Sanjeev Kumar Singh *Editor*

Innovations and Implementations of Computer Aided Drug Discovery Strategies in Rational Drug Design

Springer

Innovations and Implementations of Computer Aided Drug Discovery Strategies in Rational Drug Design

Sanjeev Kumar Singh
Editor

Innovations and Implementations of Computer Aided Drug Discovery Strategies in Rational Drug Design

Springer

Editor
Sanjeev Kumar Singh
Computer Aided Drug Design and
Molecular Modelling Lab,
Department of Bioinformatics
Alagappa University
Karaikudi, Tamil Nadu, India

ISBN 978-981-15-8935-5 ISBN 978-981-15-8936-2 (eBook)
https://doi.org/10.1007/978-981-15-8936-2

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Contents

About the Editor

Sanjeev Kumar Singh is currently Professor at Department of Bioinformatics, Alagappa University, Karaikudi, Tamil Nadu, India. He received his graduation, post graduation and Doctoral degree in Applied Theoretical chemistry from the CSJM University, Kanpur. Prof. Singh possess extensive exposure in the area of Computer Aided Drug Design and Molecular Modelling and achieved distinct landmarks in Computational Drug Discovery to identify several therapeutics against Viruses, Bacterial pathogens and Cancer. He received "ICMR - Lala RAM Chand Kandhari Award" from Indian coouncil of Medical Research, Govt. of India. He also honoured with 'Fellow of Royal Society of Biology', UK and 'Biotech research Society of India'. He is the Secretary General of Bioinformatics and Drug Discovery Society (BIDDS), and Member of various prestigious societies like 'The National Academy of Sciences', Indian Biophysical Society, Indian Science Congress Association etc. He also published more than 170 research articles in highly reputed journals.

Chapter 1
CADD: Some Success Stories from *Sanjeevini* and the Way Forward

Ankita Singh, Shashank Shekhar, and B. Jayaram

Abstract *Sanjeevini*, a comprehensive drug design software suite, has been developed for lead molecule discovery taking off from protein and DNA as potential targets. *Sanjeevini* is a culmination of multiple modules of significance including detection of active sites in proteins and DNA in an automated manner, screening of a library of a million compounds for hit molecule identification, docking and scoring using all atom energy based algorithms and various other utilities which assist in designing hit molecules with desired affinities and specificities against the given targets. A few of the modules of *Sanjeevini* software suite along with some of the success stories are illustrated in this chapter.

A. Singh
Supercomputing Facility for Bioinformatics and Computational Biology (SCFBio), Indian Institute of Technology Delhi, New Delhi, India
e-mail: ankita@scfbio-iitd.res.in; http://www.scfbio-iitd.res.in

S. Shekhar
Supercomputing Facility for Bioinformatics and Computational Biology (SCFBio), Indian Institute of Technology Delhi, New Delhi, India

School of Interdisciplinary Research (SIRe), Indian Institute of Technology Delhi, New Delhi, India
e-mail: shashank@scfbio-iitd.res.in; http://www.scfbio-iitd.res.in

B. Jayaram (✉)
Supercomputing Facility for Bioinformatics and Computational Biology (SCFBio), Indian Institute of Technology Delhi, New Delhi, India

Kusuma School of Biological Sciences, Indian Institute of Technology Delhi, New Delhi, India

Department of Chemistry, Indian Institute of Technology Delhi, New Delhi, India
e-mail: bjayaram@chemistry.iitd.ac.in; http://www.scfbio-iitd.res.in

S. K. Singh (ed.), *Innovations and Implementations of Computer Aided Drug Discovery Strategies in Rational Drug Design*,
https://doi.org/10.1007/978-981-15-8936-2_1

1.1 Introduction to Computer Aided Drug Design (CADD)

Drug discovery is considered to be a lengthy, expensive and technically intricate process that has a limited number of matches in the entire R&D ecosystem. Computer-aided drug design (CADD) methodologies are being routinely adopted now by the pharmaceutical sector with an eye on both economy of scale and economy of scope to accelerate the processes involved in the pipeline. On an average, it takes approximately 10–15 years of time and 2.6 billion US dollars to launch a new drug into the market from conception, with production and testing of lead molecules contributing significantly to the entire process (Daina et al. 2017) (Fig. 1.1). Thus, it is apt to apply methodologies related to computational approaches

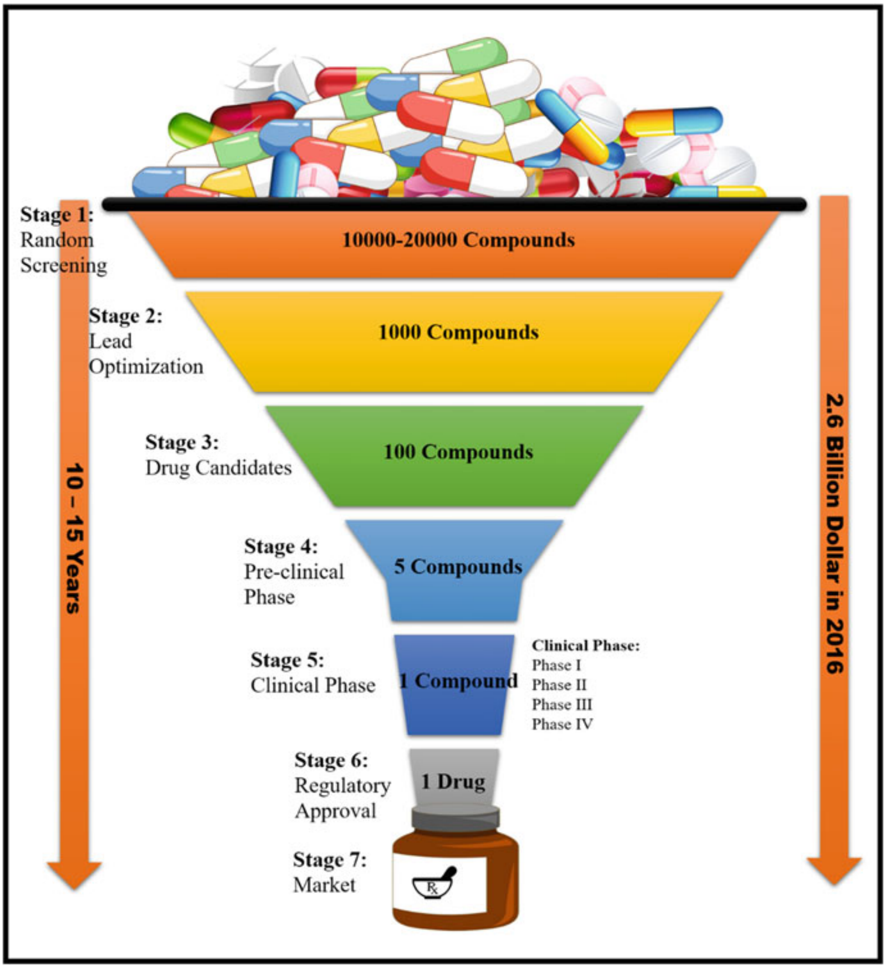


Fig. 1.1 Conventional process of drug discovery and development (Daina et al. 2017)

to identify hits and to eventually optimize the leads developed from the hits. This exercise covers a wide chemical space while narrowing down the list of compounds to be synthesized for further *in vitro* testing. Structure based docking analyses and energy profiles for hit analogs, searching for new compounds using ligand-based screening approach which have similarities in terms of chemical structures or having improved predicted chemical affinities or better biological activities, estimation and optimization of drug metabolism and pharmacokinetics (DMPK) and achieving desirable absorption, distribution, metabolism, excretion, and toxicity (ADMET) profiles (Shaikh et al. 2007) are all intrinsic parts of identification and optimization of hit compounds using computational techniques. One of the major factors which led to the wide range of acceptability and acceleration of the entire drug discovery pipeline is the adoption of advancements and developments of technological innovations in both software and hardware. In the context of qualities of the computational techniques, the contribution of computer aided drug discovery to the area of life sciences is invaluable (Trott and Olson 2010; Yamada and Itai 1993; Khanna and Ranganathan 2011; Xiang et al. 2012).

1.2 Active Site Prediction

1.2.1 Introduction

Active site is the region of a protein, an enzyme for instance, where the binding partner or the substrate binds and performs the biological function. Active site has a pocket/cavity located deep into the enzyme or at the interface between multimeric enzymes (Kahraman et al. 2008). The site is lined up with a number of amino acid residues which form chemical bonds or interactions with the substrate which undergoes a chemical reaction, catalysis for instance. An active site can be easily understood by considering a dynamic lock and key hypothesis; implying that an enzyme has only an active site which can accommodate only a specific kind of substrate. Bigger the pocket, more are the interactions polar and non-polar. Knowledge of the active site inside the protein molecule is crucial to developing a blueprint for a candidate lead/drug molecule in order to control the activity of that particular protein.

1.2.2 Active Site Prediction Servers

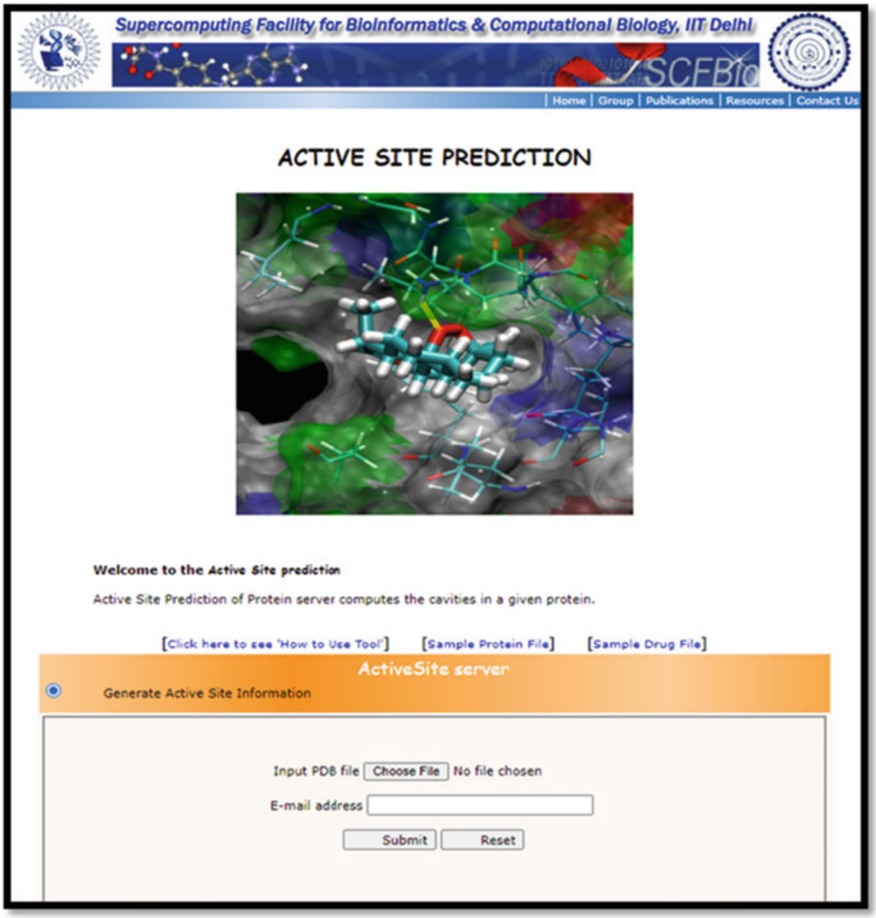
Functionally relevant binding pocket or active site prediction from the tertiary structure of proteins is one of the key steps in drug designing process (Szarecka and Dobson 2019). Experimental evidence of the binding pocket of the target protein molecule is beneficial but in case the experimental information on the active site is not available, a large number of tools are freely accessible which can identify the cavities with high accuracies. Here we discuss our web-server AADS for active site prediction.

AADS

AADS is a robust active site recognition, docking and scoring tool for proteins (Singh et al. 2011). It predicts all cavities or potential binding pockets on the target protein and ranks them based on a fuzzy score which utilizes the information and properties of functional groups lining the active sites of the proteins. Physico-chemical principles constitute the underlying properties. AADS was validated on ~620 proteins with varying amino acid lengths and known active sites and its accuracy was 100% when the top ten cavities were considered. If the user wishes to proceed further for automated docking, all top most 10 binding sites predicted are then shortlisted for protein-ligand docking. An all atom energy-based Monte Carlo method has been adopted for the development of the docking tool. Energy optimized docked structures presenting varied positions and alignments of the candidate drug molecule are preserved from different cavity points resulting in 80 docked structures overall which are further optimized using an efficient free energy function to select 5 best docked structures (Fig. 1.2).

1.3 Virtual High Throughput Screening

Computational drug discovery relies heavily on screening of molecules against a desired target (Lill 2013). The objective of virtual screening is to identify molecules which bind to a macromolecular target with high affinity and specificity. With the advent of faster computing devices, it is now possible to scan millions of compounds in a relatively short period of time. Various techniques have been described using which virtual screening can be achieved. Broadly these techniques can be divided into two types: (a) on the basis of Ligand information and (b) on the basis of Structural information. The success of these techniques can be measured in the diversity they produce in identifying a small library of molecules which can potentially bind to the biomolecule of interest. Ligand based methodologies rely on the structural and binding information of previously known ligands. Perhaps the oldest and the most popular method is of scanning using a pharmacophore. This technique involves generation of a pharmacophore model using structures of bioactive molecules against a target of interest. Scanning for a molecule with similar pharmacophore is achieved using molecule libraries. Another methodology involves identification of similar molecules using 2D comparisons. Techniques go a bit further to compute 3D similarities and report similar molecules which could be potentially active. Some methodologies involve breaking bioactive molecules to


Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi
SCFBio
| Home | Group | Publications | Resources | Contact Us
[Click here to see 'How to Use Tool']
[Sample Protein File]
[Sample Drug File]
Choose File
Submit
Reset

ACTIVE SITE PREDICTION

Welcome to the Active Site prediction

Active Site Prediction of Protein server computes the cavities in a given protein.

ActiveSite server

● Generate Active Site Information

Input PDB file: No file chosen

E-mail address:

Fig. 1.2 Screen shot of the front-end of Automated Version of Active Site Prediction (AADS). Web link: http://www.scfbio-iitd.res.in/dock/ActiveSite_new.jsp

sub-structures and scanning a database of molecules for similar substructures to eventually assemble newer molecules. In recent times, machine learning techniques are also being harnessed to look for bioactive molecules (Macalino et al. 2015; Hoque et al. 2017). Structure based methodologies use molecular docking programs to dock every molecule to a known active site. These should ideally be the most accurate but unfortunately are not applicable on large databases containing molecules in excess of million. These methodologies are better suited for a smaller set of molecules generated by other methods such as ligand-based methodologies (Imam and Gilani 2017; Anderson 2003).

