Challenges and Advances
in Computational Chemistry and Physics 27
Series Editor: Jerzy Leszczynski

C. Gopi Mohan *Editor*

Structural Bioinformatics: Applications in Preclinical Drug Discovery Process

Springer

Challenges and Advances in Computational Chemistry and Physics

Volume 27

Series editor

Jerzy Leszczynski
Department of Chemistry and Biochemistry
Jackson State University, Jackson, MS, USA

This book series provides reviews on the most recent developments in computational chemistry and physics. It covers both the method developments and their applications. Each volume consists of chapters devoted to the one research area. The series highlights the most notable advances in applications of the computational methods. The volumes include nanotechnology, material sciences, molecular biology, structures and bonding in molecular complexes, and atmospheric chemistry. The authors are recruited from among the most prominent researchers in their research areas. As computational chemistry and physics is one of the most rapidly advancing scientific areas such timely overviews are desired by chemists, physicists, molecular biologists and material scientists. The books are intended for graduate students and researchers.

All contributions to edited volumes should undergo standard peer review to ensure high scientific quality, while monographs should be reviewed by at least two experts in the field. Submitted manuscripts will be reviewed and decided by the series editor, Prof. Jerzy Leszczynski.

More information about this series at http://www.springer.com/series/6918

C. Gopi Mohan
Editor

Structural Bioinformatics: Applications in Preclinical Drug Discovery Process

Springer

Editor
C. Gopi Mohan
Amrita Centre for Nanosciences
and Molecular Medicine
Amrita Institute of Medical Sciences
and Research Centre
Kochi, India

ISSN 2542-4491 ISSN 2542-4483 (electronic)
Challenges and Advances in Computational Chemistry and Physics
ISBN 978-3-030-05281-2 ISBN 978-3-030-05282-9 (eBook)
https://doi.org/10.1007/978-3-030-05282-9

Library of Congress Control Number: 2018962784

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Human society has immense faith in the potential of drugs. Our belief towards therapeutically safer drugs to alleviate the symptoms of different types of diseases is accelerating nowadays. The twenty-first century witnessed tremendous progress in the scientific and technical aspects in several therapeutic domains, such as viral, bacterial, cancer and other metabolic and infectious diseases. Further, bioinformatics and computational biology disciplines are integrated into all levels of medicine and health care. Future breakthroughs will depend on the strong collaborations between experimental and computational biologists. Areas such as building predictive models of the cell, organelles, and organs, understanding ageing, designing enzymes, and improving drug design and target validation are becoming crucial for the drug discovery programme.

The main concept of the present book includes computer-aided molecular modelling and protein/enzyme design in preclinical discovery towards understanding the molecular mechanisms of different diseases. This technique can be successfully employed in different areas of medical research, including rare and neglected diseases. Different case studies integrated with the experimental research as well the future plan for clinical aspects are described effectively. The present 12 chapters of the book have been contributed by leading and internationally recognized scientists. It addresses computer simulation techniques for studying biological phenomena from the perspective of both methodology and applications. The chapters are organized on the methodology of molecular simulations and its applications, chemoinformatics methods and its use of experimental information in computational simulations. Selected applications of structural biology and structure-based drug design, focussing towards druggable targets, and its physiological molecular mechanisms of actions are critically addressed.

The first five chapters are devoted to theories and methodologies, which form the backbone of the structure-based drug design concepts as well as different molecular modeling techniques in computer-aided drug design. Chapter "Structure-Based Drug Design of *Pf*DHODH Inhibitors as Antimalarial Agents" describes the latest theories and computational methodologies in structure-based drug design for the development of inhibitors against key druggable target *Plasmodium falciparum* dihydroorotate

dehydrogenase. Chapter "Recent Advancements in Computing Reliable Binding Free Energies in Drug Discovery Projects" is dedicated to understanding the protein–ligand binding affinities and different concepts and methods towards free energy calculations for the drug discovery projects. Next chapter (Chapter "Integrated Chemoin–formatics Approaches Towards Epigenetic Drug Discovery") addresses the epigenetics molecular mechanism and its key targets involved in different diseases by efficiently employing different chemoinformatics strategies. Chapter "Structure-Based Drug Design with a Special Emphasis on Herbal Extracts" directly deals with the natural products, a component of Ayurinformatics, and its emphasis on the application of structure-based drug design. Chapter "Impact of Target-Based Drug Design in Anti-bacterial Drug Discovery for the Treatment of Tuberculosis" is devoted completely towards tuberculosis drug discovery and the role of three-dimensional druggable targets in the structure-based anti-tuberculosis design. The role of big data and high-performance computing is prevalent nowadays in different fields, and the concept and algorithms presented in Chapter "Turbo Analytics: Applications of Big Data and HPC in Drug Discovery" directly address its importance and application towards the preclinical drug discovery aspects. Finally, Chapter "Single-Particle cryo-EM as a Pipeline for Obtaining Atomic Resolution Structures of Druggable Targets in Preclinial Structure-Based Drug Design" is devoted towards the latest technique in structural biology, i.e. single-particle cryo-EM to solve the atomic structures of single and multi-protein druggable targets and which is key to the structure-based drug design studies.

