neoref.ils.unc.edu/tamal	Haplotype-tagging SNPs, nonsynonymous amino acid SNP effect, splice site SNP identification, intronic/upstream, downstream regulatory region, or undiscovered exon SNP identification
PicSNP [118]	http://plaza.umin.ac.jp/hchang/picsnp/	Catalog of nonsynonymous SNPs in the Human Genome, gene-centric mutation annotation,
TopoSNP [119]	http://gila.bioengr.uic.edu/snp/toposnp/	Protein structural annotations of SNPs
MutDB [109]	http://www.mutdb.org/	A topographic database of non-synonymous SNPs, protein structural information of SNPs
PARSESNP [120]	http://www.proweb.org/parsesnp/	Predict the locations, effects of SNPs and severity of missense changes
LS-SNP [30, 108]	http://alto.compbio.ucsf.edu/LS-SNP/	A genomic scale software pipeline to annotate nsSNPs
PMUT [48]	http://mmb2.pcb.ub.es:8080/PMut/	Annotation of pathological mutations on proteins
ESEfinder [121]	http://rulai.cshl.edu/tools/ESE/	Rapid analysis of exon sequences to identify putative ESEs
RESCUE-ESE [122]	http://genes.mit.edu/burgelab/rescue-ese/	Identify candidate exonic splicing enhancers in vertebrate exons
PipMaker [123]	http://pipmaker.bx.psu.edu/cgi-bin/pipmaker	Identify conserved segments and for producing informative, high-resolution displays of the resulting alignments

(continued)

Table 17.8 (continued)

Web resources	URL	Comments
Vista [124]	http://genome.lbl.gov/vista/index.shtml	Visualizing global DNA sequence alignments of arbitrary length. Intronic/upstream, downstream regulatory region or undiscovered exon SNP identification
ECR browser [125]	http://ecrbrowser.dcode.org/	A tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. Intronic/upstream, downstream regulatory region or undiscovered exon SNP identification
rVISTA [126]	http://rvista.dcode.org	Evolutionary analysis of transcription factor binding sites. Intronic/upstream, downstream regulatory region SNP identification
Promolign [110]	http://polly.wustl.edu/promolign/main.html	A database for upstream region analysis and SNPs. Upstream (promoter) region SNP identification
MAPPER[a] [127]	http://bio.chip.org/mapper/mapper-top	A platform for the computational identification of transcription factor binding sites (TFBSs) in multiple genomes. Upstream (promoter) region SNP identification
Match* [128]	http://www.gene-regulation.com/pub/programs.html#match	Search potential binding sites for transcription factors (TF binding sites) nucleotide sequences. Intronic/upstream, downstream regulatory region SNP identification Intronic/upstream, downstream regulatory region SNP identification
BLAST [129]	http://www.ba.itb.cnr.it/BIG/Blast/BlastUTR.html	Sequence analysis tool, UTR regulatory region SNP identification

All the web resources and tools in this table are free charged

[a] The online tool needs register before using

protein function. SIFT has recently been applied to SNPs in both DNA repair genes and separately to BRCA1 [56].

Sunyaev et al. [57] included protein structural features in their study to classify and survey non-synonymous SNPs. They compared disease-associated mutations in orthologous genes and human cSNPs. Both protein structural information (such as solvent accessibility, location within beta strands or active sites) and evolutionary information (evolutionary conservation) were taken into account to assess local functionality for a given position. The authors found that approximately 70 % of disease-associated mutations were in protein structural sites described above and most likely to affect protein function.

Saunders and Baker [45] assessed different features for prediction of intolerant mutations by following the analysis of Sunyaev et al. and Chasman and Adams. Decision trees and a linear logistic regression were applied to find that a protein structure-derived solvent accessibility term (C density) and an evolutionary term

derived from a PSSM matrix (SIFT) were the most accurate terms for prediction. They found that in both human alleles and in vitro cases, the SIFT and Cdensity terms classified the best, and that the normalised B-factor and Sunyaev derived structural rules did not improve classification accuracy when incorporated with the former terms in a combined analysis. Cai et al. applied a Bayesian method for predicting disease-associated SNPs and obtained relatively low false positive error rates, in exchange for a relatively high false negative rate [58].

Among recent researches, classification tools based on SVMs or decision trees and the best features for classification based on structural and evolutionary properties showed better performance. Structurally, solvent accessibility has consistently been shown to be important in determining whether a mutation will be tolerated [41, 45, 59, 60]. Evolutionarily, nontolerated mutations inferred using a PSSM matrix are generally better than using positional conservation approaches [45].

The widely accepted and easy to use tools and web resources provided for functional annotation of variation are PolyPhen55 and SIFT.50. SIFT uses conservation in a multiple sequence alignment as its sole feature, and experimental mutations as its training data. PolyPhen based on human allele data and includes protein structure data as well as other features. More recently, other methods have been developed and deployed online, including SNPs3D [31], LS-SNP [30], PMut (http://mmb2.pcb.ub.es:8080/PMut/) [48], the SAP prediction method (http://sapred.cbi.pku.edu.cn/) [61], Screening for Nonacceptable Polymorphisms (SNAP, http://cubic.bioc.columbia.edu/services/SNAP/) [62], Predicting the Amino Acid Replacement Probability (Parepro http://www.mobioinfor.cn/parepro/) [63] and Protein Analysis Through Evolutionary Relationships (PANTHER, http://www.pantherdb.org/) [64]. LS-SNP and the method SNAP are two more recent additions to this library of tools that are the SVM utilized and have web sites available for prediction.

Synonymous variation has been shown to be functional as well. Mutations can occur in splicing factor binding sites such as intron–exon splice sites [65, 66], especially the exonic splicing enhancers (ESEs), which are short sequences that occur in exons, and encourage exon recognition by the cell's splicing machine. Mutation in ESEs may affect mRNA splicing and causing exon skipping [22]. Furthermore, it has been shown that mutations that affect mRNA splicing are the most common type of mutations in neurofibromatosis type 1 [67]. A recent review highlights the importance of splicing function on genetic disease [68]. Disease associated variation that disrupts ESEs were found in the breast cancer-associated genes BRCA1 [69, 70] and BRCA2 [71].

Fairbrother et al. reported the original approach to the analysis of variation that disrupts ESEs. They aligned SNPs that are in predicted ESE sites and selected out nearly 20 % of the polymorphisms that are most notable near splicing sites [72]. Nowadays, several tools available for annotation of splicing effects caused by synonymous variation, including Polymorphism Mining and Annotation Programs (PolyMAPr http://pharmacogenomics.wustl.edu/) [73], PupaSuite (http://pupasuite.bioinfo.cipf.es/) [74] and the SNP Function Portal [75]. Motif or

position specific scoring matrix (PSSM) are generally used in these resources to predict of splicing signals or known sites in humans or comparative sites in model organisms such as ESEFinder [76].

17.6.2.2 Predicting Functional SNPs in Non-coding Sequence

SNPs affecting transcription processing (e.g., splice site recognition) may also occur within the intronic portion of a splice site or within the 5′- or 3′-UTR of an exon. Introns are important, although they do not code for proteins, because they contain sequences that dictate other attributes of how the protein is made. Introns, exonic UTRs, and noncoding regions upstream and downstream of genes are known to contain various regulatory elements important for transcription and translation [77–79]. Continued advancements in the genome have made it clear that noncoding regions are far from unimportant.

It is difficult to annotate and predict non-coding SNPs, so only a few studies that have attempted to examine the relationship between gene expression and variation. Most of the projects combined computational methods with experimental analysis of gene expression levels using microarrays. Cowles et al. focused their studies on the expression levels in an F1 hybrid mouse and addressed the problem of removing trans-acting factors [80]. Wittkopp et al. found that most of the genes with significant expression level differences had cis-regulatory differences by comparing differences in gene expression between closely related Drosophila species [81]. Hoogendoorn et al. have screened different promoter variants to identify haplotypes that are likely to affect gene expression [82–84]. Later, Buckland et al. found that approximately 18 % of the variants altered expression levels with the ability of 20 variant promoters on chromosome 21 [85].

Very little bioinformatics research has been performed to build predictors of variation that is likely to affect gene expression levels. Currently, computational way of roughly estimating whether a variant will affect expression levels is identifying whether the polymorphism sites in a known regulatory motif. Consite is a method that predicts transcription factor binding sites [86, 87]. PupaSNP Finder is a tool for identifying SNPs that could have an effect on transcription [74, 88]. Using Ensembl, the authors map SNPs in dbSNP to transcription factor binding sites, intron/exon border consensus sequences, ESE sequences and variations that are nonsynonymous. rSNP_Guide is another resource which contains annotations of SNPs based on potential effects to regulation [89, 90].

Accurate prediction of genetic regulatory networks appears to be in its infancy because transacting regulation appears to be more complicated [91–94]. Recently, sequence based prediction of expression was shown to be feasible in Drosophila employing the sequences of transcription factor binding sites [95]. However, this approach was failed to work for changes as small as a SNP.

17.7 Conclusion

Genetic variations are exponentially increasing volume of sequence data in public and private databases. As databases of genome data continue to grow, our understanding of the genome grows as well. SNPs, the widespread genetic changes in the genome, are important markers in many studies that link sequence variations to phenotypic changes. SNPs provide opportunities to find detrimental mutations which related to a variety of diseases, and serve as a powerful tool for the discovery of high-risk people, disease gene identification, drug design, and fundamental biology research.

Bioinformatics play an important role in SNPs discovery and analysis. Powerful approaches for investigating the molecular basis of disease were provided. And we addressed human SNPs prediction from DNA sequence by developing an effective approach with the extract sequence information. Besides, tools including both computational procedures for data analysis as well as methods to efficiently store and retrieve information were developed.

In future, with the progress of sequencing projects and increase SNP- related research, more data will be gained for analysis and integrate. Bioinformatics, which provides scientists with access to the genomic information, will exhibits tremendous potential for playing a major role in the SNP discovery and functional analysis and become an integral part of genetics. Many sophisticated, extremely valuable, easily-accessible and user-friendly tools will be developed. At the same time, further insights and difficult data management still challenges bioinformatics scientists. The integration of bioinformatics tools and recourses for further genomic biomedical research are urgently needed.

References

1. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409(6822):928–933
2. Collins FS, Brooks LD, Chakravarti A (1998) A DNA polymorphism discovery resource for research on human genetic variation. Genome Res 8(12):1229–1231
3. Collins FS, Morgan M, Patrinos A (2003) The human genome project: lessons from large-scale biology. Science 300(5617):286–290
4. International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437(7063):1299–1320
5. Rocha D, Gut I, Jeffreys AJ, Kwok PY, Brookes AJ, Chanock SJ (2006) Seventh international meeting on single nucleotide polymorphism and complex genome analysis: ‘ever bigger scans and an increasingly variable genome’. Hum Genet 119(4):451–456

6. Brookes AJ (1999) The essence of SNPs. Gene 234(2):177–186
7. Mullikin JC, Hunt SE, Cole CG, Mortimore BJ, Rice CM, Burton J, Matthews LH, Pavitt R, Plumb RW, Sims SK, Ainscough RM, Attwood J, Bailey JM, Barlow K, Bruskiewich RM, Butcher PN, Carter NP, Chen Y, Clee CM, Coggill PC, Davies J, Davies RM, Dawson E, Francis MD, Joy AA, Lamble RG, Langford CF, Macarthy J, Mall V, Moreland A, Overton-Larty EK, Ross MT, Smith LC, Steward CA, Sulston JE, Tinsley EJ, Turney KJ, Willey DL, Wilson GD, McMurray AA, Dunham I, Rogers J, Bentley DR (2000) An SNP map of human chromosome 22. Nature 407(6803):516–520
8. Mooney S (2005) Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. Brief Bioinform 6(1):44–56
9. Clifford RJ, Edmonson MN, Nguyen C, Scherpbier T, Hu Y, Buetow KH (2004) Bioinformatics tools for single nucleotide polymorphism discovery and analysis. Ann NY Acad Sci 1020:101–109
10. Yan R, Boutros PC, Jurisica I, Penn LZ (2007) Comparison of machine learning and pattern discovery algorithms for the prediction of human single nucleotide polymorphisms. In: 2007 IEEE international conference on granular computing, pp 452–457
11. Karinen S, Heikkinen T, Nevanlinna H, Hautaniemi S (2011) Data integration workflow for search of disease driving genes and genetic variants. PLoS One 6(4):e18636
12. Takeuchi F, Kobayashi S, Ogihara T, Fujioka A, Kato N (2011) Detection of common single nucleotide polymorphisms synthesizing quantitative trait association of rarer causal variants. Genome Res 21(7):1122–1130
13. Yaspan BL, Bush WS, Torstenson ES, Ma D, Pericak-Vance MA, Ritchie MD, Sutcliffe JS, Haines JL (2011) Genetic analysis of biological pathway data through genomic randomization. Hum Genet 129(5):563–571
14. Yuan X, Zhang J, Wang Y (2011) Simulating linkage disequilibrium structures in a human population for SNP association studies. Biochem Genet 49(5–6):395–409
15. Shoemaker R, Deng J, Wang W (2010) Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. Genome Res 20:884–889
16. Kaminsky ZA, Tang T, Wang SC, Ptak C, Oh GH, Wong AH, Feldcamp LA, Virtanen C, Halfvarson J, Tysk C, McRae AF, Visscher PM, Montgomery GW, Gottesman II, Martin NG, Petronis A (2009) DNA methylation profiles in monozygotic and dizygotic twins. Nat Genet 41(2):240–245
17. Zhao Z, Zhang F (2006) Sequence context analysis in the mouse genome: single nucleotide polymorphisms and CpG island sequences. Genomics 87(1):68–74
18. Zhao Z, Zhang F (2006) Sequence context analysis of 8.2 million single nucleotide polymorphisms in the human genome. Gene 366(2):316–324
19. Xie H, Wang M, Bischof J, Bonaldo Mde F, Soares MB (2009) SNP-based prediction of the human germ cell methylation landscape. Genomics 93(5):434–440
20. Derya Ubeyli E (2008) Analysis of EEG signals by combining eigenvector methods and multiclass support vector machines. Comput Biol Med 38(1):14–22
21. Keerthi SS, Lin CJ (2003) Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Comput 15(7):1667–1689
22. Blencowe BJ (2000) Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends Biochem Sci 25(3):106–110
23. Laskowski RA, Thornton JM (2008) Understanding the molecular machinery of genetics through 3D structures. Nat Rev Genet 9(2):141–151
24. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome Res 12(6):996–1006
25. URL: http://www.ensembl.org/. Access on 17 May 2011
26. URL: http://snpper.chip.org/. Access on 17 May 2011
27. Riva A, Kohane IS (2004) A SNP-centric database for the investigation of the human genome. BMC Bioinform 5:33

28. Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA, Wagner L, Rapp BA (2001) Database resources of the national center for biotechnology information. Nucleic Acids Res 29(1):11–16
29. Hammond MP, Birney E (2004) Genome information resources—developments at Ensembl. Trends Genet 20(6):268–272
30. Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, Eswar N, Haussler D, Sali A (2005) LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics 21(12):2814–2820
31. Yue P, Melamud E, Moult J (2006) SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinform 7:166
32. Singh A, Olowoyeye A, Baenziger PH, Dantzer J, Kann MG, Radivojac P, Heiland R, Mooney SD (2008) MutDB: update on development of tools for the biochemical analysis of genetic variation. Nucleic Acids Res 36(Database issue):D815–D819
33. Jegga AG, Gowrisankar S, Chen J, Aronow BJ (2007) PolyDoms: a whole genome database for the identification of non-synonymous coding SNPs with the potential to impact disease. Nucleic Acids Res 35(Database issue):D700–D706
34. Pieper U, Eswar N, Webb BM, Eramian D, Kelly L, Barkan DT, Carter H, Mankoo P, Karchin R, Marti-Renom MA, Davis FP, Sali A (2009) MODBASE, a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 37(Database issue):D347–D354
35. Collins FS, Green ED, Guttmacher AE, Guyer MS (2003) A vision for the future of genomics research. Nature 422(6934):835–847
36. Timofeeva MN, Kropp S, Sauter W, Beckmann L, Rosenberger A, Illig T, Jager B, Mittelstrass K, Dienemann H, Bartsch H, Bickeboller H, Chang-Claude JC, Risch A, Wichmann HE (2009) CYP450 polymorphisms as risk factors for early-onset lung cancer: gender-specific differences. Carcinogenesis 30(7):1161–1169
37. Li Y, Bezemer ID, Rowland CM, Tong CH, Arellano AR, Catanese JJ, Devlin JJ, Reitsma PH, Bare LA, Rosendaal FR (2009) Genetic variants associated with deep vein thrombosis: the F11 locus. J Thromb Haemost 7(11):1802–1808
38. Konstantou J, Ioannou PC, Christopoulos TK (2007) Genotyping of single nucleotide polymorphisms by primer extension reaction and a dual-analyte bio/chemiluminometric assay. Anal Bioanal Chem 388(8):1747–1754
39. Bickeboller H, Goddard KA, Igo RP Jr, Kraft P, Lozano JP, Pankratz N, Balavarca Y, Bardel C, Charoen P, Croiseau P, Guo CY, Joo J, Kohler K, Madsen A, Malzahn D, Monsees G, Sohns M, Ye Z (2007) Issues in association mapping with high-density SNP data and diverse family structures. Genet Epidemiol 31(Suppl 1):S22–S33
40. Bowie JU, Reidhaar-Olson JF, Lim WA, Sauer RT (1990) Deciphering the message in protein sequences: tolerance to amino acid substitutions. Science 247(4948):1306–1310
41. Chasman D, Adams RM (2001) Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. J Mol Biol 307(2):683–706
42. Pirooznia M, Yang JY, Yang MQ, Deng Y (2008) A comparative study of different machine learning methods on microarray gene expression data. BMC Genom 9(Suppl 1):S13
43. Krishnan VG, Westhead DR (2003) A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. Bioinformatics 19(17):2199–2209
44. Care MA, Needham CJ, Bulpitt AJ, Westhead DR (2007) Deleterious SNP prediction: be mindful of your training data! Bioinformatics 23(6):664–672
45. Saunders CT, Baker D (2002) Evaluation of structural and evolutionary contributions to deleterious mutation prediction. J Mol Biol 322(4):891–901
46. Cai Z, Tsung EF, Marinescu VD, Ramoni MF, Riva A, Kohane IS (2004) Bayesian approach to discovering pathogenic SNPs in conserved protein domains. Hum Mutat 24(2):178–184

47. Ng PC, Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res 11(5):863–874
48. Ferrer-Costa C, Gelpi JL, Zamakola L, Parraga I, de la Cruz X, Orozco M (2005) PMUT: a web-based tool for the annotation of pathological mutations on proteins. Bioinformatics 21(14):3176–3178
49. Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30(17):3894–3900
50. Sunyaev S, Ramensky V, Koch I, Lathe 3rd W, Kondrashov AS, Bork P (2001) Prediction of deleterious human alleles. Hum Mol Genet 10(6):591–597
51. Huang T, Wang P, Ye ZQ, Xu H, He Z, Feng KY, Hu L, Cui W, Wang K, Dong X, Xie L, Kong X, Cai YD, Li Y (2010) Prediction of deleterious non-synonymous SNPs based on protein interaction network and hybrid properties. PLoS ONE 5(7):e11900
52. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22(3):231–238
53. Li S, Xi L, Li J, Wang C, Lei B, Shen Y, Liu H, Yao X, Li B (2011) In silico prediction of deleterious single amino acid polymorphisms from amino acid sequence. J Comput Chem 32(7):1211–1216
54. Herrgard S, Cammer SA, Hoffman BT, Knutson S, Gallina M, Speir JA, Fetrow JS, Baxter SM (2003) Prediction of deleterious functional effects of amino acid mutations using a library of structure-based function descriptors. Proteins 53(4):806–816
55. Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31(13):3812–3814
56. Zhu Y, Spitz MR, Amos CI, Lin J, Schabath MB, Wu X (2004) An evolutionary perspective on single-nucleotide polymorphism screening in molecular cancer epidemiology. Cancer Res 64(6):2251–2257
57. Sunyaev S, Ramensky V, Bork P (2000) Towards a structural basis of human non-synonymous single nucleotide polymorphisms. Trends Genet 16(5):198–200
58. Cai Z, Tsung EF, Marinescu VD, Ramoni MF, Riva A, Kohane IS (2004) Bayesian approach to discovering pathogenic SNPs in conserved protein domains. Hum Mutat 24(2):178–184
59. Liu YH, Li CG, Zhou SF (2009) Prediction of deleterious functional effects of non-synonymous single nucleotide polymorphisms in human nuclear receptor genes using a bioinformatics approach. Drug Metab Lett 3(4):242–286
60. Wang Z, Moult J (2001) SNPs, protein structure, and disease. Hum Mutat 17(4):263–270
61. Ye ZQ, Zhao SQ, Gao G, Liu XQ, Langlois RE, Lu H, Wei L (2007) Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (SAP). Bioinformatics 23(12):1444–1450
62. Bromberg Y, Rost B (2007) SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res 35(11):3823–3835
63. Tian J, Wu N, Guo X, Guo J, Zhang J, Fan Y (2007) Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines. BMC Bioinform 8:450
64. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, Guo N, Muruganujan A, Doremieux O, Campbell MJ, Kitano H, Thomas PD (2005) The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res 33(Database issue):D284–D288
65. Santoro A, Cannella S, Trizzino A, Bruno G, De Fusco C, Notarangelo LD, Pende D, Griffiths GM, Arico M (2008) Mutations affecting mRNA splicing are the most common molecular defect in patients with familial hemophagocytic lymphohistiocytosis type 3. Haematologica 93(7):1086–1090

66. Defesche JC, Schuurman EJ, Klaaijsen LN, Khoo KL, Wiegman A, Stalenhoef AF (2008) Silent exonic mutations in the low-density lipoprotein receptor gene that cause familial hypercholesterolemia by affecting mRNA splicing. Clin Genet 73(6):573–578
67. Ars E, Serra E, Garcia J, Kruyer H, Gaona A, Lazaro C, Estivill X (2000) Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. Hum Mol Genet 9(2):237–247
68. Wang GS, Cooper TA (2007) Splicing in disease: disruption of the splicing code and the decoding machinery. Nat Rev Genet 8(10):749–761
69. Orban TI, Olah E (2001) Purifying selection on silent sites—a constraint from splicing regulation? Trends Genet 17(5):252–253
70. Liu HX, Cartegni L, Zhang MQ, Krainer AR (2001) A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. Nat Genet 27(1):55–58
71. Fackenthal JD, Cartegni L, Krainer AR, Olopade OI (2002) BRCA2 T2722R is a deleterious allele that causes exon skipping. Am J Hum Genet 71(3):625–631
72. Fairbrother WG, Holste D, Burge CB, Sharp PA (2004) Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS Biol 2(9):E268
73. Freimuth RR, Stormo GD, McLeod HL (2005) PolyMAPr: programs for polymorphism database mining, annotation, and functional analysis. Hum Mutat 25(2):110–117
74. Conde L, Vaquerizas JM, Santoyo J, Al-Shahrour F, Ruiz-Llorente S, Robledo M, Dopazo J (2004) PupaSNP finder: a web tool for finding SNPs with putative effect at transcriptional level. Nucleic Acids Res 32(Web Server issue):W242–W248
75. Wang P, Dai M, Xuan W, McEachin RC, Jackson AU, Scott LJ, Athey B, Watson SJ, Meng F (2006) SNP Function Portal: a web database for exploring the function implication of SNP alleles. Bioinformatics 22(14):e523–e529
76. Smith PJ, Zhang C, Wang J, Chew SL, Zhang MQ, Krainer AR (2006) An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. Hum Mol Genet 15(16):2490–2508
77. Pyle AM (2010) The tertiary structure of group II introns: implications for biological function and evolution. Crit Rev Biochem Mol Biol 45(3):215–232
78. Mattick JS (1994) Introns: evolution and function. Curr Opin Genet Dev 4(6):823–831
79. Sonenberg N (1994) mRNA translation: influence of the 5' and 3' untranslated regions. Curr Opin Genet Dev 4(2):310–315
80. Cowles CR, Hirschhorn JN, Altshuler D, Lander ES (2002) Detection of regulatory variation in mouse genes. Nat Genet 32(3):432–437
81. Wittkopp PJ, Haerum BK, Clark AG (2004) Evolutionary changes in cis and trans gene regulation. Nature 430(6995):85–88
82. Hoogendoorn B, Coleman SL, Guy CA, Smith K, Bowen T, Buckland PR, O'Donovan MC (2003) Functional analysis of human promoter polymorphisms. Hum Mol Genet 12(18):2249–2254
83. Li C, Wu W, Liu J, Qian L, Li A, Yang K, Wei Q, Zhou J, Zhang Z (2006) Functional polymorphisms in the promoter regions of the FAS and FAS ligand genes and risk of bladder cancer in south China: a case-control analysis. Pharmacogenet Genomics 16(4):245–251
84. Hoogendoorn B, Coleman SL, Guy CA, Smith SK, O'Donovan MC, Buckland PR (2004) Functional analysis of polymorphisms in the promoter regions of genes on 22q11. Hum Mutat 24(1):35–42
85. Buckland PR, Coleman SL, Hoogendoorn B, Guy C, Smith SK, O'Donovan MC (2004) A high proportion of chromosome 21 promoter polymorphisms influence transcriptional activity. Gene Expr 11(5–6):233–239
86. Sandelin A, Wasserman WW, Lenhard B (2004) ConSite: web-based prediction of regulatory elements using cross-species comparison. Nucleic Acids Res 32(Web Server issue): W249–W252
87. URL: http://www.phylofoot.org/consite/. Access on 17 May 2011
88. URL: http://pupasnp.bioinfo.cnio.es/. Access on 17 May 2011