1.3.1 RASPD

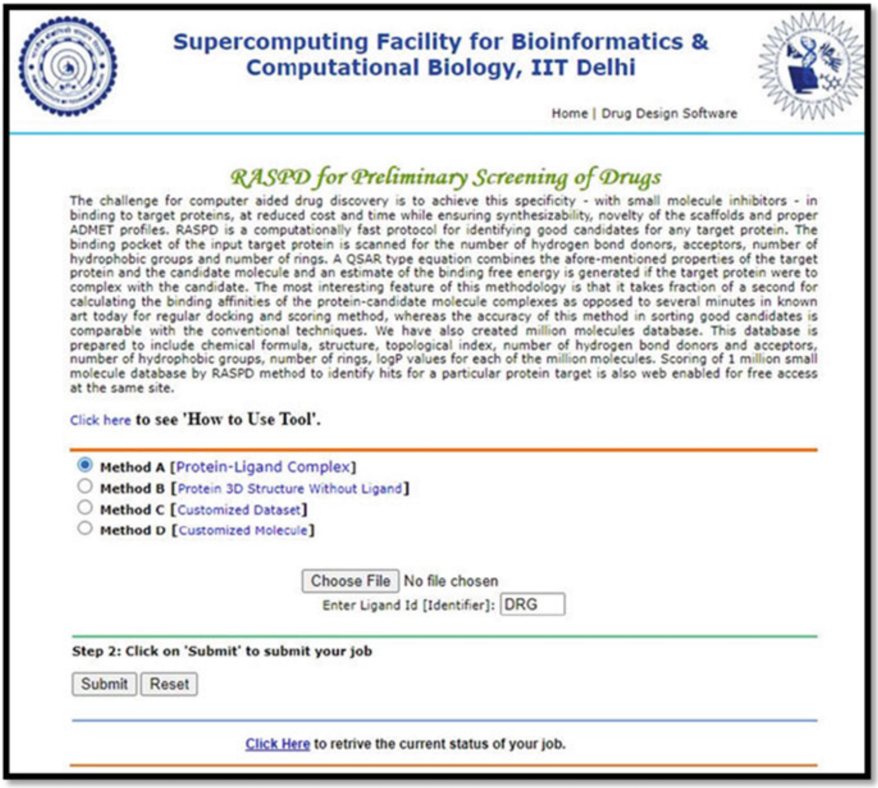
To recall, the aim of computer-aided drug design is to attain desired affinities and specificities with inhibitors in the form of small molecules which bind to target proteins. Additionally, it must be cost-effective and time efficient while investigating scaffolds in terms of their novelty and ADMET characteristics. In pursuit of these objectives, we have developed RASPD, a fast and efficient computational screening protocol which can suggest probable candidates for a given target protein (Mukherjee and Jayaram 2013). The active site cavity where the candidate molecule binds to the target protein is scanned for hydrogen bond acceptors and donors, hydrophobic groups and number of rings. Similarly, a large database of small molecules is developed with pre-stored information of these physico-chemical descriptors. The protein-small molecule complementarily is set up via a QSAR type equation in the descriptor space for binding energy estimation of each molecule in the database without actually docking these molecules. The exciting feature of this technique is that it takes 10–15 min to compute the binding affinities of a million candidate molecules with the target protein, in contrast to the many minutes for a single protein-ligand molecule taken in the conventional methods of docking and scoring. The precision of this technique is analogous to traditional methods aimed at selecting good candidates. The RASPD web-server (Fig. 1.3) is freely accessible for scoring a million small molecules to find hit molecules for a specified protein target.

1.3.2 BAITOC

Druggable biomolecules are limited but number of small molecules capable of moderating their activities is huge. Instead of searching for a molecule for a given target, a new approach of finding a target for a known molecule can also be utilized. There are many molecules which are synthesized in laboratories across the globe but are never tested for their bioactivity. Also, cases exist where bioactive compounds are known but their biomolecular targets are unknown. BAITOC is an application software (Fig. 1.4) which aims to fill this gap, by a quick examination against a databank of pathogen's protein structures. The application screens thousands of protein structures at a time against input molecules using the RASPD logic and provides information on potential protein targets for molecules under investigation.

1.4 Docking and Scoring

In computational approaches, to identify hit compounds and to further optimize them, the conventional and the most common method employed is docking the small molecule in the active site of the macromolecular target and then scoring the pose in



Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi

Home | Drug Design Software

RASPD for Preliminary Screening of Drugs

The challenge for computer aided drug discovery is to achieve this specificity - with small molecule inhibitors - in binding to target proteins, at reduced cost and time while ensuring synthesizability, novelty of the scaffolds and proper ADMET profiles. RASPD is a computationally fast protocol for identifying good candidates for any target protein. The binding pocket of the input target protein is scanned for the number of hydrogen bond donors, acceptors, number of hydrophobic groups and number of rings. A QSAR type equation combines the afore-mentioned properties of the target protein and the candidate molecule and an estimate of the binding free energy is generated if the target protein were to complex with the candidate. The most interesting feature of this methodology is that it takes fraction of a second for calculating the binding affinities of the protein-candidate molecule complexes as opposed to several minutes in known art today for regular docking and scoring method, whereas the accuracy of this method in sorting good candidates is comparable with the conventional techniques. We have also created million molecules database. This database is prepared to include chemical formula, structure, topological index, number of hydrogen bond donors and acceptors, number of hydrophobic groups, number of rings, logP values for each of the million molecules. Scoring of 1 million small molecule database by RASPD method to identify hits for a particular protein target is also web enabled for free access at the same site.

Click here to see 'How to Use Tool'.

● Method A [Protein-Ligand Complex]
○ Method B [Protein 3D Structure Without Ligand]
○ Method C [Customized Dataset]
○ Method D [Customized Molecule]

Choose File No file chosen

Enter Ligand Id [Identifier]: DRG

Step 2: Click on 'Submit' to submit your job

Click Here to retrive the current status of your job.

Fig. 1.3 Screen shot of the front end of RASPD webserver. Web link: http://www.scfbio-iitd.res.in/software/drugdesign/raspd2.jsp

order to realize potential complementarities to binding sites (Meng et al. 2011). Numerous drugs have been developed in recent times adopting structure based virtual screening (SBVS) strategies. Conformational information of the ligand along with its orientation in the targeted binding site is obtained using the docking process (Evanthia et al. 2014; Pinzi and Rastelli 2019).

1.4.1 Protein-Ligand Docking Server

ParDOCK

ParDOCK is a docking software which has been developed using an all atom energy-based Monte Carlo procedure. It utilizes Monte Carlo based pose generation and evaluation method to generate the most probable configurations of the ligand in the active site. The method is comprehensively validated on 226 protein-ligand

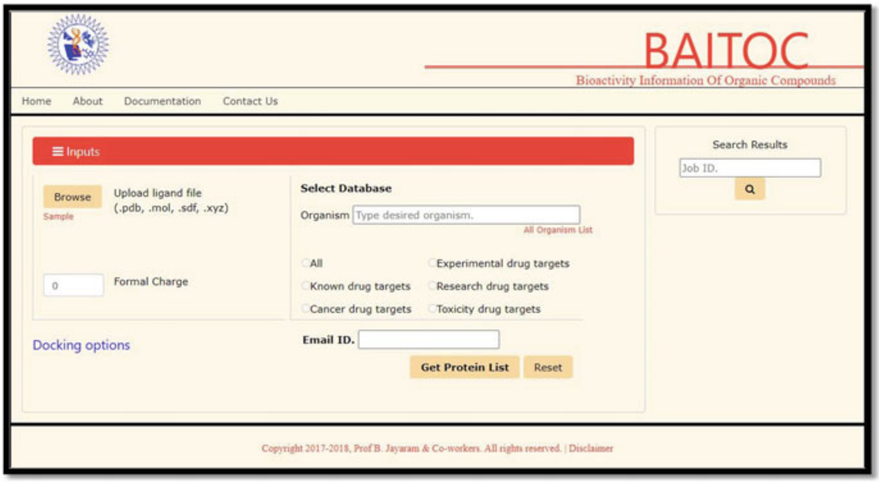


Fig. 1.4 Screen shot of the front-end of BAITOC: Bioactivity information of organic compounds. Web link: http://www.scfbio-iitd.res.in/baitoc

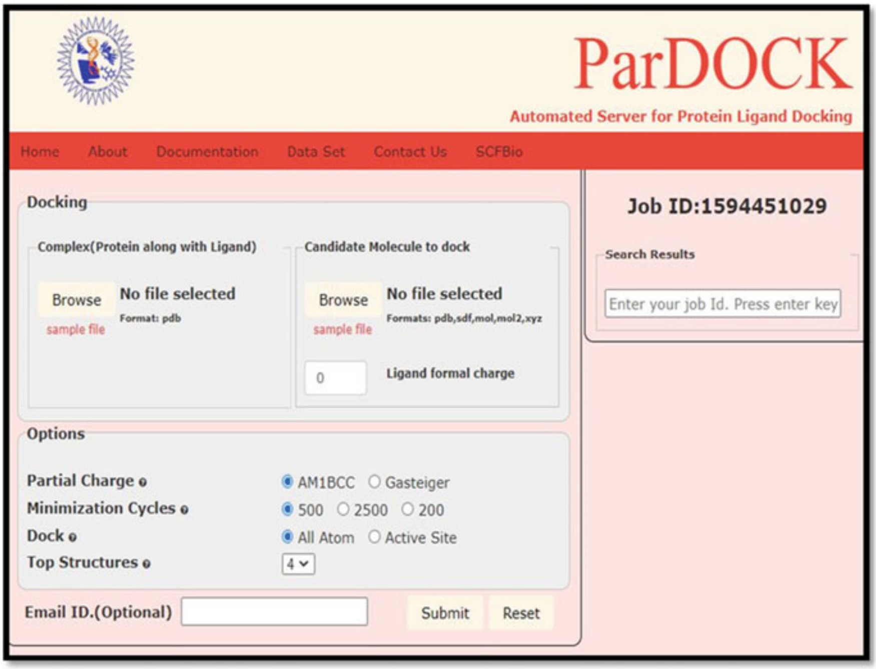
complexes, resulting in a mean RMSD of around 0.53 Å. ParDOCK outputs best 8 energetically favorable poses. This is followed by 2500 steps of protein-ligand complex energy minimization. The energy minimized complex is then evaluated using BAPPL scoring function. The resulting top four poses are provided to the user as top results (Gupta and Sharma 2007) (Fig. 1.5).

BAPPL

BAPPL is a binding affinity prediction software. It evaluates the binding energy of a docked protein-ligand complex. The prediction is based on electrostatics, van der Waals, entropic and hydrophobic characteristics (Jain and Jayaram 2005). The predicted binding energy is expressed in kcal/mol. Workflow for BAPPL is shown in Figs. 1.6 and 1.7.

Bappl+

More recently, we developed Bappl+, an improved methodology of BAAPL for predicting the binding affinities of protein-ligand and metalloprotein-ligand complexes. It computes binding affinity based on the most important energetic contributors such as electrostatics, van der Waals, hydrophobicity and entropy of protein and ligand. For metalloprotein-ligand complexes, it uses the explicitly-derived quantum-optimized charges for various metal ions (Zn, Mn, Mg, Ca and Fe). It uses a Random Forest algorithm to derive the final score. This methodology (implemented in web server mode (Fig. 1.8)) is widely tested on the PBD bind



Docking

Complex(Protein along with Ligand)

No file selected

Format: pdb

Candidate Molecule to dock

No file selected

Formats: pdb,sdf,mol,mol2,xyz

0 Ligand formal charge

Options

Partial Charge ● AM1BCC ○ Gasteiger

Minimization Cycles ● 500 ○ 2500 ○ 200

Dock ● All Atom ○ Active Site

Top Structures 4

Email ID.(Optional)

Job ID:1594451029

Search Results

Enter your job Id. Press enter key

Fig. 1.5 Screen shot of the front-end of ParDOCK webserver. Web link: http://www.scfbio-iitd.res.in/pardock

datasets of 2007, 2013 and 2016 releases. Bappl+ achieves strong correlations with respect to experimental affinities with all the core datasets with low standard deviations and works better than most of the known state-of-the-art scoring functions (Soni et al. 2020).

1.5 *Sanjeevini*

Sanjeevini is a collection of several modules, some of which are shown in Fig. 1.9 with linkages, to assist in lead molecule design. A user can input a bimolecular (protein) target and a candidate drug molecule and upload them on the *Sanjeevini* web portal. User may also decide to input a self-drawn molecule or opt to scan a million compound library or a natural product library. The sub module of the software AADS, predicts the potential active sites, whereas modules such as ParDOCK dock and score the candidate drugs and gives back four best optimized structures for each candidate molecule bound to the protein target. Also, estimates of binding free energies can be obtained using one of the modules named BAPPL/BAPPL+. The predicted four structures in the docked form represent the best possible poses of the candidate drug molecule in the active site of the protein

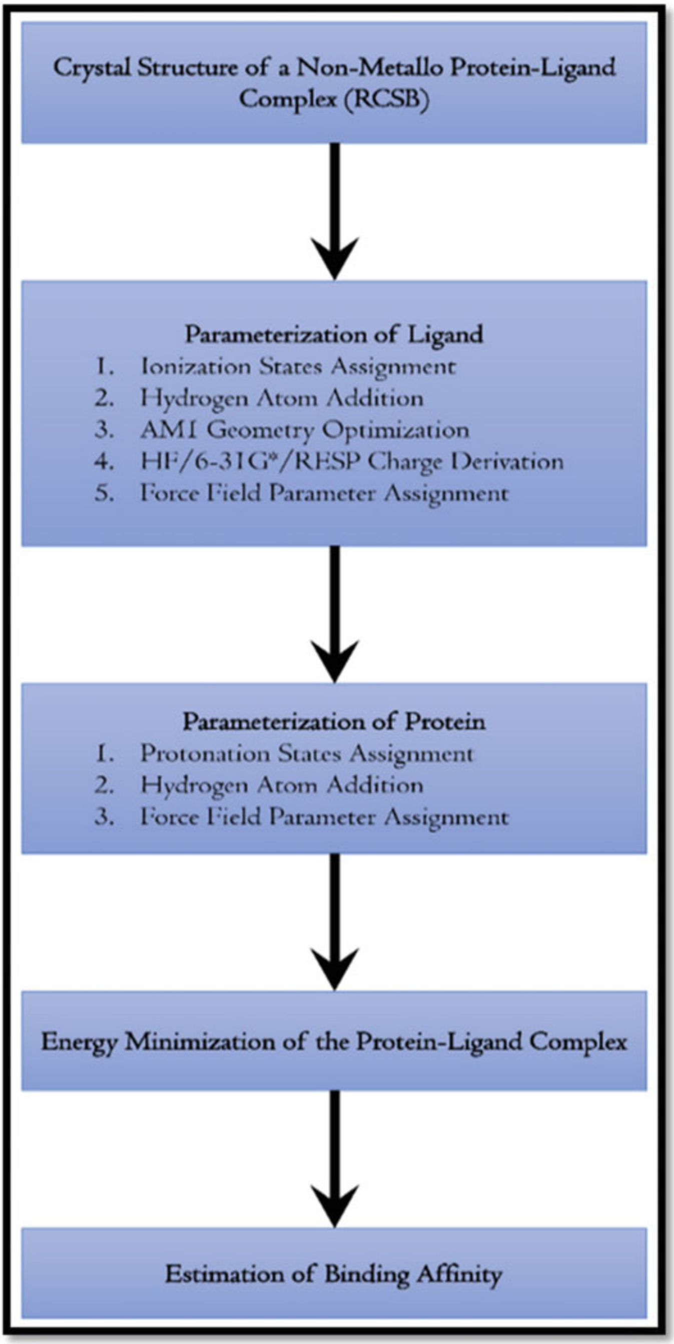
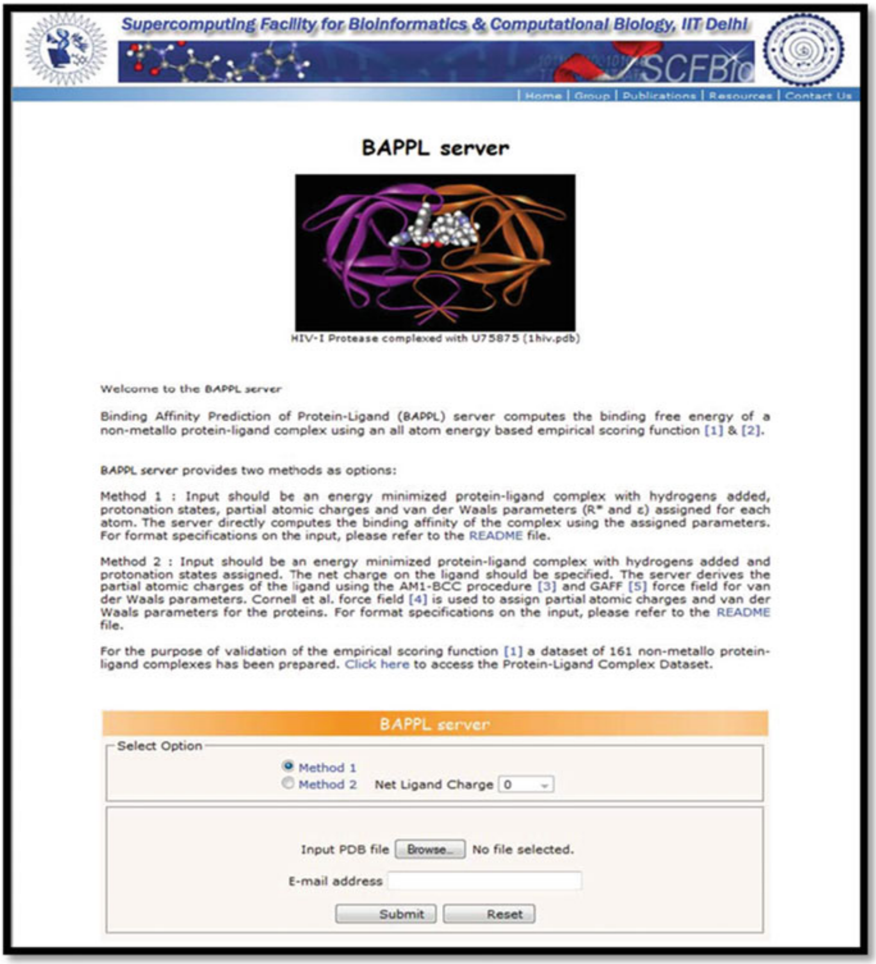

