# 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

© Shanghai Jiao Tong University Press, Shanghai and Springer Science+Business Media Dordrecht 2015,

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 $\beta$-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.