89. Ponomarenko JV, Merkulova TI, Orlova GV, Fokin ON, Gorshkova EV, Frolov AS, Valuev VP, Ponomarenko MP (2003) rSNP_Guide, a database system for analysis of transcription factor binding to DNA with variations: application to genome annotation. Nucleic Acids Res 31(1):118–121
90. URL: http://www.mgs.bionet.nsc.ru/mgs/systems/rsnp/. Accessed on 17 May 2011
91. Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, Kruglyak L (2003) Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. Nat Genet 35(1):57–64
92. Li J, Yuan Z, Zhang Z (2010) Revisiting the contribution of cis-elements to expression divergence between duplicated genes: the role of chromatin structure. Mol Biol Evol 27(7):1461–1466
93. Doniger SW, Fay JC (2007) Frequent gain and loss of functional transcription factor binding sites. PLoS Comput Biol 3(5):e99
94. Tirosh I, Reikhav S, Sigal N, Assia Y, Barkai N (2010) Chromatin regulators as capacitors of interspecies variations in gene expression. Mol Syst Biol 6:435
95. Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U (2008) Predicting expression patterns from regulatory sequence in Drosophila segmentation. Nature 451(7178):535–540
96. Smigielski EM, Sirotkin K, Ward M, Sherry ST (2000) dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res 28(1):352–355
97. Sherry ST, Ward M, Sirotkin K (1999) dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res 9(8):677–679
98. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Graf S, Haider S, Hammond M, Holland R, Howe K, Jenkinson A, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Rios D, Schuster M, Slater G, Smedley D, Spooner W, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wilder S, Zadissa A, Birney E, Cunningham F, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Kasprzyk A, Proctor G, Smith J, Searle S, Flicek P (2009) Ensembl 2009. Nucleic Acids Res 37(Database issue):D690–D697
99. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, Diekhans M, Dreszer TR, Giardine BM, Harte RA, Hillman-Jackson J, Hsu F, Kirkup V, Kuhn RM, Learned K, Li CH, Meyer LR, Pohl A, Raney BJ, Rosenbloom KR, Smith KE, Haussler D and Kent WJ (2011) The UCSC Genome Browser database: update 2011. Nucleic Acids Res 39(Database issue):D876–D882
100. Hirakawa M, Tanaka T, Hashimoto Y, Kuroda M, Takagi T, Nakamura Y (2002) JSNP: a database of common gene variations in the Japanese population. Nucleic Acids Res 30(1):158–162
101. Fredman D, Munns G, Rios D, Sjoholm F, Siegfried M, Lenhard B, Lehvaslaiho H, Brookes AJ (2004) HGVbase: a curated resource describing human DNA variation and phenotype relationships. Nucleic Acids Res 32(Database issue):D516–D519
102. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN (2003) Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 21(6):577–581
103. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31(1):365–370
104. Clifford R, Edmonson M, Hu Y, Nguyen C, Scherpbier T, Buetow KH (2000) Expression-based genetic/physical maps of single-nucleotide polymorphisms identified by the cancer genome anatomy project. Genome Res 10(8):1259–1265
105. Riva A, Kohane IS (2002) SNPper: retrieval and analysis of human SNPs. Bioinformatics 18(12):1681–1685
106. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR,

Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Res 12(10):1611–1618
107. Staats B, Qi L, Beerman M, Sicotte H, Burdett LA, Packer B, Chanock SJ, Yeager M (2005) Genewindow: an interactive tool for visualization of genomic variation. Nat Genet 37(2):109–110
108. Ryan M, Diekhans M, Lien S, Liu Y, Karchin R (2009) LS-SNP/PDB: annotated non-synonymous SNPs mapped to Protein Data Bank structures. Bioinformatics 25(11):1431–1432
109. Mooney SD, Altman RB (2003) MutDB: annotating human variation with functionally relevant data. Bioinformatics 19(14):1858–1860
110. Zhao T, Chang LW, McLeod HL, Stormo GD (2004) PromoLign: a database for upstream region analysis and SNPs. Hum Mutat 23(6):534–539
111. Conde L, Vaquerizas JM, Dopazo H, Arbiza L, Reumers J, Rousseau F, Schymkowitz J, Dopazo J (2006) PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. Nucleic Acids Res 34(Web Server issue):W621–W625
112. Kim BC, Kim WY, Park D, Chung WH, Shin KS, Bhak J (2008) SNP@Promoter: a database of human SNPs (single nucleotide polymorphisms) within the putative promoter regions. BMC Bioinform 9(Suppl 1):S2
113. Yue P, Moult J (2006) Identification and analysis of deleterious human SNPs. J Mol Biol 356(5):1263–1274
114. Reumers J, Schymkowitz J, Ferkinghoff-Borg J, Stricher F, Serrano L, Rousseau F (2005) SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. Nucleic Acids Res 33(Database issue):D527–D532
115. Conde L, Vaquerizas JM, Ferrer-Costa C, de la Cruz X, Orozco M, Dopazo J (2005) PupasView: a visual tool for selecting suitable SNPs, with putative pathological effect in genes, for genotyping purposes. Nucleic Acids Res 33(Web Server issue):W501–W55
116. Xu H, Gregory SG, Hauser ER, Stenger JE, Pericak-Vance MA, Vance JM, Zuchner S, Hauser MA (2005) SNPselector: a web tool for selecting SNPs for genetic association studies. Bioinformatics 21(22):4181–4186
117. Hemminger BM, Saelim B, Sullivan PF (2006) TAMAL: an integrated approach to choosing SNPs for genetic studies of human complex traits. Bioinformatics 22(5):626–627
118. Chang H, Fujita T (2001) PicSNP: a browsable catalog of nonsynonymous single nucleotide polymorphisms in the human genome. Biochem Biophys Res Commun 287(1):288–291
119. Stitziel NO, Binkowski TA, Tseng YY, Kasif S, Liang J (2004) topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. Nucleic Acids Res 32(Database issue):D520–D522
120. Taylor NE, Greene EA (2003) PARSESNP: A tool for the analysis of nucleotide polymorphisms. Nucleic Acids Res 31(13):3808–3811
121. Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR (2003) ESEfinder: a web resource to identify exonic splicing enhancers. Nucleic Acids Res 31(13):3568–3571
122. Fairbrother WG, Yeo GW, Yeh R, Goldstein P, Mawson M, Sharp PA, Burge CB (2004) RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. Nucleic Acids Res 32(Web Server issue):W187–W190
123. Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W (2000) PipMaker—a web server for aligning two genomic DNA sequences. Genome Res 10(4):577–586
124. Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS, Dubchak I (2000) VISTA : visualizing global DNA sequence alignments of arbitrary length. Bioinformatics 16(11):1046–1047
125. Ovcharenko I, Nobrega MA, Loots GG, Stubbs L (2004) ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. Nucleic Acids Res 32(Web Server issue):W280–W286
126. Loots GG, Ovcharenko I (2004) rVISTA 2.0: evolutionary analysis of transcription factor binding sites. Nucleic Acids Res 32(Web Server issue):W217–W221

127. Marinescu VD, Kohane IS, Riva A (2005) MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. BMC Bioinform 6:79
128. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E (2003) MATCH: a tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res 31(13):3576–3579
129. McGinnis S, Madden TL (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res 32(Web Server issue):W20–W25

Chapter 18
An Application of QM/MM Simulation: The Second Protonation of Cytochrome P450

Peng Lian and Dongqing Wei

Abstract The multiscale model strategy, hybrid quantum mechanics and molecular mechanics (QM/MM), has become more and more prevalent in the theoretical study of enzymatic reactions. It combines both the efficiency of the Newtonian molecular calculations and the accuracy of the quantum mechanical methods. Simulation using QM/MM multiscale model may be one of the most promising approaches that could further narrow the gap between the theoretical models and the real problems. It is capable of dealing with not only the conformational changes of biomacromolecules, but also the catalytic reactions. Herein, we reviewed some of our recent work to demonstrate the application of the QM/MM simulations in exploring the enzymatic reactions.

Keywords QM/MM · Conformational changes · Catalytic reactions · Enzymatic reactions

18.1 Introduction

Ever since the hybrid quantum mechanics and molecular mechanics (QM/MM) method has been introduced by Warshel and Levitt in 1976 [1], it has become a valuable tool in investigating organic/inorganic, solid-state, or reactions in explicit solvents. In the studies of biological systems, the QM/MM approach has been widely applied after Field, Bash and Karplus developed a combined potential of semi-empirical QM and the CHARMM MM force field in 1990 [2]. Nowadays, it becomes one of the most popular approaches in studying the mechanism of enzymatic reactions.

P. Lian · D. Wei (✉)
State Key Laboratory of Microbial Metabolism, College of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China
e-mail: dqwei@sjtu.edu.cn

D. Wei et al. (eds.), *Advance in Structural Bioinformatics*, Advances in Experimental Medicine and Biology 827, DOI 10.1007/978-94-017-9245-5_18

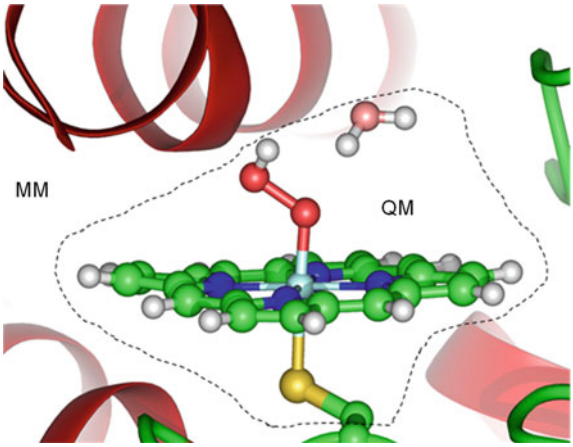


Fig. 18.1 A schematic diagram of QM/MM method

QM/MM method combines both the accuracy of the quantum mechanical methods and the speed of the force field based molecular calculations. The fundamental idea is that the system can be partitioned into different regions and then be described by different levels of approximations. As is shown in Fig. 18.1, for an enzymatic reaction, the ligand and substrate in the active site are usually be treated as QM region which is then calculated by the more accurate quantum approaches, including semi-empirical, *ab initio*, density functional theory (DFT) quantum mechanics etc. While the other part of the enzyme and the solvent environment are partitioned as MM region, which could take advantage of the high speed of the force field based molecular calculations. The quantum mechanical treatment of the QM region allows for modeling the electronic rearrangement in chemical bond making and breaking, while the force field modeling of the MM region allows for exploring the effects of conformation change of the enzyme and solvent environment on the enzymatic reactions.

18.2 The QM/MM Method

There are two general types of QM/MM, the additive approach and the subtractive. For the additive QM/MM approach, the energy of the entire system could be written as bellow,

$$E_{Total} = E_{QM} + E_{MM} + E_{QM-MM} \tag{18.1}$$

The total energy (E_{Total}) consists of three parts, the energy of the QM region (E_{QM}), the energy of the MM region (E_{MM}) and the interaction energy between QM and MM region (E_{QM-MM}).

Different from the additive QM/MM approach, the energy function of the subtractive scheme is shown as bellow,

$$E_{Total} = E_{Total}(MM) + E_{QM}(QM) - E_{MM}(QM) \tag{18.2}$$

In this function, $E_{Total}(MM)$ is the total energy of the entire system from molecular mechanics level calculations. $E_{QM}(QM)$ represents the quantum level calculations of the inner QM region of the whole system. $E_{MM}(QM)$ is the molecular mechanics level energy of the QM region of the system. For subtractive QM/MM, the entire scheme is actually the MM calculation of the entire system with a certain region of the system has been cut out and treated at the QM level. Therefore, this method is straightforward and there is no need to deal with the QM–MM interaction explicitly. However, a calculation of the QM region at both the molecular level and quantum level is required, which means a complete set of MM parameters for the inner region is necessary. Generally, MM parameters for a chemical reaction are difficult to obtain, which limits the application of the subtractive QM/MM method.

In order to balance the accuracy and the required computational resource, a QM region with a proper size is critical for an effective QM/MM calculation. For small molecule reactions, if the reactant could be fully partitioned to the QM region, the cutting-covalent bond problem can be avoided. Otherwise, special treatment of the cutting-covalent bond is ineluctable. Generally, there are two wildly used boundary schemes, the link-atom approach, and the localized-orbital approach. For the link-atom scheme, the basic idea is to introduce an additional atom L in the cutting-covalent bond to cap the free valence of the QM region atom. This link atom L is usually a hydrogen atom; it is not a real part of the system, only to saturate the free valence of the QM atoms. In order to cap the QM region, the localized-orbital schemes are trying to put a set of hybrid orbitals at the boundary. Some of these orbitals are kept frozen so that they do not participate in the SCF iteration. This method could date back to the birth of the QM/MM scheme [1]. There are many elaborated schemes, i.e., Local Self-Consistent Field (LSCF) [3], Frozen Orbitals [4], and Generalized Hybrid Orbitals (GHO) [5]. No matter for which schemes, the cutting covalent bond should better be unipolar and not involved in conjugative interactions, for instance, a single C–C bond.

18.3 QM/MM Study on the Mechanism of the Second Protonation of P450cam

Cytochrome P450 catalyzes various stereospecific and regioselective processes of oxygenation of the substrates. It is a superfamily of mono-oxygenases with an immense biological impact of drug metabolism [6–10]. According to the previous

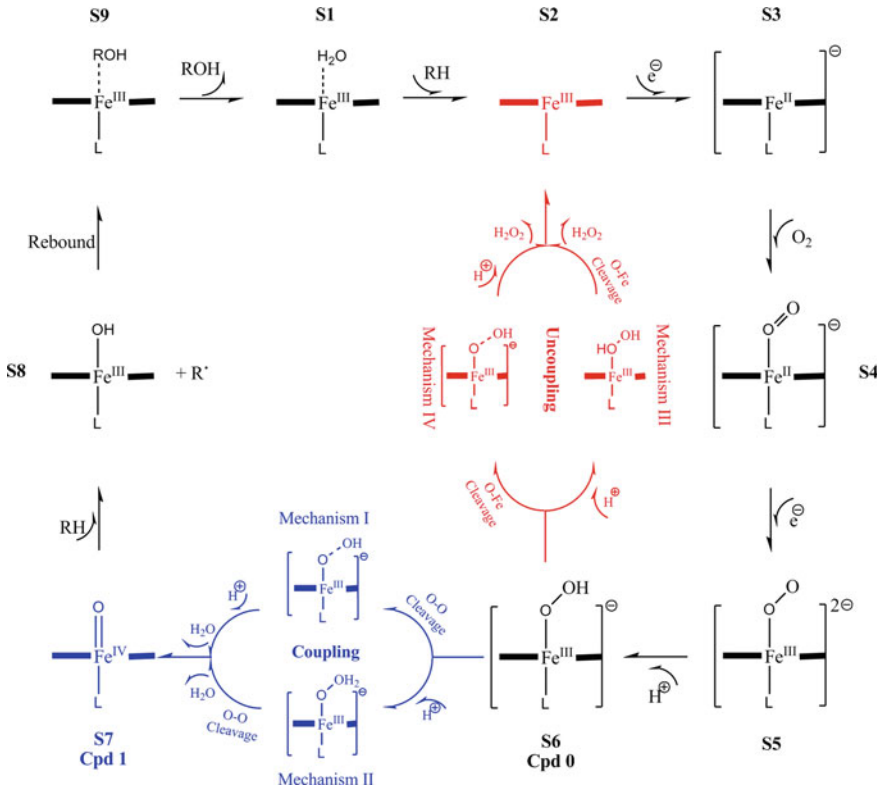


Fig. 18.2 The catalytic cycle of P450. Possible mechanisms [16] for coupling and uncoupling reactions are shown in *blue* and *red*, respectively

studies, the catalytic cycle of P450 enzymes could be summarized in Fig. 18.2 [10–15].

There are two protonation steps in this cycle. The protonation of the ferric peroxo (or ferric superoxo, **S5**) species and the subsequent proton transfer that produces the iron (IV) oxo porphyrin π-cation radical intermediate (Compound I, Cpd I). Cpd 0 is a precursor of Cpd I, and, it was not well characterized until Naruta and co-workers synthesized the similar ferric hydroperoxo porphyrin complex successfully [17]. The second proton transfer, which relates to the generation of Cpd I from Cpd 0 is of great interest [13, 16, 18–25].

The second proton transfer would lead to either the coupling process or the uncoupling one. In coupling, the proton is transferred to the distal oxygen (denoted as O(d). See Fig. 18.3) of Cpd 0, then the O–O bond is cleaved and Cpd I is generated. In uncoupling, the proton is delivered to the proximal oxygen [denoted as O(p)] of Cpd 0, then a hydrogen peroxide is produced and the enzyme goes back to its resting state. These two oxygen atoms (O(d) and O(p)) compete against each other for protons and affect the turnover rate of the catalysis of substrates.

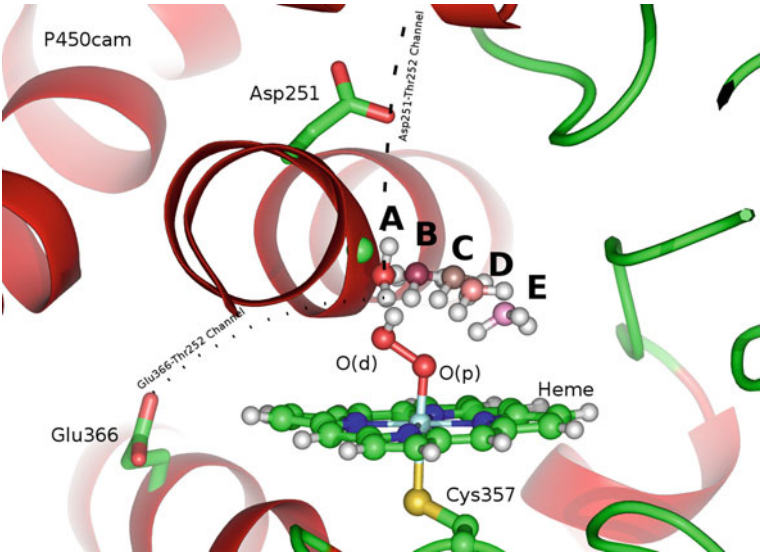


Fig. 18.3 Superposition of the five models A–E. These models are different only in the position of the hydronium probe (H_3O^+). The prophyrin, the hydroperoxo moiety and the hydronium probe are represented in *ball*-and-*stick*. The P450cam enzyme is shown in carton. The Asp251-Thr252 channel (*dash line*) and the Glu366-Thr252 channel (*dot line*) are shown schematically

However, the factors that determine the catalytic cycle of P450 goes to coupling or uncoupling remain to be clarified.

According to the previous studies, proton transfer channels seem to play a role in the second protonation of P450cam. There are two main channels in P450cam, Glu366-Thr252 and Asp251-Thr252 [10, 16, 19, 26, 27]. Glu366-Thr252 channel provides a proton to the active site via the Poulos–Kraut mechanism [15, 20]. However, Glu366 is difficult to get re-protonated after offering its proton, that's because for Glu366, there isn't any direct connections to bulk water; therefore, the function of Glu366 in the second protonation is questionable. In contrast, the Asp251-Thr252 channel connects the active site to the bulk solvent and shuttles the proton via the Asp251-Arg186 salt bridge [15, 20]. Moreover, according to the previously computation, the energy barrier for proton transfer in Glu366-Thr252 channel is 6.6 kcal/mol, while it is essentially barrierless in Asp251-Thr252 channel [20]. Thus, in this study, we consider on the Asp251-Thr252 channel exclusively.

Asp251 plays a vital role during the proton transfer [15, 19, 22, 25]. Experimental results on the activities of P450cam and its mutants are summarized in Table 18.1. It shows that mutations on Asp251 reduce the catalytic rate significantly, while having little effect on the product formation. That's probably because Asp251 controls the proton shuttle through a "carboxylate switch" mechanism [15, 27]. However, mutations on Thr252 exhibit interesting effects [16]. If it was a

Table 18.1 Experimental observations on coupling and uncoupling

P450cam mutation	Connolly solvent excluded volume of Sidechain [28] (A^3)	Hydropathy index [29] of amino acid	O_2 consuming rate (μM/min)	5-exohy-droxycamphor coupling (%)	H_2O_2 uncoupling (%)	Ref.
D251G			21	99	2	[30]
D251A			3	89	12	[30]
D251N			6	90	–	[31]
D251N + T252A			5	52	64	[32]
T252T(WT)[a]	48.6	−0.7	1,370	100	2	[33]
T252T(WT)	48.6	−0.7	1,330	96	5	[33]
T252T(WT)	48.6	−0.7	1,350	97	3	[30]
T252T(WT)[b]	48.6	−0.7	1,340	100	–	[34]
T252T(WT)	48.6	−0.7	1,350	100	–	[32]
T252S	31.4	−0.8	1,100	81	15	[33]
T252S	31.4	−0.8	830	85	15	[30]
T252T-OMe[b]	64.9	–	410	100	–	[34]
T252 V	58.5	4.2	420	22	45	[33]
T252I	75.9	4.5	277	44	40	[32]
T252A	24.3	1.8	1,100	6	83	[33]
T252A	24.3	1.8	1,150	5	89	[30]
T252G	8.9	−0.4	1,090	3	88	[30]

[a] Wild type from *Pseudomonas putida*
[b] Synthesized in vitro
Others were expressed in *Escherichia coli*

medium-sized residue with hydrophilic side-chain, like Thr and Ser, the coupling product, 5-exo-hydroxycamphor could be produced quickly. Otherwise, if it were replaced by an amino acid with large hydrophobic side-chain, like Val and Ile, both the reaction rate and the ratio of coupling product would decrease significantly. For mutants with a short side-chain, like Ala and Gly, uncoupling could be observed. It seems the subsequent effect of the Thr252 mutations depends on the volume and the hydrophobicity of its side-chain.

Recently, theoretical studies [16] on P450cam at QM/MM level have shown that proton transfer pathways which are constructed by key residues and water molecules may be responsible for the competition between coupling and uncoupling. Possible pathways, e.g. Asp251-Wat901-Thr252-FeOOH, Asp251-Wat901-FeOOH and Asp251-Wat901-WatS-FeOOH, were explored. It has been found that with an extra water molecule (WatS) in active site, the uncoupling barriers will be reduced around 10 kcal/mol for T252X (X = V, A and G). In addition, the low barrier of Grotthuss mechanism (around 2–3 kcal/mol [35]) indicates that the proton could be transferred freely between water molecules. Therefore, there may be multiple pathways for the proton being transferred to the hydroperoxo unit. Thus, the observed coupling/uncoupling in Thr252 mutants may be owing to the change of the proton-transfer pathways.

Herein, we explored the relationship between proton-transfer pathways and the competition of coupling/uncoupling. Five models, A–E, representing possible pathways of the second proton transfer were built and studied via CPMD/MM dynamics simulations.

18.4 Methods

The crystal structure, PDB ID 1DZ8 [15], was used as the initial coordinates. The protocols developed by Thiel's group [16, 36–39] were employed for protonation and solvation procedures. In order to obtain enough space in the active site for studying possible proton transfer pathways in all mutants and the wild-type enzyme T252G mutant was introduced. To make sure the intrinsic proton affinities of O(p) and O(d) are investigated the substrate camphor was removed. Then the deMon package [40] were used to optimize the geometry of the HOO moiety, heme unit and the axial Cys ligand in gas phase. After that, an energy minimization was performed on the whole system. During the minimization the heme unit, the sulfur and the C_β of Cys357 were fixed. The OPLS all-atoms force field and GROMACS package were used.

Five models (A–E) shown in Fig. 18.3 were built with the previously optimized system and different probes (H_3O^+). In model A which is for the wild-type enzyme, both the oxygen atom [O(aq)] and the to-be-transferred proton [H^+(aq)] of the probe were kept the same positions as in the hydroxyl group of Thr252 in Wang et al.'s study [39]. B and D were built to simulate the T252S and T252 V mutants, respectively [16]. Model C served as a reference. In this model, both

O(aq) and H^+(aq) of the probe were placed equidistantly to O(d) and O(p). Meanwhile, four atoms, O(aq), H^+(aq) from the probe, O(d) and O(p) of the hydroperoxo unit, were kept on the same plane with H^+(aq) closer to the peroxide group. The other two protons of the probe were optimized symmetrically according to the orientation of the peroxide bond and the O(aq)–H^+(aq) bond of the probe. E is to cover Oprea et al.'s two-state mode water channel [41]. Different from the other channels, this channel is made up by waters and starting from the down side of the porphyrin ring (we define the peroxide side as up).

Each model was dealt with the same procedure below. The probe, the heme unit and the side chain of Cys357 (49 atoms in total) were defined as QM region and calculated by CPMD program (http://www.cpmd.org/). For QM calculations, the GGA DFT method BLYP [42] and Vanderbilt ultrasoft pseudopotentials (USPPs) [43–45] were used. The cutoff of the wave function was 25.0 Rydberg. The rest of each system was the MM part, and was described by the OPLS all-atom force field [46, 47] as implemented in GROMACS [48–50]. For the MM part, atoms within 15 Å of the QM region and all the polar residues and ions in the system were chosen as inner layer. The s-wave partial wave expansion method which was provided by the GROMACS-CPMD QMMM interface [51] was used to take account for the polarization effect of the inner layer atoms. The link atom scheme was used as the boundary of QM/MM [52].

Preconditioned conjugate gradients method [53] was used to perform the QM/MM geometry optimizations. The convergence of geometry was set to 5×10^{-3} Å. During the geometry optimization, coordinates of O(aq) and H^+(aq) of the probe, O(d) and O(p) in the hydroperoxo moiety, the iron, the sulfur and the C_β atom of Cys357 were fixed. The MM part was equilibrated to 300 K in 5 ps using the Berendsen algorithm [54] with the QM region fixed. After that the QM/MM *ab initio* MD was performed. For QM region, the Car-Parrinello molecular dynamics [55] with a time step of 5 a.u. and a fictitious electron mass of 400 a.u. was employed. For MM region, the leap-frog algorithm with the time step of 1 fs and the cutoff for non-bond calculations as 1.0 nm was used. The total *ab initio* simulation time was 37.5 ps. Similar approaches have been successfully applied to other P450 enzymes and various bio-macromolecular systems [56–59].

18.5 Results and Discussion

According to our simulations, coupling is thermodynamically more favorable than uncoupling by about 77 kcal/mol. However, among all of the five models, only two of them, A and B, lead to coupling. The other three, including the reference model C, generate uncoupling products. Both coupling and uncoupling were exothermic and proceeded rapidly and without a barrier (or too small to be observed). It agrees with former studies on the exothermic characters [13, 18, 21, 60].

18.5.1 The Bond Properties

The progress of a reaction could be indicated from the dynamical changes of bond lengths and bond orders. In this study, we captured the bond lengths of key atoms during the first 200 fs and the related Mayer bond orders during the reaction (50 fs) (Fig. 18.4).