> In the future, Computers will design, discover, people will verify—John Rumble
>
> Science knows no country, because knowledge belongs to humanity, and is the torch which illuminates the world—Louis Pasteur
>
> Science is beautiful when it makes simple explanations of phenomena or connections between different observations. Examples include the double helix in biology and the fundamental equations of physics—Stephen Hawking

The purpose of this book is to explore the theoretical strategies involved in drug discovery and development by proper integration with the experimental concepts as well. Further, the book is intended to deliver the reader with an overview of multifaceted, challenging and rapidly evolving field. We feel that the scientific material covered herein will provide the reader with an excellent overview in preclinical drug discovery programme.

Ämrita Vishwa Vidyapeetham, Kochi, India C. Gopi Mohan
October 2018

Contents

Editor and Contributors

About the Editor

Dr. C. Gopi Mohan is Incharge, Bioinformatics and Computational Biology Laboratory, Center for Nanosciences and Molecular Medicine, Amrita Vishwa Vidyapeetham, Kochi. He graduated with a Ph.D. degree from Banaras Hindu University, Varanasi. He gained experience as Postdoctoral Fellow from Molecular Biophysics Unit at IISc, Bangalore, and as Research Officer from Department of Biology and Biochemistry, University of Bath, UK. Further, he worked as Associate Researcher of CNRS, University Henri Poincare, Nancy, France. During his research career, he visited different countries which include UK, Canada, France, Finland, and USA.

He has supervised many Ph.D. and postgraduate students and completed different research and industrial consultancy projects. He has published more than 80 peer-reviewed research papers and chapters and is Active Reviewer of international/national journals, thesis, and grants. His research interests are Computational Biology & Structural Bioinformatics, Structure-Based Drug Design, Protein Crystallography, and Nanoinformatics. He is Member of the Indian Biophysical Society and the American Chemical Society. He was Invited Speaker for different international conferences and recently was awarded ICMR-Senior Biomedical Scientist International Fellowship.

Contributors

Avinash Bayaskar High Performance Computing-Medical & Bioinformatics Applications Group, Centre for Development of Advanced Computing (C-DAC), Savitribai Phule Pune University Campus, Pune, India

Shweta Bhagat Department of Medicinal Chemistry, National Institute of Pharmaceutical Education and Research (NIPER), S.A.S. Nagar, Punjab, India

Prasad V. Bharatam Department of Medicinal Chemistry, National Institute of Pharmaceutical Education and Research (NIPER), S.A.S. Nagar, Punjab, India

Atanu Bhattacharjee Department of Biotechnology & Bioinformatics, North-Eastern Hill University, Shillong, India

Raja Biswas Center for Nanosciences and Molecular Medicine, Amrita Institute of Medical Sciences and Research Centre, Amrita Vishwa Vidyapeetham, Kochi, Kerala, India

Chinmayee Choudhury Center for Molecular Modelling, Indian Institute of Chemical Technology, Hyderabad, India; Department of Biochemistry, All India Institute of Medical Sciences, Basni, Jodhpur, Rajasthan, India

Evans C. Coutinho Molecular Simulations Group, Department of Pharmaceutical Chemistry, Bombay College of Pharmacy, Mumbai, India

Anuj Gahlawat Department of Pharmacoinformatics, National Institute of Pharmaceutical Education and Research (NIPER), S.A.S. Nagar, Punjab, India

Vivek Gavane High Performance Computing-Medical & Bioinformatics Applications Group, Centre for Development of Advanced Computing (C-DAC), Savitribai Phule Pune University Campus, Pune, India

Indira Ghosh School of Computational and Integrative Sciences (SCIS), Jawaharlal Nehru University, New Delhi, India

Pawan Gupta Advanced Protein Science Building, Institute of Microbial Technology, Chandigarh, India

Vinod Jani High Performance Computing-Medical & Bioinformatics Applications Group, Centre for Development of Advanced Computing (C-DAC), Savitribai Phule Pune University Campus, Pune, India

Rajendra R. Joshi High Performance Computing-Medical & Bioinformatics Applications Group, Centre for Development of Advanced Computing (C-DAC), Savitribai Phule Pune University Campus, Pune, India

Vikrant Karmahapatra Advanced Protein Science Building, Institute of Microbial Technology, Chandigarh, India

Shruti Koulgi High Performance Computing-Medical & Bioinformatics Applications Group, Centre for Development of Advanced Computing (C-DAC), Savitribai Phule Pune University Campus, Pune, India

N. H. V. Kutumbarao CAS in Crystallography and Biophysics, University of Madras, Chennai, Tamil Nadu, India

Saurabh Loharch Advanced Protein Science Building, Institute of Microbial Technology, Chandigarh, India

Rethi Madathil Advanced Protein Science Building, Institute of Microbial Technology, Chandigarh, India

Rashmi Mahajan High Performance Computing-Medical & Bioinformatics Applications Group, Centre for Development of Advanced Computing (C-DAC), Savitribai Phule Pune University Campus, Pune, India

Sandeep Malviya High Performance Computing-Medical & Bioinformatics Applications Group, Centre for Development of Advanced Computing (C-DAC), Savitribai Phule Pune University Campus, Pune, India

Elvis A. F. Martis Molecular Simulations Group, Department of Pharmaceutical Chemistry, Bombay College of Pharmacy, Mumbai, India

C. Gopi Mohan Center for Nanosciences and Molecular Medicine, Amrita Institute of Medical Sciences and Research Centre, Amrita Vishwa Vidyapeetham, Kochi, Kerala, India

N. Arul Murugan Department of Theoretical Chemistry and Biology, School of Engineering Sciences in Chemistry, Biotechnology and Health, Royal Institute of Technology, Stockholm, Sweden