Crystal Structure of a Non-Metallo Protein-Ligand Complex (RCSB)
Parameterization of Ligand
1. Ionization States Assignment
2. Hydrogen Atom Addition
3. AM1 Geometry Optimization
4. HF/6-31G*/RESP Charge Derivation
5. Force Field Parameter Assignment
Parameterization of Protein
1. Protonation States Assignment
2. Hydrogen Atom Addition
3. Force Field Parameter Assignment
Energy Minimization of the Protein-Ligand Complex
Estimation of Binding Affinity

Fig. 1.6 A computational flowchart used for the calculation of binding affinities of protein–ligand complexes


Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi
SCFBio
| Home | Group | Publications | Resources | Contact Us
Browse...
Submit
Reset

BAPPL server

HIV-I Protease complexed with U75875 (1hiv.pdb)

Welcome to the BAPPL server

Binding Affinity Prediction of Protein-Ligand (BAPPL) server computes the binding free energy of a non-metallo protein-ligand complex using an all atom energy based empirical scoring function [1] & [2].

BAPPL server provides two methods as options:

Method 1 : Input should be an energy minimized protein-ligand complex with hydrogens added, protonation states, partial atomic charges and van der Waals parameters (R* and ε) assigned for each atom. The server directly computes the binding affinity of the complex using the assigned parameters. For format specifications on the input, please refer to the README file.

Method 2 : Input should be an energy minimized protein-ligand complex with hydrogens added and protonation states assigned. The net charge on the ligand should be specified. The server derives the partial atomic charges of the ligand using the AM1-BCC procedure [3] and GAFF [5] force field for van der Waals parameters. Cornell et al. force field [4] is used to assign partial atomic charges and van der Waals parameters for the proteins. For format specifications on the input, please refer to the README file.

For the purpose of validation of the empirical scoring function [1] a dataset of 161 non-metallo protein-ligand complexes has been prepared. Click here to access the Protein-Ligand Complex Dataset.

BAPPL server

Select Option

● Method 1
○ Method 2 Net Ligand Charge 0

Input PDB file Browse... No file selected.

E-mail address

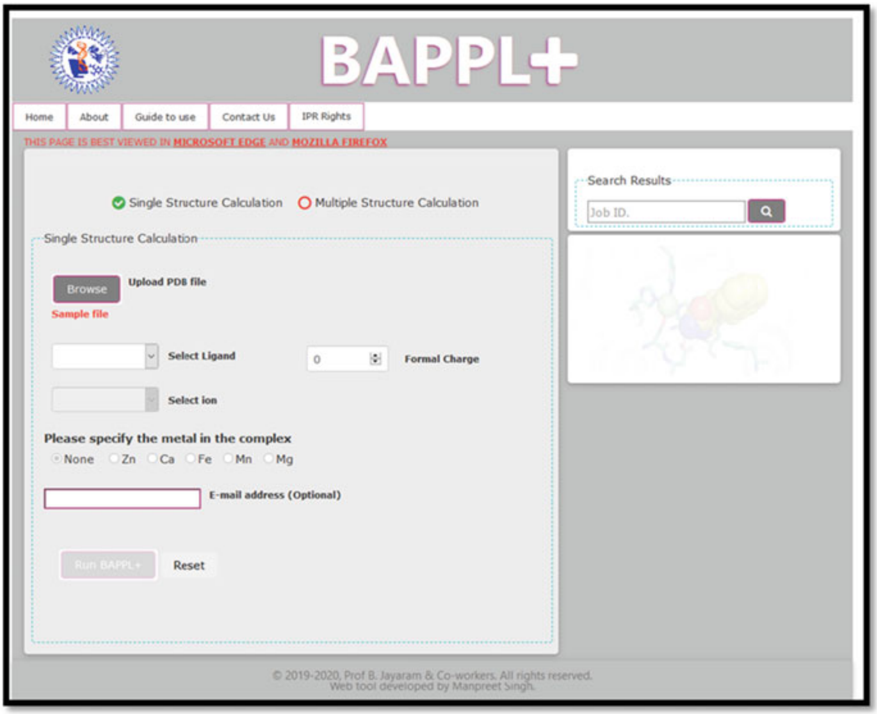
Fig. 1.7 Screen shot of the front-end of BAPPL web server. Web link: http://www.scfbio-iitd.res.in/software/drugdesign/bappl.jsp

biomolecule (Jayaram et al. 2012). Versatility of the methodologies of *Sanjeevini* are partly captured in Fig. 1.9.

1.5.1 Sanjeevini's *Success Stories*

(a) *Anti-Cancer molecules*

Breast cancer, in recent times, is considered as the most common form of cancer found in women. Estrogen receptor (ER) is a popular drug target to discover


BAPPL+
Home
About
Guide to use
Contact Us
IPR Rights
THIS PAGE IS BEST VIEWED IN MICROSOFT EDGE AND MOZILLA FIREFOX
Search Results
Job ID.
Browse
Upload PDB file
Sample file
Select Ligand
Formal Charge
Select ion
E-mail address (Optional)
Run BAPPL+
Reset
© 2019-2020, Prof B. Jayaram & Co-workers. All rights reserved.
Web tool developed by Manpreet Singh.

● Single Structure Calculation ○ Multiple Structure Calculation

Single Structure Calculation

Formal Charge: 0

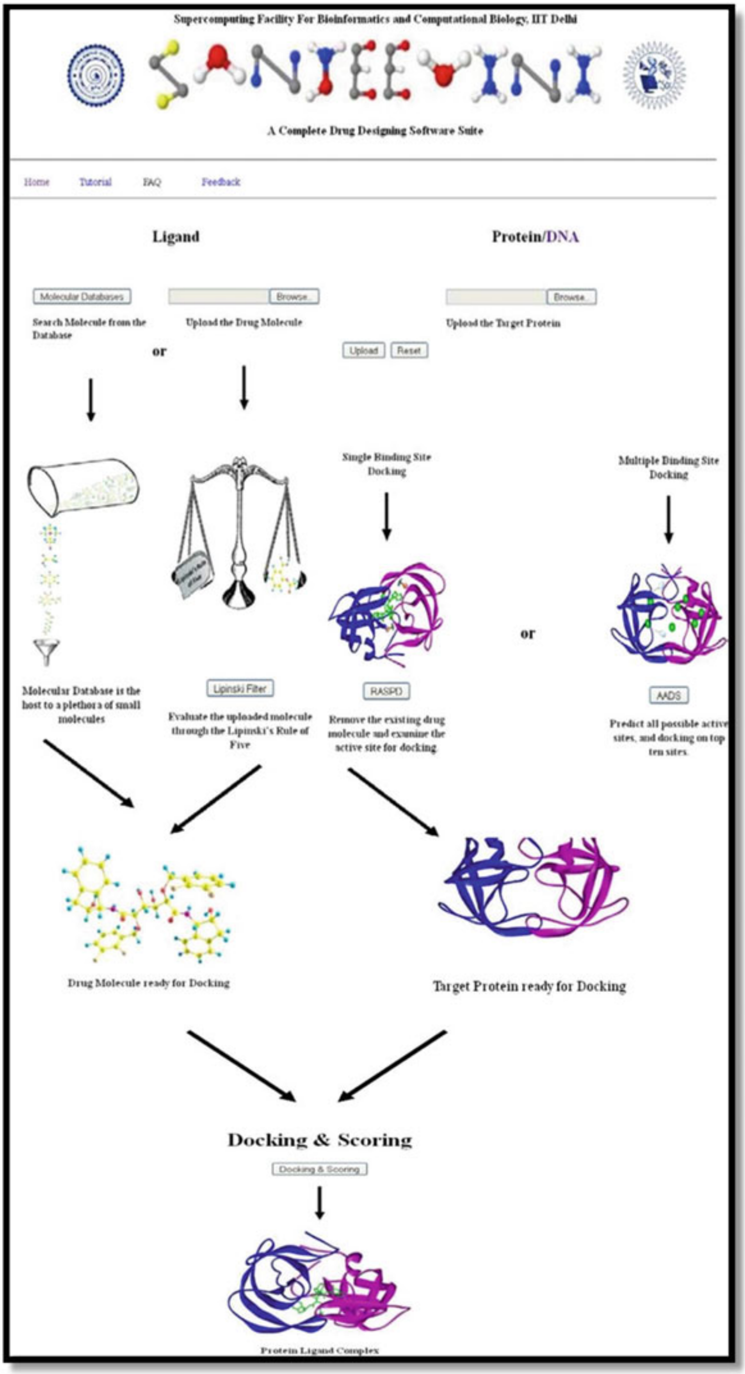
Please specify the metal in the complex
● None ○ Zn ○ Ca ○ Fe ○ Mn ○ Mg

Fig. 1.8 Screen shot of the front-end of BAPPL+ web server. Web link: http://www.scfbio-iitd.res.in/bappl+

agents known as selective estrogen receptor modulators (SERMs). Existing known drugs show severe adverse effects and some have developed resistance over time. Thus, there is a need for better therapeutic profiles with newer agents. ERα and ERβ are two isoforms which share 56% similarity but show distinct physiological functions and expressions in a number of tissues. With the help of computational docking and molecular dynamics (MD) simulations, a few biphenyl derivatives were designed, synthesized and successfully tested presenting potential chemical agents for selectively targeting ERα (Bhatnagar et al. 2017).

(b) *Anti-Alzheimer's agents*

In the progression of Alzheimer's disease (AD), Acetylcholinesterase (AChE) is considered as a vital enzyme and as a result it has been subjected to intense drug discovery programmes. Using a multi-dimensional approach by employing computational, chemical and biological pathways, a few derivatives of pyrimidine with triazolopyrimidine based hybrid scaffold were developed as AChE inhibitors for the treatment of AD disease (Kumar et al. 2018).


Supercomputing Facility For Bioinformatics and Computational Biology, IIT Delhi
SANJEEVINI
A Complete Drug Designing Software Suite
Home
Tutorial
FAQ
Feedback
Ligand
Protein/DNA
Molecular Databases
Browse..
Browse..
or
Upload
Reset
Single Binding Site Docking
Multiple Binding Site Docking
or
Lipinski Filter
RASPD
AADS
Drug Molecule ready for Docking
Target Protein ready for Docking
Docking & Scoring
Docking & Scoring
Protein Ligand Complex

Search Molecule from the Database

Upload the Drug Molecule

Upload the Target Protein

Molecular Database is the host to a plethora of small molecules

Evaluate the uploaded molecule through the Lipinski's Rule of Five

Remove the existing drug molecule and examine the active site for docking.

Predict all possible active sites, and docking on top ten sites.

Fig. 1.9 Screen shot of the front-end of *Sanjeevini* webserver

(c) *Anti-Malarials*

Plasmodium falciparum caused malaria is one of the dreadful diseases particularly because of the ineffectiveness of the available drugs for curing malaria and emergence of extensive resistance. A few triazine based novel derivatives were designed and synthesized based on docking studies against P*f*PMT target, a protein unique to the pathogen, and were found to be good inhibitors. These molecules against malaria causing parasites showed activity in the range of low μ-molar to sub μ-molar. The strategies based on Molecular dynamics simulations have developed insights on inhibition based mechanism of action of the new inhibitors of PfPMT (Shandilya et al. 2017).

(d) *Anti-virals*

Hepatitis A Virus (HAV) belongs to Picornaviridae family of viruses and can be transmitted feco-orally. A viable drug target for HAV is 3C protease, which is also common to other picorna viruses, for post-translational proteolysis of viral polyproteins and for inhibiting host innate immune pathways. Chemical synthesis and experimental validation after computational screening resulted in identification of a few low micromolar compounds which could inhibit HAV 3C activity. Further experimental testing with comparable results were obtained when these compounds were tested against the 3C protease of Human Rhinovirus which is a member of the same Picornaviridae family (Banerjee et al. 2019). More recently, *Sanjeevini* protocols have also been used to repurpose drug molecules against CHIKV (Tripathi et al. 2020)

(e) *Anti-Fungal*

Fungal resistance to existing azole drugs has been a cause of severe concern. However, due to the absence of required molecular level knowledge of the mechanisms of activation of genes by transcription activators associated with the disease, the therapeutic targeting efforts have been hindered. After further investigations on this, it was discovered that the feasibility of blocking transcription factor-binding site in mediator with small-molecule could bring out a unique therapeutic strategy to deal with anti-fungal resistance (Nishikawa et al. 2016).