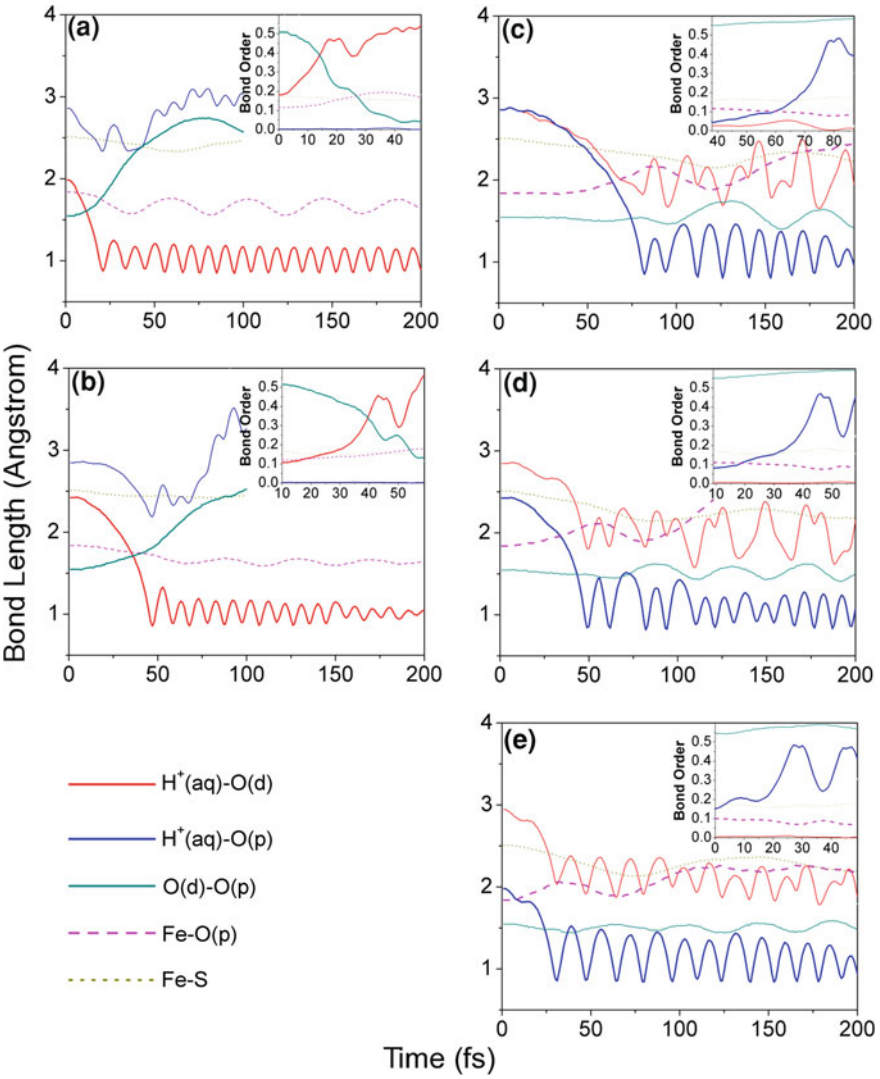


Fig. 18.4 The time-dependent bond length and bond order evolution in simulations

In coupling cases (Model A and B), although the probe was at different pathways, both show a strong negative correlation between bond H^+(aq)–O(d) (decreasing) and O(d)–O(p) (increasing). The proton transfer and the O(d)–O(p) cleavage are coupled. This agrees well with former studies which show that the O(d)–O(p) cleavage is assisted by the second proton transfer [13, 19, 22]. In uncoupling cases (Model C, D and E), the Fe–O(p) bond didn't break immediately, but was weakened significantly. That is probably because the unoccupied orbital of the iron atom was taken up by the lone pair electrons of one oxygen atom of H_2O_2. The formation of H^+(aq)–O(p) bond and the stretching of Fe–O(p) bond were also proceeding simultaneously.

Based on these observations, we conclude that for both coupling and uncoupling the proton transfer is associated with the O(d)–O(p) cleavage or the Fe–O(p) stretching.

18.5.2 Electronic Structural Properties

During the reaction the evolution of spin densities and natural charges of key atoms and groups were monitored. NBO population analysis [61] was performed at B3LYP/BS level (BS: LANL2DZ for Fe and 6-31G* for other atoms) with Gaussian09 [62]. The spin densities and natural charges for coupling (Model A) were shown in Fig. 18.5a, b, and uncoupling (Model E) in Fig. 18.5c, d, respectively.

In coupling the increase of negative charge on O(d) is accompanied by the decrease of the charge on porphyrin. Meanwhile, the spin density on O(d) was not affected, while that on O(p) and porphyrin increases. It suggests that there is an electron transfer from porphyrin to the hydroperoxo unit along with the proton transfer. This electron flaw was probably to neutralize the positive charge of the incoming proton. After the heterolytic cleavage of the O(p)–O(d) bond, a water molecule, Wat903, is produced. Thus, in this step it follows a proton-coupled electron transfer (PCET) mechanism [20]. In the uncoupling reaction, spin density on the hydrogen peroxide diminishes during the reaction. The electron that is used to neutralize the charge on the proton is pumped from Fe and Cys ligand. Porphyrin does not contribute much to the electron transfer in uncoupling.

The porphyrin ring plays different roles in both reactions. During coupling, the heterolytic cleavage of the O(p)–O(d) bond generates a ferryl species with a high valent Fe atom, which is stabilized by the charge transfer from the porphyrin. In uncoupling, after the Fe(III)–O(p) bond is broken the enzyme goes back to its resting state. It seems in uncoupling it is not required to pump additional electrons into the hydroperoxo unit. It shows that although porphyrin can stabilize the ferryl species formed in the coupling process, it does not have preference on either the coupling or the uncoupling in the second proton transfer of P450cam.

In this application, our study shows that the coupling process is associated with the proton delivery pathways represented in Model A and B, while the uncoupling

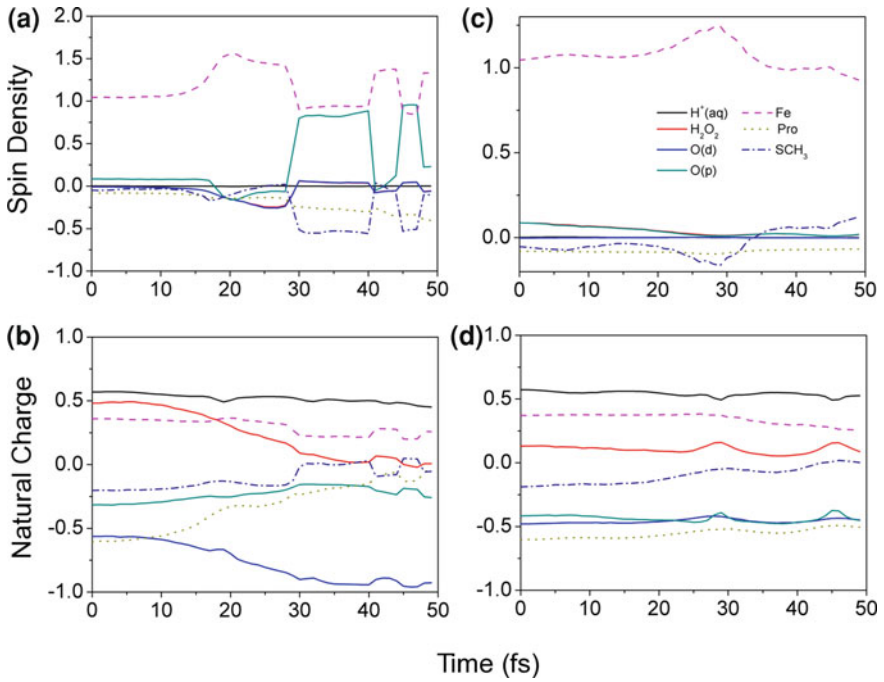


Fig. 18.5 Spin densities (*upper panels*) and natural charges (*lower panels*) during the reactions of both coupling (**a**, **b**) and uncoupling(**c**, **d**)

with those in Model C, D and E. This suggests that besides the intrinsic proton affinities of O(d) and O(p) [The natural charge on O(d) is about −0.56e, while −0.32e on O(p)], the proton delivery pathway plays an vital role in determine the coupling and uncoupling as well. It implies the importance of the topology of active site. Therefore, it explains the tricks of the wild-type enzyme. In WT, the proton is delivered through pathway the Asp251-Thr252 channel (Model A), which leading to a coupling process exclusively. When this channel is destroyed by mutation, for instance, when Thr252 is mutated to Ala or Gly, it increases the probability of the proton undergoing other pathways, for example, pathways in Model C, D and E, thus leads to uncoupling reactions.

18.6 Conclusions

QM/MM method is a powerful tool for tackling the mechanism of the enzymatic reaction. In this article, we briefly reviewed the main idea of the methodology at first. Then, the mechanism of the second protonation of P450cam was taken as an example to demonstrate the application of QM/MM. In this example, we built five models (A–E) which represent five possible proton transfer pathways to explore

the key factors that determine the coupling and uncoupling reactions. It is found that two of them (A and B) led to the coupling, while the other three (C–E) generated uncoupling products. During each reaction, bond properties and the electronic structure of QM region show that in coupling, O(d)–O(p) cleavage, proton transfer and electron delivery are coupled, while in uncoupling, stretching of the Fe–O(p) bond and proton transfer take place spontaneously. Moreover, the proton transfer pathway seems to play a vital role in the determination of the reactions of the second proton transfer. The enzyme is likely to keep a high coupling rate by maintaining a specific proton transfer channel, the Asp251–Thr252 channel, through which the second proton is transferred to the ideal position for coupling reaction.

References

1. Warshel A, Levitt M (1976) Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. J Mol Biol 103(2):227–249
2. Field MJ, Bash PA, Karplus M (1990) A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. J Comput Chem 11(6):700–733
3. Théry V et al (1994) Quantum mechanical computations on very large molecular systems: the local self-consistent field method. J Comput Chem 15(3):269–282
4. Murphy RB, Philipp DM, Friesner RA (2000) A mixed quantum mechanics/molecular mechanics (QM/MM) method for large-scale modeling of chemistry in protein environments. J Comput Chem 21(16):1442–1457
5. Gao J et al (1998) A generalized hybrid orbital (GHO) method for the treatment of boundary atoms in combined QM/MM calculations. J Phys Chem A 102(24):4714–4721
6. Ortiz de Montellano PR (2005) Cytochrome P450: structure, mechanism, and biochemistry. Springer, Heidelberg
7. Dawson JH, Sono M (1987) Cytochrome P-450 and Chloroperoxidase—thiolate-ligated heme enzymes—spectroscopic determination of their active-site structures and mechanistic implications of thiolate ligation. Chem Rev 87(5):1255–1276
8. Sono M et al (1996) Heme-containing oxygenases. Chem Rev 96(7):2841–2887
9. Hawkes DB et al (2002) Cytochrome P450cin (CYP176A), isolation, expression, and characterization. J Biol Chem 277(31):27725–27732
10. Denisov IG et al (2005) Structure and chemistry of cytochrome P450. Chem Rev 105(6):2253–2277
11. de Visser SP, Valentine JS, Nam W (2010) A biomimetic ferric hydroperoxo porphyrin intermediate. Angew Chem Int Ed 49(12):2099–2101
12. Coon MJ (2005) Cytochrome P450: nature's most versatile biological catalyst. Annu Rev Pharmacol Toxicol 45:1–25
13. Guallar V, Friesner RA (2004) Cytochrome P450CAM enzymatic catalysis cycle: a quantum mechanics/molecular mechanics study. J Am Chem Soc 126(27):8501–8508
14. Gunsalus IC, Pederson TC, Sligar SG (1975) Oxygenase-catalyzed biological hydroxylations. Annu Rev Biochem 44:377–407
15. Schlichting I et al (2000) The catalytic pathway of cytochrome P450cam at atomic resolution. Science 287(5458):1615–1622
16. Altarsha M et al (2009) How is the reactivity of cytochrome P450cam affected by Thr252X mutation? A QM/MM study for X = serine, valine, alanine glycine. J Am Chem Soc 131(13):4755–4763

17. Liu JG et al (2009) Spectroscopic characterization of a hydroperoxo-heme intermediate: conversion of a side-on peroxo to an end-on hydroperoxo complex. Angew Chem Int Ed 48(49):9262–9267
18. Ogliaro F et al (2002) Searching for the second oxidant in the catalytic cycle of cytochrome P450: a theoretical investigation of the iron (III)-hydroperoxo species and its epoxidation pathways. J Am Chem Soc 124(11):2806–2817
19. Kumar D et al (2005) New features in the catalytic cycle of cytochrome P450 during the formation of compound I from compound 0. J Phys Chem B 109(42):19946–19951
20. Zheng JJ et al (2006) QM/MM study of mechanisms for compound I formation in the catalytic cycle of cytochrome P450cam. J Am Chem Soc 128(40):13204–13215
21. Harris DL, Loew GH (1998) Theoretical investigation of the proton assisted pathway to formation of cytochrome P450 compound I. J Am Chem Soc 120(35):8941–8948
22. Hata M et al (2004) Theoretical study on compound i formation in monooxygenation mechanism by cytochrome P450. J Phys Chem B 108(30):11189–11195
23. Shaik S et al (2005) Theoretical perspective on the structure and mechanism of cytochrome P450 enzymes. Chem Rev 105(6):2279–2328
24. Sen K, Hackett JC (2009) Molecular Oxygen Activation and Proton Transfer Mechanisms in Lanosterol 14a-Demethylase Catalysis. J. Phys. Chem. B 113(23):8170–8182
25. Altarsha M et al (2010) Coupling and uncoupling mechanisms in the methoxythreonine mutant of cytochrome P450cam: a quantum mechanical/molecular mechanical study. J Biol Inorg Chem 15(3):361–372
26. Taraphder S, Hummer G (2003) Protein side-chain motion and hydration in proton-transfer pathways. Results for cytochrome P450cam. J Am Chem Soc 125(13):3931–3940
27. Vidakovic M et al (1998) Understanding the role of the essential Asp251 in cytochrome P450cam using site-directed mutagenesis, crystallography, and kinetic solvent isotope effect. Biochemistry 37(26):9211–9219
28. Richmond TJ (1984) Solvent accessible surface area and excluded volume in proteins: analytical equations for overlapping spheres and implications for the hydrophobic effect. J Mol Biol 178(1):63–89
29. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157(1):105–132
30. Shimada H et al (1990) Mechanism of oxygen activation by cytochrome P-450cam. In: International symposium on oxygenases and oxygen activation: yamada conference XXVII. Yamada Science Foundation, Osaka, Japan
31. Gerber NC, Sligar SG (1992) Catalytic mechanism of cytochrome P-450: evidence for a distal charge relay. J Am Chem Soc 114(22):8742–8743
32. Hishik T et al (2000) X-ray crystal structure and catalytic properties of Thr252Ile mutant of cytochrome P450cam: roles of Thr252 and water in the active center. J Biochem 128(6):965–974
33. Imai M et al (1989) Uncoupling of the cytochrome P-450cam monooxygenase reaction by a single mutation, threonine-252 to alanine or valine: possible role of the hydroxy amino acid in oxygen activation. Proc Natl Acad Sci USA 86(20):7823–7827
34. Kimata Y et al (1995) Role of Thr-252 in cytochrome P450(Cam)—a study with unnatural amino-acid mutagenesis. Biochem Biophys Res Commun 208(1):96–102
35. Agmon N (1995) The grotthuss mechanism. Chem Phys Lett 244(5–6):456–462
36. Schoneboom JC et al (2002) The elusive oxidant species of cytochrome P450 enzymes: characterization by combined quantum mechanical/molecular mechanical (QM/MM) calculations. J Am Chem Soc 124(27):8142–8151
37. Schoneboom JC, Thiel W (2004) The resting state of P450cam: a QM/MM study. J Phys Chem B 108(22):7468–7478
38. Altun A, Thiel W (2005) Combined quantum mechanical/molecular mechanical study on the pentacoordinated ferric and ferrous cytochrome P450cam complexes. J Phys Chem B 109(3):1268–1280

39. Wang D et al (2008) Quantum and molecular mechanical study of the first proton transfer in the catalytic cycle of cytochrome P450cam and its mutant D251N. J Phys Chem B 112(16):5126–5138
40. Koster AM et al (2003) DeMon 2003 code. NRC, Canada
41. Oprea TI, Hummer G, García AE (1997) Identification of a functional water channel in cytochrome P450 enzymes. Proc Natl Acad Sci USA 94(6):2133–2138
42. Becke AD (1988) Density-functional exchange-energy approximation with correct asymptotic behavior. Phys Rev A 38(6):3098–3100
43. Vanderbilt D (1990) Soft self-consistent pseudopotentials in a generalized eigenvalue formalism. Phys Rev B 41(11):7892–7895
44. Laasonen K et al (1991) Implementation of ultrasoft pseudopotentials in *ab initio* molecular dynamics. Phys Rev B 43(8):6796–6799
45. Laasonen K et al (1993) Car-parrinello molecular dynamics with vanderbilt ultrasoft pseudopotentials. Phys Rev B 47(16):10142–10153
46. Jorgensen WL, Tirado-Rives J (1988) The OPLS force field for proteins. energy minimizations for crystals of cyclic peptides and crambin. J Am Chem Soc 110(6):1657–1666
47. Jorgensen WL, Maxwell DS, Tirado-Rives J (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 118(45):11225–11236
48. Berendsen HJC, van der Spoel D, van Drunen R (1995) GROMACS: a message-passing parallel molecular dynamics implementation. Comput Phys Commun 91(1–3):43–56
49. Lindahl E, Hess B, van der Spoel D (2001) GROMACS 3.0: a package for molecular simulation and trajectory analysis. J Mol Model 7(8):306–317
50. van der Spoel D et al (2005) GROMACS: fast, flexible, and free. J Comput Chem 26(16):1701–1718
51. Biswas PK, Gogonea V (2005) A regularized and renormalized electrostatic coupling hamiltonian for hybrid quantum-mechanical-molecular-mechanical calculations. J Chem Phys 123(16):164114–164122
52. Das D et al (2002) Optimization of quantum mechanical molecular mechanical partitioning schemes: gaussian delocalization of molecular mechanical charges and the double link atom method. J Chem Phys 117(23):10534–10547
53. Hestenes MR, Stiefel E (1952) Methods of conjugate gradients for solving linear systems. J Res Natl Bur Stand 49(6):409–436
54. Berendsen HJC et al (1984) Molecular-dynamics with coupling to an external bath. J Chem Phys 81(8):3684–3690
55. Car R, Parrinello M (1985) Unified approach for molecular-dynamics and density-functional theory. Phys Rev Lett 55(22):2471–2474
56. Lian P et al (2010) Tethered-hopping model for protein-DNA binding and unbinding based on Sox2-Oct1-Hoxb1 ternary complex simulations. Biophys J 98(7):1285–1293
57. Lian P WD-Q, Wang J-F, Chou K-C (2011) An allosteric mechanism inferred from molecular dynamics simulations on phospholamban pentamer in lipid membranes. PLoS One 6(4):e18587
58. Lian P et al (2013) Catalytic mechanism and origin of high activity of cellulase TmCel12A at high temperature: a quantum mechanical/molecular mechanical study. Cellulose 1–13
59. Lian P et al (2013) Car-parrinello molecular dynamics/molecular mechanics (CPMD/MM) simulation study of coupling and uncoupling mechanisms of cytochrome P450cam. J Phys Chem B 117(26):7849–7856
60. Kamachi T et al (2003) Does the hydroperoxo species of cytochrome P450 participate in olefin epoxidation with the main oxidant, compound I? Criticism from density functional theory calculations. Bull Chem Soc Jpn 76(4):721–732
61. Glendening ED, Reed AE, Carpenter JE, Weinhold F NBO Version 3.1
62. Frisch MJ et al (2009) Gaussian 09, revision A.1. Gaussian Inc, Wallingford, CT

Part V
Application of Structural Bioinformatics in Drug Design

Chapter 19
Recent Progress on Structural Bioinformatics Research of Cytochrome P450 and Its Impact on Drug Discovery

Tao Zhang and Dongqing Wei

Abstract Cytochrome P450 is predominantly responsible for human drug metabolism, which is of critical importance for drug discovery and development. Structural bioinformatics focuses on analysis and prediction of three-dimentional structure of biological macromolecules and elucidation of structure-function relationship as well as identification of important binding interactions. Rapid advancement of structural bioinformatics has been made over the last decade. With more information available for CYP structures, the methods of structural bioinformatics may be used in the CYP field. In this review, we demonstrate three previous studies on CYP using the methods of structural bioinformatics, including the investigation of reasons for decrease of enzymatic activity of CYP1A2 caused by a peripheral mutation, the construction of a pharmacophore model specific to active site of CYP1A2 and the prediction of the functional consequences of single residue mutation in CYP. By illustrating these studies we attempt to show the potential role of structural bioinformatics in CYP research and help better understanding the importance of structural bioinformatics in drug designing.

Keywords Cytochrome P450 · Structural bioinformatics · Drug designing

T. Zhang (✉) · D. Wei (✉)
State Key Laboratory of Microbial Metabolism, College of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China
e-mail: taozhang0912@gmail.com

D. Wei
e-mail: dqwei@sjtu.edu.cn

T. Zhang
School of Biomedical Engineering, Tianjin Medical University, Tianjin, China

D. Wei et al. (eds.), *Advance in Structural Bioinformatics*, Advances in Experimental Medicine and Biology 827, DOI 10.1007/978-94-017-9245-5_19

19.1 Introduction

Drug discovery is a time-consuming, costly and complicated process involving many stages from initial target identification and validation, high throughput screening of compound library, lead identification and optimization to final selection of candidates for clinical evaluation [17]. Many advanced technologies are applied in these component stages, such as microscopic imaging techniques [4], embryonic stem cell technology [38], lab-on-a-chip technology [27], SNP technology [44] and so on. More recently, the impressive progress in genome sequencing, protein expression and high-throughput crystallography and NMR has significantly accelerated drug development [19]. In addition, protein structure is considered to play influential role in each stage throughout whole drug development process [21]. The methods of structural bioinformatics focused on macromolecular structure particularly protein structure also efficiently facilitated target identification and lead discovery [3].

Generally bioinformatics can be broadly divided into two branches, namely sequential bioinformatics and structural bioinformatics [7]. Compared with sequential bioinformatics, structural bioinformatics mainly focuses on macromolecular structure and reveals potential structure-function relationship. The main advantage of structural bioinformatics over other sequence-based methods lies in the fact that it can provide more detailed insights into the mechanisms by which biological events occur. Structural bioinformatics, therefore, can be considered as a crucial tool for deciphering the biological insights from macromolecular structure [6]. The methods of structural bioinformatics have been used to investigate many important biological processes such as blood coagulation [43], ion transfer through channel [5] and diverse functions of membrane proteins [29].

Cytochromes P450 (CYP) is an important family of oxidative enzymes that exists in many species and involves in the biosynthesis of endogenous substances and the metabolism of exogenous compounds. In particular, CYPs play a central role in the phase I-dependent metabolism of drugs and xenobiotics. It is estimated that CYPs can metabolize approximately 80 % drugs and other xenobiotics that are present in the human body [23]. Due to the great importance of CYPs for drug metabolism, CYPs have attracted substantial attention from the researchers in the field of drug design and development [13]. For example, whether a drug candidate can be properly metabolized by CYPs has become an important consideration in the process of drug development. In addition, CYP genes contain a large number of genetic polymorphisms, which are related to inter-individual variation in drug-metabolizing ability [54]. As a result, the genetic polymorphism of CYPs may cause unexpected serious clinic consequences. Now, taking into account of individual differences of CYPs' activity has also become a critical consideration in drug design [1].

Over the past several decades, there have been many research groups dedicated to CYPs research for understanding the process of CYP-mediated metabolism and unraveling the reaction mechanisms of CYPs. However, many important questions are still largely unanswered due to the complexity of CYPs. These questions

include how to distinguish their substrates, how to allosterically modulate other CYP activity and how to lead to individual variation in drug disposition and metabolism for the genetic polymorphisms of CYPs, etc. With more information available for CYP structures, the methods of structural bioinformatics may be used to address these questions and may provide possible solutions in which CYP protein structure and interaction with small molecules will be considered as the basis for comprehending these issues.

In this review, three previous studies in our laboratory are presented to demonstrate how the methods of structural bioinformatics have been applied to understand the effects of a surface mutation on CYP activity [51] and investigate the interaction of drugs bound within CYP active site with key residues [52] and construct predictive model for the likely outcomes of nsSNP [47]. By illustrating these studies we attempt to show the potential role of structural bioinformatics playing in the CYP research field that is an essential component for designing better drugs [9].

19.2 Study of Long-Range Effects of Peripheral Mutation in CYP1A2

19.2.1 CYP1A2 Structure and F186L Mutant

CYP1A2 is a commonly studied member of CYP family since it can metabolize about 20 % of clinical drugs [46]. CYP1A2 crystal structure bound with naphthoflavone has been determined using experimental method [37]. Similar to the structures of other CYP proteins, CYP1A2 crystal structure contains 12 alpha-helices and 4 beta-sheets (Fig. 19.1a). These secondary structural elements comprise both conserved and distinct regions among CYP structures. The conserved regions are associated with the proximal binding sites for heme prosthetic group and other redox partners such as cytochrome P450 reductase and cytochrome b5 [24]. The distinct regions constitute the distal surfaces of the substrate binding cavity of CYP1A2.

In vitro and in vivo experiments have found that some genetic variants significantly resulted in changes in CYP1A2 activity. To better understand the mechanism by which genetic variants affect the CYP1A2 activity, the methods of structural bioinformatics may be suitable to investigate the change in protein structure caused by the genetic variants. In the first demonstration we will show how to identify potential reasons for decrease in CYP1A2 activity caused by a mutation of F186L at the level of protein structure. We chose to study the mutation, F186L, due to three reasons. Firstly, the in vitro experiment showed that the O-deethylation reaction rates of 7-ethoxyresorufin and phenacetin were dramatically decreased to about 28 and 12.5 % of the wild-type, respectively. However, the F186L mutation did not perturb CYP 1A2 protein expression [26]. Secondly, based on the crystal structure of CYP1A2, F186 residue is situated on the flexible loop between helice D and E near

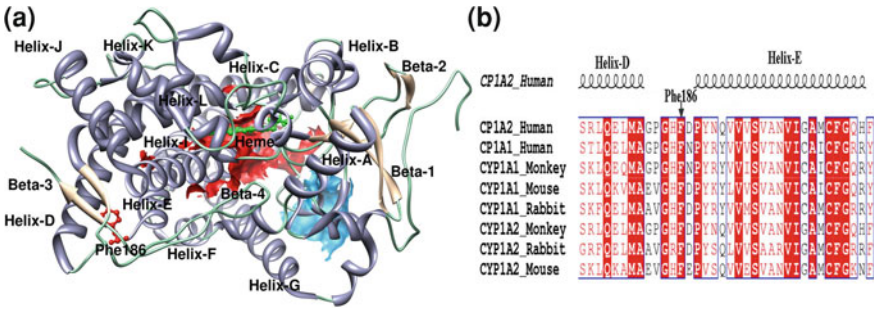


Fig. 19.1 Three dimensional structure of CYP1A2 and the conservation of F186 among eight CYP1A family proteins in four species. *Panel a* shows CYP1A2 crystal structure, in which 12 alpha-helices and 4 beta-sheets are marked and two conserved regions, i.e. the proximal binding sites for heme prosthetic group and redox partners, are colored *red* and *blue*, respectively. F186 residue is also shown in CYP1A2 structure and depicted in ball-and-stick model. *Panel b* is the partial multiple sequence alignment involving eight CYP1A family proteins

the surface of the enzyme, at about 26 Å away from the heme iron atom embedded inside the active site of CYP1A2 as shown in Fig. 19.1a. Thirdly, the multiple sequences alignment of CYP1A subfamily shown in Fig. 19.1b demonstrated 100 % conservation of the F186 residue, indicating its importance in maintaining the normal catalytic function of CYP1A2 [22]. In short, the strong effect on protein enzymatic activity, the large distance from active site and the high sequence conservation of residue contribute to the potential role of this mutation F186L as a long-range effector. Thus, the study of such peripheral mutation may elucidate the paths of long-range communication between F186L mutation site and active site and may also provide a promising direction for future studies of CYPs that have potential applications in related drug design.