G. Narahari Sastry Center for Molecular Modelling, Indian Institute of Chemical Technology, Hyderabad, India

Ramanathan Natesh School of Biology, Indian Institute of Science Education and Research Thiruvananthapuram (IISER-TVM), Trivandrum, Kerala, India

Debnath Pal Department of Computational and Data Sciences, Indian Institute of Science, Bangalore, India

Shailesh Kumar Panday School of Computational and Integrative Sciences (SCIS), Jawaharlal Nehru University, New Delhi, India

Sudhir Pandey High Performance Computing-Medical & Bioinformatics Applications Group, Centre for Development of Advanced Computing (C-DAC), Savitribai Phule Pune University Campus, Pune, India

Raman Parkesh Advanced Protein Science Building, Institute of Microbial Technology, Chandigarh, India

Vasanthanathan Poongavanam Department of Physics, Chemistry, Pharmacy, University of Southern Denmark, Odense M, Denmark

U. Deva Priyakumar CCNSB, International Institute of Information Technology, Gachibowli, Hyderabad, India

Anju Choorakottayil Pushkaran Center for Nanosciences and Molecular Medicine, Amrita Institute of Medical Sciences and Research Centre, Amrita Vishwa Vidyapeetham, Kochi, Kerala, India

E. P. Ramakrishnan High Performance Computing-Medical & Bioinformatics Applications Group, Centre for Development of Advanced Computing (C-DAC), Savitribai Phule Pune University Campus, Pune, India

Chandan Raychaudhury Department of Computational and Data Sciences, Indian Institute of Science, Bangalore, India

Md. Imbesat Hassan Rizvi Department of Computational and Data Sciences, Indian Institute of Science, Bangalore, India

Amit Saxena High Performance Computing-Medical & Bioinformatics Applications Group, Centre for Development of Advanced Computing (C-DAC), Savitribai Phule Pune University Campus, Pune, India

Neeru Sharma High Performance Computing-Medical & Bioinformatics Applications Group, Centre for Development of Advanced Computing (C-DAC), Savitribai Phule Pune University Campus, Pune, India

Dmitrii Shcherbinin Laboratory of Structural Bioinformatics, Institute of Biomedical Chemistry, Moscow, Russia

Uddhavesh Sonavane High Performance Computing-Medical & Bioinformatics Applications Group, Centre for Development of Advanced Computing (C-DAC), Savitribai Phule Pune University Campus, Pune, India

Mallikarjunachari Uppuladinne High Performance Computing-Medical & Bioinformatics Applications Group, Centre for Development of Advanced Computing (C-DAC), Savitribai Phule Pune University Campus, Pune, India

D. Velmurugan CAS in Crystallography and Biophysics, University of Madras, Chennai, Tamil Nadu, India

Alexander Veselovsky Laboratory of Structural Bioinformatics, Institute of Biomedical Chemistry, Moscow, Russia

V. Viswanathan CAS in Crystallography and Biophysics, University of Madras, Chennai, Tamil Nadu, India

Free Energy-Based Methods to Understand Drug Resistance Mutations

Elvis A. F. Martis and Evans C. Coutinho

Abstract In this chapter, we present an overview of various computational methods, particularly, those that are used to compute the free energy of binding to understand target site mutations that will enable us to foresee mutations that could significantly affect drug binding. We begin by looking at the driving forces that lead to drug resistance and throw some light on the various mechanisms by which drugs can be rendered ineffective. Next, we studied molecular dynamic simulations and its use to understand the thermodynamics of protein–ligand interactions. Building on these fundamentals, we discuss various methods that are available to compute the free energy binding, their mathematical formulations, the practical aspects of each these methods and finally their use in understanding drug resistance.

Keywords Molecular dynamics · Drug resistance · MM-PB(GB)-SA Free energy perturbation · Linear interaction energy · Computational mutational scanning · Thermodynamic integration

1 Drug Resistance Problem

Every organism attempts to survive in hostile conditions by making minor modifications in its life cycle. Though these modifications are observed phenotypically, genetic reshuffling and alterations are the underlying cause of these changes. Although we are unable to accurately explain this phenomenon and its initiation, we have been able to use this observed knowledge and empirically derive explanations for such modifications. However, it may not always be necessary to know all the details regarding genetic modifications, so long as we can correctly, at least empirically, understand such observations, and put it to effective use to predict and understand the drug resistance problem. Often the enzymes in the biochemical

E. A. F. Martis · E. C. Coutinho (✉)
Molecular Simulations Group, Department of Pharmaceutical Chemistry, Bombay College of Pharmacy, Kalina, Santacruz [E], Mumbai 400098, India
e-mail: evans.coutinho@bcp.edu.in

C. G. Mohan (ed.), *Structural Bioinformatics: Applications in Preclinical Drug Discovery Process*, Challenges and Advances in Computational Chemistry and Physics 27, https://doi.org/10.1007/978-3-030-05282-9_1

pathways undergo mutations to improve the survival rate of the organism by either improving the protein function or catalytic efficiency and stability to escape the inhibitory action of the drug. In the latter case, the motive for modifying the drug target is to ensure that drug binding is weakened. Moreover, the mutations are such that substrate binding is unaffected or minimally affected. Most of the computational methods employed to study the mechanism of drug resistance, attempt to understand the differences in the binding patterns of the substrate and the drug molecule, i.e. understanding the "**substrate-envelope hypothesis**". Here, we present an overview of those computational methods that employ free energy of binding as a tool to gauge the differences in the binding of the substrate and the drug molecule before and after mutation.