1.6 *Dhanvantri*

Dhanvantri is a pipeline incorporating several novel scientific methods and highly efficient algorithms that combine the principles of chemistry and biology with information technology for accomplishing drug design as shown in Fig. 1.10. The pipeline covers all aspects of the genes, proteins, the active sites etc. from the genome to the proposed final lead molecule. Automation at each stage facilitates use of the pipeline (with default parameters) that provides scientifically significant output in most cases of the disease. Also, users are allowed to make any necessary

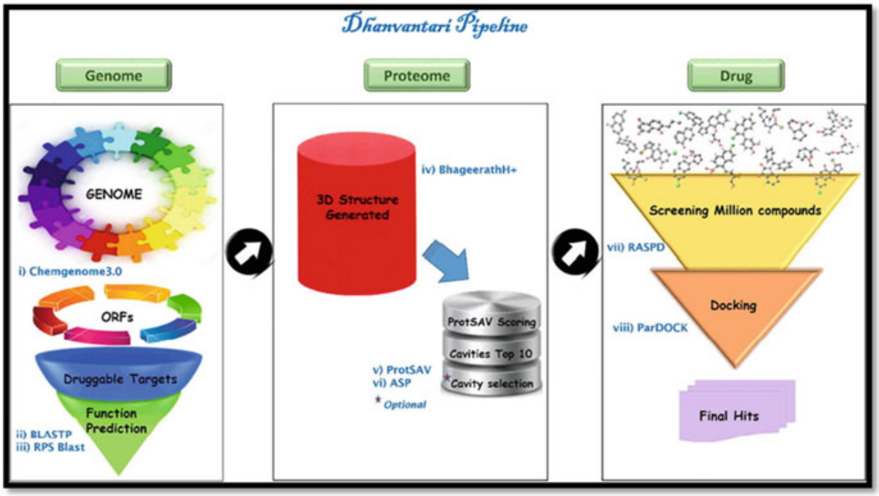
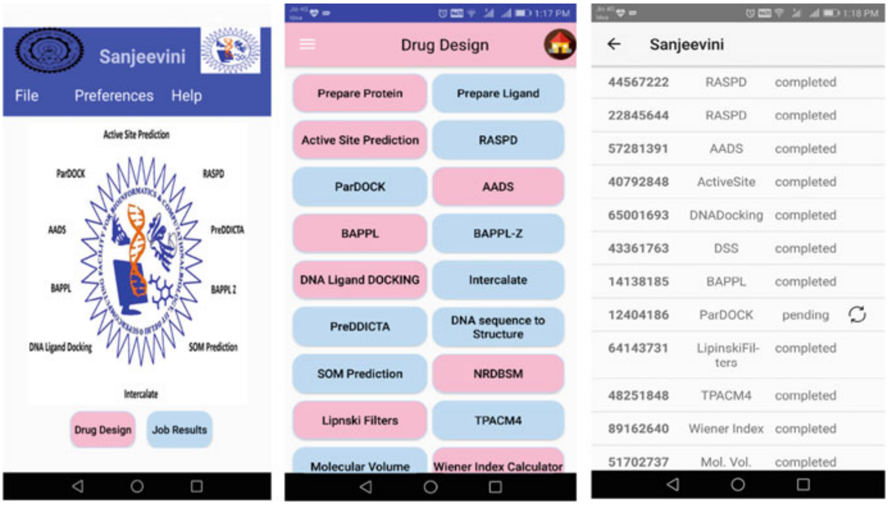


Fig. 1.10 Pipeline of the *Dhanvantari* Software suite. Web link: http://www.scfbio-iitd.res.in/software/dhanvantari_new/Home.html

changes to the pipeline parameters and enter at any stage of the pipeline as per their requirement. The computational time required is ~6–12 h for the entire pipeline. In short, this pipeline bridges the gap between diseases and their potential cures (Bhat et al. 2020).

1.7 *Sanjeevini* Application in Android Mode

A mobile application developed for *Sanjeevini* software suite could be considered as the portable version of our efforts in the area of computer aided drug discovery (CADD) apart from the Web server version. The mobile application commonly termed as *Sanjeevini* app comprises a variety of drug discovery modules for active site prediction inside a protein molecule, screening hundreds or thousands of ligands in little or no time, protein and DNA target based docking and scoring etc. It also serve as a medium to connect to a million-molecule database developed on the basis of physico-chemical properties. The application in its current form has an option to use in-house developed software modules like ParDOCK, RASPD, AADS, DNA Docking, Intercalator, BAPPL, BAPPLZ, SOM, TPACM4, PreDDICTA which are already available in the Web server form for several years backed up by scientific credibility though publications in peer reviewed journals (Fig. 1.11).



44567222	RASPD	completed
22845644	RASPD	completed
57281391	AADS	completed
40792848	ActiveSite	completed
65001693	DNADocking	completed
43361763	DSS	completed
14138185	BAPPL	completed
12404186	ParDOCK	pending
64143731	LipinskiFilters	completed
48251848	TPACM4	completed
89162640	Wiener Index	completed
51702737	Mol. Vol.	completed

Fig. 1.11 Screenshot of *Sanjeevini* mobile app. The application is freely accessible from google play store and can be installed by searching for "Sanjeevini – SCFBio - CADD". Link to download the Sanjeevini mobile application on the google play: https://play.google.com/store/apps/details?id=com.sanjeevini&hl=en; Application Web link on SCFBio's Web portal: http://www.scfbio-iitd.res.in/sanjapp/webSearch/Sanjeevini_webpage.html

1.8 Conclusions

New drug discovery is an important process to ensure human health. Recent developments in this field have made it much faster, inexpensive and extremely successful because of the amalgamation of computational techniques and biology directed to the emergence of multiple tools and softwares. *Sanjeevini* provides a viable option for structure-based computer aided drug discovery (CADD). The software suite in its present form has been implemented on an 80-processor cluster and is made freely accessible to the user community. The higher accuracies of the individual modules, faster results through a nice graphical user interface enable the users to design new lead compounds without much hassle. Building in toxicity filters and further stringent experimental validations on a large number of targets could lead to iterative improvements in *Sanjeevini* and in CADD in general.

Acknowledgement The authors are grateful to the National Supercomputing Mission administered through CDAC and MEITY, Govt. of India for support to the Supercomputing Facility for Bioinformatics and Computational Biology (SCFBio), IIT Delhi.

References

Anderson, A. C. (2003). The process of structure-based drug design. *Chemistry and Biology, 10*(9), 787–797.

Banerjee, K., Bhat, R., Rao, V. U. B., Nain, A., Rallapalli, K. L., Gangopadhyay, S., et al. (2019). Toward development of generic inhibitors against the 3C proteases of picornaviruses. *The FEBS Journal, 286*(4), 765–787. https://doi.org/10.1111/febs.14707

Bhat, R., Kaushik, R., Singh, A., Das Gupta, D., Jayaraj, A., Soni, A., et al. (2020). A comprehensive automated computer-aided discovery pipeline from genomes to hit molecules. *Chemical Engineering Science, 222*, 115711. https://doi.org/10.1016/j.ces.2020.115711

Bhatnagar, S., Soni, A., Kaushik, S., Rikhi, M., Retnabai, T., Kumar, S., et al. (2017). Indian patent entitled "1, 3-Diacetyl Biphenyl Analogs, and their Derivatives" with Application number: 3126/DEL/2012. *Chemical Biology and Drug Design*. https://doi.org/10.1111/cbdd.13126

Daina, A., Blatter, M. C., Gerritsen, V. B., Palagi, P. M., Marek, D., Xenarios, I., et al. (2017). Drug design workshop: A web-based educational tool to introduce computer-aided drug design to the general public. *Journal of Chemical Education, 94*(3), 335–344.

Evanthia, L., Spyrou, G., Vassilatis, D. K., & Cournia, Z. (2014). Structure-based virtual screening for drug discovery: Principles, applications and recent advances. *Current Topics in Medicinal Chemistry, 14*(16), 1923–1938.

Gupta, A., & Sharma, P. (2007). Gandhimathi and Jayaram B: ParDOCK: An all atom energy-based monte carlo docking protocol for protein-ligand complexes. *Protein and Peptide Letters, 14*, 632–646.

Hoque, I., Chatterjee, A., Bhattacharya, S., & Biswas, R. (2017). An approach of computer-aided drug design (CADD) tools for in silico pharmaceutical drug design and development. *International Journal of Advanced Research in Biological Sciences, 4*(2), 60–71.

Imam, S. S., & Gilani, S. J. (2017). Computer aided drug design: A novel loom to drug discovery. *Organic and Medicinal Chemistry, 1*(4), 1–6.

Jain, T., & Jayaram, B. (2005). An all atom energy based computational protocol for predicting binding affinities of protein-ligand complexes. *FEBS Letters, 579*, 6659–6666.

Jayaram, B., Singh, T., Mukherjee, G., Mathur, A., Shekhar, S., & Shekhar, V. (2012). Sanjeevini: A freely accessible web-server for target directed lead molecule discovery. *BMC Bioinformatics, 13*(17), S7. https://doi.org/10.1186/1471-2105-13-S17-S7

Kahraman, A., Thornton, J. M., Schwede, T., & Peitsch, M. (2008). Methods to characterize the structures of enzyme binding sites. In *Computational structural biology: Methods and applications* (pp. 189–221). London: World Scientific Publishing.

Khanna, V., & Ranganathan, S. (2011). *In silico* approach to screen compounds active against parasitic nematodes of major socio-economic importance. *BMC Bioinformatics, 12*, S25.

Kumar, J., Marziya, S., Asim, G., Anju, S., Shandilya, A., Jameel, E., et al. (2018). Pyrimidine-triazolo pyrimidine and pyrimidine-pyridine hybrids as potential acetylcholinesterase inhibitors for Alzheimer's disease. *Chemistry Select: Medicinal Chemistry & Drug Discovery, 3*(2), 736–747. https://doi.org/10.1016/j.jmgm.2016.10.022

Lill, M. (2013). Virtual screening in drug design. In *Silico models for drug discovery. Methods in molecular biology* (Vol. 993, pp. 1–12). Springer Protocols.

Macalino, S. J., Gosu, V., Hong, S., & Choi, S. (2015). Role of computer-aided drug design in modern drug discovery. *Archives of Pharmacol Research, 38*(9), 1686–1701.

Meng, X. Y., Zhang, H. X., Mezei, M., & Cui, M. (2011). Molecular docking: a powerful approach for structure-based drug discovery. *Current Computer Aided Drug Design, 7*(2), 146–157. https://doi.org/10.2174/157340911795677602

Mukherjee, G., & Jayaram, B. (2013). A rapid identification of hit molecules for target proteins via physico-chemical descriptors. *Physical Chemistry Chemical Physics, 15*, 9107–9116.

Nishikawa, J. L., Boeszoermenyi, A., Vale-Silva, L. A., Torelli, R., Posteraro, B., Sohn, Y. J., et al. (2016). Inhibiting fungal multidrug resistance by disrupting an activator-Mediator interaction. *Nature, 530*(7591), 485–489. https://doi.org/10.1038/nature16963

Pinzi, L., & Rastelli, G. (2019). Molecular docking: Shifting paradigms in drug discovery. *International Journal of Molecular Sciences, 20*(18), 4331.

Shaikh, S., Jain, T., Sandhu, G., Latha, N., & Jayaram, B. (2007). A physico-chemical pathway from targets to leads. *Current Pharmaceutical Design, 13*, 3454–3470.

Shandilya, A., Hoda, N., Khan, S., Jameel, E., Kumar, J., & Jayaram, B. (2017). De novo lead optimization of triazine derivatives identifies potent antimalarials. *Journal of Molecular Graphics & Modelling, 71*, 96–103. https://doi.org/10.1016/j.jmgm.2016.10.022

Singh, T., Biswas, D., & Jayaram, B. (2011). AADS - An automated active site identification, docking and scoring protocol for protein targets based on physico-chemical descriptors. *Journal of Chemical Information and Modeling, 51*, 2515–2527.

Soni, A., Bhat, R., & Jayaram, B. (2020). Improving the binding affinity estimations of protein-ligand complexes using machine-learning facilitated force field method. *Journal of Computer-Aided Molecular Design*. https://doi.org/10.1007/s10822-020-00305-1

Szarecka, A., & Dobson, C. (2019). Protein structure analysis: introducing students to rational drug design. *The American Biology Teacher, 81*(6), 423–429. https://doi.org/10.1525/abt.2019.81.6.423. ISSN 0002-7685, electronic ISSN 1938-4211.

Tripathi, P. K., Anjali, S., Singh, Y. S. P., Kumar, A., Gaurav, N., Siva, R. B., et al. (2020). Evaluation of novobiocin and telmisartan for anti-CHIKV activity. *Virology, 548*, 250–260. https://doi.org/10.1016/j.virol.2020.05.010

Trott, O., & Olson, A. J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *Journal of Computational Chemistry, 31*, 455–461.

Xiang, M., Cao, Y., Fan, W., Chen, L., & Mo, Y. (2012). Computer-aided drug design: Lead discovery and optimization. *Combinatorial Chemistry & High Throughput Screening, 15*(4), 328–337.

Yamada, M., & Itai, A. (1993). Development of an efficient automated docking method. *Chemical and Pharmaceutical Bulletin, 41*, 1200–1202.

Chapter 2
Virtual Screening: Practical Application of Docking, Consensus Scoring and Rescoring Using Binding Free Energy

Sunita Gupta, Mohd. Waseem, Naveen Kumar Meena, Roopa Kuntal, Andrew M. Lynn, and Smriti Mishra

Abstract The chapter describes the development of a virtual screening workflow involving open-source docking algorithms, consensus scoring, and rescoring using binding energy estimation to identify hits in the drug discovery process. Plasmodium falciparum dihydrofolate reductase (PfDHFR-TS, PDB 1J3I) is used as an example to describe the Virtual Screening (VS) workflow.

The open-source docking algorithms included in the workflow are Autodock4, Autodock Vina, and DOCK6. Post docking analysis is done using consensus scoring to provide a standard scoring scheme across all methods. Sum rank, Sum Score, and Reciprocal rank methods are evaluated. The results are visualized using enrichment plots. Besides, the application of the molecular dynamics simulations to estimate binding energy and rescoring is described.

2.1 Introduction

Virtual screening is a high-throughput computational technique adopted in a drug discovery program to screen a library of ligands against a biological target and identify the hit compounds. The screening method involves the evaluation of the binding activity of the compounds to a target. Since it involves a virtual chemical library, the chemical library to be screened in the experiment is easy to synthesize and manage. Moreover, virtual screening reduces the time and cost to identify hits in the drug discovery process. Virtual screening encompasses a plethora of docking algorithms. Docking algorithms such as Autodock4, Autodock Vina, DOCK6,

Electronic Supplementary Material The online version of this chapter (https://doi.org/10.1007/978-981-15-8936-2_2) contains supplementary material, which is available to authorized users.

S. Gupta · M. Waseem · N. K. Meena · R. Kuntal · A. M. Lynn (✉) · S. Mishra (✉)
School of Computational and Integrative Sciences, Jawaharlal Nehru University, Delhi, India
e-mail: andrew@jnu.ac.in; s_mishra@jnu.ac.in

S. K. Singh (ed.), *Innovations and Implementations of Computer Aided Drug Discovery Strategies in Rational Drug Design*,
https://doi.org/10.1007/978-981-15-8936-2_2

GOLD, and GLIDE are widely used in the virtual screening of small molecules (Kitchen et al. 2004; Warren et al. 2006). Amongst them, Autodock4, Autodock Vina, and DOCK6 are open-source docking algorithms. Moreover, Autodock4 and Autodock Vina and DOCK6 are frequently used in academia. Besides, algorithms such as Autodock Vina exhibit good performance and are preferred to commercial docking algorithms.

Open source tools are free to use and involve less strict license requirements. Besides, the application source code of open source tools is easily accessible, and modification of the code is possible. Open source algorithms also exhibit good and comparable performance to commercial algorithms (Pagadala et al. 2017). However, commercial tools are accessible on payment, involve stringent license requirements, and may be costly. Unlike open-source tools, modification of the code may not be easy. Hence, open-source docking algorithms are encouraged to use in the virtual screening process.

Docking algorithms, in conjunction with consensus scoring, improves docking results. Consensus scoring function is performed to screen, rank, and rescore the library of ligands in early drug discovery. Moreover, consensus scoring is preferred to overcome the discrepancy in results associated with different scoring functions. Besides consensus scoring, the estimation of binding energy in the virtual screening process overcomes the limitations of the docking results (Feher 2006). Refinement and rescoring using binding energy improves the docking results and generate reasonable scores and conformations.

The performance of VS docking approaches is tested using a benchmarking dataset. Benchmarking dataset available for testing of VS approaches include Directory of Useful Decoys (DUD), DUD-Enhanced (DUD-E), Virtual Decoy Sets (VDS), G protein-coupled receptors (GPCRs) ligand library (GLL) and GPCRs Decoy Database (GDD), Demanding Evaluation Kits for Objective in Silico Screening (DEKOIS) and DEKOIS 2.0 Nuclear Receptors Ligands and Structures Benchmarking DataBase (NRLiSt BDB), DUD LIB VS 1.0, reproducible virtual screens database (REPROVIS-DB) and Maximum Unbiased Validation (MUV) (Xia et al. 2015; Lagarde et al. 2015).