19.2.2 Long-Range Effects of F186L Mutation

In this study we investigated this peripheral mutation by carrying out molecular dynamics simulation and related structural analyses including protein conformation analysis and access channel analysis. Based on results from these analyses, we compared overall and local structures between wild type and mutant and dynamically explored the formation and collapse of important access channels in both structures.

The structural comparison clearly showed that F186L mutation did not perturb the global protein conformation of CYP1A2, but increased the structural flexibility of the protein. Based on the detailed analysis of local structure, we found that the high flexibility of protein structure was attributed to a collective protein motion involving mostly the D, E, F helices and their inter-helical loops.

Furthermore, some of residues involved in such collective protein motion constituted the entrance to several main substrate access channels. We performed further investigation on access channel analysis and the results showed that the collective protein motion significantly affected the state of access channels. According to the state of access channels, the F186L mutant was found to exist in two sub-populations of conformational states. Two conformations correspond to the substrate access channel being either open or closed. Closure of the main access channel would lead to a decrease in enzymatic activity of the protein. Thus we proposed an "access mechanism" to rationally explain the long-range effects of the peripheral mutation F186L on the enzymatic activity of CYP1A2. Additionally, our results also demonstrate that F186L mutation may serve as an allosteric mutation [10] and the long-range effects of F186L are through structural flexibility change and population change of protein conformations.

19.3 Pharmacophore Model for Active Site of CYP1A2

19.3.1 CYP Catalytic Cycle

Since the first CYP catalytic cycle was proposed in 1973 [14], the mechanism of reaction catalyzed by CYPs has been extensively investigated using experimental and computational methods [2, 40]. Now it is generally accepted that several sequential steps consist of complete CYP reaction cycle including substrate binding, electron transformation, compound-I formation, hydrogen and electron extraction as well as products generation [25].

Substrate binding is generally considered as a rather complicated process in which several sub-steps are contained, such as substrate recognizing and entering as well as binding within active site [11]. In addition, substrate binding is also the initial event in the reaction cycle of CYP and consequently can trigger subsequent steps by changing the spin state of iron and reducing the redox potential of heme [48]. As a result, the study of substrate binding is of great importance to CYP-mediated drug metabolism.

19.3.2 CYP Substrate Binding

With more and more structures available for CYPs, many computational methods especially the methods of structural bioinformatics have been applied to investigate substrate binding [8, 30, 36]. For instance, few main channels transporting substrates and water molecules have been identified based on available crystal structures [45]. The key interactions between substrate and residues have been observed using molecular dynamics simulation and molecular docking in dynamic

manner [18, 42]. Other structure-based methods such as quantitative structure-activity relationship (QSAR) [35] and 3D pharmacophore mode [39] as well as the similarity analysis of fingerprint between protein and ligand [12] have also been applied to explore the relation between CYPs and their substrates.

19.3.3 Pharmacophore Model for Active Site of CYP1A2

In the second demonstration, we integrate three methods, namely molecular dynamics simulation, molecular docking and comparative molecular filed analysis (CoMFA), together to develop a pharmacophore model of CYP1A2 active site in terms of the role of residues in substrate binding. There are also three reasons for choosing CYP1A2 and its substrates as research subject. Firstly, as the major isoform among all CYP enzymes, CYP1A2 can not only metabolize many drugs but also active a number of procarcinages to carcinges [15, 46]. Thus, constructing such pharmacophore model can be quite useful to comprehensively understand the catalytic mechanism of CYP1A2. Secondly, CYP1A2 crystal structure demonstrates a narrow and planar active site with relatively small volume [37]. So it is feasible to conclude the role of key residues in substrates binding by superimposing different substrate-binding conformations. Last, but most important, the typical CYP1A2 substrates have also the planar aromatic group, so that they can properly fit the unique shape of CYP1A2 active site [55]. When other compounds with similar structure to CYP1A2 substrates are bound within the binding cavity of CYP1A2, the similar interactions between functional groups and key residues may occur. Therefore, the model developed from known substrates will be suitable for other compounds that are likely to be metabolized by CYP1A2.

In order to construct the pharmacophore model, 17 CYP1A2 substrates shown in Fig. 19.2 were collected together with related information about molecular name and the experimentally determined metabolite site as well as their Michaelis constant (Km) value.

Based on CYP1A2 protein and other molecular structures, we searched for possible binding conformation of these substrates within the active site of CYP1A2 using molecular dynamics simulation and docking method. We then defined the likely binding conformations from a large number of conformations. After aligning all selected binding conformations for 17 substrates, we performed CoMFA to determine the relation between substrate binding and molecular structure. On the basis of results from CoMFA we finally constructed the pharmacophore model specific to CYP1A2 active site, which is depicted in Fig. 19.3b and c.

All results of site-directed mutagenesis experiments and calculation of binding energy are well agreed with the model. From this model we can obtain important insights into drug metabolism and drug design. Moreover, this study indicates that this strategy of combining structural bioinformatics with other structure-based methods can also be applied into other CYP proteins or enzymes to obtain the detailed information about protein-ligand interactions.

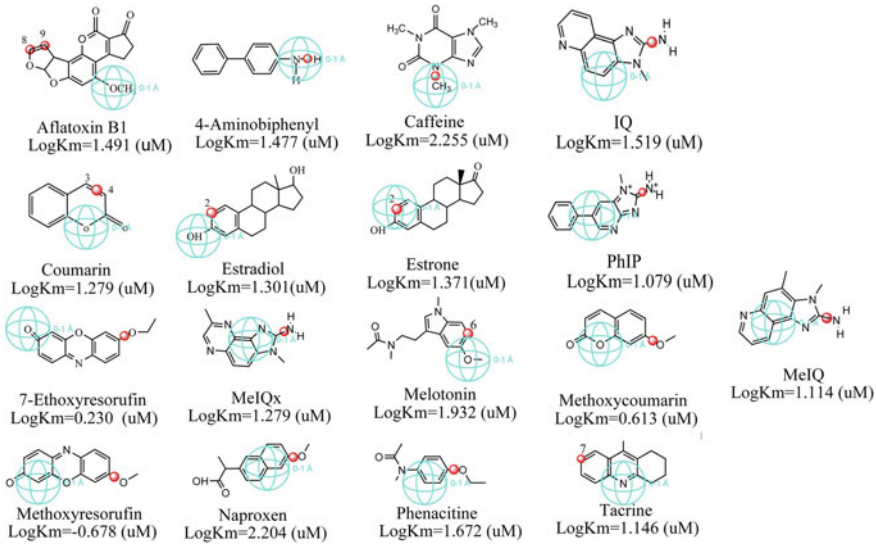


Fig. 19.2 The structure of 17 typical CYP1A2 substrates. Their molecular name and the logarithm value of Michaelis constant (Km) value are also listed below structure. *Red ball* represents the metabolite site of each substrate for CYP1A2 and *blue ball* is the exclusive sphere of molecule

19.4 Predicting Functional Consequence of Single Residue Mutation

19.4.1 Single Nucleotide Polymorphism

Single nucleotide polymorphisms (SNPs) are firmly associated with differences in phenotypes and disease susceptibility [49]. Typically, SNPs in protein coding regions may be considered to be more important because of their potential effects on protein structure and function [41]. These important SNPs are valuable for understanding the mechanism of disease. Therefore, connecting structural effects of residue mutations to their functional outcomes is also a major topics in structural bioinformatics [34]. Many computational tools have been developed to investigate the associations between amino acid mutations and disease in the context of protein structure [20, 28].

19.4.2 SNP and Drug Metabolism

As discussed previously, CYPs play an essential role in drug metabolism and can metabolize almost all clinically used drugs. Consequently, the genetic

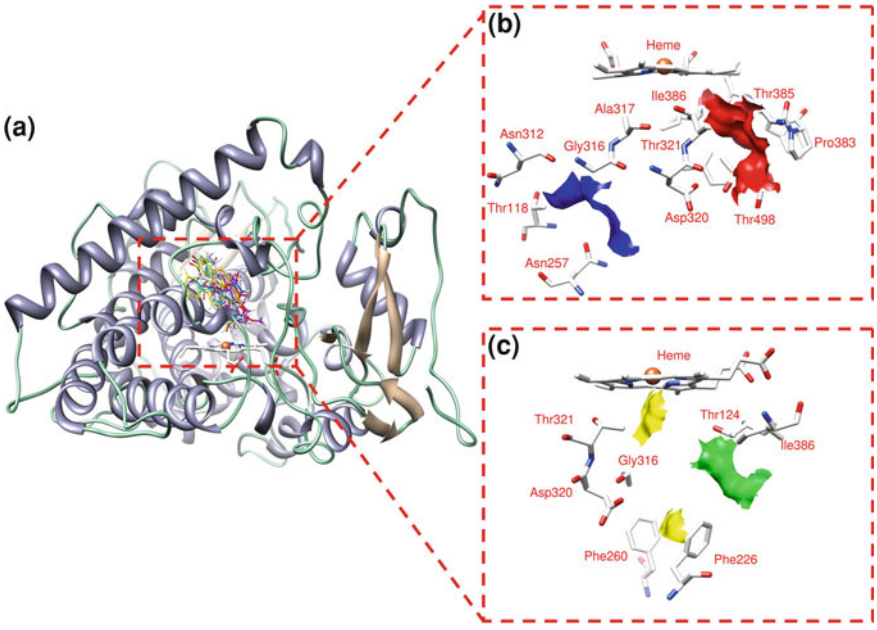


Fig. 19.3 The pharmacophore model specific to CYP1A2 active site. *Panel a* demonstrates the binding model of 17 substrates into the active site of CYP1A2. The electrostatic and steric maps in the pharmacophore model of CYP1A2 active site are depicted in *Panel b* and *c*, respectively. In *Panel b*, the regions favorable for negative and positive charged group are colored in *blue* and *red*, whereas in *Panel c* the *green* and *yellow* surfaces are the region favorable for *smaller* and *bulky* group

polymorphisms in CYP genes may be an important causative factor for the inter-individual differences in drug metabolism. Many investigations have proved that non-synonymous SNPs (nsSNPs) of CYP genes could significantly alter drug efficacy and pharmacokinetics [32]. For instance, the nsSNP enhancing metabolic activity of CYP can lead to rapid clearance of drugs from the body. Thus these drugs will not exert their therapeutic effects due to insufficient concentration in the blood. On the contrary, if the nsSNPs can abolish or reduce enzymatic activity of CYP, the drugs metabolized by such CYP cannot be metabolized safely and the associated side effects may occur. Although there are some database for the effects of nsSNPs on CYPs activity [33, 53], estimating such effects of nsSNPs is also important for investigating drug responses among different patients and predicting clinical implication of the novel genetic variants in CYP genes.

19.4.3 Predicting Functional Consequence of nsSNP

Alterations of CYP activity resulting from mutations are generally considered in the context of protein structure. Regarding possible structural mechanisms by which nsSNP and resultant mutation alter CYP activity, there are mainly several possible interpretations such as altering the physicochemical and geometric properties of CYP active site and disrupting the stability and folding of CYP enzyme [21]. Although the computational methods for predicting the effect of nsSNPs on protein function can be classified into two major types, sequence-based and structure-based methods, the latter type of method seems more rational. In addition, the enzyme kinetic results demonstrate that some nsSNPs may affect CYP activity in a substrate-dependant manner [16, 31]. Therefore, the information about structural characteristics of drugs that can be metabolized by CYP enzymes is also useful for predicting the effects of nsSNPs on CYP activity.

In the third demonstration presented here, for more accurate prediction of the effects of nsSNP on CYP activity we calculated two types of descriptors representing characteristics of CYP enzyme and specific drugs metabolized by such CYP, respectively. As shown in Fig. 19.4, the former descriptors mainly reflect the

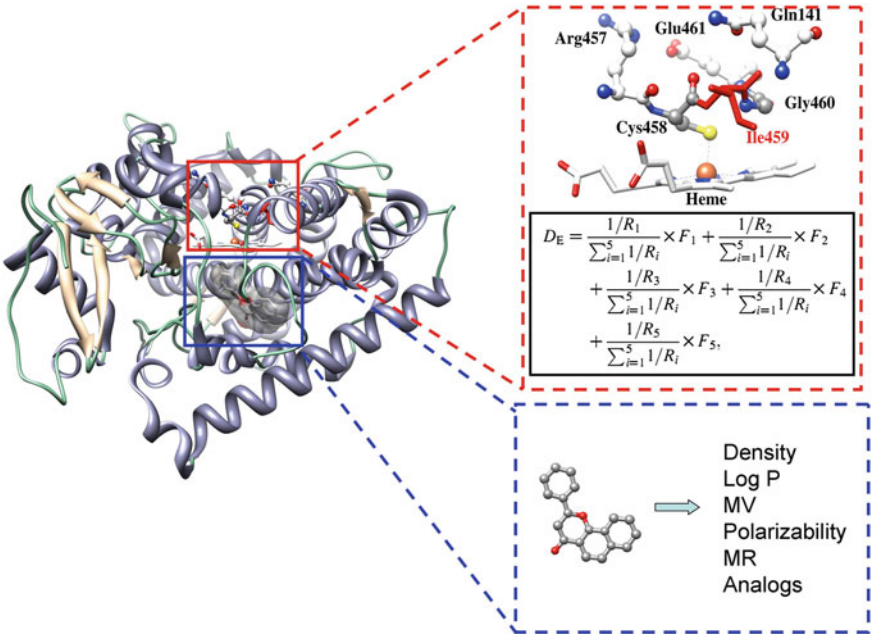


Fig. 19.4 Two type of descriptors calculated to predict the consequence of nsSNP. The protein-based descriptors reflected physiochemical properties changes are calculated using the weighted average features of five amino acids around mutated site according to CYP structure. The substrate-based descriptors are obtained from the 2D and 3D molecular structure

alternations of physiochemical properties caused by amino acid substitution. They include the features difference between original and substituted amino acid as well as the weighted average features of five amino acids around mutated site according to CYP structure. On the other hand, the latter descriptors represent the structural and chemical characteristics of drugs calculated from their 2D and 3D molecular structure.

Furthermore, we used these two classes of descriptors to develop the model for predicting the functional effect of nsSNPs on CYP activity. For constructing such model, total 134 nsSNPs from nine CYP enzymes were collected together with the relevant data from the drug metabolism studies. In addition, support vector machine (SVM) and genetic algorithm (GA) were also applied to select the relevant descriptors from all calculated descriptors. After multi-round selection, five protein descriptors and eight substrate descriptors were finally chosen to construct the computational model, by which we can roughly estimate the likely impact of nsSNPs on the CYP activity metabolizing a given substrate. Therefore, this model will be useful to facilitate the functional analysis of nsSNPs in CYP. Prior to experimental observation, we can firstly carry out preliminary investigation to examine whether the unknown single amino acid substitution can cause the functional changes in CYP enzymatic activity.

19.5 Conclusion

Currently structural bioinformatics has become an invaluable tool for CYPs research by offering the efficient tools and servers that enable visualization and analysis of CYP structure and interaction with drugs. Thus, at the atomic level, we can more deeply understand how CYP activity is affected by genetic, disease and environmental factor, and we also can make initial prediction whether the new chemical entity can be metabolized by CYPs [50]. Integrating these information into the drug development process will facilitate identification of the drug candidates with excellent metabolic profile. With the development and application of more powerful computational methods and tools, we believe that structural bioinformatics will play an increasingly important role in the CYPs research field and ultimately accelerate the course of drug development.

Refrences

1. Arınç E (2010) The role of polymorphic cytochrome P450 enzymes in drug design, development and drug interactions with a special emphasis on phenotyping. J Mol Catal B Enzym 64(3):120–122
2. Bathelt CM, Ridder L, Mulholland AJ, Harvey JN (2004) Mechanism and structure-reactivity relationships for aromatic hydroxylation by cytochrome P450. Org Biomol Chem 2(20):2998–3005

3. Blundell TL, Sibanda BL, Montalvão RW, Brewerton S, Chelliah V, Worth CL, Harmer NJ, Davies O, Burke D (2006) Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery. Philos Trans Roy Soc B: Biol Sci 361(1467):413–423
4. Bullen A (2008) Microscopic imaging techniques for drug discovery. Nat Rev Drug Discovery 7(1):54–67
5. Capener CE, Kim HJ, Arinaminpathy Y, Sansom MS (2002) Ion channels: structural bioinformatics and modelling. Hum Mol Genet 11(20):2425–2433
6. Chandra N, Anand P, Yeturu K (2010) Structural bioinformatics: deriving biological insights from protein structures. Interdisc Sci: Comput Life Sci 2(4):347–366
7. Chou K-C (2006) Structural bioinformatics and its impact to biomedical science and drug discovery. Front Med Chem 3(1):455–502
8. Cojocaru V, Winn PJ, Wade RC (2007) The ins and outs of cytochrome P450 s. Biochimica et Biophysica Acta (BBA)-General Subjects 1770(3):390–401
9. De Groot MJ (2006) Designing better drugs: predicting cytochrome P450 metabolism. Drug Discov Today 11(13):601–606
10. del Sol A, Tsai C-J, Ma B, Nussinov R (2009) The origin of allosteric functional modulation: multiple pre-existing pathways. Structure 17(8):1042–1050
11. Denisov IG, Makris TM, Sligar SG, Schlichting I (2005) Structure and chemistry of cytochrome P450. Chem Rev 105(6):2253–2278
12. Freitas R, Bauab R, Montanari C (2010) Novel application of 2D and 3D-similarity searches to identify substrates among cytochrome P450 2C9, 2D6, and 3A4. J Chem Inf Model 50(1):97–109
13. Guengerich FP (2002) Cytochrome P450 enzymes in the generation of commercial products. Nat Rev Drug Discov 1(5):359–366
14. Guengerich FP, Macdonald TL (1984) Chemical mechanisms of catalysis by cytochromes P-450: a unified view. Acc Chem Res 17(1):9–16
15. Gunes A, Dahl M-L (2008) Variation in CYP1A2 activity and its clinical implications: influence of environmental factors and genetic polymorphisms. Pharmacogenomics 9(5):625–637
16. Han S, Choi S, Chun Y-J, Yun C-H, Lee CH, Shin HJ, Na HS, Chung MW, Kim D (2012) Functional characterization of allelic variants of polymorphic human cytochrome P450 2A6 (CYP2A6* 5,* 7,* 8,* 18,* 19, and* 35). Biol Pharm Bull 35(3):394–399
17. Hughes J, Rees S, Kalindjian S, Philpott K (2011) Principles of early drug discovery. Br J Pharmacol 162(6):1239–1249
18. Ito Y, Kondo H, Goldfarb PS, Lewis DF (2008) Analysis of CYP2D6 substrate interactions by computational methods. J Mol Graph Model 26(6):947–956
19. Kalyaanamoorthy S, Chen Y-PP (2011) Structure-based drug design to augment hit discovery. Drug Discov Today 16(17):831–839
20. Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4(7):1073–1081
21. Lahti JL, Tang GW, Capriotti E, Liu T, Altman RB (2012) Bioinformatics and variability in drug response: a protein structural perspective. J R Soc Interface 9(72):1409–1437
22. Lewis DF (2004) 57 varieties: the human cytochromes P450. Pharmacogenomics 5(3):305–318
23. Lewis DF, Ito Y (2010) Human CYPs involved in drug metabolism: structures, substrates and binding affinities. Expert Opin Drug Metab Toxicol 6(6):661–674
24. Mestres J (2005) Structure conservation in cytochromes P450. Proteins: Struct, Funct, Bioinf 58(3):596–609
25. Meunier B, De Visser SP, Shaik S (2004) Mechanism of oxidation reactions catalyzed by cytochrome P450 enzymes. Chem Rev 104(9):3947–3980
26. Murayama N, Soyama A, Saito Y, Nakajima Y, Komamura K, Ueno K, Kamakura S, Kitakaze M, Kimura H, Y-i Goto (2004) Six novel nonsynonymous CYP1A2 gene

polymorphisms: catalytic activities of the naturally occurring variant enzymes. J Pharmacol Exp Ther 308(1):300–306
27. Neuzi P, Giselbrecht S, Lange K, Huang TJ, Manz A (2012) Revisiting lab-on-a-chip technology for drug discovery. Nat Rev Drug Discov 11(8):620–632
28. Ng PC, Henikoff S (2006) Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet 7:61–80
29. Nugent T, Jones DT (2012) Membrane protein structural bioinformatics. J Struct Biol 179(3):327–337
30. Otyepka M, Skopalík J, Anzenbacherová E, Anzenbacher P (2007) What common structural features and variations of mammalian P450 s are known to date? Biochimica et Biophysica Acta (BBA)-General Subjects 1770(3):376–389
31. Palma BB, e Sousa MS, Vosmeer C, Lastdrager J, Rueff J, Vermeulen N, Kranendonk M (2010) Functional characterization of eight human cytochrome P450 1A2 gene variants by recombinant protein expression. Pharmacogenomics J 10(6):478–488
32. Pilgrim J, Gerostamoulos D, Drummer OH (2011) Review: pharmacogenetic aspects of the effect of cytochrome P450 polymorphisms on serotonergic drug metabolism, response, interactions, and adverse effects. Forensic Sci Med Pathol 7(2):162–184
33. Preissner S, Kroll K, Dunkel M, Senger C, Goldsobel G, Kuzman D, Guenther S, Winnenburg R, Schroeder M, Preissner R (2010) SuperCYP: a comprehensive database on Cytochrome P450 enzymes including a tool for analysis of CYP-drug interactions. Nucleic Acids Res 38(suppl 1):D237–D243
34. Reumers J, Schymkowitz J, Rousseau F (2009) Using structural bioinformatics to investigate the impact of non synonymous SNPs and disease mutations: scope and limitations. BMC Bioinform 10(Suppl 8):S9
35. Ringsted T, Nikolov N, Jensen GE, Wedebye EB, Niemelä J (2009) QSAR models for P450 (2D6) substrate activity. SAR QSAR Environ Res 20(3–4):309–325
36. Rydberg P, Rod TH, Olsen L, Ryde U (2007) Dynamics of water molecules in the active-site cavity of human cytochromes P450. J Phys Chem B 111(19):5445–5457
37. Sansen S, Yano JK, Reynald RL, Schoch GA, Griffin KJ, Stout CD, Johnson EF (2007) Adaptations for the oxidation of polycyclic aromatic hydrocarbons exhibited by the structure of human P450 1A2. J Biol Chem 282(19):14348–14355
38. Sartipy P, Bjorquist P, Strehl R, Hyllner J (2007) The application of human embryonic stem cell technologies to drug discovery. Drug Discov Today 12(17):688–699
39. Schuster D (2011) 3D pharmacophores as tools for activity profiling. Drug Discov Today: Technol 7(4):e205–e211
40. Shaik S, Kumar D, de Visser SP, Altun A, Thiel W (2005) Theoretical perspective on the structure and mechanism of cytochrome P450 enzymes. Chem Rev 105(6):2279–2328
41. Shastry BS (2009) SNPs: impact on gene function and phenotype. In: Single nucleotide polymorphisms. Springer, Berlin pp 3–22
42. Sun H, Scott DO (2010) Structure-based drug metabolism predictions for drug design. Chem Biol Drug Des 75(1):3–17
43. Villoutreix BO (2002) Structural bioinformatics: methods, concepts and applications to blood coagulation proteins. Curr Protein Pept Sci 3(3):341–364
44. Voisey J, Morris CP (2008) SNP technologies for drug discovery: a current review. Curr Drug Discov Technol 5(3):230–235
45. Wade RC, Winn PJ, Schlichting I (2004) A survey of active site access channels in cytochromes P450. J Inorg Biochem 98(7):1175–1182
46. Wang B, Zhou S-F (2009) Synthetic and natural compounds that interact with human cytochrome P450 1A2 and implications in drug development. Curr Med Chem 16(31):4066–4218
47. Wang Y, Zhou Q, Dai H, Zhang T, Wei D-Q (2012) Prediction of the functional consequences of single amino acid substitution in human cytochrome P450. Mol Simul 38(14–15):1297–1307

48. Werck-Reichhart D, Feyereisen R (2000) Cytochromes P450: a success story. Genome Biol 1(6):3003.3001–3003.3009
49. Zhang F, Gu W, Hurles ME, Lupski JR (2009) Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet 10:451–481
50. Zhang T, Chen Q, Li L, Angela Liu L, Wei D-Q (2011) In silico prediction of cytochrome P450-mediated drug metabolism. Comb Chem High Throughput Screening 14(5):388–395
51. Zhang T, Liu LA, Lewis DF, Wei D-Q (2011) Long-range effects of a peripheral mutation on the enzymatic activity of cytochrome P450 1A2. J Chem Inf Model 51(6):1336–1346
52. Zhang T, Wei D-Q, Chou K-C (2012) A pharmacophore model specific to active site of CYP1A2 with a novel molecular modeling explorer and CoMFA. Med Chem 8(2):198–207
53. Zhang T, Zhou Q, Pang Y, Wang Y, Jin C, Huo J, Liu LA, Wei D (2012) CYP-nsSNP: A specialized database focused on effect of non-synonymous SNPs on function of CYPs. Interdisc Sci: Comput Life Sci 4(2):83–89
54. Zhou S-F, Liu J-P, Chowbay B (2009) Polymorphism of human cytochrome P450 enzymes and its clinical impact. Drug Metab Rev 41(2):89–295
55. Zhou S-F, Yang L-P, Zhou Z-W, Liu Y-H, Chan E (2009) Insights into the substrate specificity, inhibitors, regulation, and polymorphisms and the clinical impact of human cytochrome P450 1A2. AAPS J 11(3):481–494

Chapter 20
Human Cytochrome P450 and Personalized Medicine

Qi Chen and Dongqing Wei

Abstract Personalized medicine has become a hot topic ascribed to the development of Human Genome Project. And currently, bioinformatics methodology plays an essential role in personal drug design. Here in this review we mainly focused on the basic introduction of the SNPs of human drug metabolic enzymes and their relationships with personalized medicine. Some common bioinformatics analysis methods and latest progresses and applications in personal drug design have also been discussed. Thus bioinformatics studies on SNPs of human CYP450 genes will contribute to indicate the most possible genes that are associated with human diseases and relevant therapeutic targets, identify and predict the drug efficacy and adverse drug response, investigate individual gene specific properties and then provide personalized and optimal clinic therapies.