In the Sect. 1, we discuss the driving force for resistant mutations and throw some light on the different mechanisms by which drug resistance can occur. In Sect. 2, we present a brief overview of molecular dynamics, thermodynamics of protein–ligand binding, and various methods for computing the free energy of binding. The last section, Sect. 3, has a detailed discussion on various free energy-based methods used to understand and predict the target site mutations leading to loss in drug binding.

1.1 Overview of the Mechanisms of Drug Resistance

The drug-induced selection pressure [1–4] is the major driving force for infectious organisms to try to evade the effects of drugs. One of the primary moves that any organism will adopt is to disrupt the action of drug molecules by one or more possible mechanisms. To show its effect, the drug must enter the cells and find its target protein. As a primary defence mechanism against drugs, the organism may down regulate the expression of influx channels that enable the entry of the drug, resulting in a decreased concentration build-up within the cell. Another strategy that hinders the build-up of the drug inside the cell is the upregulation of the expression of efflux channels/pumps that facilitate the egress of the drug molecules. These strategies are often very difficult to understand owing to the complicated pathways involved in the upregulation or downregulation of various proteins associated in the regulation of traffic to and from the cell. This attribute is difficult to study using computational techniques that use free energy-based methods. Target site mutations [5–8] that lead to disruption in the drug binding without significant loss of the protein function [9, 10] is another mechanism of drug resistance. Such mutations can be studied using computer simulations that enable us to estimate the free energy difference between the drug binding to the mutant and the wild-type protein. An essential factor to consider while understanding target site mutation is the fitness cost associated with the mutational change. This can be estimated by the change in the free energy of binding of the natural ligands/substrates; for example, a drop in their binding energy indicates that substrate binding is impeded, which this leads to increased fitness cost. This means the enzyme now must expend more energy to

carry out the same reaction. Hence, we can assume that such mutations are seldom seen, and if at all they occur, a compensatory mutation(s) will be seen to counter the detrimental effects of those mutations [11, 12]. Another strategy adopted by organisms is to increase the production of drug-metabolizing enzymes that modify the drugs to their inactive form eventually leading to their elimination. A classic example of this is the inactivation of penicillin by the enzyme β-lactamase.

1.2 *Overview of Computational Methods to Study Drug Resistance*

Broadly, computer-assisted methods used to study drug resistance can be classified into two categories based on the information they require and the output they return. The first category of methods requires only 1D sequence data as input and the output is generally a classification type, i.e. the test sequence is classified as a resistant or a non-resistant sequence. Thus, the methods grouped under this class are collectively called as "sequence-based" methods [13]. The workflow of these methods is akin to machine learning or QSAR type classification methods. In a nutshell, sequence-based methods require sequences with the corresponding biological activity data (K_i or IC_{50} or any other suitable numerical value) for the drug under study. Such data can be curated from databases like HIVDB (for HIV resistance, curated and maintained by Stanford University; [14, 15]) CancerDR (for cancer resistance, curated by CSIR Institute of Microbial Technology and OSDD, India; [16]), tuberculosis resistance mutation database (curated and maintained by various departments and schools with Harvard University; [17], and many other such databases. The data is then split into training and test sets to develop and validate the predictive models. The advantage of such methods is that it is not necessary to know the tertiary structure of the protein or the drug-receptor interactions. Therefore, sequence-based methods are computationally inexpensive and large amount of data can be trained to obtain decent quality predictive models in a short time. However, they suffer from two major drawbacks; (1) a lot of a priori information on drug-resistant mutations is needed to train/develop predictive models and (2) no mechanistic insights or atomistic details can be obtained.

The drawbacks seen in the sequence-based methods are efficiently overcome by structure-based methods [13, 18, 19]. Further, structure-based methods are the methods of choice when atomistic details are desired. However, these additional details come at an added computational cost and require high-resolution protein structures to be able to make accurate and reliable predictions. However, unlike the sequence-based methods, they do not require large a priori information on mutations; on the contrary, they can be applied to systems where no data on mutation is available. To assess the binding stability which is the basis for predictions, these methods employ either empirical scoring functions that implicitly try to reflect the free energy of binding or use techniques that compute the free energy of binding

per se. Molecular docking-based methods use empirical scoring functions to find the best docking conformations, and these methods are computationally less expensive. Therefore, they can be applied to assess many protein–ligand complexes. The ligand can be docked to various mutant proteins to predict their binding strength before and after mutations, and this will allow one to understand the effect of the mutation on the binding strength. The accuracy of docking-based methods relies on the accuracy of the scoring function, and they are best suited for rank ordering of compounds rather than computing the absolute free energy of binding. The major issue with docking-based methods is that most docking programs treat proteins as rigid entities, and therefore, mutations in highly flexible protein–ligand systems are poorly understood [19]. However, in recent times there have been several attempts to incorporate protein flexibility in molecular docking [20]. This has largely improved the enrichment scores. Due to the limited scope of this chapter, such docking methods will not be discussed here and have been treated elsewhere [21–25]. Molecular dynamics-based methods can incorporate flexibility in the protein–ligand complexes, and in most cases, are the methods of choice as a conformational sampling tool to explore the phase space accessible to the system under study. The conformations sampled are used to compute the free energy change. However, the drawback of MD-based methods is the computational cost, which is several magnitudes higher compared to docking-based methods.