In this chapter, the application of Autodock4, Autodock Vina, and DOCK6, consensus scoring, and rescoring of the binding free energy in the virtual screening process is emphasized.

2.1.1 Docking Algorithms: Autodock4, Autodock Vina, DOCK6

Docking algorithms are designed to predict the biological activity of small molecules in the drug discovery process. A typical docking program includes algorithms to search the conformations and score using scoring functions. Different search algorithms and scoring functions are used to sample the conformations and score the receptor-ligand interactions. The search methods are designed specifically to manage ligand flexibility and, to some extent, protein flexibility. Ligand flexibility may be treated using systematic, random/stochastic and simulations search methods. The systematic search method involves incremental construction, conformational search, and databases. Stochastic involves Monte Carlo, genetic algorithms, and tabu (conformational space) search approach. Simulations involve molecular dynamics and energy minimization. To score the receptor-ligand interactions force-field based, knowledge-based and empirical scoring functions are integrated into the docking tools (Kitchen et al. 2004). The search methods and scoring functions integrated into Autodock4, Autodock Vina, and DOCK6 docking algorithms are summarized below.

Autodock4

Autodock is a grid-based automated docking algorithm. Autodock4 is a new version of Autodock and integrates the Lamarckian genetic search algorithm and empirical force field scoring function of Autodock3. Besides the characteristics of Autodock3, Autodock4 incorporates limited flexibility in the receptor to assess the covalent bound ligands. Autodock4 also includes simulated annealing and traditional genetic search algorithms.

Lamarckian genetic algorithm (LGA) selects conformations with the lowest binding energy. Autodock4 uses empirical force field-based scoring function to evaluate the receptor-ligand interactions involving the binding energy of the ligand molecules. It evaluates unbound and bound energy states of the ligand, protein, and complex to predict the binding free energy. Autodock4 also includes a charge based desolvation method. The desolvation method involves a typical set of atom types and charges. The algorithm also includes protein flexibility involving side chains of the amino acids (Morris et al. 2009; Forli et al. 2016).

Autodock Vina

In comparison to Autodock, Autodock Vina is easy to use, exhibits better performance, more accurate in the prediction of binding mode, includes automatic grid map calculation, and is compatible with Autodock tools. The docking tool is fast and reduces the processing time involving multithreading approach on the multiple cores of the machine (*AutoDock Vina—molecular docking and virtual screening program*, http://vina.scripps.edu/). The algorithm includes iterated local search global optimizer and conformational dependent scoring function. Vina scoring function is inspired by the X score. Iterated local search global optimizer involves a local optimization method (quasi-Newton method). Besides, Autodock Vina includes protein flexibility involving amino acid side chains (Trott and Olson 2009; Jaghoori et al. 2016).

DOCK6

The DOCK algorithm involves geometric matching and incremental construction to generate conformers of the ligand. DOCK6 integrates the previous versions of the DOCK algorithm. The algorithm has improved sampling and scoring functions. The scoring function integrated into DOCK6 is energy scoring (Allen et al. 2015).

2.1.2 Consensus Scoring

Consensus scoring is used to enrich the number of active inhibitors at an early stage of screening. The approach combines the individual scores, and the errors are balanced. Consensus scoring integrates the results of the individual docking algorithm. The method involves an average of the score or rank of each molecule generated using the docking program. Popular consensus scoring approaches include voting, coarse quantiles voting, rank voting, simple sum ranks, deprecated sum-ranks, worst-best ranks, weighted sum ranks, regression schemes, and multivariate methods (Feher 2006; Palacio-Rodríguez et al. 2019).

2.1.3 Rescoring Using Binding Free Energy

The development of better algorithms and high-end computation accelerates the process of estimating the binding (ΔG_{bind}), contrary to the experimental procedure, which is a time consuming and costly process in drug discovery. Therefore, improved computational methods have been developed to reduce the cost and time to discover new drugs. In structure-based drug design, molecular mechanics energies along with Poisson-Boltzmann or generalized Born and Surface area continuum solvation (MMPBSA and MMGBSA) are routinely adopted to estimate the free energies of binding ($\Delta G_{\text{binding}}$) of small molecules to biological macromolecules.

Binding free energy is estimated using the equation $\Delta G_{\text{binding}} = G_{\text{comlex}} - G_{\text{receptor}} - G_{\text{ligand}}$. The free energy of each of the system are calculated as a sum of molecular mechanics energy (E_{MM}, the sum of the internal energy of the molecule plus the electrostatic and van der Waals interactions in vacuo), polar contribution to the solvation energy of the molecule (G_{psolv}) and nonpolar solvation free energy (G_{npsolv}) as shown in the equation, $G = (E_{\text{MM}}) + (G_{\text{psol}}) + (G_{\text{npsolv}})$. G_{psolv} is calculated by solving the Poisson-Boltzmann (PB) and Generalized-Born (GB) equations for MMPBSA and MMGBSA methods. In contrast, G_{npsolv} is calculated using the equation $G_{\text{npsolv}} = \gamma * \text{SASA} + b$, where SASA is the solvent-accessible surface area calculated using the linear combinations of pairwise overlaps or Molsurf methods.

MMPBSA approach is adopted to study relative free energies of macromolecules, binding free energies between protein-protein, protein-ligand, and nucleic acid systems (Kollman et al. 2000). However, the MMGBSA model involves Atomic Born radii to calculate the approximations to the electrical polarization component (G_{psolv}) of solvation free energy. The MMGBSA approach is reported to give comparable free energy with experimental results. In the estimation of binding free energy, MMPBSA and MMGBSA involve a well-equilibrated system of small-molecule-receptor complexes and explicit molecular dynamics (MD) simulations (Genheden and Ryde 2015; Wang et al. 2018). The methods exhibit an accurate estimation of binding energy. But, the methods are time-consuming and cannot be routinely used in the drug discovery process.

Hence, to reduce computational cost and time, Rastelli et al. 2009 developed a method to rescore using binding energy (Rastelli et al. 2009). In this chapter, the implementation of the method is described (Gupta et al. 2018).

2.2 Virtual Screening and Rescoring Workflow

2.2.1 Computational Requirements

All Computational studies are performed on a personal computer with Centos Linux 7 operating system involving Intel® Core™ i7-3770 CPU @3.40 GHz processor. Molecular visualization is performed with Chimera 1.10.2 (*UCSF Chimera Home Page*, http://www.cgl.ucsf.edu/chimera/). Open source tools—Autodock (*AutoDock—AutoDock*, http://autodock.scripps.edu/), Autodock Vina (AutoDock Vina—molecular docking and virtual screening program), Dock6 (*UCSF DOCK*, http://dock.compbio.ucsf.edu/), MGLtools (*MGLTools Website—Welcome — MGLTools*, http://mgltools.scripps.edu/), Putative Active Sites and Spheres (PASS) (*overview.html*, http://www.ccl.net/cca/software/UNIX/pass/overview.shtml; Brady and Stouten 2000), OpenBabel (*Open Babel—Browse/openbabel/2.3.2 at SourceForge.net*, https://sourceforge.net/projects/openbabel/files/openbabel/2.3.2/) Mayachemtools *MayaChemTools: Home*, http://www.mayachemtools.org/), (*About MODELLER*, https://salilab.org/modeller/; Webb and Sali 2016) and enrichvs (*CRAN—Package enrichvs*, https://cran.r-project.org/web/packages/enrichvs/index.html) were downloaded and locally installed in the system. Download and installation instructions are described in the supplementary document (S1). R 3.2.3 is used to plot enrichment curves (*R: The R Foundation*, https://www.r-project.org/foundation/).

2.2.2 Virtual Screening Using Docking

Virtual screening using docking involves a practical application of the Autodock4, Autodock Vina, and Dock6 open-source docking programs. *Plasmodium falciparum* dihydrofolate reductase PfDHFR-TS (PDB 1J3I) (Yuvaniyama et al. 2003) complex consisting of NADPH and two water molecules (W1249 and W1250) are selected to demonstrate the virtual screening workflow (Fig. 2.1). The virtual screening workflow is described in the following sections. The steps and the scripts involved in the execution of the workflow can be obtained from the Supplementary.

Briefly, docking with Autodock4 involving a grid size of 50 × 50 × 50 with 0.375 grid spacing, and −0.1465 dielectric constant is constructed. Lamarckian Genetic Algorithm (LGA) is used as the search method. The Auto Grid program

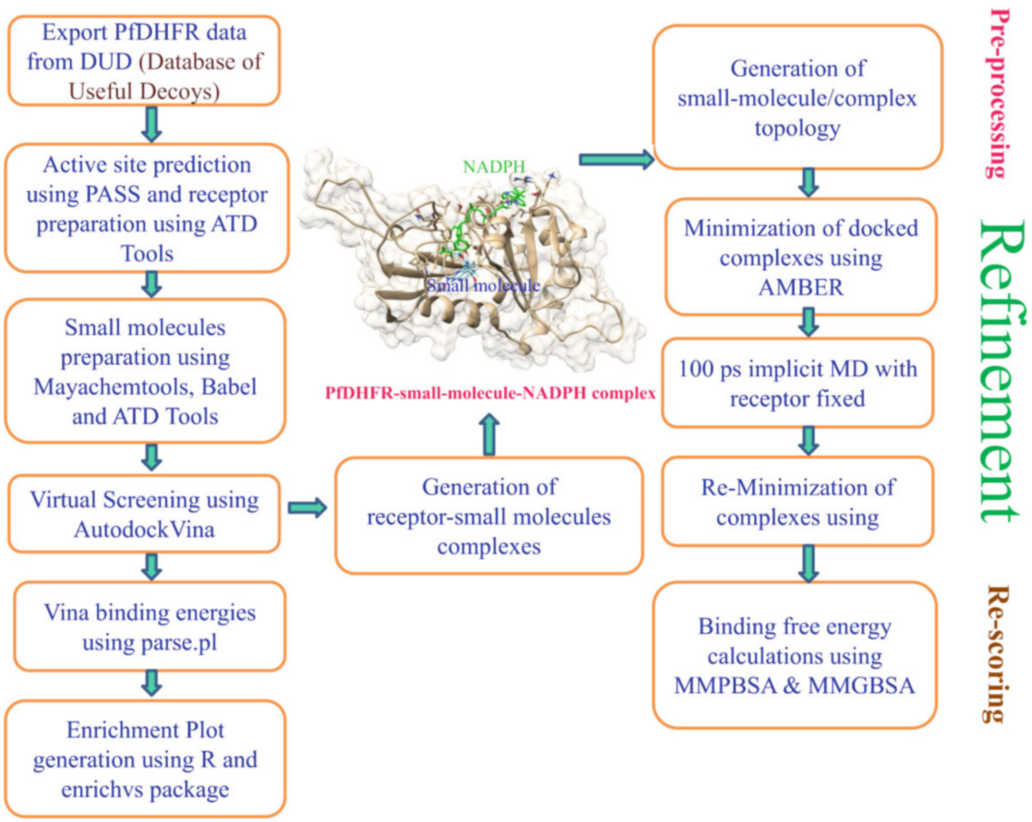

Pre-processing
Refinement
Re-scoring
Export PfDHFR data from DUD (Database of Useful Decoys)
Active site prediction using PASS and receptor preparation using ATD Tools
Small molecules preparation using Mayachemtools, Babel and ATD Tools
Virtual Screening using AutodockVina
Vina binding energies using parse.pl
Enrichment Plot generation using R and enrichvs package
Generation of receptor-small molecules complexes
NADPH
Small molecule
PfDHFR-small-molecule-NADPH complex
Generation of small-molecule/complex topology
Minimization of docked complexes using AMBER
100 ps implicit MD with receptor fixed
Re-Minimization of complexes using
Binding free energy calculations using MMPBSA & MMGBSA

Fig. 2.1 Workflow of virtual screening and post-processing method

determines the affinity maps of grids. The docking parameters included a number of genetic algorithm (GA) runs: 10, individual population size: 150, the maximum number of energy evaluation: 2,50,0000, and the maximum number of generation: 27,000.

Perl wrappers are used to execute the python codes for the library of ligand and decoy. The PDBQT format of transformed receptor and dataset is involved in docking with Autodock4 and Autodock Vina. The workflow for Autodock4 and Autodock Vina is shown in Fig. 2.2.

Flexible docking is performed with Dock6. DMS program of Dock6 determines the molecular surface and active binding site of the transformed receptor. Sphgen tool generates overlapping spheres to describe the shape of the molecule or molecular surfaces. The active site is determined by using a 10 Angstrom radius with reference to WR1092. Grid is generated using a Showbox tool. The Showbox tool involves an energy scoring method to generate a grid. The Mol2 format of the transformed ligand library is used for docking.

Preprocessing of the Receptor and Ligand Library

Preprocessing of receptor and ligand involved the transformation of the receptor and ligand to 0.0 0.0 0.0 coordinates, adding hydrogens and charge to the PDB structure and small molecules. Charges are required to enable interactions between the ligand and receptor. Modeling of missing residues is performed using Modeller.

Receptor

The PDB structure of PfDHFR-TS PDB 1J3I (Yuvaniyama et al. 2003) is retrieved from RCSB Protein Data Bank. Chain B of PfDHFR-TS (PDB 1J3I) is used for docking studies. Modeller models the protein and ten homology models are generated. The best model based on the DOPE score is validated using PROCHECK (PROCHECK—DOE-MBI Structure Lab UCLA, https://servicesn.mbi.ucla.edu/PROCHECK/). All water molecules are removed from the protein except W1249 and W1250. W1249 and W1250 are in proximity to Asp54 residue of the active site. The co-crystallized ligand WR99210 is removed from the complex. However, the NDP cofactor is retained.