Keywords Personalized medicine · SNPs · CYP450

20.1 Introduction of Human Cytochrome P450

With the rapid development of Human Genome Project and International HapMap Project, the whole sequencing of human genome has completed, which has greatly enriched the analysis datasets and experiment references for the researches of human diseases and drug design. In 1959, Vogel proposed an idea of Pharmacogenetics that mainly focused on the researches of genetic variations of human beings and resulted different drug responses using large amount of genomics information and methods, so as to improve and discover better methods for new drug design and medical therapy [1]. Since then, continuously studies about

Q. Chen · D. Wei (✉)
State Key Laboratory of Microbial Metabolism, College of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China
e-mail: dqwei@sjtu.edu.cn

D. Wei et al. (eds.), *Advance in Structural Bioinformatics*, Advances in Experimental Medicine and Biology 827, DOI 10.1007/978-94-017-9245-5_20

pharmacogenetics both from experimental and computational respects have been held and reported in literatures.

According to the pharmacodynamics theory, drugs go thought absorption, distribution, metabolism and excretion (ADME) after they are taken into human bodies; the ADME properties are essential for efficient investigations of new drugs or candidates. Among the four steps, metabolism is the most important and meaningful one owes to its inclusion of nearly 40 % of the pharmacodynamics interactions. The metabolic systems of drugs can be regarded as two phases, phase I and phase II. Oxidation, reduction, and hydrolysis of xenobiotics are involved in phase I, while phase II contains synthesis and conjugation of phase I products [2]. Different kinds of drug metabolic enzymes participate during the whole process, such as cytochrome P450, N-acetyl Transferases, Flavin Monooxygenases, Esterase, Alcohol dehydrogenase (ADH) that work as phase I enzymes and Methyltransferase, Glutathione S-transferases, Sulfotransferases, N-acetyltransferases, UDP-glucuronosyltransferases that work as phase II enzymes.

Among all of phase I drug metabolic enzymes, cytochrome P450 superfamily (officially abbreviated as CYP, also named CYP450) acts as an important role responsible for the oxidation of endogenous substrates and xenobiotic compounds like fatty acids, steroids, toxins and 90 % currently used drugs [3–5]. The catalytic reaction can be summarized as $RH + O_2 + 2H^+ + 2e^- \rightarrow ROH + H_2O$, where RH represents different kinds of substrates. CYP450 is named because these enzymes have a property to form a complex with CO (Carbon monoxide) and produce a spectrally identifiable absorption peak at 450 nm. CYP450 enzymes can be found widely distributed in intestines, liver, lung, kidney and brain of human body; as well as other species like animals, plants, microorganisms. All the CYP genes can be divided mainly into CYP1, CYP2, and CYP3 three subfamilies according to the sequence similarity of their amino acid [3]. There are more than 50 CYP450 enzymes, among which CYP1A2, CYP2C9, CYP2C19, CYP2D6, CYP3A4, and CYP3A5 are the major drug metabolizing isoforms that metabolize 90 % of drugs. Figure 20.1 demonstrates the main distribution of drug metabolism by different families of CYP enzymes.

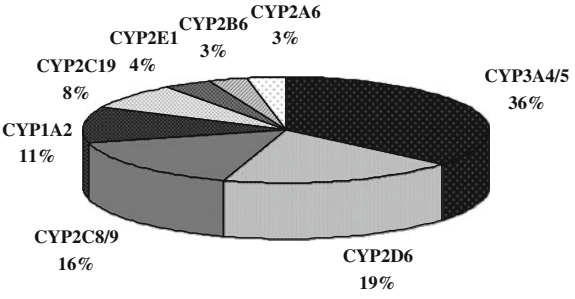


Fig. 20.1 Distribution of drugs metabolized by different CYP450s

20.2 Introduction of Personalized Medicine

In current clinical treatments, the metabolism situation of drugs will be of variety for different patients, as well as the drug efficiency and drug side effects. Drugs will not have expected efficacy to 30–40 % patients in some circumstance, besides some adverse drug reactions can be observed in some patients to whom sometimes much worse effect will happen.

Take Warfarin for example, Warfarin (Coumadin) was introduced into clinical use in the 1950s to treat and prevent thromboembolic disease. However it is very difficult to manage the therapy well as the therapeutic index of this drug is very narrow and patient responses are different individually. Relative studies have already been taken and it is now clear that CYP2C9 is strongly related to Warfarin responsiveness [6, 7].

The fact that we can't predict the exact efficacy and side effect of a drug preclude the development of pharmacy industry, more and more research institutions and companies have paid attention to this area recently. Several factors are responsible for various drug responses, such as age, sex, weight and genetic factors. Numerous researches about CYP450 genomics represent that the mutations occur in part of the CYP450 genes are mainly contribute to why different people has different drug reactions.

These pheromones can be named as the Single Nucleotide Polymorphism (SNP), which is the mutation of a single nucleotide like substitution, insertion or missing with in a DNA sequence. As known to all, human DNA is composed of four kinds of nucleotide which are A (adenine), G (guanine), C (cytosine), and T (thymine), SNP occurs when one of them is replaced by the others. Among all three billion DNA base pairs, 99.9 % are identical for all the human beings and the rest 0.1 % make everyone unique in the world. If over 85 % of these differences are SNPs, which will be potentially 3 million base pairs total.

Up to March, 25, 2010, 23,653,737 human SNPs have been identified and updated to Single Nucleotide Polymorphism (dbSNP; http://www.ncbi.nlm.nih.gov/projects/SNP; dbSNP Build 131) in NCBI Database. SNPs will affect the gene transcription/translation or structure of proteins, thus it will change the function of CYP450s in some situations. For instance, among all the CYP450 enzymes, CYP2C19 and CYP2D6 represent the most significant individual difference in humans. 5–10 % in a Caucasian population will act as poor metabolizers for anti-arrhythmia agent like Metoprolol and Propafenone because of gene mutations in CYP2D6. Some patients with CYP2C19 gene variant will be sensitive to Phenytoin and Cyclobarbital, or even cause toxicity reactions [8].

Abundant of high throughput experimental methods have been proposed for SNPs identification such as single-strand conformation polymorphism (SSCP), conformation-sensitive gel electrophoresis (CSGE), chemical cleavage of mismatch (CCM), allele specific PCR (AS-PCR), allele specific oligonucleotide hybridization (ASO), DNA chip, pyrosequencing for SNP genotyping, matrix assisted laser desorption ionization-time of flight mass spectrometry (MAIDI-

TOF), denaturing high performance liquid chromatography (DHPLC) [9, 10]. On the other hand, computational methods have been utilized as a faster and cheaper tool for detecting SNPs such as multiple sequence alignments based on expressed sequence tag (EST) or sequence tagged sited (STS), multiple linear regression (MLR), support vector machines (SVM), learning machine an so on [11–14].

Variations of SNPs in human CYP450 genes will cause different drug effects. Studies on SNPs can be used as predictive markers from different aspects in medical care area, including disease-causing genes, drug efficacy and even adverse effect of various drugs. As there are several different factors that will cause human disease such as environment, lifestyle or genes, it is very difficult to apply screening test methods to most diseases like cardiovascular diseases, Alzheimer's disease, and diabetes. SNPs studies will provide some fundamental understanding about these diseases, which will indicate the most possible genes that are associated with a disease and relevant therapeutic targets.

As we all know, large percent of patients will have positive responses after taking a drug while some others will not benefit from it or even die of it. So another purpose for SNPs studies is to identify and predict the drug efficacy and adverse drug response from pharmacogenomics point of view. This idea leads us a new way in medical therapy named as "personalized medicine". Personalized medicine (also named as personalized therapy) is a new idea based on individual pharmacogenetics and pharmacogenomics information. As variances of human genes will lead different sensibility of disease in individual people, personalized medicine will focus on investigating individual gene specific properties and then providing personalized and optimal clinic therapy. For example, Genetic variation in human CYP genes CYP2D6, CYP2C19, and CYP2C9 will cause metabolism influences of neuroleptic drugs. These information has been applied for antipsychotics usage decisions which can reduce the drug side effects by almost 20 % [15, 16].

20.3 Data Resources for Researches on SNPs and CYP450s

Computer-Aided Drug Design (CADD) plays a very important role in pharmacy not only because it is faster, cheaper and much more efficient, but also because it leads a new research direction to screen important micromolecules which are essential to human disease and drug design. With the rapid development of bioinformatics theory, more and more bioinformatics efforts have been tried and used in pharmacogenomics such as predicting gene variations which are likely to have some functional or genotypic effects and classifying associated downstream molecular effects.

To start researches on SNPs and CYP450s for drug design through bioinformatics method, some particular and reliable database are necessary. There are several famous bioinformatics database distributed in the USA, Europe and Japan. Table 20.1 lists the most important database used in bioinformatics research and drug design [17, 18]. Besides the traditional databases, some new specified data

Table 20.1 Common databases

Nucleotide sequence database	Genbank	http://www.ncbi.nlm.nih.gov/Genbank/
	EMBL (European Molecular Biology Laboratory)	http://www.ebi.ac.uk/embl.html
	DDBJ (DNA Data Bank of Japan)	http://www.ddbj.nig.ac.jp
Protein sequence database	SWISS-PROT and TrEMBL	http://www.expasy.org/sprot/
	PIR and PSD	http://pir.georgetown.edu/
	OWL	http://bioinfman.ac.uk/dbbrowser/OWL/
	NRL3D	http://www.gdb.org/Dan/proteins/nrl3d.html
Protein domain databases	PROSITE	http://www.expasy.org/prosite
	BLOCKS	http://www.blocks.fhcr.org/blocks/
	PRINTS	http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/PRINTS.html
3D structure database of proteins	PDB	http://www.pdb.org/
	NDB (Nucleic Acid Databank)	http://ndbserver.rutgers.edu/
	BioMagResBank	www.bmrb.wisc.edu
	Protein structure database	http://www.rcsb.org/pdb
	CCSD (Complex carbohydrate structure database)	http://www.boc.chem.uu.nl/sugabase/carbbank.html
The data on human CYP genes	Entrez gene on the NCBI web site	http://www.ncbi.nlm.nih.gov/sites/entrez
	HUGO gene database	http://www.genenames.org/
Phenotype of SNPs of human CYP genes	PubMed	http://www.ncbi.nlm.nih.gov/PubMed/
	OMIM (Online Mendelian Inheritance in Man)	http://www.ncbi.nlm.nih.gov/omim
	UniProtKB/Swiss-Prot databases	http://ca.expasy.org/sprot/
	Human gene mutation database	http://www.hgmd.cf.ac.uk

sources have been developed and published along with various web resources such as Nelsons Homepage [19], PubChem [20], the P450 Knowledgebase designed by The Center for Molecular [21], and so on. More details can be found in Table 20.2.

Information about CYP450 interactions usually can be found in huge amounts of publications. Thus another essential way to gain useful data for future researches is to extract meaningful information from latest biomedical literature. Several computational approaches have been developed for data extracting [22]. A natural

Table 20.2 Specified data sources

Nelsons homepage	http://drnelson.uthsc.edu/nelsonhomepage.html
Flockharts interaction table	http://www.medicine.iupui.edu/Flockhart/table.htm
University of Maryland's drug checker	http://www.umm.edu/adam/drug_checker.htm
PubChem	http://pubchem.ncbi.nlm.nih.gov/
The cytochrome P450 homepage	http://drnelson.utmem.edu/CytochromeP450.html
PharmGKB (The Pharmacogenomics Knowledge Base)	http://www.pharmgkb.org/
SuperCYP	http://bioinformatics.charite.de/supercyp/
The human CYP450 allele nomenclature committee	http://www.imm.ki.se/CYPalleles/
The directory of P450-containing systems	http://www.icgeb.trieste.it/
The P450 knowledgebase	http://cpd.ibmh.msk.su/
The P450s in PROMISE	http://metallo.scripps.edu/PROMISE/P450.html

language processing (NLP) based method has been proposed and applied for extracting relationships information from literatures [23, 24]. However, the NLP systems can't be directly performed as an analysis tool from genes and proteins to other fields like interactions between exogenous chemicals and biomolecules. A NLP-based specialized system was carried out for extracting information about chemical-enzyme interactions from the literature by Mitsuru Hashida et al.. The system was concluded to be very feasible and powerful by a test research on extracting relationships between chemicals and CYP3A4. Those gained dataset is important for predicting CYP450 associated drug-drug interactions and subsequent drug discovery and clinical applications. Xia Yang performed a systematic analysis of CYP450 enzyme activities in human liver, genetics, gene expression, and enzyme activity measurements were integrated and investigated using systems biology approaches. Human liver transcriptional network structure was then defined using a weighted coexpression network and a Bayesian regulatory network. Several activity SNPs which are strongly associated with CYP450 enzyme activities were identified. This review provides comprehensive information about the functionality, genetic control, and interactions of CPY450s [25].

20.4 Bioinformatics Applications

So far, polymorphisms in CYP genes have been studied widely for investigating their potential implication of human disease.

Polymorphisms in cytochrome P450 1A1 (CYP1A1) gene have been analyzed in the context of oral and pharyngeal cancer, lung cancer, prostate cancer, esophageal cancer and breast cancer. Using the same idea of meta-analysis, by searching the MEDLINE bibliographical database for suitable articles and performing sensitivity analysis, three SNPs in CYP1B1 have been studied to check whether they are associated with breast cancer risk [26]. Genetic polymorphisms in CYP2E1 were reviewed and evaluated for discover their affects on CYP2E1 function as they were proved linked to altered susceptibility to hepatic cirrhosis. According to the known knowledge about the effects of the polymorphisms and their frequency in the population, a population distribution of CYP2E1 activity can be constructed which will be very useful for further SNPs research [27]. Congenital adrenal hyperplasia (CAH) is one of the most common metabolic disorders due to 21-hydroxylase deficiency [28]. CYP21A2 is proved to be a steroid 21-hydroxylase enzyme while a series of deleterious mutations can be found. Concolino et al. (2010) reviewed and updated the recently CYP21A2 mutations based on the CYP21A2 database created by the Human Cytochrome P450 Allele Nomenclature Committee. The molecular and genetic diagnosis of 21-hydroxylase deficiency was reported as well for later studies [29].

Polymorphisms in CYP genes will also affect the clinical usage of drugs. SNPs in CYP genes encoding can lead to lower enzyme activity which will decrease the enzymatic metabolism of particular drugs that are substrates of these enzymes. SNPs in CYP2D6 and CYP2C19 have already been proved to cause different adverse drug effects during the treatment cardiovascular diseases, psychiatric disorders and cancer. CYP2D6*4 in Caucasians has been concluded to be the most common variant allele that acted "poor metabolism" of CYP2D6 substrates [30]. Relationships between CYP2D6*4, CYP3A5*3 and ABCB1 3435T polymorphisms and drug related falls have been investigated using multivariate logistic regression method [31]. Sofi F et al. focused on the association with the loss-of-function CYP2C19*2 (or 681 G > A) polymorphism and coronary artery disease (CAD) in patients taking clopidogrel, the results turned out that the CYP2C19*2 polymorphism is associated with the cardiovascular disease and stent thrombosis in an increased risk [32]. Testing the SNPs in human CYP450 gene can further be used for evaluating the clinical effectiveness and cost-effectiveness of specific drugs. Fleeman et al. reviewed the analytical validity, clinical validity and clinical utility as well as economic evaluations of CYP polymorphisms testing in patients with schizophrenia treated with antipsychotics [33].

20.5 Structure-Based Analysis

As more and more 3-D structure of the proteins have been identified, structure-based analysis has been developed and applied for analysis of target protein for drug design. Normally after getting the 3-D structure of the target protein, computational approaches are used to identify the initial drug candidates. Virtual

screening is one of the most widely used approaches which perform the docking analysis of the drug candidates on the active site of the target protein. Virtual screening will provide a scoring function about the computational estimation of binding free energy, binding constant, docking score and so on, drug candidates which have high binding affinity are selected and tested in Vitro and in vivo later [34, 35].

Significant effects have been made in CYP450 three-dimensional (3D) structure and mechanism investigation during the last few years [36–38]. Abundant of structure, activity and regulation information about CYP450 contribute to the development of computational model approaches for predicting CYP-related metabolism properties, detecting CYP450 SNPs, and identifying their associated implications for drug design. Those modeling methodologies can be classified as ligand-based, structure/protein-based, and ligand–protein interaction based approaches according to the recent review [39].

Ligand-based approaches use structure information of the molecules interacting with the target of interest such as CYP450s, while in structure-based methods docking techniques are performed to find the possible binding modes of a ligand to a receptor based on the structure information.

For the ligand–protein interaction based approaches both ligand and protein information is involved [40]. Ligand-based approaches include various QSAR (quantitative structure activity relationships) methods based on the ligand structure information. QSAR can be performed through 2D or 3D structures and have been widely used in drug discovery. As a methodology based on the assumption that compounds can be mathematically defined as the distribution of molecular descriptors, QSAR has been proved to be a powerful virtual-model based tool for predicting pharmacodynamic, pharmacokinetic or toxicological properties as well as quantities like binding affinity and molecules' toxic potential. Furthermore, chemical structure and metabolic properties are linked together quantitatively in QSAR [41]. 3D-QSAR was induced for analyzing quantitative biological activity of 3D structures of the ligands and electrostatic, steric, hydrophobic and hydrogen bond fields. 3D-QSAR approach is applicable to more heterogeneous datasets. QSAR modeling method contains several steps as follows: (1) dataset collection (2) molecular descriptors calculation (3) model generation and optimization (4) data updating [42].

By combining molecular biology, molecular dynamics, quantum chemistry and graphical display system together, Structure-based Drug Design (SBDD) was proposed based on the structures of receptors for revealing the molecule interaction mode between ligands and receptors and guiding computational drug design. SBDD methods contains analysis procedures about receptor and associated ligands based on either an X-ray or NMR structure of the ligand, QM or QM/MM methods, homology modeling, energy decomposition methods are all involved [43].

The methodology to combine molecular docking operation with the molecular dynamics (MD) simulations together to predict possible binding sites of the SNPs was wildly utilized. MD simulations can solve the classical equations of motions for a system formed by target protein (SNPs) and small ligands such as drug

molecular Fluvoxamine, Lescol and Ticlopidine under specified ensembles. In this research, two SNPs of CYP2C19 which are named as W120R and I331 V were chosen as the investigation targets [44]. Then docking simulation was performed for W120R and I331 V binding reactions with selected ligands to find the most possible binding site of these two SNPs. Docking results were verified by molecular dynamics simulations.

Docking and molecular dynamics have also been widely used in researches about cancer targets for identifying the recognition processes between ligands and targets at the atomic level as well as affinity or conformational changes of the molecular complexes. It could be very helpful for corresponding drug design and development of improved drug efficacy for individuals [45].

20.6 Conclusion

Personalized medicine has become a hot topic ascribed to the development of Human Genome Project. One key point to achieve personalized medicine personalized drug design is researches about the variance genetic polymorphisms in human drug metabolism enzymes. Great efforts have been taken both in experimental and computational aspects, large abundant of datasets haven been published and gathered.

Among all the methods, bioinformatics methodology plays an essential role in personal drug design. Here in this review we mainly focused on the basic introduction of the SNPs of human drug metabolic enzymes and their relationships with personalized medicine. Some common bioinformatics analysis methods and latest progresses and applications in personal drug design have also been discussed.

Thus bioinformatics studies on SNPs of human CYP450 genes will contribute to indicate the most possible genes that are associated with human diseases and relevant therapeutic targets, identify and predict the drug efficacy and adverse drug response, investigate individual gene specific properties and then provide personalized and optimal clinic therapies.

References

1. Vogel F (1959) Moderne probleme der humangenetik. Ergebn Inn Med Kinderheilk 12:52–60
2. Tanmay SP, Shaun FN, Sanish D, Vishal MS, Nilima AK, Nithya JG (2006) Evaluation of the activity of CYP2C19 in Gujrati and Marwadi subjects living in Mumbasi (Bombay). BMC Clin Pharm 6(8):1–5
3. Lewis DFV (1998) The CYP2 family: models, mutants and interactions. Xenobiotica 28(77):617–661
4. Sheweita SA (2000) Drug-metabolizing enzymes: mechanisms and functions. Curr Drug Metab 1(2):107–132

5. Nebert DW, Russell DW (2002) Clinical importance of the cytochrome P450. Lancet 360(9340):1155–1162
6. Rettie AE, Tai G (2006) The pharmocogenomics of warfarin closing in on personalized medicine. Mol Interv 6(4):223–227
7. Yang S, Xu L, Wu HM (2010) Rapid genotyping of SNPs influencing warfarin drug response by SELDI-TOF mass spectrometry. J Mol Diagn 12(2):162–168
8. Sadée W, Dai Z (2005) Pharmacogenetics/genomics and personalized medicine. Hum Mol Genet 14(2):207–214
9. Kim S, Misra A (2007) SNP genotyping: technologies and biomedical applications. Annu Rev Biomed Eng 9:289–320
10. Ding C, Jin S (2009) High-throughput methods for SNP genotyping. Methods Mol Biol 578:245–254
11. Matukumalli LK, Grefenstette JJ, Hyten DL, Choi IY, Cregan PB, Van Tassell CP (2006) Application of machine learning in SNP discovery. BMC Bioinf 7:4
12. Chorley BN, Wang X, Campbell MR, Pittman GS, Noureddine MA, Bell DA (2008) Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: current and developing technologies. Mutat Res 659(1–2):147–157
13. Mooney SD, Krishnan VG, Evani US (2010) Bioinformatic tools for identifying disease gene and SNP candidates. Methods Mol Biol 628:307–319
14. Wang J, Zou Q, Guo MZ (2010) Mining SNPs from EST sequences using filters and ensemble classifiers. Genet Mol Res 9(2):820–834
15. Arranz MJ, de Leon J (2007) Pharmacogenetics and pharmacogenomics of schizophrenia: a review of last decade of research. Mol Psychiatry 12(8):707–747
16. Arranz MJ, Kapur S (2008) Pharmacogenetics in psychiatry: are we ready for widespread clinical use? Schizophr Bull 34(6):1130–1144
17. Bairoch A, Apweiler R (1999) The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. Nucleic Acids Res 27(1):49–54
18. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28(1):235–242
19. Nelson DR (2006) Cytochrome P450 nomenclature, 2004. Methods Mol Biol 320:1–10
20. Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, He S, Liu C, Shi W, Bryant SH (2010) The NCBI biosystems database. Nucleic Acids Res 38:492–496
21. Sim SC, Ingelman-Sundberg M (2006) The human cytochrome P450 allele nomenclature committee web site: submission criteria, procedures, and objectives. Methods Mol Biol 320:183–191
22. Thomas J, Milward D, Ouzounis C, Pulman S, Carroll M (2000) Automatic extraction of protein interactions from scientific abstracts. Pac Symp Biocomput 5:541–552
23. Horn F, Lau AL, Cohen FE (2004) Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. Bioinformatics 20(4):557–568
24. Hu ZZ, Narayanaswamy M, Ravikumar KE, Vijay-Shanker K, Wu CH (2005) Literature mining and database annotation of protein phosphorylation using a rule-based system. Bioinformatics 21(11):2759–2765
25. Yang X, Zhang B, Molony C, Chudin E, Hao K, Zhu J, Gaedigk A, Suver C, Zhong H, Leeder JS, Guengerich FP, Strom SC, Schuetz E, Rushmore TH, Ulrich RG, Slatter JG, Schadt EE, Kasarskis A, Lum PY (2010) Systematic genetic and genomic analysis of cytochrome P450 enzyme activities in human liver. Genome Res 20(8):1020–1036
26. Economopoulos KP, Sergentanis TN (2010) Three polymorphisms in cytochrome P450 1B1 (CYP1B1) gene and breast cancer risk: a meta-analysis. Breast Cancer Res Treat 122(2):545–551
27. Neafsey P, Ginsberg G, Hattis D, Johns DO, Guyton KZ, Sonawane B (2009) Genetic polymorphism in CYP2E1: population distribution of CYP2E1 activity. J Toxicol Environ Health B Crit Rev 12(5–6):362–388

28. White PC, Speiser PW (2000) Congenital adrenal hyperplasia due to 21-hydroxylase deficiency. Endocr Rev 21:245–291
29. Concolino P, Mello E, Zuppi C, Capoluongo E (2010) Molecular diagnosis of Congenital adrenal hyperplasia due to 21-hydroxylase deficiency: an update of new CYP21A2 mutations. Clin Chem Lab Med 48(8):1057–1062
30. Bradford LD (2002) CYP2D6 allele frequency in European Caucasians, Asians, Africans and their descendants. Pharmacogenomics 3(2):229–243
31. Blonk MI, van der Velde N, van den Bemt PM, van Schaik RH, van der Cammen TJ (2010) CYP2D6*4, CYP3A5*3 and ABCB1 3435T polymorphisms and drug-related falls in elderly people. Pharm World Sci 32(1):26–29
32. Sofi F, Giusti B, Marcucci R, Gori AM, Abbate R, Gensini GF (2010) Cytochrome P450 2C19(*)2 polymorphism and cardiovascular recurrences in patients taking clopidogrel: a meta-analysis. Pharmacogenomics J (Epub ahead of print)
33. Fleeman N, McLeod C, Bagust A, Beale S, Boland A, Dundar Y, Jorgensen A, Payne K, Pirmohamed M, Pushpakom S, Walley T, de Warren-Penny P, Dickson R (2010) The clinical effectiveness and cost-effectiveness of testing for cytochrome P450 polymorphisms in patients with schizophrenia treated with antipsychotics: a systematic review and economic evaluation. Health Technol Assess 14(3):1–157
34. Jorgensen WL (2004) The many roles of computation in drug discovery. Science 303(5665):1813–1818
35. Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discovery 3(11):935–949
36. Schlichting I, Berendze J, Chu K, Stoch AM, Maves SA, Benson DE, Sweet RM, Ringe D, Petsko GA, Sligar SG (2000) The catalytic pathway of cytochrome P450cam at atomic resolution. Science 287:1615–1622
37. Wang JF, Chou KC (2010) Molecular modeling of cytochrome P450 and drug metabolism. Curr Drug Metab 11(4):342–346
38. Wang JF, Zhang CC, Chou KC, Wei DQ (2009) Structure of cytochrome P450s and personalized drug. Curr Med Chem 16(2):232–244
39. de Graaf C, Vermeulen NPE, Feenstra A (2005) Cytochrome P450 in silico: an integrative modeling approach. J Med Chem 48(8):2725–2755
40. Crivori P, Poggesi I (2006) Computational approaches for predicting CYP-related metabolism properties in the screening of new drugs. Eur J Med Chem 41(7):795–808
41. Lill MA (2007) Multi-dimensional QSAR in drug discovery. Drug Discov Today 12(23–24):1013–1017
42. Khan MT (2010) Predictions of the ADMET properties of candidate drug molecules utilizing different QSAR/QSPR modelling approaches. Curr Drug Metab 11(4):285–295
43. Raha K, Peters MB, Wang B, Yu N, Wollacott AM, Westerhoff LM, Merz KM Jr (2007) The role of quantum mechanics in structure-based drug design. Drug Discov Today 12(17–18):725–731
44. Wang JF, Wei DQ, Li L, Zheng SY, Li YX, Chou KC (2007) 3D structure modeling of cytochrome P450 2C19 and its implication for personalized drug design. Biochem Biophys Res Commun 355(2):513–519. Erratum in: (2007) Biochem Biophys Res Commun 357(1):330. (2009) Biochem Biophys Res Commun 384(3):399
45. Sano E, Li W, Yuki H, Liu X, Furihata T, Kobayashi K, Chiba K, Neya S, Hoshino T (2009) Mechanism of the decrease in catalytic activity of human cytochrome P450 2C9 polymorphic variants investigated by computational analysis. J Comput Chem 31(15):2746–2758

Chapter 21
The α7 nAChR Selective Agonists as Drug Candidates for Alzheimer's Disease

Huaimeng Fan, Ruoxu Gu and Dongqing Wei

Abstract The nicotinic acetylcholine receptors (nAChRs) are ion channels distribute in the central or peripheral nervous system. They are receptors of the neurotransmitter acetylcholine and activation of them by agonists mediates synaptic transmission in the neuron and muscle contraction in the neuromuscular junction. Current studies reveal relationship between the nAChRs and the learning and memory as well as cognation deficit in various neurological disorders such as Alzheimer's disease, Parkinson's disease, schizophrenia and drug addiction. There are various subtypes in the nAChR family and the α7 nAChR is one of the most abundant subtypes in the brain. The α7 nAChR is significantly reduced in the patients of Alzheimer's disease and is believed to interact with the Aβ amyloid. Aβ amyloid is co-localized with α7 nAChR in the senile plaque and interaction between them induces neuron apoptosis and reduction of the α7 nAChR expression. Treatment with α7 agonist in vivo shows its neuron protective and procognation properties and significantly improves the learning and memory ability of the animal models. Therefore, the α7 nAChR agonists are excellent drug candidates for Alzheimer's disease and we summarized here the current agonists that have selectivity of the α7 nAChR over the other nAChR, introduced recent molecular modeling works trying to explain the molecular mechanism of their selectivity and described the design of novel allosteric modulators in our lab.