Another critical issue that must be addressed about the structure-based methods is, how fast predictions can be made, in addition to how reliable are the predictions. These methods find application in drug discovery programs, wherein additional filters can be placed to weed out molecules likely to encounter a high level of resistance or assist in suitably modifying leads to inhibit the mutant proteins. Drug discovery itself is an extremely lengthy and expensive process, and an additional filter like resistance should be economical in terms of time as well as money. Moreover, such methods should also assist medicinal chemists during lead optimization stages to identify potential groups that will help evade drug resistance and avoid late-stage failures that lead to huge financial losses.

2 Molecular Dynamics Simulations and Free Energy Calculations

2.1 Overview of MD and Conformational Sampling Methods

Computer simulations are very useful in predicting changes in molecular properties brought about by alterations in an atom or a group of atoms, particularly, amino acid residues. Therefore, they find good application in predicting the effect of mutations on drug binding at the active site or elsewhere. Protein design experiments clarify the effect of a mutation on drug or substrate binding, thereby facilitating prediction of drug-resistant mutations. This way the program can be used to

select all mutations wherein drug binding is hampered and substrate binding is either improved or [26].

In case of free energy calculations, molecular dynamics (MD) simulations are the most commonly used technique to generate conformational ensembles. Hence, it is rightly called as one of the main toolkits for theoretically studying biological molecules (Hansson et al. [27], Binder et al. [28]. MD calculates the time-dependent behaviour of particles or atoms, by numerical integration of Newton's second law of motion and predicts the future positions and momenta. MD simulations have provided detailed information on the fluctuations and conformational changes of proteins and nucleic acids upon drug/substrate binding. As a result, it is now routinely used to investigate the structure, dynamics and thermodynamics of biological molecules and their complexes. MD simulations have an advantage in that, starting from an X-ray or NMR solved structure, it can provide insights into the dynamic nature of biomolecules that are inaccessible to experiments. To accurately simulate the behaviour of molecules, one must be able to account for the thermal fluctuations and the environment-mediated interactions arising in diverse and complex systems (e.g., a protein-binding site or bulk solution). This depends on how accurately the force fields represent the atoms and treats the non-bonded interactions. A complete account of force fields can be found in the review by Pissurlenkar et al. [29]. However, most of the biological events occur at timescales that are not routinely reachable by classical MD simulations, for example, protein folding occurs in the timescale of few seconds, whereas drug binding and unbinding occur in the timescale of few microseconds to milliseconds. The routine timescale that is feasible using high-end servers equipped with graphic processing units [30–32] and distributed grid computing [33, 34], is few tens of microseconds, that is nearly 1/100th of the timescale required to study protein folding. Conventional MD suffers from the severe limitation that it is extremely difficult to sample high-energy regions and surmount energy barriers, leading to inaccuracies in free energy calculations.

The limitations of classical MD simulations have motivated the development of new conformational sampling algorithms that facilitate the sampling of conformational space that is inaccessible to classical MD simulation. The simplest way to encourage the system to sample the high-energy regions on the phase space is to increase the target temperature [35]. This leads to increased kinetic energy of the system that enables it to surmount these barriers. However, it has been argued by many, that such elevated temperatures ($\sim$400 K and above) lead to physiologically unrealistic states that may severely distort the results; however, such methods have been found to be advantageous in improving the sampling efficiency during MD simulations. Another method that uses elevated temperature to enhance the sampling is the replica-exchange molecular dynamics (parallel tempering, [36, 37]). In this approach, several replicas are simulated in parallel at different temperatures. At appropriate intervals, the replicas switch temperatures with the nearest replica, and this exchange is governed by the Metropolis acceptance criteria. However, all these methods do not prohibit the system from revisiting the same conformational space. This problem was resolved by adding the memory concept in molecular dynamics

(local elevation method [38] Metadynamics [39]) uses Gaussian potentials that discourage the system from sampling the same conformational space. These are few of the most commonly used methods to tackle sampling problems in molecular dynamics, a complete account on enhanced sampling algorithms can be found elsewhere [40–44].

2.2 An Overview of Thermodynamics of Protein–Ligand Binding

Molecular interactions, between the ligand and receptor, are primarily non-covalent in nature and governed by attractive and repulsive forces. In drug design experiments, the goal is always to optimize the attractive interactions and reduce the repulsive ones [45–47]. Moreover, these associations are temporary, and the lifespan of such complexes are governed by the off rates (K_{off}) or the dissociation constant (K_d), both of which indicate the binding strength of a ligand to its protein counterpart. In the realm of thermodynamics, binding is governed by enthalpic and entropic components [48] given by Eq. 1.

$$\Delta G = \Delta H - T\Delta S \tag{1}$$

where ΔG is the binding free energy; ΔH is enthalpy; ΔS is entropy and T is the temperature in Kelvin.

The association is favourable, i.e. spontaneous when the ΔG_{Gibbs} is negative and unfavourable otherwise. All the binding and pre-binding (recognition and pre-organization) events in biomolecular associations are either enthalpy (ΔH) driven or entropy (ΔS) driven. The enthalpic component represents several types of non-covalent interactions like electrostatic, van der Waals, ionic, hydrogen bonds and halogen bonds, while the entropic components reflect the contribution to binding due the dynamics or flexibility of the system. Computing the enthalpic component of binding has reached far heights, in terms of methods available for calculating the aforementioned type of interactions. However, till date, calculation of the entropic component is extremely difficult, and the algorithms are computationally very demanding.