Ligand Library

The data set of PfDHFR-TS (PDB 1J3I) consists of ligands and decoys. It is obtained from the directory of useful decoys in the SDF format (dhfr_ligands.sdf.gz and dhfr_decoys.sdf.gz) (*DUD—A Directory of Useful Decoys*, http://dud.docking.org/r2/). The *dhfr_ligands.sdf* file is split, and duplicates are removed from the library using *SplitSDFiles.pl* script of Mayachemtools. The *dhfr_ligands.sdf* file is split into

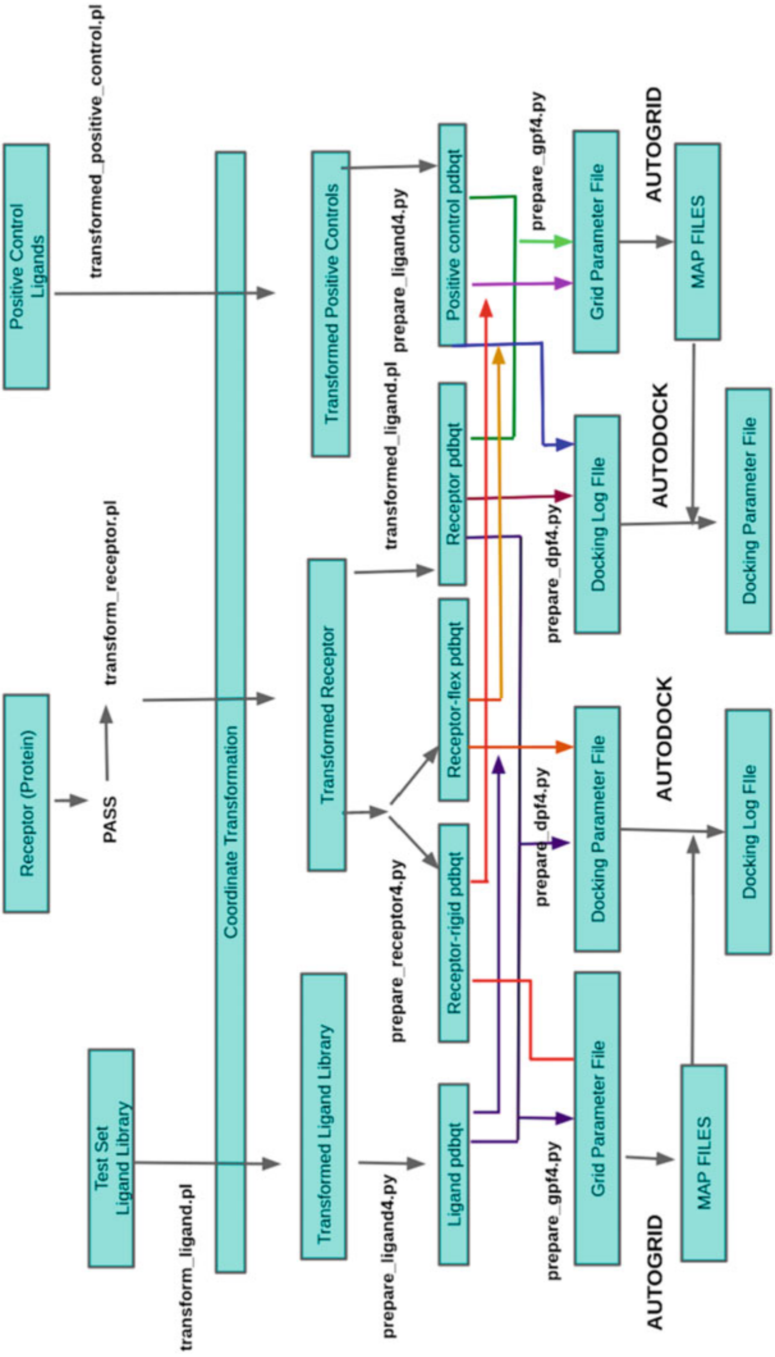


Fig. 2.2 Workflow for Autodock4 and Autodock Vina

new SDF files of the ligands, and the name of each new SDF file corresponds to the ligand name. Subsequently, hydrogen atoms and gasteiger charges are added to the molecule using Openbabel. Openbabel also generates the 3D structure of the ligands.

Transformation of the Receptor and Ligand Library

The transformation of the receptor and ligand library is a significant step included in the virtual screening protocol. It involves positioning the coordinate center of active site points to the center of the active cavity. Thus, coordinates are positioned to 0.0 0.0 0.0 of the predicted active site.

PASS algorithm determines the active site point (*receptor.asps*) in the receptor. The script *transform-receptor.py* transforms the receptor to 0,0,0 coordinates.

However, the transformation of the ligand library involves the Perl script (*transform-ligand.pl*). The prepped transformed files of the receptor (*receptor_transformed.pdb*), and ligand *(ligand_transformed.pdb*) may be used for processing with different docking algorithms (see "Electronic Supplementary Material").

Processing

The transformed receptor and ligand are processed using Autodock4, Autodock Vina, and Dock6 (Forli et al. 2016).

Preparing the transformed receptor and ligand in PDBQT format for processing with Autodock and Autodock Vina

The transformed receptor and ligand library is prepared in the PDBQT (*receptor_transformed.pdbqt* and *ligand name_transformed.pdbqt*) format using python (*prepare_receptor4.py*) and Perl (*prepare_ligand.pl*) wrapper, respectively (see "Electronic Supplementary Material").

Autodock4

Processing with Autodock4 involves the preparation of the grid parameter file, calculation of grid maps for each atom type in the ligand using autogrid4, and the docking parameter file (specifies the files and parameters for the docking calculation).

The grid parameter file is obtained using *prepare_gpf.pl* Perl wrapper. The size of the grid (*npts* = '*50,50,50*') is mentioned in the Perl wrapper. Subsequently, the grid maps are generated and include all the possible atoms types in the ligand. The final step of the processing involves the preparation (*prepare_dpf.pl*) of the docking parameter file (*.dpf*). The docking parameter files (*.dpf*) are listed and on the execution of the script *prepare_dlg.pl*, docking log files (*.dlg*) are generated (see "Electronic Supplementary Material").

The results obtained using Autodock are evaluated using Perl script. Two directories (*hi* and *hna*) are created. The *hi* directory contains a copy of the docking log files (*.dlg*). *summarize_result.pl* script. On the execution of the script *summarize_result.pl*, a summary file including the root mean square deviation (rmsd), Genetic Algorithm runs and calculated lowest energy is generated as an output. The docking log files (*.dlg*) are processed and are converted to PDBQT format using the *prepare_write_complexes.pl*. The *prepare_pdbqt_to_pdb.pl* script may be used to convert PDBQT files to PDB format.

Autodock Vina

The receptor (*receptor_transformed. pdbqt*) and the ligand files (*ligand name_transformed.pdbqt*) are processed using AutoDock Vina. A *vina-folder* directory is created. The *vina-folder* contains a copy of the *recetor_transformed.pdbqt* and a directory *lig_library*. The *lig_library* contains a copy of all the ligands in the PDBQT format (*ligand name_transformed.pdbqt*). The *vina-folder* also contains *vina_VS_input_final.sh* and *conf_input_final.py* python script. A *conf.txt* file is generated and mentions the grid parameters. Besides, a *lig_lib* is obtained containing the docked files (see “Electronic Supplementary Material”).

DOCK6

DOCK6 algorithm compatible receptor and ligands are prepared. The transformed receptor (*receptor_transformed.pdb)* and transformed ligand library in the PDB format (*ligand_transformed.pdb)* is used as the input file. Receptor and ligands are prepared as described in the DOCK6 tutorial. The transformed ligand library in PDB format (*ligand_transformed.pdb*) is converted to MOL2 format and combined to a single file (*transformed_ligand.mol2)*.

Grid generation and processing are performed as described in the DOCK6 tutorial (*UCSF DOCK*, http://dock.compbio.ucsf.edu/DOCK_6/tutorials/index.htm).

Consensus Scoring

Post docking analysis is performed by ranking all the docked ligand and decoy molecules obtained by Autodock Vina, Autodock, and Dock6 in the descending order (highest to lowest score). In the early recognition of active compounds, consensus scoring is done using the sum score, sum rank, and reciprocal rank method (Feher 2006).

Sum Rank
sum rank (i) = {1/(autodock_rank(i)/total no. of docked molecules) + 1/(vina_rank (i)/total no. of docked molecules) + 1/(dock6_rank(i)/total no. of docked molecules)}/total no of docking method

Reciprocal Rank
reciprocal rank (i) = 1/{((1/(autodock_rank(i)/total no. of docked molecules) + 1/ (vina_rank(i)/total no. of docked molecules) + 1/(dock6_rank(i)/total no. of docked molecules)/total no of docking method))/total no of docking method}

Sum Score
sum_score (i) = {(autodock_score[i]/autodock_score_min) + (vina_score [i]/vina_score_min) + (dock6_score[i]/dock6_score_min) }/total no. of docking software

Enrichment Plot

Enrichment plots are drawn between the percentage of top-ranked database and percentage of active compounds using enrichVS package. The performance of VS open-source docking tools and consensus scoring approaches were evaluated using the benchmarking dataset (ligands and decoys for PfDHFR-TS (PDB 1J3I). Sum rank, sum score, and reciprocal rank were calculated using the docking scores of Autodock, Autodock Vina, and Dock6. The performance of the open-source docking tools and consensus scoring approaches were compared with GOLD and GLIDE.

In the early recognition of the actives, the overall performance involving the consensus scoring approach in the workflow was better than the open-source docking tools. The open source tools in conjunction with consensus scoring approach exhibits comparable performance. While the overall performance of GLIDE and GOLD on the complete dataset is superior, the use of consensus scoring actually provides better results for the detection of early hits. The number of actives determined using the consensus scoring approach was approximately 65% in the top 10% of the ranked database. Amongst open-source docking tools, the performance of AutodockVina was better than Autodock and Dock6 (Fig. 2.3).

.

2.2.3 *Rescoring Using MMPBSA/MMGBSA*

Rescoring using binding energy has emerged as an efficient tool to post-process the docked poses. The approach was developed by Rastelli et al. 2009 and is popular as Binding Estimation After Refinement (BEAR) (Rastelli et al. 2009). BEAR involves a three-tier refinement (MM/MD/MM) process and rescores the ligands using accurate scoring functions MM(PB/GB)SA (Rastelli et al. 2009; Anighoro and

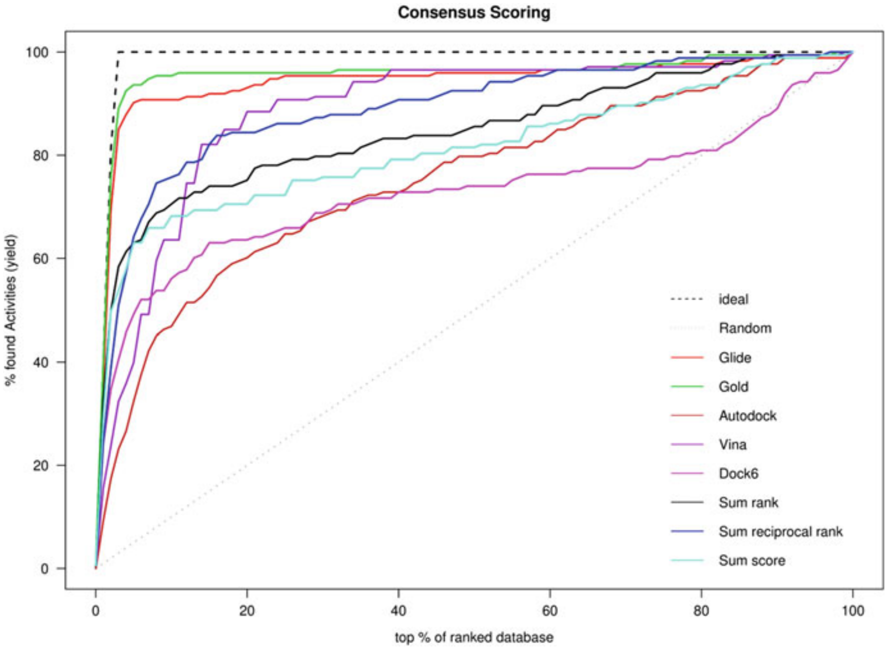


Fig. 2.3 Comparison of open source docking algorithms and open source docking algorithms in combination with consensus scoring methods using the virtual screening workflow described in the chapter. Y-axis represents the yield (percentage of the activities). X-axis represents the top percentage of ranked database (pfDHFR data)

Rastelli 2013; Rastelli and Pinzi 2019). The refinement includes energy minimization of the complexes followed by MD simulations. The approach allows the movement of ligands and involves restrain on the receptor. The final step is the re-minimization of the entire complexes (Degliesposti et al. 2011).

Preparation of the System: Receptor-Ligand Library Complex

Molecular docking yields receptor-ligand complexes. Post-processing is required to filter the false negatives from the docking ensembles and to rank the top hits using the binding free energy obtained via MM(PB/GB)SA (Sgobba et al. 2012). The BEAR algorithm involves the *pre-processing* of docked complexes, *refinement of the complexes*, and the *estimation of binding free energy*. The pre-processing includes the parameterization of the receptor, ligands, and generation of complex topology. Amber Molecular Dynamics package is used in the pre-processing, refinement of the complexes, and estimation of the binding free energy (*The Amber Molecular Dynamics Package*, http://ambermd.org/).

Parameterization of Receptor and Ligand Library

The pre-processing step includes the addition of hydrogen atoms to the receptor. The Leap module assigns Amber atom types and charges to the receptor using the ff99SBILDN force field. The ligand library is parameterized and atomic charges are calculated using the AM1-BCC model (Jakalian et al. 2002). The Antechamber module builds molecular mechanics parameters of the ligands using Generalized Amber Force Field (GAFF) (Wang et al. 2004). Parmchck assigns the missing forcefield parameters.

Refinement of the Receptor-Ligand Complex

The refinement of the receptor-ligand complex involves the three-tier refinement procedure. The initial MM energy minimization is performed on the entire receptor-ligand complexes. It is followed by a short MD simulation involving restrain on the receptor and the unrestrained movement of the ligand. In the last step, the entire complex is re-minimized. The energy minimization of the complexes is performed using the sander module of AMBER16 (Case et al. 2016) for 2000 steps. It involves distance-dependent dielectric constant $\varepsilon = 4r$ and a cutoff of 12 Å. Subsequently, 100 ps MD simulation is performed and the ligand unrestrained. The final step is again re-minimization and the entire complex is re-minimized for 2000 steps. The MD is performed at a temperature 300 K. and involves SHAKE algorithm (Ryckaert et al. 1977) to constrain the hydrogen atoms (time step of 2 fs).

Rescoring

A single conformation of each refined complexes is selected for estimating the free energy of binding using MMPBSA.py script available in AMBER module. An automated script including parameterization, system preparation, refinement, and rescoring, can be obtained from the supplementary.

2.3 Conclusion

The VS workflow integrates open-source docking programs (Autodock4, Autodock Vina and DOCK6), consensus scoring (sum rank, sum score, and reciprocal rank) and rescoring using binding energy. The performance of the docking programs is assessed by docking the PfDHFR-TS (PDB 1J3I) with a benchmarking dataset downloaded from DUD. Consensus Scoring was done using sum rank, sum score, and reciprocal method. The number of actives in early recognition was higher with the consensus scoring method than with open source programs. Besides, rescoring using binding energy estimation enables the efficient sampling of receptor-ligand complexes. It also allows us to accurately rescore the docking poses at reasonable time. Hence, it can be routinely used in the drug discovery process.

References

Allen, W. J., Balius, T. E., Mukherjee, S., et al. (2015). DOCK 6: Impact of new features and current docking performance. *Journal of Computational Chemistry, 36*, 1132–1156. https://doi.org/10.1002/jcc.23905

Anighoro, A., & Rastelli, G. (2013). BEAR, a molecular docking refinement and rescoring method. *Computational Molecular Bioscience, 03*, 27–31. https://doi.org/10.4236/cmb.2013.32004

Brady, G. P., & Stouten, P. F. W. (2000). Fast prediction and visualization of protein binding pockets with PASS. *Journal of Computer-Aided Molecular Design, 14*, 383–401. https://doi.org/10.1023/A:1008124202956

Case, D. A., Walker, R. C., Junmei, D., & Wang, T. (2016). *Amber 2016 Reference Manual Principal contributors to the current codes* (pp. 1–923). San Francisco, CA: University of California.

Degliesposti, G., Portioli, C., Parenti, M. D., & Rastelli, G. (2011). BEAR, a novel virtual screening methodology for drug discovery. *Journal of Biomolecular Screening, 16*, 129–133. https://doi.org/10.1177/1087057110388276

Feher, M. (2006). Consensus scoring for protein–ligand interactions. *Drug Discovery Today, 11*, 421–428. https://doi.org/10.1016/j.drudis.2006.03.009

Forli, S., Huey, R., Pique, M. E., et al. (2016). Computational protein-ligand docking and virtual drug screening with the Auto Dock suite. *Nature Protocols, 11*, 905–919. https://doi.org/10.1038/nprot.2016.051

Genheden, S., & Ryde, U. (2015). The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery, 10*, 449–461.