Keywords Nicotinic acetylcholine receptor · Alzheimer's disease · α7 nAChR · Agonist · Selectivity · Allosteric modulator

H. Fan · R. Gu · D. Wei (✉)
State Key Laboratory of Microbial Metabolism, College of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China
e-mail: dqwei@sjtu.edu.cn

D. Wei et al. (eds.), *Advance in Structural Bioinformatics*, Advances in Experimental Medicine and Biology 827, DOI 10.1007/978-94-017-9245-5_21

21.1 Introduction

The nicotinic acetylcholine receptors (nAChRs) are ligand gated cation channels that distribute throughout nearly all human tissues, especially on the neuromuscular junction and the pre- or post-synaptic of neuron of the peripheral and central nervous systems [1]. Binding of endogenous neurotransmitter acetylcholine as well as exogenous agonists such as nicotine to the nAChRs invokes ion flux through the membrane bilayer that responsible for a neuron exciting and synaptic communication [2]. The nAChRs in the central nervous system mainly distribute in the hippocampus, temporal cortex, and basal forebrain, which are areas responsible for memory and learning. Current studies have correlated the nAChRs with various neuron disorders such as Alzheimer's Disease, schizophrenia, Parkinson's Disease as well as drug addiction [3].

21.2 α7 nAChR is a Target of Alzheimer's Disease

Five subunits are required to constitute a functional nAChRs. Neuron nAChRs are constituted by several kinds of subunits including α2–α7, α9, α10, and β2–β4 in mammals, whereas other kinds of subunits such as α8 are found in birds and invertebrates. Many kinds of neuron nAChRs have been described with the α4β2 and α7 subtypes being the most abundant ones in the brain [4]. The α7 nAChR is a homo-pentameric and is different from the α4β2 subtype by being more permeable for Ca^{2+} and desensitizes more quickly than the α4β2 nAChR [5].

It is found in the patients' of Alzheimer's disease that the expression of α7 nAChR is reduced significantly and therefore, it is of great interest for the researchers studying the Alzheimer's disease [6–8]. It is believed that, the α7 nAChR is involved in this disease by interacting with the Aβ amyloid which is derived from the amyloid peptide precursor and assembles to form fibrils and senile plaque [9, 10]. The Aβ amyloid binds to the α7 nAChR with significantly higher affinity than the α-bungarotoxin which is a very potent α7 nAChR antagonist. It also interacts with other type nAChRs such as the α4β2 nAChR but with much lower affinity. It blocks the α7 nAChR but in specific conditions it also induces ion flux through the channel and the disturbed Ca^{2+} signal in the neuron induces cell apoptosis [11–14]. Interaction between the Aβ amyloid and α7 nAChR also induces hyperphosphorylation of the tau protein. The tau proteins constitute the microtubule and hyperphosphorylated tau proteins aggregate to form oligomerics and the microtubule as well as the cytoskeleton is damaged. The damaged cytoskeleton then results in the neurofibrillary tangle which is one of the most important pathological characterizations in the brain of the Alzheimer's disease patients. Since the microtubule is correlated with intercellular transport of proteins and cell organs, it is possible that the reduction of the α7 nAChR expression is partially because of the hyperphosphorylation of the tau protein and

the damage of the microtubule [15]. It is found that the agonist of α7 nAChR could reduce the apoptosis induced by Aβ amyloid [16]. However, the endogenous neurotransmitter acetylcholine which is a potent agonist of nAChRs is greatly reduced in the Alzheimer's disease patients. Therefore, applying the exogenous agonists that have selectivity for the α7 nAChR is of helpful for the treatment of the Alzheimer's disease and the α7 nAChR is a main target for drug design. We have summarized here the recently found α7 nAChR specific agonists which may be interest for not only the pharmacologists but also the scientists that are performing electrophysiological experiments of ion channels.

21.3 The Agonist Binding Site of α7 nAChR

The neuron nAChRs are integral transmembrane proteins that can be divided into the extracellular domain, the transmembrane domain and the cytoplasmic domain. Its transmembrane domain which is responsible for the channel activity crosses the membrane four times and the second transmembrane helix of each subunit constitutes the channel pore [17]. The agonists bind to the extracellular domain of the channel and the conformational changes deduced from agonist binding is then transferred to the transmembrane domain which is ~40 Å away and the channel is open. The channel is closed again which is known as the desensitized state after long exposure to the agonists [18].

The nAChRs are homo- or hetero-pentamerics that usually in a stoichiometry of $(\alpha)_5$, $(\alpha)_2(\beta)_3$ or $(\alpha)_3(\beta)_2$ [19, 20]. The agonist binding site is located at the interfacial region of one α subunit and one non-α subunit (or the interfacial region of two α subunits in the homo-pentameric) as shown in Fig. 21.1. The α subunit and the non-α subunit (or the corresponding α subunit in the homo-pentameric) of the binding cavity is known as the primary component and the complementary component, respectively. Therefore, there are five binding sites on the homo-pentemric whereas only two or three binding sites on the hetero-pentamerics [21, 22].

21.4 The Pharmocophores of the nAChR Agonist

It is found that two pharmacophores are usually important for the activity of an agonist, one is a cation center which is usually a positively charged nitrogen atom, whereas the other is a pyridine nitrogen which acts as an hydrogen bond donor. The cation center resides in the aromatic cage formed by five aromatic residues of the binding cavity and formed cation-π interaction with the aromatic rings. The pyridine form hydrogen bonds with two residues of the complementary component via water molecules, as shown in Fig. 21.2 [23, 24].

The agonists that have selectivity for the α7 nAChR are excellent drug candidates for treating cognition impacts in the neurological disorders such as the

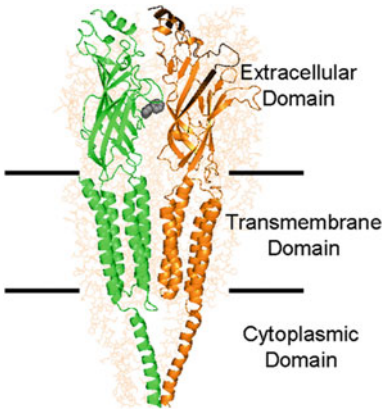


Fig. 21.1 The 3-D structure of the nicotinic acetylcholine receptor (*nAChR*). The nAChR is constituted by five subunits and two of them are shown in cartoon model in *green* and *orange*, respectively, whereas the other three are shown in *wheat lines* for clarify. The nAChR is divided in to the extracellular domain, the transmembrane domain and a cytoplasmic domain. The agonist binding site is located at the interfacial region of the extracellular domain. One agonist is shown in *gray* space filling model

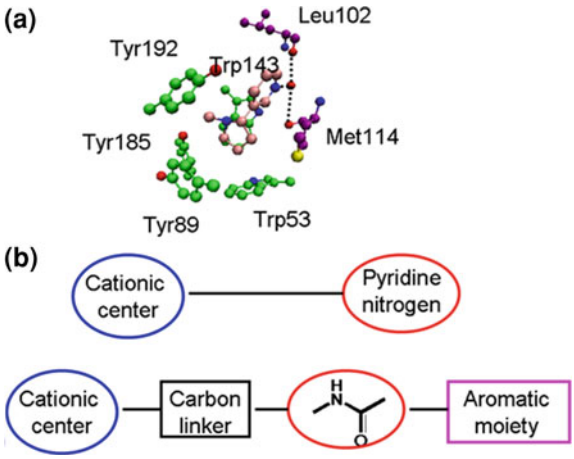


Fig. 21.2 **a** Interaction between the nicotinic and the acetylcholine binding protein (*AChBP*), a homology of the extracellular domain of the nAChRs. The cationic center resides in the aromatic cage of the binding cavity and formed cation-π interaction with the aromatic ring whereas the pyridine nitrogen form water mediated hydrogen bonds with residues form the complementary component. The carbon atoms of the aromatic residues, the two residues from the complementary component, and nicotine are shown in *green*, *pink* and *purple*, respectively, and the water mediated hydrogen bonds are shown in *black dashed line*. The oxygen and nitrogen atoms are shown in *red* and *blue*, respectively. **b** Pharmacophore models of the nicotine (*top*) and α7 nAChR selectivity agonist (*bottom*)

Alzheimer's disease and the schizophrenia. At present, various α7 nAChR selective agonists have been discovered and SEN12333/WAY-317538 is one example of them [25]. The structure activity relation has been investigated for this series compounds and a pharmacophore model is proposed. As shown in Fig. 21.2b, one cation center is connected to an amine group by a carbon linker and an aromatic moiety is connected to the amine group on the other side. The invert of the amine group direction doesn't affect the agonist activity [25, 26]. The amine group in this model is corresponding to the pyridine nitrogen of nicotine and the significant differences between the pharmacophore model of the α7 nAChR selective agonist and that of those without subtype selectivity is the aromatic moiety connected to the amine group (Fig. 21.2).

21.5 Full Agonist of α7 nAChR

A full agonist of an ion channel is defined as a ligand that can activate the channel as efficiently as the endogenous agonist does. For the α7 nAChR, the full agonists refer to those invoke inward ion flux in an efficient of >75 % of acetylcholine does. The current full agonists that have selectivity for the α7 nAChR including AR-R17779, ABBF, ABT-107, SEN12333/WAY-317538, WYE-103914/SEN34625, TC-5619, compound 7aa, and PNU-282987 (Fig. 21.3) and their interaction properties with nAChRs and some in vitro or in vivo functions are summarized here.

AR-R17779. AR-R17779 is the first reported full agonist that selective for the α7 nAChR. The (-)-AR-R17779 binds with the α7 nAChR with much higher affinity (Ki = 92 nM) than with the α4β2 subtype (Ki = 16,000 nM) and it is twice as potent as nicotine for the α7 nAChR (EC_{50} of ~21 mM and ~43 mM for AR-R17779 and nicotine respectively). It is an agonist of the 5-HT_3 receptor [27]. In vivo experiments have proved that the AR-R17779 improves the social recognition memory of the rats by interacting with the α7 nAChR, implying the relationship between the α7 nAChR and the learning and memory abilities [28].

ABBF. The ABBF is a full agonist of the human and rat α7 nAChR but acts as a competitive antagonist of the other nAChR subtypes such as the α3β4, α4β2 and muscle type nAChR at high concentration. It is also a competitive antagonist of the 5-HT3 receptor. The ABBF has much higher binding affinity with the α7 nAChR (Ki = 62 nM) labeled by radioligand [3H]methyllycaconitine than nicotinic (Ki = 770 nM) and acetylcholine (Ki = 3 μM). The EC_{50} value of 3 μM of ABBF shows more potent activity of this agonist for α7 nAChR than the acetylcholine (EC_{50} = 170 μM). Behavioral experiments show that it could improve the learning and memory ability of both rats and mice in vivo [29].

ABT-107. ABT-107 is highly selectivity full agonist for the human and rat α7 nAChRs which shows no activation for the α3β4 nAChR and very low agonist activity for the α4β2 nAChR. The binding affinity and EC_{50} value of ABT-107 for the human α7 nAChR are ~0.22 and ~10.4 nM, respectively [30]. Initial in vivo

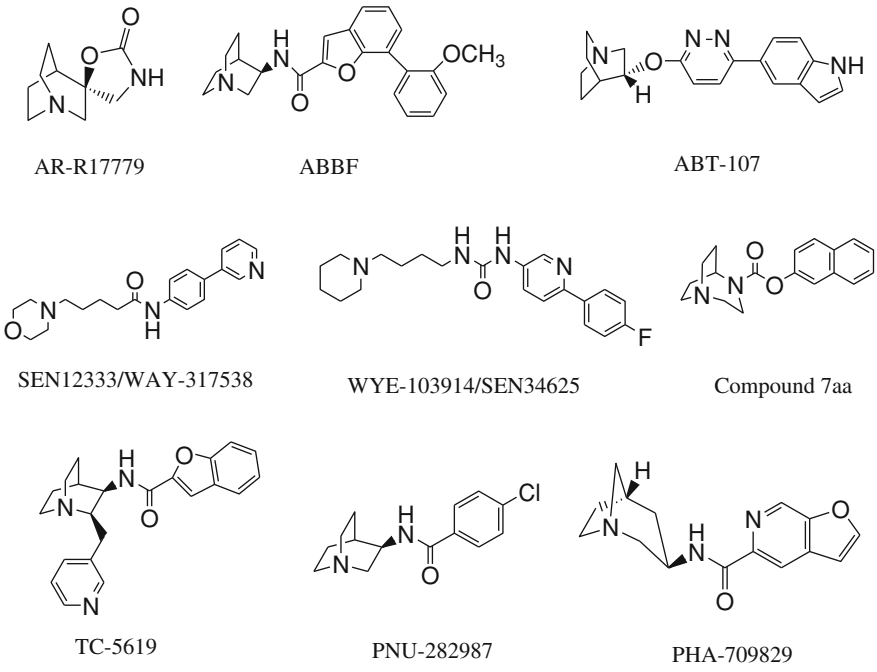


Fig. 21.3 Structures of full agonists of the α7 nAChR

pharmacokinetic and safety experiments in healthy human volunteers imply it is a good candidate for further development [31].

SEN12333/WAY-317538. SEN12333/WAY-317538 is a full agonist of α7 nAChR with an EC_{50} value of ~1.6 μM and it is also selective for the α7 nAChR (Ki = 260 nM) over other nAChRs such as the α1, α3 containing subtypes as well as the 5-HT_3 receptor [25]. In vivo and in vitro experiments have show that it has excellent pharmacokinetic profiles, brain penetration ability, as well as oral bioavailability. It has been proved to improve the learning and cognation ability and is neuroprotective for the experimental animals by interacting with the α7 nAChR [32].

WYE-103914/SEN34625. WYE-103914/SEN34625 is a full agonist (EC_{50} = 130 nM) of the human α7 nAChR and it binds to the α7 nAChR with significant higher affinity (Ki = 44 nM) than the α1, α3 containing subtypes, the α4β2 nAChR as well as the 5-HT3 receptor. It shows excellent brain penetration ability and oral bioavailability as well as cognition improvements in the in vitro and in vivo tests [26].

TC-5619. TC-5619 is a full agonist of α7 nAChR with an EC_{50} value of ~33 nM and it is also selective for the α7 nAChR (Ki = 1 nM) over the α4β2 nAChR (Ki = 2,800 nM). It has very low agonist activity for the α3β4 and muscle type nAChR [33]. TC-5619 show significant memory and cognation improvement on

rats and mice and it is now under Phase II investigation for attention deficit/hyperactivity disorder (ADHD) and cognitive dysfunction in schizophrenia (CDS).

Compound 7aa. Compound 7aa is a subtype selective full agonist of the α7 nAChR with a binding affinity of ~23 nM. Initial experiments show excellent pharmacokinetic properties of this compound [34].

PNU-282987. PNU-282987 is a potent α7 nAChR agonist with an affinity of 27 nM and an EC_{50} value of 154 nM for the α7-5HT3 chimera [35]. However, it is not suitable for development as a drug candidate because of its significant activity for the human hERG channel [36]. The later found compound PHA-709829 is highly selective for the α7 nAChR (Ki = 3 nM and EC_{50} = 46 nM for the α7-5HT3 chimera) over the other subtypes such as the α4β2 nAChR, α3β4 nAChR, and the muscle type nAChR. Its toxicity targeting the hERG channel is significantly improved compared with the PNU-282987 [37].

EVP-6124, EVP-6124 is a selective agonist for the α7 nACh and is being developed by En Vivo for potential cognitive enhancement in both schizophrenia and Alzheimer's patients. It has been shown to have excellent CNS penetration, oral bioavailability, pharmacokinetics and metabolic profile. Two previous studies with EVP-6124 in Alzheimer's disease patients demonstrated that it was well tolerated and produced significant effects on a variety of cognitive measures of brain function such as attention, memory and executive function (complex thinking tasks) (Table 21.1).

21.6 Partial Agonist of α7 nAChR

A partial agonist of an ion channel is defined as a ligand that activate the channel in an efficacy lower than the endogenous agonist does. Many partial agonists of the nAChRs are found (Fig. 21.4) and those have selectivity for the α7 nAChR are listed below.

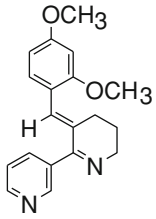
DMXB-A. The DMXB-A, a derivative of the anabaseine, is considered as "functional selective" for it is a partial agonist (EC_{50} = 81 μM, Emax = 50 %) of the α7 nAChR but exhibits antagonist activity for other nAChR subtypes as well as the 5-HT3 receptor. It binds to the α4β2 nAChRs (*K*i = 20 nM) 100-fold more potently than to the human α7 nAChRs [38]. It is now under Phase II investigation for the Alzheimer's Disease, Praksion Disease and schizophrenia.

JN403. JN403 is a partial agonist of the α7 nAChR with an Emax of 55 % in Xenopus oocytes expressing human α7 nAChR. It binds to the α7 nAChR with much higher affinity (*K*i = 200 nM) compared with other nAChR subtypes such as the α3β4 nAChR (*K*i = 6,309 nM), the α4β2 nAChR (*K*i = 158,489 nM) as well as the 5-HT3 receptor (*K*i = 12,589 nM) [39]. The agonist activity of JN403 on the nAChRs is blocked by MLA, a very potent competitive antagonist for the α7 nAChR. In vivo tests show that the JN403 facilitates the improvement of learning and memory abilities of experimental animals [40].

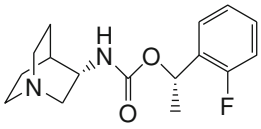
Table 21.1 Full agonists of α7 nAChR

Agonists	Binding affinity for α7 nAChR (Ki) (nM)	Efficacy (EC_{50})	Activity on other receptors	Clinical trials
AR-R17779	92	21 nM	Low affinity with the α4β2 nAChR (Ki = 16,000 nM), agonist of the 5-HT_3 receptor	No
ABBF	62	170 μM	Competitive antagonist of the α3β4, α4β2, muscle type nAChR, the 5-HT_3 receptor	No
ABT-107	0.22	10.4 nM	No activation for the α3β4 nAChR, very low agonist activity for the α4β2 nAChR	No
SEN12333/WAY-317538	260	1.6 μM	Low binding affinity with the α1, α3 containing subtypes and the 5-HT_3 receptor	No
WYE-103914/SEN34625	44	130 nM	Low binding affinity with the α1, α3 containing subtypes, the α4β2 nAChR and the 5-HT3 receptor	No
TC-5619	1	33 nM	Low agonist activity for the α3β4, α4β2 and muscle type nAChR, Ki of 2800 nM with α4β2 nAChR	Phase II investigation for attention deficit/hyperactivity disorder (ADHD) and cognitive dysfunction in schizophrenia (CDS)
Compound 7aa	23	175 %	–	No
PNU-282987	27	154 nM for the α7-5HT_3 chimera	Antagonist of the 5-HT_3 receptor (IC_{50} = 4,541 nM), low binding affinity with the α4β2 subtype	Discontinued in phase II clinical trials because of activity for the human hERG channel
PHA-709829	3	46 nM for the α7-5HT_3 chimera	–	No

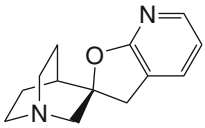
AZD0328. AZD0328 binds to the human and rat α7 nAChR (Ki = 3.0 and 4.7 nM for human and rat α7 nAChR, respectively) and 5-HT_3 receptor (Ki = 12.0 and 25.0 nM for human and rat 5-HT_3 receptor, respectively), but has only moderate binding affinity with the rat α4β2 nAChR (Ki = 140 nM) and very



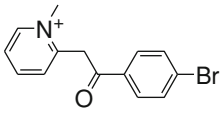
DMXB-A



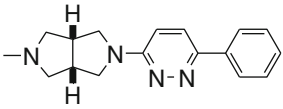
JN403



AZD0328



S 24795



A-582941

Fig. 21.4 Structures of partial agonists of the α7 nAChR

low binding affinities with the α3 containing nAChRs and the muscle type nAChR. It is a partial agonist of the human α7 nAChR with an EC_{50} value of 338 nM and Emax of 65 %, however, it has very low efficacy on the mouse 5-HT3 receptor and the human α4β2 nAChR and no activity on the human α3β4 nAChR [41]. Applying AZD0328 in animal models shows immediate and sustained improvement of the cognation ability as well as the improvement of the working memory [42]. AZD-0328 was under phase II trial in schizophrenia in 2008 but was discontinued in 2009 [43].

S 24795. S 24795 is a partial agonist of the rat α7 nAChR and has no affinities on the α4β2 nAChR, the α3β4 nAChR as well as the muscle type nAChR. It active the α7 nAChR with an EC_{50} of 34 nM and an efficacy of ~10 % of the acetylcholine does [44]. Applying the S 24795 in vitro or in vivo reduces the $A\beta_{42}$-α7 nAChR complex and helps the release of $A\beta_{42}$ from the $A\beta_{42}$-α7 nAChR complex [45].

A-582941. The A-582941 exhibits high affinity (Ki = 16.7 nM) [46], and partial agonist activity (EC_{50} = 4.26 μM; activity = 52 %) for the human α7 nAChR. Initial in vitro and in vivo experiments show its appropriate pharmacokinetic properties and excellent brain penetration. A-582941 improves the cognition abilities in monkey, rat, and mouse and rescues the sensory gating deficits induced by the potent α7 nAChR antagonist MLA [47].

R3487/MEM3454. The R3487/MEM3454 are antagonist and partial agonist of the human 5-HT_3 receptor and the rat α7 nAChR, respectively, and it shows similar high affinities for both receptors (Ki = 6 nM for rat α7 nAChR and Ki = 2 nM for human 5-HT3 receptor). It activates the monkey recombinant α7 nAChR with an EC_{50} of 0.4 nM [48]. It is now under development for treating

Table 21.2 Partial agonists of the α7 nAChR

Agonist	Binding affinity for α7 nAChR (Ki)	Efficacy (EC_{50}, Emax)	Activity on other receptors	Clinical trials
DMXB-A	2,000 nM	81 μM, 50 %	Antagonist for other nAChR subtypes, higher binding affinity with α4β2 nAChR (Ki = 20 nM)	Phase II investigation for the Alzheimer's disease, Praksion disease and schizophrenia
JN403	200 nM	55 %	Binding affinities for the α4β2, α3β4 and 5-HT3 receptors are 158, 489, 6,309 and 12,589 nM	No
AZD0328	3.0 nM	338 nM, 65 %	Binding affinities for the α4β2 and 5-HT3 receptors are 140 and 12.0 nM, very low binding affinities with the α3 containing subtypes and the muscle type nAChR	Discontinued in phase II trial in schizophrenia in 2009
S 24795		34 nM, 10 %	No affinity for the α4β2, α3β4, and muscle type nAChRs	No
A-582941	16.7 nM	4.26 μM, 52 %	Weak binding affinities (Ki > 100 μM) with the α4, α3 and α1 containing subtypes	No
R3487/ MEM3454	6 nM for rat α7 nAChR	0.4 nM	Antagonist of the 5-HT_3 receptor, binding affinity of 2 nM	Phase IIb trial for Alzheimer's disease

Alzheimer's disease and schizophrenia [49]. A structural related molecular named as MEM-63908 (R-4996) was developed and entered phase I investigation for Alzheimer's disease in March 2009 (Table 21.2).