The Gibbs equation is more relevant in biochemistry for calculating the free energy and is given by Eq. 2:

$$\Delta G_{Gibbs} = -RT \ln K_d \tag{2}$$

where ΔG_{Gibbs} is Gibbs free energy, R is universal gas constant, T is the temperature in Kelvin, K_d is the dissociation constant. Equations 1 and 2, along with the Born–Haber cycle [46] (Fig. 1) form the basis for the development of the methods used to compute the free energy binding. The two main methods are Free energy perturbation (FEP) and Thermodynamics Integration (TI), both of which will be

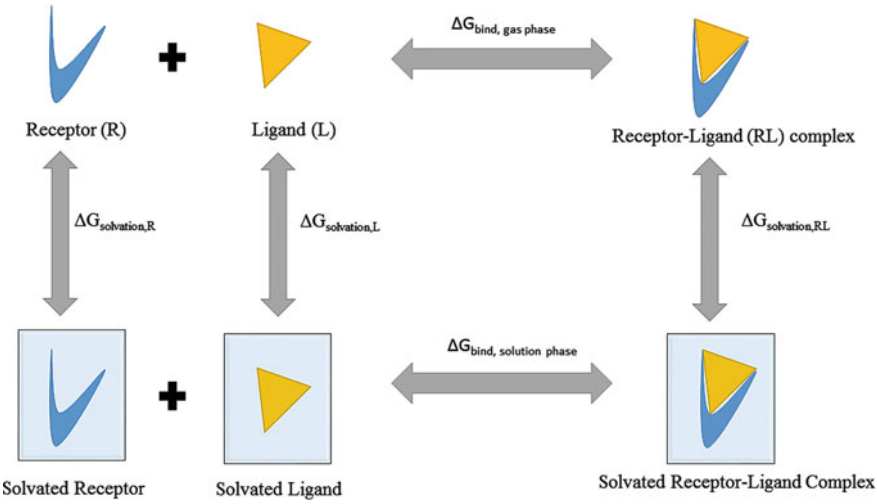

Fig. 1 Thermodynamic or Born–Haber cycle for the receptor-ligand binding

dealt with in the subsequent Sect. 2.3.2. However, measuring the dissociation constants from simulations is a daunting task; nevertheless, computing the partition functions from the molecular simulations is relatively easy. Hence, the ratios of the partition functions can be used to estimate the free energy of binding, which is given by Eq. 2a,

$$\Delta G = -k_{\mathrm{B}} T \ln \frac{Q_{\mathrm{PL}}}{Q_{\mathrm{P}} Q_{\mathrm{L}}} \tag{2a}$$

where k_{B} is the Boltzmann constant, T is the temperature in Kelvin, Q is the partition function with subscripts PL, P and L indicating protein–ligand complex, protein, and ligand, respectively. This section presents a summary of thermodynamics, which is imperative for understanding the application and methods developed to compute binding free energy. More elaborate discussions on the thermodynamics of protein–ligand binding can be found in the reviews by Bronowska [48], and Homans [46].

2.3 *Methods to Compute Free Energy Binding*

Free energy is a quantity that can be measured for systems such as liquids or flexible macromolecules with several minimum energy configurations separated by high-energy barriers. However, its computation is far from trivial and the associated quantities such as entropy and chemical potential are also difficult to calculate. More so, the free energy cannot be accurately determined from classical molecular

dynamics or Monte Carlo simulations due to their inability to sample adequately from the high-energy regions of the phase space, which also make important contributions to the free energy. However, the free energy differences ($\Delta\Delta G$) are rather simple to compute. The free energy binding for the non-covalent association of two molecules (protein and ligand in this case) may be written as follows:

$$\Delta G_{\text{bind}} = G_{\text{complex}} - \left(G_{\text{protein}} + G_{\text{ligand}}\right) \tag{3}$$

The binding event is an additive interaction of many events [49–52], for example solvation energy (G_{sol}), conformational energy (G_{conf}), energy due to interaction with residues in the vicinity (G_{int}), and energy associated with different types of motions (translational, rotational and vibrational, G_{motion}). The classical binding free energy equation now can be rewritten as follows:

$$\Delta G_{\text{bind}} = G_{\text{sol}} + G_{\text{conf}} + G_{\text{int}} + G_{\text{motion}} \tag{4}$$

Directly computing the free energy from an MD or MC simulation is not trivial; hence, the following methods have been formulated. Broadly, the methods used for computing free energy are classified as partitioning-based methods or end-state free energy methods and non-partitioning-based methods. The partitioning-based methods partition the binding energy into various components as shown in Eq. 4; however, this method has been highly criticized [53] stating that it is physically unreal to partition the free energy into components.

2.3.1 End-State Free Energy Methods or Partitioning-Based Methods

The human body majorly comprises of water; hence, it is imperative to carefully include the solvation effects while computing the free energy of binding. More importantly, water plays a crucial role in ligand recognition and in the binding phenomenon. In computational chemistry, the methods for incorporation of solvent are divided into three groups: (i) continuum electrostatic methods/implicit solvent, (ii) explicit solvent models with microscopic detail and (iii) hybrid approaches. Historically, the continuum electrostatic methods were among the first to consider the solvent effect, and they still represent very popular approaches to evaluate solvation free energies, especially in quantum chemistry. Polarizable continuum model (PCM, [54]), COnductor-like Screening MOdel (COSMO, [55]) and SMD solvation model [56] are few popular models for treating solvent effects implicitly in quantum chemistry. Continuum solvation methods are computationally economical; however, the frictional drag of the solvent is highly underestimated and as a consequence may drive the system to non-physical states. Moreover, solvent–solvent and solute–solvent interactions are inadequately treated, posing a danger of underestimating the effects of such interactions. The explicit treatment of solvent enables one to consider the solvent–solvent and solute–solvent interactions. This prohibits the systems from visiting non-physical states due to the inclusion of the

dampening effect shown by the solvent atoms. The principal drawback of explicit solvent models is the number of atoms to be considered in the system leading to increased computational cost. However, with the help of GPU-based acceleration, this drawback, now, is hardly any cause for worry.