Gupta, S., Lynn, A. M., & Gupta, V. (2018). Standardization of virtual-screening and post-processing protocols relevant to in-silico drug discovery. *3 Biotech, 8*, 1–7. https://doi.org/10.1007/s13205-018-1523-5

Jaghoori, M. M., Bleijlevens, B., & Olabarriaga, S. D. (2016). 1001 ways to run AutoDock Vina for virtual screening. *Journal of Computer-Aided Molecular Design, 30*, 237–249. https://doi.org/10.1007/s10822-016-9900-9

Jakalian, A., Jack, D. B., & Bayly, C. I. (2002). Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *Journal of Computational Chemistry, 23*, 1623–1641. https://doi.org/10.1002/jcc.10128

Kitchen, D. B., Decornez, H., Furr, J. R., & Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nature Reviews Drug Discovery, 3*, 935–949. https://doi.org/10.1038/nrd1549

Kollman, P. A., Massova, I., Reyes, C., et al. (2000). Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Accounts of Chemical Research, 33*, 889–897. https://doi.org/10.1021/ar000033j

Lagarde, N., Zagury, J. F., & Montes, M. (2015). Benchmarking data sets for the evaluation of virtual ligand screening methods: Review and perspectives. *Journal of Chemical Information and Modeling, 55*, 1297–1307.

Morris, G. M., Huey, R., Lindstrom, W., et al. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry, 30*, 2785–2791. https://doi.org/10.1002/jcc.21256

Pagadala, N. S., Syed, K., & Tuszynski, J. (2017). Software for molecular docking: A review. *Biophysical Reviews, 9*, 91–102.

Palacio-Rodríguez, K., Lans, I., Cavasotto, C. N., & Cossio, P. (2019). Exponential consensus ranking improves the outcome in docking and receptor ensemble docking. *Scientific Reports, 9*, 1–14. https://doi.org/10.1038/s41598-019-41594-3

Rastelli, G., Degliesposti, G., Del Rio, A., & Sgobba, M. (2009). Binding estimation after refinement, a new automated procedure for the refinement and rescoring of docked ligands in virtual screening. *Chemical Biology & Drug Design, 73*, 283–286. https://doi.org/10.1111/j.1747-0285.2009.00780.x

Rastelli, G., & Pinzi, L. (2019). Refinement and rescoring of virtual screening results. *Frontiers in Chemistry, 7*, 498. https://doi.org/10.3389/fchem.2019.00498

Ryckaert, J. P., Ciccotti, G., & Berendsen, H. J. C. (1977). Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *Journal of Computational Physics, 23*, 327–341. https://doi.org/10.1016/0021-9991(77)90098-5

Sgobba, M., Caporuscio, F., Anighoro, A., et al. (2012). Application of a post-docking procedure based on MM-PBSA and MM-GBSA on single and multiple protein conformations. *European Journal of Medicinal Chemistry, 58*, 431–440. https://doi.org/10.1016/j.ejmech.2012.10.024

Trott, O., & Olson, A. J. (2009). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry, 31*(2), 455–461. https://doi.org/10.1002/jcc.21334

Wang, C., Greene, D., Xiao, L., et al. (2018). Recent developments and applications of the MMPBSA method. *Frontiers in Molecular Biosciences, 4*, 87.

Wang, J., Wolf, R. M., Caldwell, J. W., et al. (2004). Development and testing of a general Amber force field. *Journal of Computational Chemistry, 25*, 1157–1174. https://doi.org/10.1002/jcc.20035

Warren, G. L., Andrews, C. W., Capelli, A. M., et al. (2006). A critical assessment of docking programs and scoring functions. *Journal of Medicinal Chemistry, 49*, 5912–5931. https://doi.org/10.1021/jm050362n

Webb, B., & Sali, A. (2016). Comparative protein structure modeling using MODELLER. *Current Protocols in Bioinformatics, 2016*, 5.6.1–5.6.37. https://doi.org/10.1002/cpbi.3

Xia, J., Tilahun, E. L., Reid, T.-E., et al. (2015). Benchmarking methods and data sets for ligand enrichment assessment in virtual screening. *Methods, 71*, 146–157. https://doi.org/10.1016/j.ymeth.2014.11.015

Yuvaniyama, J., Chitnumsub, P., Kamchonwongpaisan, S., et al. (2003). Insights into antifolate resistance from malarial DHFR-TS structures. *Nature Structural & Molecular Biology, 10*, 357–365. https://doi.org/10.2210/PDB1J3I/PDB

Chapter 3
Aspects of Protein Structure, Function, and Dynamics in Rational Drug Designing

Daliah Michael, Namrata Bankoti, Ansuman Biswas, and K. Sekar

> *If we were to name the most powerful assumption of all, which leads one on and on in an attempt to understand life, it is that all things are made of atoms, and that everything that living things do can be understood in terms of the jigglings and wigglings of atoms.*
>
> *-Richard Feynman, Nobel Prize in Physics, 1965*

Abstract The availability of the first three-dimensional protein structure of myoglobin, ever since, has changed the way drug designers approach the protein-drug binding problem. Thereafter, the dogma shifted from the "lock and key" to the "induced fit" and later the "conformational selection" model. This shift could be attributed to the various experimental techniques used to solve the protein's three-dimensional structure and its function. The basis of this new ideology lies in the fact that the atoms of the protein are not static but in constant motion. Furthermore, due to the folding of the protein's secondary structural elements to arrange themselves spatially, there is a flexibility associated with the whole protein structure. In addition, computational methodologies such as molecular docking and molecular dynamics simulations have proved to be a boon to the drug designing process. This chapter explains, in a nut shell, how protein dynamics and computer-aided drug design play important roles in rational drug designing.

D. Michael · N. Bankoti · K. Sekar (✉)
Department of Computational and Data Sciences, Indian Institute of Science, Bangalore, India
e-mail: sekar@iisc.ac.in

A. Biswas
National Center for Biological Sciences, Bangalore, India

S. K. Singh (ed.), *Innovations and Implementations of Computer Aided Drug Discovery Strategies in Rational Drug Design*,
https://doi.org/10.1007/978-981-15-8936-2_3

3.1 Introduction

Proteins can be perceived as nanodevices/nanomachines that play many critical roles in the body. They play a pivotal role in maintaining coordination in the structure, function, and regulation of the cells, tissues, and organs through their function. Proteins are made up of smaller units called amino acids, which are attached in a linear sequence of peptide bonds to form long chains. Twenty naturally occurring amino acids combine to form a protein. This linear arrangement of amino acids is referred to as the *primary structure* of the protein. The sequence of amino acids in a protein helps in determining its unique three-dimensional structure (Anfinsen and Haber 1961) and its specific function. The different patterns adopted by these chains based on the hydrogen bonds formed between the amino acids constitute the *secondary structure* of the protein molecule. There are two major types of secondary structures, alpha helix and beta-pleated sheets. It is interesting to note that long before the first protein structure was made available, Linderstrom Lang (1952) had postulated that protein sequences can form distinct structural motifs that will eventually fold into a three-dimensional structure. These motifs are nothing but the secondary structural elements of the protein (alpha-helix and beta-pleated sheets). *Tertiary structure* is when these features fold and arrange into three-dimensional structures, which are actually the functional forms of the protein. Furthermore, proteins can form larger macromolecule complexes to stabilize their structure and function. This higher level of arrangement is referred to as the *quaternary structure* of the protein. The best example to give here is that of the hemoglobin protein structure (Fig. 3.1), a popular example that can be found in most textbooks. The structure of hemoglobin was solved by Max Perutz in 1959 (Perutz et al. 1960) by X-ray crystallography. Hemoglobin has a quaternary structure whose function is to transport oxygen in the blood. It has two alpha chain and two beta chain subunits.

Based on structure, proteins can be broadly classified as fibrous and globular proteins. *(1) Fibrous proteins* usually comprise a particular type of secondary structural element in which the polypeptide units are arranged in strands or sheets. Their function mainly lies in providing strength and support to vertebrates. Collagen and keratin are two such proteins. (2) *Globular proteins* are made up of different secondary structural elements, which are arranged in a globular shape. Enzymes and most other functional proteins fall in the category of globular proteins; for instance, hemoglobin, which isa protein that transports oxygen (Imai 1999). Globular proteins are dynamic in nature and display a close relationship between sequence-structure and function. Based on functions, proteins can fall into eight broad categories (hormone, enzyme, structural protein, defensive protein, storage protein, transport protein, receptor protein, and contractile Protein) out of which enzymes and receptor proteins are crucial for drug designing (Geronikaki 2019; Marques et al. 2017; Rekka et al. 2019; Yari et al. 2017).

From their function as detergents (Hasan et al. 2010) to complex ones (Sweeney and Holzbaur 2018) in our body, enzymes are indispensable biological molecules that are vital for a wide variety of functions in living organisms. Based on the

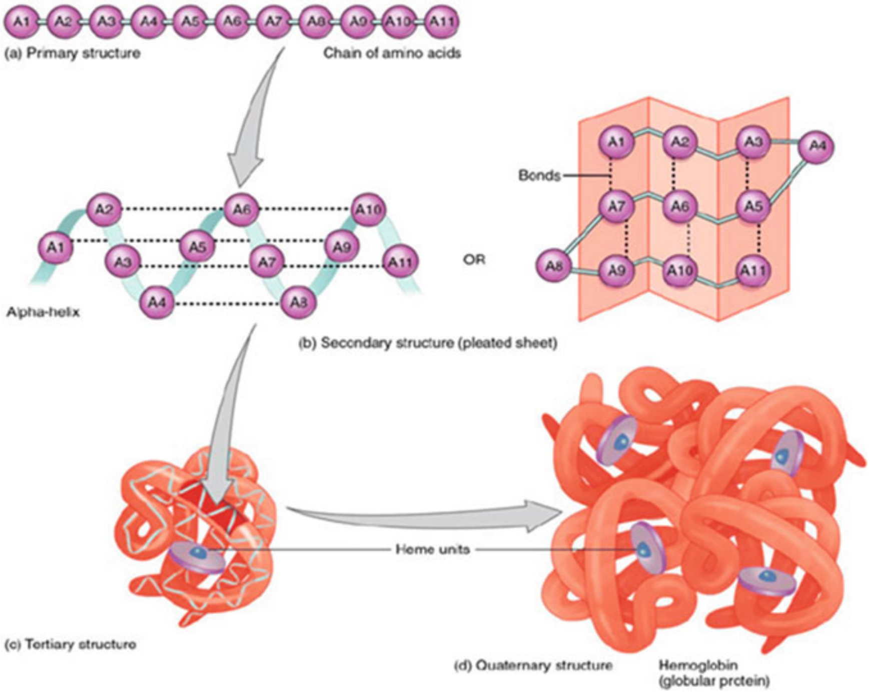


Fig. 3.1 Depicting primary, secondary, tertiary, and quaternary structures of hemoglobin. (Source: OpenStax College, http://commons.wikimedia.org/wiki/File:225_Peptide_Bond-01.jpg, cc by 3.0.)

Enzyme Commission (EC) number (Webb 1992), enzymes are classified into six different categories (Koolman and Roehm 2005). These proteins are biocatalysts and constitute different metabolic pathways in the cells.

In evolutionary related studies, the first impressions were that proteins having high sequence similarity reflect high homology between proteins of different organisms and thus have similar functions or have probably evolved from a common ancestor. Later, after the relationship between sequence and structure was revealed (Anfinsen and Haber 1961), it was established that if the sequences of proteins do not share similarity, their three-dimensional structures were highly similar (Chothia and Lesk 1986; Rost 1999; Sander and Schneider 1991). Such revelations led to the affirmation that it was after all the structure of the protein that plays an important role in determining its function. Proteins are dynamic molecules; they depict interdomain, secondary structure level movements (Biswas et al. 2017a, 2017b; Chaudhary et al. 2018; Deocaris et al. 2009; Jana et al. 2011; Putri et al. 2019) for their function. Homologous proteins display conservation in sequence, structure, and dynamics levels and subsequently demonstrate similarity in function space. However, there are subtle differences in all these levels across homologous proteins, which may lead to subtle changes in function as well.

Often, proteins from pathogens are specifically targeted using small molecule inhibitors in an attempt to design drugs against those pathogenic infections (Kumar et al. 2019; Okuhira et al. 2017; Peixoto et al. 2017). Inhibitors are designed in such a way that they do not target the homologous protein from the host. In this attempt, the subtle differences in sequence, structure, and dynamics space across the homologues are taken into account (Śledź and Caflisch 2018).

G-protein coupled receptors, more popularly known as GPCRs, form major part of the human membrane proteins and constitute more than 1% of the human genome. They comprise seven transmembrane helices that are conserved secondary structure across all GPCRs. A recent survey revealed that approximately 40% of available drugs have GPCRs as targets (Brink et al. 2004; Drews 2000; Klabunde and Hessler 2002). These drugs cover a variety of diseases ranging from diabetes, immune disorders, cancer, and cardiovascular disorders. Furthermore, it is known that specific conformational changes in GPCRs activate them to cause intracellular signals (Latorraca et al. 2017). Please refer (Lee et al. 2019) and (Lee et al. 2018) for further reading on GPCRs. Thus, in recent years, this family of proteins has been involved in various pharmacology studies (Wacker et al. 2017), especially in rational drug designing.

3.1.1 Rational Drug Designing

Till around half a century ago, drug discovery was set upon the basis of "trial and error" methods and the availability of basic materials. Ever since the three-dimensional structures of proteins have been made available, over the years, the basis of drug designing lies in studying the structure and function of the protein target. The aim is to design drugs that are "tailor-made" compounds with high specificity for the target protein, thereby specifically inhibiting the protein function to obtain desired therapeutic effects. This procedure is known as "*rational drug designing*." This strategy requires the following steps: (1) a receptor molecule/enzyme related to a disease, which is the target for the drug; (2) complete information on the structure, function, and dynamics of the protein target; and (3) designing potential and highly specificligands, which are suitable to bind to the target. The structure and dynamics of the protein may decipher different aspects of ligand binding—interacting residues and the types of interactions among their sidechains and the small molecules, electrostatic properties of the binding site, and residue level movements of the binding site (Guo et al. 2015; Naderi et al. 2019). These parameters are crucial in designing effective small molecule inhibitors. Current trends in rational drug designing include computer-aided drug designing (CADD) techniques as an integral part of the overall drug designing process.

An article, dating all the way back to 1986 (Hol 1986), states that ultimately it is the precise three-dimensional structure of a protein complexed with many "lead" molecules that provides the basis of drug design. The author states that taking into account even the slightest conformational changes of the protein will result in the small molecules/ligands binding differently to the protein. This will eventually give rise to active and effective "tailor-made" drugs, with a high degree of specificity. However, obtaining a comprehensive view of protein dynamics requires the three-dimensional structure of a protein as snapshots along the dynamic trajectory. Such structures, as in the case of adenylate kinase, along the reaction coordinate solved through X-ray crystallography, provide deep insights into the functionally important movements during catalysis (Jana et al. 2011). Even a single structure comes with the b-factor statistics of the atoms, which is a metric for atomic movements. Several biophysical techniques, such as small-angle X-ray scattering (SAXS) and wide-angle X-ray scattering (WAXS) (Lamba 2016; Walther et al. 2000), are able to probe different protein dynamics in solution. Protein structure determination has often been faced with the bottleneck of obtaining diffraction quality crystals. This has partially been overcome by the advent of cryo-electron microscopy, which does not require the crystalline state of the molecule for structure determination (Callaway 2015).