References

1. Graham AJ, Ray MA, Perry EK, Jaros E, Perry RH, Volsen SG, Bose S, Evans N, Lindstrom J (2003) Differential nicotinic acetylcholine receptor subunit expression in the human hippocampus. J Chem Neuroanat 25(2):97–113
2. Lloyd GK, Williams M (2000) Neuronal nicotinic acetylcholine receptors as novel drug targets. J Pharmacol Exp Ther 292(2):461
3. Nashmi R, Dickinson ME, McKinney S, Jareb M, Labarca C, Fraser SE, Lester HA (2003) Assembly of α4β2 nicotinic acetylcholine receptors assessed with functional fluorescently

labeled subunits: effects of localization, trafficking, and nicotine-induced upregulation in clonal mammalian cells and in cultured midbrain neurons. J Neurosci 23(37):11554–11567
4. Dani JA, Bertrand D (2007) Nicotinic acetylcholine receptors and nicotinic cholinergic mechanisms of the central nervous system. Annu Rev Pharmacol Toxicol 47:699–729
5. Albuquerque EX, Pereira EFR, Alkondon M, Rogers SW (2009) Mammalian nicotinic acetylcholine receptors: from structure to function. Physiol Rev 89(1):73–120
6. Wevers A, Monteggia L, Nowacki S, Bloch W, Schütz U, Lindstrom J, Pereira EFR, Eisenberg H, Giacobini E, De Vos RAI (1999) Expression of nicotinic acetylcholine receptor subunits in the cerebral cortex in Alzheimer's disease: histotopographical correlation with amyloid plaques and hyperphosphorylated-tau protein. Eur J Neurosci 11(7):2551–2565
7. Guan ZZ, Zhang X, Ravid R, Nordberg A (2000) Decreased protein levels of nicotinic receptor subunits in the hippocampus and temporal cortex of patients with Alzheimer expression of nicotinic acetylcholine receptor subunits in the cerebral cortex in Alzheimer's disease. J Neurochem 74(1):237–243
8. Wevers A, Witter B, Moser N, Burghaus L, Banerjee C, Steinlein OK, Schütz U, De Vos RAI, Jansen Steur ENH, Lindstrom J (2000) Classical Alzheimer features and cholinergic dysfunction: towards a unifying hypothesis? Acta Neurol Scand 102:42–48
9. Luheshi LM, Tartaglia GG, Brorsson AC, Pawar AP, Watson IE, Chiti F, Vendruscolo M, Lomas DA, Dobson CM, Crowther DC (2007) Systematic in vivo analysis of the intrinsic determinants of amyloid β pathogenicity. PLoS Biol 5(11):e290
10. Roychaudhuri R, Yang M, Hoshi MM, Teplow DB (2009) Amyloid β-protein assembly and Alzheimer disease. J Biol Chem 284(8):4749
11. Oneill MJ, Murray TK, Lakics V, Visanji NP, Duty S (2002) The role of neuronal nicotinic acetylcholine receptors in acute and chronic neurodegeneration. Curr Drug Targets CNS Neurol Disord 1 (4):399–411
12. Grassi F, Palma E, Tonini R, Amici M, Ballivet M, Eusebi F (2003) Amyloid β1-42 peptide alters the gating of human and mouse α-bungarotoxin-sensitive nicotinic receptors. J Physiol 547(1):147–157
13. Dineley KT, Bell KA, Bui D, Sweatt JD (2002) β-Amyloid peptide activates α7 nicotinic acetylcholine receptors expressed in Xenopus oocytes. J Biol Chem 277(28):25056–25061
14. Dougherty JJ, Wu J, Nichols RA (2003) β-Amyloid regulation of presynaptic nicotinic receptors in rat hippocampus and neocortex. J Neurosci 23(17):6740–6747
15. Wang HY, Li W, Benedetti NJ, Lee DHS (2003) α7 nicotinic acetylcholine receptors mediate β-amyloid peptide-induced tau protein phosphorylation. J Biol Chem 278(34):31547
16. Freir DB, Herron CE (2003) Nicotine enhances the depressive actions of Aβ1-40 on long-term potentiation in the rat hippocampal CA1 region in vivo. J Neurophysiol 89(6):2917–2922
17. Wildman SS, Marks J, Churchill LJ, Peppiatt CM, Horisberger JD, King BF, Unwin RJ (2005) Molecular interactions between cloned epithelial sodium channels and ATP-gated P2X receptors. FASEB J 19(5):A1177
18. Miyazawa A, Fujiyoshi Y, Unwin N (2003) Structure and gating mechanism of the acetylcholine receptor pore. Nature 423(6943):949–955
19. Karlin A (2002) Emerging structure of the nicotinic acetylcholine receptors. Nat Rev Neurosci 3(2):102–114
20. Arias HR (2009) Is the inhibition of nicotinic acetylcholine receptors by bupropion involved in its clinical actions? Int J Biochem Cell Biol 41(11):2098–2108
21. Arias HR (2000) Localization of agonist and competitive antagonist binding sites on nicotinic acetylcholine receptors. Neurochem Int 36(7):595–645
22. Arias HR (2010) Molecular interaction of bupropion with nicotinic acetylcholine receptors. J Pediatr Biochem 1(2):185–197
23. Gu RX, Zhong YQ, Wei DQ (2011) Structural basis of agonist selectivity for different nAChR subtypes: insights from crystal structures, mutation experiments and molecular simulations. Curr Pharm Des 17(17):1652–1662
24. Tondera JE, Olesen PH, Hansen JB, Begtrup M, Pettersson I (2001) An improved nicotinic pharmacophore and a stereoselective CoMFA-model for nicotinic agonists acting at the

central nicotinic acetylcholine receptors labelled by [3H]-N-methylcarbamylcholine. J Comput Aided Mol Des 15(3):247–258
25. Haydar SN, Ghiron C, Bettinetti L, Bothmann H, Comery TA, Dunlop J, La Rosa S, Micco I, Pollastrini M, Quinn J (2009) SAR and biological evaluation of SEN12333/WAY-317538: novel alpha 7 nicotinic acetylcholine receptor agonist. Bioorg Med Chem 17(14):5247–5258
26. Ghiron C, Haydar SN, Aschmies S, Bothmann H, Castaldo C, Cocconcelli G, Comery TA, Di L, Dunlop J, Lock T (2010) Novel alpha-7 nicotinic acetylcholine receptor agonists containing a urea moiety: identification and characterization of the potent, selective, and orally efficacious agonist 1-[6-(4-Fluorophenyl) pyridin-3-yl]-3-(4-piperidin-1-ylbutyl) urea (SEN34625/WYE-103914). J Med Chem 53(11):4379–4389
27. Mullen G, Napier J, Balestra M, DeCory T, Hale G, Macor J, Mack R, Loch Iii J, Wu E, Kover A (2000) (-)-Spiro [1-azabicyclo [2.2. 2] octane-3, 5'-oxazolidin-2'-one], a conformationally restricted analogue of acetylcholine, is a highly selective full agonist at the α7 nicotinic acetylcholine receptor. J Med Chem 43(22):4045–4050
28. Kampen M, Selbach K, Schneider R, Schiegel E, Boess F, Schreiber R (2004) AR-R 17779 improves social recognition in rats by activation of nicotinic a 7 receptors. Psychopharmacology 172(4):375–383
29. Boess FG, De Vry J, Erb C, Flessner T, Hendrix M, Luithle J, Methfessel C, Riedl B, Schnizler K, van der Staay FJ (2007) The novel α7 nicotinic acetylcholine receptor agonist N-[(3R)-1-azabicyclo [2.2. 2] oct-3-yl]-7-[2-(methoxy) phenyl]-1-benzofuran-2-carboxamide improves working and recognition memory in rodents. J Pharmacol Exp Ther 321(2):716
30. Malysz J, Anderson DJ, Gr Nlien JH, Ji J, Bunnelle WH, H Kerud M, Thorin-Hagene K, Ween H, Helfrich R, Hu M (2010) In vitro pharmacological characterization of a novel selective α7 neuronal nicotinic acetylcholine receptor agonist ABT-107. J Pharmacol Exp Ther 334(3):863
31. Othman AA, Lenz RA, Zhang J, Li J, Awni WM, Dutta S (2011) Single-and multiple-dose pharmacokinetics, safety, and tolerability of the selective α7 neuronal nicotinic receptor agonist, ABT-107, in healthy human volunteers. J Clin Pharmacol 51(4):512
32. Roncarati R, Scali C, Comery TA, Grauer SM, Aschmi S, Bothmann H, Jow B, Kowal D, Gianfriddo M, Kelley C (2009) Procognitive and neuroprotective activity of a novel α7 nicotinic acetylcholine receptor agonist for treatment of neurodegenerative and cognitive disorders. J Pharmacol Exp Ther 329(2):459
33. Hauser TA, Kucinski A, Jordan KG, Gatto GJ, Wersinger SR, Hesse RA, Stachowiak EK, Stachowiak MK, Papke RL, Lippiello PM (2009) TC-5619: an alpha7 neuronal nicotinic receptor-selective agonist that demonstrates efficacy in animal models of the positive and negative symptoms and cognitive dysfunction of schizophrenia. Biochem Pharmacol 78(7):803–812
34. O'Donnell CJ, Peng L, O'Neill BT, Arnold EP, Mather RJ, Sands SB, Shrikhande A, Lebel LA, Spracklin DK, Nedza FM (2009) Synthesis and SAR studies of 1, 4-diazabicyclo [3.2. 2] nonane phenyl carbamates-subtype selective, high affinity [alpha] 7 nicotinic acetylcholine receptor agonists. Bioorg Med Chem Lett 19(16):4747–4751
35. Bodnar AL, Cortes-Burgos LA, Cook KK, Dinh DM, Groppi VE, Hajos M, Higdon NR, Hoffmann WE, Hurst RS, Myers JK (2005) Discovery and structure-activity relationship of quinuclidine benzamides as agonists of α7 nicotinic acetylcholine receptors. J Med Chem 48(4):905–908
36. Hajos M, Hurst RS, Hoffmann WE, Krause M, Wall TM, Higdon NR, Groppi VE (2005) The selective α7 nicotinic acetylcholine receptor agonist PNU-282987 [N-[(3R)-1-azabicyclo [2.2. 2] oct-3-yl]-4-chlorobenzamide hydrochloride] enhances GABAergic synaptic activity in brain slices and restores auditory gating deficits in anesthetized rats. J Pharmacol Exp Ther 312(3):1213
37. Acker BA, Jacobsen EJ, Rogers BN, Wishka DG, Reitz SC, Piotrowski DW, Myers JK, Wolfe ML, Groppi VE, Thornburgh BA (2008) Discovery of N-[(3R, 5R)-1-azabicyclo [3.2. 1] oct-3-yl] furo [2, 3-c] pyridine-5-carboxamide as an agonist of the [alpha] 7 nicotinic acetylcholine receptor: in vitro and in vivo activity. Bioorg Med Chem Lett 18(12):3611–3615

38. Briggs CA, Anderson DJ, Brioni JD, Buccafusco JJ, Buckley MJ, Campbell JE, Decker MW, Donnelly-Roberts D, Elliott RL, Gopalakrishnan M (1997) Functional characterization of the novel neuronal nicotinic acetylcholine receptor ligand GTS-21 in vitro and in vivo. Pharmacol Biochem Behav 57(1–2):231–241
39. Feuerbach D, Nozulak J, Lingenhoehl K, McAllister K, Hoyer D (2007) JN403, in vitro characterization of a novel nicotinic acetylcholine receptor [alpha] 7 selective agonist. Neurosci Lett 416(1):61–65
40. Feuerbach D, Lingenhoehl K, Olpe HR, Vassout A, Gentsch C, Chaperon F, Nozulak J, Enz A, Bilbe G, McAllister K (2009) The selective nicotinic acetylcholine receptor [alpha] 7 agonist JN403 is active in animal models of cognition, sensory gating, epilepsy and pain. Neuropharmacology 56(1):254–263
41. Sydserff S, Sutton EJ, Song D, Quirk MC, Maciag C, Li C, Jonak G, Gurley D, Gordon JC, Christian EP (2009) Selective [alpha] 7 nicotinic receptor activation by AZD0328 enhances cortical dopamine release and improves learning and attentional processes. Biochem Pharmacol 78(7):880–888
42. Castner SA, Hudzik T, Maier DL, Mrzljak L, Piser T, Smith JS, Widzowski D, Williams GV (2009) A composition comprising (R)-spiro [L-Azabicyclo [2.2. 2] Octane-3, 2'(3'H)-Furo [2, 3-B] Pyridine (AZD0328) and its use in the treatment of Alzheimer's disease, ADHD or cognitive dysfunction. EP patent 2,120,937
43. Haydar SN, Dunlop J (2010) Neuronal nicotinic acetylcholine receptors-targets for the development of drugs to treat cognitive impairment associated with schizophrenia and Alzheimer's disease. Curr Top Med Chem 10(2):144–152
44. Lopez-Hernandez G, Placzek AN, Thinschmidt JS, Lestage P, Trocme-Thibierge C, Morain P, Papke RL (2007) Partial agonist and neuromodulatory activity of S 24795 for alpha7 nAChR responses of hippocampal interneurons. Neuropharmacology 53(1):134–144
45. Wang HY, Bakshi K, Shen C, Frankfurt M, Trocmé-Thibierge C, Morain P (2010) S 24795 limits [beta]-Amyloid-[alpha] 7 nicotinic receptor interaction and reduces Alzheimer's disease-like pathologies. Biol Psychiatry 67(6):522–530
46. Anderson DJ, Bunnelle W, Surber B, Du J, Surowy C, Tribollet E, Marguerat A, Bertrand D, Gopalakrishnan M (2008) [3H]A-585539[(1S,4S)-2, 2-Dimethyl-5-(6-phenylpyridazin-3-yl)-5-aza-2-azoniabicyclo [2.2. 1] heptane], a novel high-affinity α7 neuronal nicotinic receptor agonist: radioligand binding characterization to rat and human brain. J Pharmacol Exp Ther 324(1):179
47. Bitner RS, Bunnelle WH, Anderson DJ, Briggs CA, Buccafusco J, Curzon P, Decker MW, Frost JM, Gronlien JH, Gubbins E (2007) Broad-spectrum efficacy across cognitive domains by α7 nicotinic acetylcholine receptor agonism correlates with activation of ERK1/2 and CREB phosphorylation pathways. J Neurosci 27(39):10578
48. Pichat P, Bergis OE, Terranova JP, Urani A, Duarte C, Santucci V, Gueudet C, Voltz C, Steinberg R, Stemmelin J (2006) SSR180711, a novel selective α7 nicotinic receptor partial agonist: (II) efficacy in experimental models predictive of activity against cognitive symptoms of schizophrenia. Neuropsychopharmacology 32(1):17–34
49. Sabbagh MN (2009) Drug development for Alzheimer's disease: where are we now and where are we headed? Am J Geriatr Pharmacother 7(3):167–185

Chapter 22
Bayesian Analysis of Complex Interacting Mutations in HIV Drug Resistance and Cross-Resistance

Ivan Kozyryev and Jing Zhang

Abstract A successful treatment of AIDS world-wide is severely hindered by the HIV virus' drug resistance capability resulting from complicated mutation patterns of viral proteins. Such a system of mutations enables the virus to survive and reproduce despite the presence of various antiretroviral drugs by disrupting their binding capability. Although these interacting mutation patterns are extremely difficult to efficiently uncover and interpret, they contribute valuable information to personalized therapeutic regimen design. The use of Bayesian statistical modeling provides an unprecedented opportunity in the field of anti-HIV therapy to understand detailed interaction structures of drug resistant mutations. Multiple Bayesian models equipped with Markov Chain Monte Carlo (MCMC) methods have been recently proposed in this field (Zhang et al. in PNAS 107:1321, 2010 [1]; Zhang et al. in J Proteome Sci Comput Biol 1:2, 2012 [2]; Svicher et al. in Antiviral Res 93(1):86–93, 2012 [3]; Svicher et al. in Antiviral Therapy 16(7):1035–1045, 2011 [4]; Svicher et al. in Antiviral Ther 16(4):A14–A14, 2011 [5]; Svicher et al. in Antiviral Ther 16(4):A85–A85, 2011 [6]; Alteri et al. in Signature mutations in V3 and bridging sheet domain of HIV-1 gp120 HIV-1 are specifically associated with dual tropism and modulate the interaction with CCR5 N-Terminus, 2011 [7]). Probabilistically modeling mutations in the HIV-1 protease or reverse transcriptase (RT) isolated from drug-treated patients provides a powerful statistical procedure that first detects mutation combinations associated with single or multiple-drug resistance, and then infers detailed dependence structures among the interacting mutations in viral proteins (Zhang et al. in PNAS 107:1321, 2010 [1]; Zhang et al. in J Proteome Sci Comput Biol 1:2, 2012 [2]).

I. Kozyryev
Department of Physics, Harvard University, Cambridge, MA, USA

J. Zhang (✉)
Department of Statistics, Yale University, New Haven, CT, USA
e-mail: jing.maria.zhang@gmail.com

J. Zhang
Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT USA

D. Wei et al. (eds.), *Advance in Structural Bioinformatics*, Advances in Experimental Medicine and Biology 827, DOI 10.1007/978-94-017-9245-5_22

Combined with molecular dynamics simulations and free energy calculations, Bayesian analysis predictions help to uncover genetic and structural mechanisms in the HIV treatment resistance. Results obtained with such stochastic methods pave the way not only for optimization of the use for existing HIV drugs, but also for the development of the new more efficient antiretroviral medicines. In this chapter we survey current challenges in the bioinformatics of anti-HIV therapy, and outline how recently emerged Bayesian methods can help with the clinical management of HIV-1 infection. We will provide a rigorous review of the Bayesian variable partition model and the recursive model selection procedure based on probability theory and mathematical data analysis techniques while highlighting real applications in HIV and HBV studies including HIV drug resistance (Zhang et al. in PNAS 107:1321, 2010 [1]), cross-resistance (Zhang et al. in J Proteome Sci Comput Biol 1:2, 2012 [2]), HIV coreceptor usage (Svicher et al. in Antiviral Therapy 16(7):1035–1045, 2011 [4]; Svicher et al. in Antiviral Ther 16(4):A14–A14, 2011 [5]; Alteri et al. in Signature mutations in V3 and bridging sheet domain of HIV-1 gp120 HIV-1 are specifically associated with dual tropism and modulate the interaction with CCR5 N-Terminus, 2011 [7]), and occult HBV infection (Svicher et al. in Antiviral Res 93(1):86–93, 2012 [3]; Svicher et al. in Antiviral Ther 16(4):A85–A85, 2011 [6]).

Keywords Bayesian statistical modeling · Markov chain Monte Carlo · HIV

22.1 Complexity of Personalized Medicine and Genomics

Tailoring preventive and therapeutic disease treatments towards each patient at a time while fully utilizing genetic code information could become the future of medicine [8–10]. Therefore, methods for discovering disease related variants in patients' genome and treatment related mutations in illness-causing virus will have important applications in biomedicine. Such a task is immensely complicated by the presence of non-trivial multilocus interactions in these problems [11, 12].

In genetics of common diseases, the presence of non-mendelian variants interacting in complicated ways could account for a significant portion of missing heritability [13]. In a different field, interacting mutations in viral genome lead to drug resistance during treatment procedures [1, 14]. Therefore, understanding of genetic interactions biochemically and statistically becomes important for further applications.

22.2 Inferring Phenotype from Genotype

Predicting phenotype from genotype plays an important role in many areas of biomedicine; particularly, in assessing the viral drug resistance [15] and common diseases susceptibility [16]. Recently a multitude of different statistical methods have been developed towards this goal [15, 17]. Specifically, the advantages of

genotypic essays include low price, commercial test kits availability and quick turnaround times [18].

However, analysis problems can arise when phenotypical response was obtained in vitro while trying to make genotype based predictions [1]. Therefore, interpretation of results from genotypic essays is not straightforward [18]. It is crucial to notice that methods described in detail in this chapter are not centered on predicting phenotype from genotype but instead detecting resistance associated mutation patterns using just the genotype treatment data [1]. Even though analysis methods described here are centered on the understanding of HIV drug-resistance, the statistical toolbox can be applied to other similar genotype problems.

22.3 State and Goals of Antiretroviral Therapy

More than 20 million people have been killed by AIDS since 1980s [19, 20]. The disease is caused by the human immunodeficiency (HIV) virus which leads to the failure of the immune system. Currently there is no cure for AIDS available, but modern treatment therapies can successfully slow disease development [20, 21]. Thus, the goals of the modern antiretroviral therapy consist of stopping AIDS development via suppression of HIV virus in the human body through the use of appropriate drug combinations. In this review, we are mostly concentrating on the more widespread HIV-1 virus.

22.3.1 Structural Biology of HIV

While encoding merely fifteen mature proteins [21], HIV-1 virus can successfully subvert human immune systems. Because every step in the virus replication cycle could be a target for the antiviral treatment [22], we briefly review the life-cycle of HIV-1 virus.

There are 13 important steps in the HIV replication cycle [21]. Particularly, the replication cycle begins with the attachment step which results in the fusion of membranes of the cell and virus and entry into the cell [21]. The last step in the cycle includes the protease-mediated mutation [21, 22]. So far, however, approved drugs aim at only three different targets in the mentioned cycle: reverse transcriptase, protease and viral entry [22].

22.3.2 Drug Classes

Approximately thirty antiretroviral drugs for HIV-1 infection are divided into nucleoside RT inhibitors (NRTI), protease inhibitors (PIs), and non-nucleoside RT

inhibitors (NNRTI) [2, 21, 23, 24]. Currently, there is an active search for other types of drugs that are less susceptible to viral drug resistance capabilities [22].

22.3.3 HAART

As we just described there is significant research directed at development of new HIV drugs; however, another crucial research venue in the HIV studies is the optimization of existing drug combinations. Particularly, clinical applications of highly active antiretroviral therapy (HAART), during which multiple drugs are given in combination, have significantly improved the control over the development of the HIV virus [19, 20]. Therefore, due to recent treatment success, AIDS is now classified as a chronic disease [19]. Yet some drugs have certain toxic effects; moreover, the development of more affordable options is still crucial [21].

Additionally, the use of multiple drugs concurrently leads to the phenomenon of cross-resistance for the HIV virus in addition to usual emergence of drug-resistant variants [2]. We return to the problem of cross-resistance near the end of the chapter.

22.3.4 Pharmacogenomics of HIV Disease

Among the multitude of issues that need to be addressed during the anti-HIV therapy in general and HAART in particular, increasing the efficiency of the drug combination and decreasing its toxicity for the patient are the most crucial ones. For the detailed review of toxicity effects from antiretroviral drugs we refer the reader to Refs. [19, 25]. Briefly, the main types of resulting toxicity effects fall under the following categories: mitochondrial toxicity, hypersensitivity, lipodystrophy, and drug-specific effects [25]. It is important to notice that genetic predisposition probably plays an important role in the magnitude of the adverse side-effects [26].

22.4 Complexity of Mutation Interactions

As can be seen from previous descriptions, anti-HIV therapy represents a complicated science from both biological and clinical points of view. Additional, complexity arises from the necessity to comprehend high-order interactions among the drug resistant mutations of the HIV-1 virus [1, 2]. Mutations can be separated into two groups: marginally or interactively associated with drug treatment.

22.5 Bioinformatics Approach to HIV Resistance

Previously, we mentioned multiple complicated interactions among mutations that have to be considered while developing statistical models for understanding of the HIV virus drug resistance. In Fig. 22.1 we clearly summarize in the graph form all the connections between resistive mutations, epistasis, and drug treatments regimen design. The ultimate goal is to be able to accurately understand all the shown couplings in single and multiple drug treatment studies while also comprehending the biological processes in the virus that lead to resistance development. Thus, while statistical understanding is important, developing approaches that can point in the direction of the appropriate biological molecular level processes taking place is the ultimate goal for all drug resistance approaches including bioinformatics analysis described here.

22.5.1 Overview of Machine Learning Methods

One of the possible approaches for exploring the drug resistance is using methods from machine learning. Specifically, it is possible to use mutual information concept from information theory to statistically calculate the probability of each position to be associated with drug resistance [18]. Furthermore, decision trees can be used for phenotype prediction [18, 20].

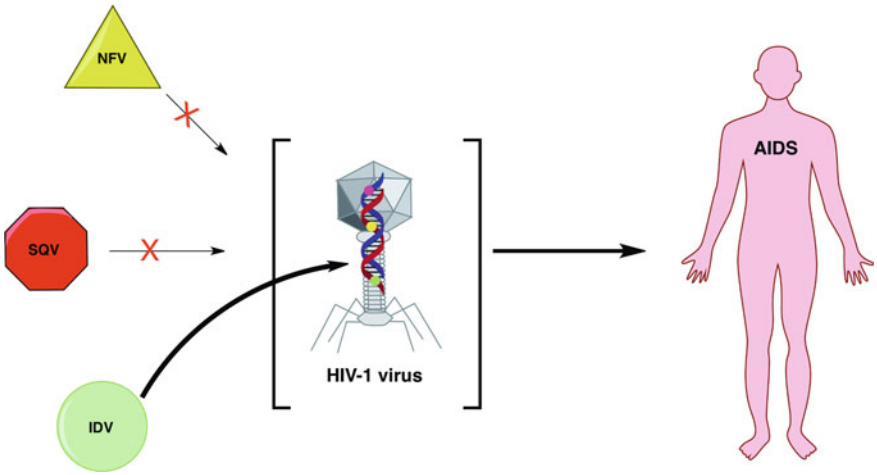


Fig. 22.1 HIV-1 drug resistance diagram. Schematic description of the drug resistance (*square brackets*) to certain HIV-1 suppressing drugs (*crossed red arrows*) arising from mutation in the viral genotype (*colored dots*). The goal is to figure out which mutations are already present to tailor the regimen to each AIDS patient individually. Notice, the diagram is just a potential situation adapted for concreteness purposes

22.5.2 Emergence of Bayesian Methods

In this chapter, however, we focus on the Bayesian approaches to the problem. As we will show, assuming Bayesian paradigm allows significant improvements in understanding of drug resistance and cross-resistance.

22.6 Statistical Background

In order to describe in detail the novel approaches to computational understanding of HIV drug resistance, it is first necessary to review the statistical background of Bayesian inference and associated graphical models. At its core, Bayesian theory provides a mathematical formulation for updating one's current knowledge about the system of interest based on presented or discovered evidence [27]. Thus, Bayesian paradigm provides a way of calculating the probability distribution over a set of hypotheses of interest utilizing the information from the collection of data/observations [28]. In the next few paragraphs we provide a more detailed introduction to the subject to form a solid foundation for introducing methods and models for HIV drug resistance.

22.6.1 Bayesian Data Analysis Paradigm

Statistical conclusions about an unknown parameter θ (or unobserved data y_{unobs}) in the Bayesian approach to parameter estimation are described utilizing probability statements which are conditional on the observed data y: $p(\theta|y)$ and $p(y_{unobs}|y)$. Implicit conditioning is performed on the values of any covariates [29]. The concept of conditioning on the observed data is what separates Bayesian statistics from other inference approaches which estimate unknown parameter over the distribution of the possible data values while conditioning on the true, yet unknown parameter value [27, 29]. Thus, Bayesian paradigm is unique in that sense.