The end-state free energy methods use the conformations extracted from an MD or MC simulation, wherein the system is simulated by explicitly defining the solvent. However, while solving the GB or PB equation, the solvent is implicitly treated by defining the external dielectric constant for water (for most drug design cases) and a suitable internal dielectric constant [57–61].

Molecular Mechanics-Poisson Boltzmann/Generalized Born Surface Area (MM-PB/GB-SA)

The MM-GBSA [62–65] approach employs molecular mechanics-based energy calculations and the generalized Born model to account for the solvation effects in the calculation of the free energy. Similarly, the MM-PBSA [66–68] approach solves the linear or nonlinear Poisson–Boltzmann equation [69–71], to account for the solvation electrostatics, whereas the MM part is calculated as in MM-GBSA from the derivative of the force field equations. Both these approaches are parameterized such that they partition the energy components into various terms, and the net free energy change is the sum of these individual terms (Coulomb, vdW, solvation, etc.). MM-PBSA has gained considerable attention for estimating the binding free energies of molecular complexes due to its exhaustive nature of computing the solvation electrostatics by iteratively solving the PB equation, whereas the GB method does not involve any rigorous and iterative procedure and hence is faster. However, this does not necessarily guarantee that the MM-PBSA method always outperforms MM-GBSA method. In MM-PB(GB)SA methods, MD- or MC-derived conformational ensembles are used to compute the "**average**" free energy of a state and this is approximated as follows:

$$\langle G \rangle = \langle E_{\mathrm{MM}} \rangle + \left\langle G_{\mathrm{PBSA/GBSA}} \right\rangle - T \langle S_{\mathrm{MM}} \rangle \tag{5}$$

where the angular bracket **<>** indicates average over the MD/MC conformations, E_{MM} is the molecular mechanics energy that typically includes bond, angle, torsion, van der Waals, and electrostatic terms (see Eqs. 7c and 7d) and is evaluated with no or extremely large (virtually infinite) non-bonded cut-off limit. The second term is solved as mentioned in the preceding stanza and it forms the crux of this method. The last term $T <S_{\mathrm{MM}}>$, is the solute entropy, which is estimated by quasi-harmonic analysis [72, 73] of the trajectory or by normal mode analysis [74–76].

The following equation (Eq. 6) shows how the binding free energy is computed from the energies of the ligand, protein, and its complex over all the MD or MC snapshots. However, the snapshots can be obtained in two possible ways—one is called the single trajectory approach and other is the multiple trajectory approach. In the single trajectory approach, only the protein–ligand complex is simulated, and

the snapshots for the protein, ligand and the complex are extracted by defining appropriate atom numbers from the parameter and coordinate file. However, in the multiple trajectory approach, three separate simulations are performed, one each for the protein, ligand and protein–ligand complex.

$$\langle \Delta G_{\text{bind}} \rangle = \langle G_{\text{complex}} \rangle - \left(\langle G_{\text{protein}} \rangle - \langle G_{\text{ligand}} \rangle \right) \tag{6}$$

Furthermore, Eq. 1 is modified to accommodate solvation electrostatics and hydrophobic terms as shown in Eq. 5. Here, Eqs. 7a–7d give the computation of the individual terms,

$$\Delta G_{\text{bind}} = \Delta E_{\text{MM}} + \Delta G_{\text{sol}} - T\Delta S \tag{7a}$$

$$\Delta G_{\text{sol}} = \Delta G_{\text{sol}-\text{elect}} + \Delta G_{\text{nonpolar}} \tag{7b}$$

$$\Delta E_{\text{MM}} = \Delta E_{\text{int}} + \Delta E_{\text{elect}} + \Delta E_{\text{vdW}} \tag{7c}$$

$$\Delta E_{\text{int}} = \Delta E_{\text{bond}} + \Delta E_{\text{angle}} + \Delta E_{\text{torsion}} \tag{7d}$$

Here, ΔE_{MM} is computed in the gas phase using classical force fields, ΔG_{sol} is computed using PBSA or GBSA method, $\Delta G_{\text{sol-elect}}$ is computed using PB or the GB method, and the $\Delta G_{\text{nonpolar}}$ is computed by the solvent accessible surface area (SA). While employing the single trajectory approach, Eq. 7d generally cancels out and hence makes negligible contribution to the binding energy.

Linear Interaction Energy (LIE)

Linear interaction energy [77–79] is similar to the MM-PB/GB-SA method with regard to the partitioning of the electrostatic and van der Waals terms (polar and non-polar contribution, respectively,); however, the use of the weighting parameter for electrostatic and van der Waals interactions, is unique to this method. LIE measures the binding energy by estimating the difference in the interaction energies of the ligand in the solvent (unbound state) and in the protein environment (bound state). Hence, to obtain these interactions, two separate MD simulations are performed. In one simulation, only the ligand is placed in the solvent (mostly water) and in the other, the protein–ligand complex is placed in the solvent. The formulation of this method is based on deriving the linear response approximation from converged ensemble interactions, most often extracted from well-equilibrated trajectories from the MD simulation of the ligand with its surroundings (solvent or protein).

The mathematical formula for computing free energies using LIE method is given in Eq. 8

$$\Delta G_{\text{bind}} = \alpha \left[\left\langle E_{\text{coul}}^{\text{L}-\text{S}} \right\rangle_{\text{PL}} - \left\langle E_{\text{coul}}^{\text{L}-\text{S}} \right\rangle_{\text{L}} \right] + \beta \left[\left\langle E_{\text{vdW}}^{\text{L}-\text{S}} \right\rangle_{\text{PL}} - \left\langle E_{\text{vdW}}^{\text{L}-\text{S}} \right\rangle_{\text{L}} \right] \tag{8}$$

where the angular bracket <> indicates ensemble over the MD trajectory, E_{coul}^{L-S} and E_{vdW}^{L-S} are electrostatic and van der Waals interactions between the ligand and its medium in the vicinity (PL—protein–ligand complex; L—ligand in solvent), and α is the weighting parameter for electrostatic interactions, which is most often set to 0.5 [78]. This value is assumed due to the linear response of the surroundings to the electrostatic field and was validated using more extensive computations on the ions (Na^+ and Ca^{2+}) in water [80]. β is the weighting parameter for van der Waals interactions and is set to 0.16−0.18 [81], which is a subject of much debate owing to the difficulty in estimating the vdW's contribution to the free energy of binding. However, these values are obtained by empirical fitting the experimental binding free energies. Moreover, the linear response of the vdW term is assumed by observing the linear trend in the interaction of the hydrocarbons with the solvent (water) that depends on the number of carbons in a hydrocarbon.

2.3.2 Non-partitioning-Based Methods

In non-partitioning methods, there is no partitioning of the free energy into various components. Statistical mechanics plays a crucial role in deriving the relationship between the free energy of a system and the ensemble average of the Hamiltonian that describes the system. These methods are far more accurate than the previously mentioned end-state free energy methods, but at the same time, are computationally very demanding. Hence, while dealing with a large dataset of molecules against a particular protein target, it is worthwhile to screen the molecules using a fast method like high-throughput virtual screening [82, 83], followed by a flexible docking-based screening, then use an end-state free energy method, and finally employ the non-partitioning methods to study few tens of molecules. Here, we will present a brief discussion on FEP and TI methods along with their mathematical treatment, and then move on to explain the idea behind alchemical free energy predictions.

Free Energy Perturbation (FEP) and Thermodynamic Integration (TI)

Most of the methods for free energy calculations are generally formulated in terms of estimating the relative free energy differences, ΔG, between two equilibrium states, or binding of two similar ligands to a common target. The free energy difference between the two states I and II can be formally obtained by Zwanzig's formula [84, 85].

$$\Delta G = G_{II} - G_{I} = \beta^{-1} \ln e_{I}^{(-\beta \Delta V)} \tag{9}$$

Here, $\beta = (k_B T)^{-1}$

This represents a sampling of the differences in potentials (ΔV) of the two states using Monte Carlo or molecular dynamics simulation over the potential of state I. To ensure the convergence of these calculations, it is recommended that the potentials of the two systems should thermodynamically overlap. For satisfying this condition, correct conformations must be selected, which is a daunting task, and hence, to achieve this, a multistep process is usually implemented. A path between the states I and II is defined by introducing a set of intermediate potential energy functions that are constructed as linear combinations of the initial (I) and final (II) state potentials and these intermediate states are non-physical states (Eq. 10).

$$V_m = (1 - \lambda_m)V_{\mathrm{I}} - \lambda_m V_{\mathrm{II}} \tag{10}$$

where the transition from one state to another is discretized into many points (m = 1,…,n), each represented by a separate potential energy function that corresponds to a given value of λ, such that λ_m varies from 0 to 1. Here, zero indicates the pure initial state of the system and one indicates pure final state of the system. The total free energy, thus, can be obtained by summing over the intermediate states along the λ variable.

$$\Delta G = G_{\mathrm{II}} - G_{\mathrm{I}} = -\beta^{-1}\sum_{m=1}^{n-1}\ln\left\langle^{[-\beta(V_{m+1}-V_m)]}\right\rangle_m \tag{11}$$

This approach is known as free energy perturbation (FEP) where $\Delta\lambda_m = \lambda_{m-1} - \lambda_m$; hence, it can be written as

$$\Delta G = -\beta^{-1}\sum_{m=1}^{n-1}\ln\langle e^{[-\beta\Delta V\Delta\lambda_m)]}\rangle_m \tag{12}$$

Since the potential difference can also be described as the derivative of the potential with respect to λ_{m}, Eq. 12 can also be written as,

$$\Delta G = -\beta^{-1}\sum_{m=1}^{n-1}\ln\langle e^{\left[-\beta\frac{\partial V_m}{\partial\lambda_m}\Delta\lambda_m)\right]}\rangle_m \tag{13}$$

Now, expansion of the Eq. 13 by the Taylor expansion series gives Eq. 14,

$$\Delta G = \sum_{m=1}^{n-1}\langle e^{\left[-\beta\frac{\partial V_m}{\partial\lambda_m}\Delta\lambda_m)\right]}\rangle_m \tag{14}$$

wherein $0 \rightarrow \lambda$ can instead be written as an integral over λ

$$\Delta G = \int_0^1\langle\beta\frac{\partial V(\lambda)}{\partial\lambda}\rangle_\lambda \mathrm{d}\lambda \tag{15}$$