At the heart of all the Bayesian approaches for detection of HIV drug resistant mutation interactions lies the concept of Bayesian inference and, specifically, Bayesian model selection. The goal is to determine the posterior distribution of all parameters in the problem (resistance association and epistatic mutation interactions), given the genetic data for the case-control study (drug treated and untreated patients) while incorporating prior beliefs about parameter values. The conditional probability of all parameters given the observed data is proportional to the product of the likelihood function of the data and prior distribution on the parameters [27]:

$$P(parameters|data) = \frac{P(data|parameters)P(parameters)}{P(data)} \tag{22.1}$$

For most of the real world applications in bioinformatics, including genome wide association studies and drug resistance studies, the marginal probability of the data, $P(data)$, cannot be explicitly calculated [30] and, therefore, $P(parameters|data)$ can be known only up to the proportionality constant as shown in Eq. 22.1. However, advanced computational techniques (iterative sampling methods) can be used to determine posterior distribution of parameters [27, 31]. The main task is to make appropriate choices of statistical models to describe $P(data|parameters)$ and also to choose appropriate prior distributions on the values of parameters: $P(parameters)$. Notice that the posterior distribution provides the probability information for all values of *parameters*.

22.6.2 Bayesian Model Selection

Instead of testing each mutation set in a stepwise manner ('frequentist' hypothesis testing), Bayesian approaches fit a single statistical model to all of the data simultaneously [30, 32, 33] allowing for increased robustness when compared to hypothesis testing methods [34, 35]. Another advantage of Bayesian approach to the problem is the ability to quantify all the uncertainties and information and to incorporate previous knowledge about each specific marker or mutation set into the statistical model through the priors [27, 30].

In the Bayesian model selection framework, we are interested in figuring out which of the set of models {M} is the most likely one given the observed data (X). In analogous way to Eq. (22.1), we can find the posterior probability for a particular model M_i given *data*, by replacing *parameters* with M_i:

$$P(M_i|X) \propto P(X|M_i)P(M_i) \tag{22.2}$$

Thus, through comparison of $P(M_i|X)$ and $P(M_j|X)$ it can be determined whether model M_i or M_j is more likely [27].

22.6.3 Graphical Models

High-dimensional biomedical data represents a complex system with intricate dependence structure arising from interactions at the molecular level [36]. Graphical models, which contain in addition to random variables also their conditional independence information, provide a way to model such complicated data in real life [28]. Particularly, Bayesian networks (BN), which represent directed

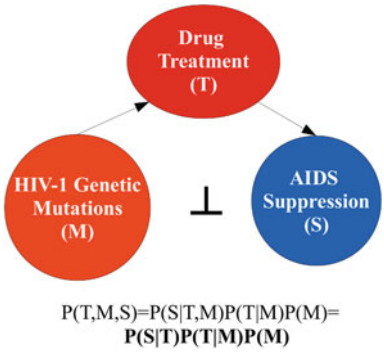


Fig. 22.2 A Bayesian network example. A simple example of using a Bayesian network (*BN*) to model independence relationships between various observables. Here a BN is used to show the effect of mutations in the HIV-1 virus on the AIDS suppression via drug resistance. Particularly, observe the presence of the conditional independence structure in the shown BN: $S \perp M|T$

acyclic graphs, find wide applicability for modeling complex dependencies among random variables [37, 38].

While a particular BN represents a joint probability distribution $p(a,b,c)$, its main advantage is that it contains the information on the conditional independence among the random variables considered [28]. Figure 22.2 shows an example of BN containing conditional independence in the data structure.

While there are a few different methods for BN structural learning from the observed data including constraint-based methods and search-and-score methods [39, 40], we will focus primarily on application of Bayesian recursive model selection algorithm [41] to BN structure inference. Next we will consider how this conceptual framework is applied in practice to extraction of mutation interactions in drug resistance studies.

22.7 Bayesian Modeling of HIV Mutations and Their Epistasis

Statistical approach to Bayesian modeling of HIV drug resistance can be divided into two distinct parts. First, it is necessary to find the mutations in HIV-1 protease or reverse transcriptase associated with drug resistance either individually or through epistasis. Then, in order to describe the structural basis of drug resistance, we need to obtain the detailed interaction structures among the involved mutations. Each of those parts is based on a set of different but equally important statistical models for virus drug resistance described in more detail below but relying on the general ideas of Bayesian data analysis paradigm we reviewed in Sect. 22.6.1.

22.7.1 Bayesian Variable Partition Model

Each marker in question is an amino acid position: HIV-1 protease amino acid sequences from drug-treated patients represent the case group data, while similar sequences from untreated individuals are the controls. BVP model allows for detection of both interacting and non-interacting drug resistance loci among a large number of mutations in the viral genome. It is an application of Bayesian model selection procedure. Particularly, all the markers are split into three non-overlapping groups: (1) mutations not associated with drug-resistance, (2) marginally resistance-associated mutations, and (3) those with interaction associated resistance effect. Thus, using the prior probabilities on the marker memberships and Markov Chain Monte Carlo (MCMC) methods, posterior probabilities for group memberships are determined by iterative sampling from the posterior distribution. Specifically, by interrogating each marker conditionally on the current status of others via MCMC method the algorithm produces posterior probabilities [30]. Particularly, the genotype counts are modeled by the multinomial distribution with the frequency parameters described by the Dirichlet prior. In order to determine the posterior probability of each marker's group membership (represented by I) the Metropolis-Hastings (MH) algorithm [31] is used to sample from $P(I|D,H)$ as given in Eq. (22.3):

$$P(I|T,U) \propto P(T1|I)P(T2|I)P(T0,U|I)P(I) \tag{22.3}$$

where T is the drug treated data set, U is the control data set (untreated HIV patients), and then T0, T1, and T2 are corresponding partitions of the treated data set into the three categories described above. The main model assumption is that case virus genotypes at the treatment associated markers will have different distributions when compared to control genotypes. Furthermore, the likelihood model assumes independence between markers in control group.

Notice that we know the posterior probability for the partition markers only up to the proportionality constant P(T,U), since summing over all possible partitions is computationally unfeasible. However, we can still determine the approximate distribution of P(I|T,U) via sampling from the posterior through advanced Monte Carlo techniques described below. First, it is necessary to specify in detail the statistical models for the likelihood distributions in Eq. (22.3). We assume that marker values for virus amino acid positions of HIV-1 virus come from a multinomial distribution with the frequency parameters described using the Dirichlet prior; moreover, marker combinations in different drug resistance influence groups are independent. If we let $a_i = (a_{i1}, a_{i2}, \ldots, a_{iL_i})$ to be the prior pseudo-counts and integrating out the nuisance parameters [28], we obtain [1]:

$$P(T1|I) = \prod_{i:I_i=1} \left[\left(\prod_{j=1}^{L_i} \frac{\Gamma(n_{ij} + a_{ij})}{\Gamma(a_{ij})} \right) \frac{\Gamma(\sum a_i)}{\Gamma(\sum n_i + a_i)} \right] \tag{22.4}$$

While in the first group we consider the mutations that marginally contribute to the drug resistance effect, in the second group we are looking at the mutations that lead to the virus drug resistance via epistatic interactions. Therefore, it is necessary to look at the mutation configurations among the markers in the group 2.

$$P(T2|I) = \left(\prod_{j=1}^{Q} \frac{\Gamma(n_j + b_j)}{\Gamma(b_j)} \right) \frac{\Gamma(\sum b_i)}{\Gamma(\sum n_i + b_j)} \tag{22.5}$$

where Q represents the number of possible mutation combinations, and $\{b_i\}$ are the prior pseudo-counts. Finally we need to specify P(T0,U|I). Here we combine together the markers not associated with the drug resistance effect and those from the control group data, we also assume that the observations are mutually independent for this group (modeling with saturated multinomial distribution can be performed similarly). By the same analysis path as above for P(T1|I), after integrating out the nuisance parameters, we obtain [1]:

$$P(T0, U|I) = \prod_{i=1}^{p} \left(\prod_{j=1}^{L_i} \frac{\Gamma(m_{ij} + a_{ij})}{\Gamma(a_{ij})} \right) \frac{\Gamma(\sum a_i)}{\Gamma(\sum m_i + a_i)} \tag{22.6}$$

where $\{m_{i1}, \ldots, m_{iL_i}\}$ are the counts of different types of mutations in the (T0 + U) combined data set at each position i. In is also important to note that utilizing the advantage of the Bayesian approach we can incorporate the prior knowledge about the mutations associated with drug resistance through the choice of the proper prior on I: P(I). One possible implementation is to use the multinomial prior to reflect the appropriate proportion of markers associated with the drug resistance. As noted previously, the results of the BVP step is the iterative separation of the markers into the three groups.

22.7.2 Recursive Bayesian Model Selection

If the interactions among the mutations play a significant role in the drug resistance capability of the HIV-1 virus, it is highly desirable to be able to know the detailed interaction patterns to figure out the structural basis of the resistance. However, results of the BVP model use the saturated model for the group 2 variables providing no details about the interactions present in the problem. While inferring a complete Bayesian network structure using one of the methods described in Sect. 6.3 is a possible approach, limited computation resources, insufficient amount of data or a large number of mutations in question might make that approach undesirable. However, limiting the choice to two simple models that provide information about conditional independence among the interacting mutations will make the computational time reasonable and provide sufficient

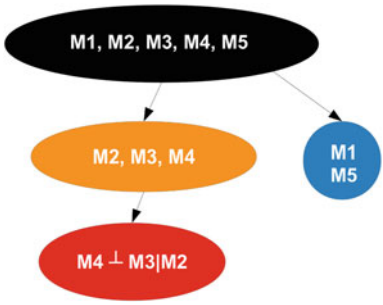


Fig. 22.3 Detailed mutation interaction structure. A diagram of the procedure for the inference of the detailed dependence structure among drug resistant mutations. In this example, five mutations (numbered M1 through M5) were assumed to be associated with the resistance. The determined independence sets within this set are singled out using *circles/oval* and *different colors*

structural details for further analysis of epistasis' contribution to drug resistance on the molecular level. Figure 22.3 shows an example of the independence structure that could potentially be obtained among five different mutations applying the recursive model selection algorithm.

22.7.2.1 Chain-Dependence Model

The first non-saturated model we consider for a group of random variables X_G is the partitioning of the index set into three non-overlapping subgroups such that $X_A \rightarrow X_B \rightarrow X_C$. Therefore, X_A and X_C are conditionally independent given variables X_B (see Fig. 22.4 for details).

Combining this information we obtain the joint distribution for all the variables together:

$$P(X_G) = P(X_A)P(X_B|X_A)P(X_C|X_B). \tag{22.7}$$

Assuming our data D consists of n independent and identically distributed observations, and Π is the set partitioning, we can write the posterior distribution of the data under the chain-dependence model as:

$$P(D, \Pi) = P(D_A|\Pi)P(D_B|D_A, \Pi)P(D_C|D_B, \Pi)P(\Pi). \tag{22.8}$$

Thus, we need to calculate the likelihood $P(D|\Pi)$. It can be shown [1] that

$$P(D_A|\Pi) = \left(\prod_{k=1}^{N_A} \frac{\Gamma\left(n_k^A + \beta_k^A\right)}{\Gamma\left(\beta_k^A\right)}\right) \frac{\Gamma\left(\sum_{k=1}^{N_A} \beta_k^A\right)}{\Gamma\left(n + \sum_{k=1}^{N_A} \beta_k^A\right)} \tag{22.9}$$

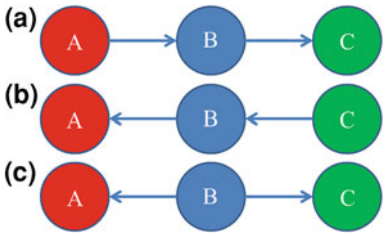


Fig. 22.4 The chain-dependence model structure. *Colors* indicate different sets of variables Three different formulizing directions (**a**–**c**) are shown, but they are all equivalent structures in BN

$$P(D_B|D_A,\Pi)=\prod_{i=1}^{N_A}\left[\frac{\Gamma\left(\vec{n}_{\cdot|i}^{B|A}+\beta_{\cdot|i}^{B|A}\right)\Gamma\left(\left|\beta_{\cdot|i}^{B|A}\right|\right)}{\Gamma\left(n_i^A+\left|\beta_{\cdot|i}^{B|A}\right|\right)\Gamma\left(\beta_{\cdot|i}^{B|A}\right)}\right] \tag{22.10}$$

$$P(D_C|D_B,\Pi)=\prod_{j=1}^{N_B}\left[\frac{\Gamma\left(\vec{n}_{\cdot|j}^{C|B}+\beta_{\cdot|j}^{C|B}\right)\Gamma\left(\left|\beta_{\cdot|j}^{C|B}\right|\right)}{\Gamma\left(n_j^B+\left|\beta_{\cdot|j}^{C|B}\right|\right)\Gamma\left(\beta_{\cdot|j}^{C|B}\right)}\right] \tag{22.11}$$

Multiplying the three equations above and the prior on the partitions we obtain the posterior of the data. Using the appropriately designed MCMC algorithm we sample from $P(D,\Pi)$.

22.7.2.2 V-dependence Model

The V-dependence model is shown in Fig. 22.5 in detail. Thus, under this model we partitioned data X_G into subsets A, B, and C such that variables in sets A and C are mutually independent. As in our description above for the chain-dependence model, we are interesting in calculating the posterior $P(D,\Pi)$, so we need to set up

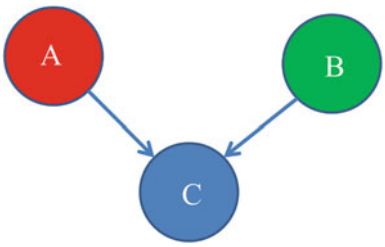


Fig. 22.5 The V-dependence model structure. *Colors* indicate different sets of variables. Notice that variables in *A* are marginally independent of the variables in *B*

the appropriate models for the likelihood $P(D|\Pi)$. First, the likelihood can be decomposed as:

$$P(D|\Pi) = P(D_A|\Pi)P(D_C|\Pi)P(D_B|D_A, D_C, \Pi). \tag{22.12}$$

First, since variables in sets A and C are marginally independent according to our model, by analogy with the results in Sect. 7.2.1 we can use multinomial-Dirichlet distributions on X_A and X_C to obtain the appropriate expressions for $P(D_C|\Pi)$ and $P(D_C|\Pi)$ which are similar to equations above. Finally, in order to find the expression for $P(D_B|D_A, D_C, \Pi)$, we use $\vec{\theta}_i^{B|AC} = \left\{\theta_{j|i}^{B|AC}, \quad j = 1, \ldots, N_B\right\}$ for the transition probabilities between $A \cup C$ and B, and $n_{j|i}^{B|AC}$ for the number of transitions from allele combination i in $A \cup C$ to allele combination j in set B. After assigning a Dirichlet prior to $\vec{\theta}_i^{B|AC}$ and integrating out the nuisance parameters, we obtain:

$$P\left(D_B|D_A, D_C, \Pi\right) = \prod_{i=1}^{N_A \times N_C} \left[\left(\prod_{j=1}^{N_B} \frac{\Gamma\left(n_{j|i}^{B|AC} + \beta_{j|i}^{B|AC}\right)}{\Gamma\left(\beta_{j|i}^{B|AC}\right)} \right) \frac{\Gamma\left(\left|\vec{\beta}_i^{B|AC}\right|\right)}{\Gamma\left(n_i^{AC} + \left|\beta_i^{B|AC}\right|\right)} \right]. \tag{22.13}$$

Combining the likelihood with the prior $P(\Pi)$, we can use an appropriate MCMC algorithm to sample from the posterior.

22.7.2.3 Sampling Strategy

Since the data D can follow either one of the above mentioned models, we need to determine the posterior:

$$P(I_{CV}, \Pi|D) \propto P(D|I_{CV}, \Pi)P(I_{CV})P(\Pi), \tag{22.14}$$

where I_{CV} is the indicator for the chain-dependence or V-dependence models. Since the constant of proportionality is unknown in expression for the posterior above, it becomes necessary to find the appropriate MCMC algorithms [31, 42] to sample from the posteriors. For a sampling from posterior given by Eq. (22.14) we use the Metropolis-Hastings (MH) algorithm. Particularly, following the description in Ref. [1], three types of proposals can be used, including (1) randomly changing two markers between the groups, (2) changing the group membership for a randomly chosen marker and finally (3) switching between the V-dependence and chain-dependence models. During the computational step of the sampling process, the acceptance decision is done according to the MH ratio of the two Gamma functions.

22.8 Applications of Bayesian Methodology to HIV Drug Resistance

Described Bayesian systematic procedure for treatment associated mutation epistasis has been already successfully applied to multiple HIV drugs in the context of both single-drug and multiple-drug treatments. Table 22.1 provides a compact summary of the results of such Bayesian analysis. Notice that multiple statistically significant interactions among resistance causing mutations have been discovered using Bayesian approaches. Results contain the analysis of multiple mutation-prone positions. It is important to observe that molecular basis of multiple interacting mutations discovered with RMS was analyzed with MD simulations and free energy calculations [2]. Therefore, this is an example of the statistical study where biological processes underling drug resistance can be extracted from the discovered independence groups. For sure, many other studies applying Bayesian methodology to other drug classes and drug combinations will follow in the future as more data become available.

Table 22.1 Results for applications of Bayesian methods to HIV drug resistance studies

Drugs	Antiretroviral effect	Discovered mutation interactions	Details
Indinavir (IDV)	Protease inhibitor	{24,47{32{46⊥54|82}}}{10,71}{73,90}	Interesting group {46,54,82}[a]
Zidovudine	Nucleoside analog RT inhibitor	{41,210,215}{67,219}{70}	Further biochemical investigations needed[b]
Nevirapine	Non-nucleoside RT inhibitor	{106}{188}{103?181}{190}	Weak interactions
IDV, SQV	Protease inhibitors	{61,71}{46,54,82}{73,90}	Other details ambiguous
IDV, NFV	Protease inhibitors	{24,54,82}{30,88}{73,90}	6 positions disappeared[c]
IDV, NFV, SQV	Protease inhibitors	{30,88}{73,90}{24,46,54,82}	Ambiguous structure in 3rd group

Epistatic mutations discovered with BVP approach are partitioned using RMS algorithm. Independence groups are enclosed in brackets. "?" indicates inconclusive result. The table contains summary of the results from Refs. [1, 2]

[a] Sequential mutation acquisition in this group leads to conditional independence. The results were confirmed by the MD simulations

[b] It is not possible to study the structural basis of mutations using MD simulations for Zidovudine

[c] When compared to single-drug treatment profiles

22.9 Extensions of the Model

As we previously referred to at the beginning of the chapter, described statistical methodology can be applied to other genotype-phenotype problems. Particularly, studies have already been performed exploring multiple-drug treatment effects on HIV-1 drug resistance. Particularly, in order to adapt the BVP and RMS algorithms in such instances Ref. [2] authors used sequential application of the algorithms to data from single-, double-, and triple-drug treated HIV patients. Comparing mutation interactions in such runs, they discovered that drug-resistant effects are not additive, but on the contrary significantly different from the independent conjecture [2]. This discovery points to the necessity to update the current clinical state of the art approaches to cross-resistance effects since they are usually ignoring significant epistasis effects discovered. Another set of applications of the presented models lies in HIV coreceptor usage [4] and occult HBV infection [3]. Therefore, applications to other infectious diseases and cancer cells of the modified Bayesian methods could follow in the near future with important medical results.

22.10 Conclusions and Future Prospects

In this chapter we described the details of the new procedure for detailed understanding of complex mutation interactions leading to HIV-1 drug resistance in single and multiple-drug treatments. Results applying described methodology can provide important information for clinical applications. Particularly, conditional independence structures discovered will aid in clinical calculations of relative risks for developing drug resistance for each patient given isolated mutation patterns.

Certain important issues need to be addressed in more detail with regard to Bayesian statistical analysis of viral drug resistance. For example, emergence of bias can result from multiple subpopulations in the data. Moreover, sensitivity of the BVP algorithm may be affected by the transmitted resistance occurrence [1].

Even though the focus of the chapter was on statistical methods of analyzing drug resistance, it is crucial to understand the importance of trying to connect discovered mutation groups and biochemical underpinnings of the HIV-1 drug resistance process. Thus, performing biochemical investigations and molecular simulations of the discovered epistatic interactions plays a significant role for clinical applications of the Bayesian results.

References

1. Zhang J et al (2010) Detecting and understanding combinatorial mutation patterns responsible for HIV drug resistance. PNAS 107:1321
2. Zhang J et al (2012) Systematic investigation on interactions for HIV drug resistance and cross-resistance among protease inhibitors. J Proteome Sci Comput Biol 1:2. doi:http://dx.doi.org/10.7243/2050-2273-1-2

3. Svicher V et al (2012) Novel HBsAg markers tightly correlate with occult HBV infection and strongly affect HBsAg detection. Antiviral Res 93(1):86–93
4. Svicher V et al (2011) Identification and structural characterization of novel genetic elements in the HIV-1 V3 loop regulating coreceptor usage. Antiviral Therapy 16(7):1035–1045
5. Svicher V et al (2011) Key-genetic elements in HIV-1 gp120 V1, V2, and C4 domains tightly and differentially modulate gp120 interaction with the CCR5 and CXCR4 N-terminus and HIV-1 antigenic potential. The 2nd international HIV and hepatitis virus drug resistance workshop. Antiviral Ther 16(4):A14–A14
6. Svicher V et al. (2011) Specific HBsAg genetic-determinants are associated with occult HBV-infection in vivo and HBsAg-detection. The 2nd international HIV and hepatitis virus drug resistance workshop. Antiviral Ther 16(4):A85–A85
7. Alteri C et al (2011) Signature mutations in V3 and bridging sheet domain of HIV-1 gp120 HIV-1 are specifically associated with dual tropism and modulate the interaction with CCR5 N-Terminus. Italian conference on AIDS and retroviruses: ICAR 2011, Florence
8. McCarthy MI et al (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9:356–369. doi:10.1038/nrg2344
9. Hall SS (2010) Revolution postponed. Sci Am 303:60–67. doi:10.1038/scientificamerican1010-60
10. Lander ES (2011) Initial impact of the sequencing of the human genome. Nature 410:187–197. doi:10.1038/nature09792
11. Cordell HJ (2009) Detecting gene–gene interactions that underline human diseases. Nat Genet 10:392–404. doi:10.1038/nrg2579
12. Kozyryev I, Zhang J (2012) Bayesian exploration of multilocus interactions on the genome-wide scale. Am J Bioinformat 1:70–78. doi:10.3844/ajbsp.2012.70.78
13. Zhang Y, Liu JS (2007) Bayesian inference of epistatic interactions in case-control studies. Nature Genet 39:1167–1173. doi:10.1038/ng2110
14. Condra JH et al (1995) In vivo emergence of HIV-1 variants resistant to multiple protease inhibitors. Nature 374:569–571
15. Ravela J et al (2003) HIV-1 protease and reverse transcriptase mutation patterns responsible for discordances between genotypic drug resistance interpretation algorithms. J Acq Immun Def Synd 33:8–14
16. WTCCC (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678. doi:10.1038/nature05911
17. Rhee SY et al (2006) Genotypic predictors of human immunodeficiency virus type 1 drug resistance. P Natl Acad Sci USA 103:17355–17360
18. Beerenwinkel N et al (2002) Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype. P Natl Acad Sci USA 99:8271–8276
19. Pirmohamed M, Back DJ (2001) The pharmacogenomics of HIV therapy. Pharmacogenomics J 1:243–253
20. Lengauer T, Sing T (2006) Bioinformatics-assisted anti-HIV therapy. Nat Rev Microbiol 4:790–797. doi:10.1038/nrmicro1477
21. Engelman A, Cherepanov P (2012) The structural biology of HIV-1: mechanistic and therapeutic insights. Nat Rev Microbiol 10:279–290. doi:10.1038/nrmicro2747
22. Flexner C (2007) HIV drug development: the next 25 years. Nature Rev Drug Discov 6:959–966
23. Shafer RW (2002) Genotypic testing for Human Immunodeficiency virus type 1 drug resistance. Clin Microbiol Rev 15:247–277
24. Shafer RW et al (2007) HIV-1 protease and reverse transcriptase mutations for drug resistance surveillance. AIDS 21:215–223
25. Carr A, Cooper DA (2000) Adverse effects of antiretroviral therapy. Lancet 356:1423–1430
26. Pirmohamed M, Park BK (2001) Genetic susceptibility to adverse drug reactions. Trends Pharmacol Sci 22:298–305
27. Rice J (2006) Mathematical statistics and data analysis, 3rd edn. ISBN 0534399428

28. Hamelryck T (2012) An overview of Bayesian inference and graphical models. In: Bayesian methods in structural bioinformatics. doi:10.1007/978-3-642-27225-7_1
29. Gelman A et al (2003) Bayesian data analysis, 2nd edn. ISBN 158488388X
30. Zhang Y, Liu JS (2007) Bayesian inference of epistatic interactions in case-control studies. Nat Genet 39:1167–1173. doi:10.1038/ng2110
31. Liu JS (2001) Monte Carlo strategies in scientific computing, 1st edn. Springer, New York. ISBN 0387952306
32. Zhang Y et al (2011) Block-based Bayesian epistasis association mapping with application to WTCCC type 1 diabetes data. Ann Appl Stat 5:2052–2077. doi:10.1214/11-AOAS469
33. Zhang J et al (2011) A Bayesian method for disentangling dependent structure of epistatic interaction. Am J Biostat 2:1–10
34. McCarthy MI et al (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9:356–369. doi:10.1038/nrg2344
35. Zhang Y (2012) A novel graphical model for genome-wide multi-SNP association mapping. Genet Epidemiol 36:36–47. doi:10.1002/gepi.20661
36. Friedman N (2004) Inferring cellular networks using probabilistic graphical models. Science 303:799–805
37. Lauritzen SL (1996) Graphical models. Clarendon Press, Oxford
38. Pearl J (2000) Causality. Cambridge University Press, Cambridge
39. Cooper GF, Herskovits E (1992) The induction of probabilistic networks from data. Mach Learn 9:309–347
40. Heckerman D et al (1995) Learning Bayesian networks: the combination of knowledge and statistical data. Mach Learn 20:197–243
41. Zhang J et al (2012) High-order interactions in rheumatoid arthritis detected by Bayesian method using genome-wide association studies data. Am Med J 3(1):56–66. doi:10.3844/amjsp.2012.56.66
42. Ferkinghoff-Borg J (2012) Monte Carlo methods for inference in high-dimensional systems. In: Bayesian methods in structural bioinformatics. doi:10.1007/978-3-642-27225-7_2