However, all these methods come with their limitations at different levels: (1) Proteins should be of high purity and desired quantity; however, at times, during the purification procedure, the protein expression becomes poor, leading to some expressed eukaryotic proteins not containing functionally important glycans. (2) Protein purification, quite often, results in some impurity being retained, which is very difficult to eradicate. (3) Highly dynamic regions of proteins often hinder the process of crystallization, and crystallization trials are attempted with the deletion mutants of protein samples, where those regions are deleted from the sequence. Therefore, the crystal structures of the deletion mutants may not provide the dynamic information of the full length protein. (4) Protein co-crystal structures with small molecules may not be determined in certain cases (as in thymidylate kinase from *Sulfolobus tokodaii*). Therefore, the specific interactions among residues and the small molecule may not be determined, and the relative movements of protein binding site residues with respect to the apo structure cannot be elucidated. (5) Expressing mammalian proteins in a bacterial system results in proteins that are devoid of glycosylation. However, glycans often play important roles in regulating protein functions in the eukaryotic system (Corfield 2017; Varki 2017). Structures without glycans may not provide full insight into protein dynamics.

CADD techniques may provide valuable insights about the system in the absence of experimental insights. Homology modeling or ab-initio protein structure determination provides predictive three-dimensional models in the absence of crystal structures. Thereafter, docking simulations can predict the possible ligand-bound protein structures. Long molecular dynamics simulations of both apo and co-structures provide possible trajectories of the functionally important movements of protein structures. However, problems regarding drug design still persist owing to the presence of experimental hurdles.

3.2 Role of Protein Structural Biology in Rational Drug Design

The first protein structure of myoglobin was solved using X-ray crystallography by Max Perutz and John Kendrew in 1958 (Kendrew et al. 1958), which won them the Chemistry Nobel Prize in 1962. Following this breakthrough, many other protein structures have been solved by X-ray crystallography, of which the structures of vitamin B12 and insulin, which won Dorothy Hodgkin the Nobel prize in chemistry in 1962, are worth mentioning.

Structural biology is the study of the biochemical and biophysical characteristics of protein structures after elucidating the three-dimensional structure. The three-dimensional structure of proteins is determined, mainly, using X-ray crystallography, amongst other techniques, and it is the atomic coordinates of the protein that shed light on the function of the protein. The detailed analysis of the three-dimensional structure of proteins facilitates the understanding of various structural and functional roles of the proteins. For instance, the catalysis process that happens at the enzyme active site, how or why two proteins interact with each other, and the molecular basis of cell signaling. Such analyses provide us with immense knowledge and open up avenues for further exploration of cellular machinery functions, such as the viral or bacterial disease cycles.

Further advancements in X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and electron microscopy (cryo-electronmicroscopy) techniques have led to the phenomenal quantity and quality of information that can be exploited to gain valuable insights into the structure, function, and dynamics of protein structures. This treasure of knowledge will pave the way for precise and specific drug designing. Therefore, since the determination of the first protein three-dimensional structure, its knowledge has become crucial for the development of drugs for protein targets.

As straightforward as it sounds, in reality, the information obtained from the crystal structure of proteins is very complicated, plainly because the atoms of the protein are not rigid/fixed but constantly “jiggle” and “wiggle.” Thus, a three-dimensional structure is not a single snapshot of the protein but rather an ensemble of different conformations. In other words, owing to the fact that protein molecules are in constant motion, the structure of a protein can adopt different conformations. This is the key ingredient to understand and study the protein conformational dynamics and its effect on protein function and binding to drug molecules (Eisenmesser et al. 2005; Frauenfelder et al. 1991; Zhuravleva and Gierasch 2015), and the drug Captopril is a successful example. Captopril is an angiotensin-converting enzyme (ACE) inhibitor, which is being used to treat hypertension or high blood pressure for decades. According to an early review article (Hol 1986), the success of this inhibitor was based on the knowledge of the three-dimensional structure of the proteins carboxypeptidase A (Rees et al. 1983) and thermolysin (Holmes and Matthews 1982).

3.3 Role of Protein Dynamics in Drug Design

Although enzymes are usually large globular protein molecules, only a part of it is involved in its catalytic functionality. The active site (catalytic site + binding site) of the enzyme is specific to its function as only a substrate with the particular shape will fit into it. The three-dimensional structure of an enzyme has an important role in determining the function of its active site, and any change in the amino acid sequence could result in a change of its three-dimensional structure and function.

The active site of a protein/enzyme is mostly a cavity in the protein where a substrate binds to activate a particular reaction (catalytic function). Active sites comprise amino acids that are important for that catalytic function and are usually highly conserved across protein families. These active site amino acids are responsible for binding to the substrate to catalyze the chemical reaction. Amino acid residues that are present on other areas of the protein also contribute to maintaining the function, features, and properties of the protein and active site (Furst et al. 2019). For instance, proteins fold and function via various interactions between the amino acids, and these interactions in evolutionary times lead to a correlation between residues. The residues that are situated around the active site of proteins are usually conserved for ligand binding and catalysis. Structural studies have shown considerable interdomain movements in the protein active site during the course of catalysis. For instance, studies on adenylate kinase (Jana et al. 2011) demonstrated how the protein attains "open" to "closed" conformations along the reaction coordinate. Combined with molecular dynamics simulations, it was also demonstrated how the protein active site attains an intermediate "half open–half closed" conformation in its dynamic trajectory (Adkar et al. 2011; Jana et al. 2011). Furthermore, several apo- and co-crystallographic structures demonstrated the role of "allosteric regulation" (described later in this chapter) in the function of Hsp70 (Stetz and Verkhivker 2017). Such studies enable us to identify the functionally important residues of the protein. For proteins from pathogens, this helps to identify residues to be targeted for highly specific inhibitor designing. Crystallographic structures of several sugar binding proteins, such as lectins, and eukaryotic protein structures solved with associated glycans demonstrate the molecular basis of cell to cell communication (Varki 2017).

Moreover, conformational changes due to the binding of substrates/ligands in these conserved sites are important factors for the study of information paths between the substrates and protein (Singh et al. 2015). They help in identifying the functionally important and correlated residues, which provide important insights into the mechanism of enzyme catalysis (Chaudhary et al. 2018). Mutation of the conserved residues during protein-protein interaction could have a detrimental effect on the function of the protein. Thus, the concept of coevolving residues could be used to study the compensatory mutation in the correlated residues, which can restore the specificity of the protein, and the effect becomes neutral when together or deleterious by itself (Adami 2004; Little and Chen 2009). Proteins undergo considerable conformational changes in the cellular environment. Sometimes these conformational changes are triggered by the binding of some small molecules, peptides, or proteins away from the site of interest. This is called allosteric regulation. In combination with the protein structure network analysis of the trajectory of protein dynamics, the functionally important residues both near the binding or active site of the protein and away from the site of interest can be determined (Chaudhary et al. 2018).

Active sites of proteins maintain a particular shape and size, which is suitable to bind to specific substrates. It is to be noted that there are other binding sites on the surface of the protein, which are shallower than the active site pocket. These binding sites facilitate the binding of the protein molecule to other macromolecules, such as protein-protein interactions and subunit associations. This is a very important step in protein dynamics related to binding with ligands and substrates.

The earliest notion of protein-ligand binding was first presented as a "*lock and key*" model (Fig. 3.2a) (Fischer 1894), where the active site of the protein is exactly the shape of the rigid ligand so that it fits perfectly like a key into a lock. However, when protein structures began to be solved by X-ray crystallography, as mentioned earlier, protein molecules were found to be in constant motion and not rigid objects. These movements can be at the atomic level, occurring in femtoseconds, or at the secondary and tertiary structure levels, ranging from seconds to hours. The latter refers to movements that occur during the event of protein folding and subunit association. The flexibility and movement of these proteins affect the dynamics of the active site, creating an ensemble of conformations and different cavity shapes. Therefore, the limitations of the "lock and key" model theory were clearly seen. Moreover, the "lock and key" model thrived during the era when protein three-dimensional structures were not yet known. Soon, a new theory was introduced by Koshland in 1958 (Koshland 1958), called the "*induced fit*" model (Fig. 3.2b). According to this model, the shape of the protein binding site/active site changes and adapts according to the ligand that binds to it. Also referred to as the "hand and glove" model (Śledź and Caflisch 2018), the specific ligand induces a conformational change in the active site such that it perfectly binds to the respective ligand. In layman's language, the hand is considered as the ligand and the glove as the protein's binding pocket. When the hand is put into the glove, the latter adjusts its shape such that it snugly fits over the former. Another model describes proteins possessing several different conformations, and the ligand selects the most suitable conformation and binds to it. This form of selective binding is called the "*conformational selection*" model (Fig. 3.2c). This was first observed when conformational changes in the protein binding/active site were noticed consequential to the allosteric binding of a ligand (Monod et al. 1965) (allosteric binding will be talked about later in this chapter). Moreover, the "conformational selection" model has been widely observed and studied in various research (Changeux and Edelstein 2011; Vogt and Di Cera 2012). Coming back to protein dynamics, it is to be noted that both intrinsic protein flexibility and the active/binding site conformational changes are contributory factors to the protein-ligand binding process. Several studies have shown that the magnitude of these movements affects the "time scale" on which they occur, which in turn influences the choice of binding model (induced fit or conformational

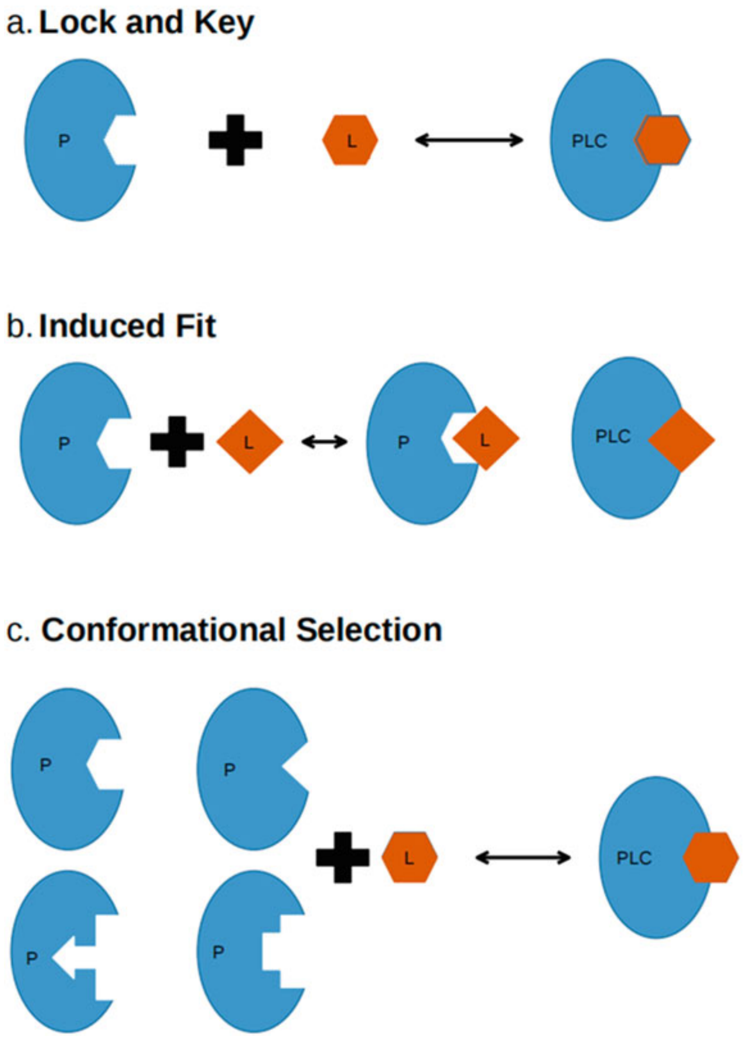


Fig. 3.2 Diagrammatic representation of the three types of binding models. (**a**) Represents the "lock and key" model, (**b**) represents the "induced fit" model, and (**c**) represents the "conformational selection" model. Here, *P* Protein, *L* Ligand, *PL* Protein-Ligand complex

selection) during the protein-ligand binding process (Gianni et al. 2014; Hammes et al. 2009; Zhou 2010). Readers are urged to read this excellent article on "Protein binding pocket dynamics"(Stank et al. 2016) for further understanding. The article clearly mentions that a binding pocket should possess certain features to suit the set of ligands that is capable of binding to it. *Firstly*, the volume of the cavity should be proportionate to the size of the ligand, so as to be able to house it. *Secondly*, the physiochemical properties of the active site should be suitable for the potential ligands to bind. These properties include electrostatics, hydrophobicity, and other interactions of the active site residues (Henrich et al. 2010; Schreyer and Blundell 2009). At this juncture, it would be appropriate to mention that each distinct conformation is associated with a different energy; the lower the energy value, the higher the stability, thereby making the conformations with lower energy values more favourable for binding. Proteins transition between different conformations and the chances of transitions highly depend on the energy variation between the two conformational states; hence, the lesser the variation, the more probable is the transition. Another important factor to be considered in the protein-ligand binding process is the "*druggability*" of the active/binding site (Hopkins and Groom 2002). Druggability may be defined as the ability of the protein's active site to bind to a drug, which also takes into account its physiochemical properties. Having realized the utmost criticality and underlying role of dynamics in protein function, a research group (Hensen et al. 2012) proposed the concept of "*protein dynasome*" based on the hypothesis that proteins with similar functions display comparable dynamic properties and thus possess a common "*dynamic fingerprint*." Interested readers are encouraged to refer to the mentioned article for further reading.

An excellent article (Csermely et al. 2010) elucidating the different binding models highlights a fourth model, which is "*extended conformational selection.*" Several studies (Grunberg et al. 2004; Wlodarski and Zagrovic 2009) have demonstrated that after following the "conformational selection" mode of binding, a "*conformational adjustment*" was observed in the final protein-ligand complex. "Conformational adjustment" is nothing but the "induced fit" action, which takes place after the ligand selected the specific representation of the same. Ubiquitin's binding site is a perfect example of this mode of binding (Wlodarski and Zagrovic 2009), wherein the side chains of the residues in the binding site tend to adopt the "induced fit" model, and the rest of the protein settles for the "conformational selection" type. Another classic study (Silva et al. 2011) also demonstrates how the LAO (Lysine-, Arginine-, Ornithine-binding) protein first adopts the "conformational selection" approach and then finishes off with the "induced fit" model during ligand binding. This study includes computational methods that are used in drug discovery and also talks about the effect of allostery in binding.

Allosteric binding is an important factor in the dynamics of the protein-ligand binding process and is especially significant in drug designing (Nussinov and Tsai 2015; Stank et al. 2016). Allosteric binding refers to the binding of a ligand or a macromolecule to the receptor protein at a site other than the active/binding site. Allosteric binding contributes to the dynamics of the protein's active/binding site causing further conformational changes, which ultimately influence the type of binding: induced fit, conformational selection, or a combination of both. How does this work? Well, when a ligand binds to an allosteric site on a protein, it causes a conformational change in the active site such that only a specific inhibitor or an activator molecule will bind, such as the ligand-induced allosteric effects on myosin V (Coureux et al. 2004) and Hsp70 (General et al. 2014). Similarly, during protein-protein interaction of either the same or discrete macromolecules, the movement caused in the entire protein complex affects the dynamics of active/binding site, thereby making it a target or receptor for specific ligands. Please refer (Nussinov and

Tsai 2015) for further reading. Protein (i.e., the target) flexibility is one of the main factors that affect receptor-drug binding kinetics. For further reading, please see reviews by Pan et al., (Pan et al. 2013) and Klebe et al., (Klebe 2015).

3.4 Role of Computer-Aided Drug Design (CADD)

The last few decades have seen a sudden stir in pharmaceutical and life science companies where everyone is talking about "drug designing," wanting to make their mark in the field, and why not? The advent of computational methodologies in the drug discovery pipeline has accelerated the process of the designing of drugs (De Vivo 2011; Jorgensen 2004) and optimizing drug candidates. Methods such as molecular dynamics (MD) and molecular docking are being used very frequently and are becoming indispensable CADD tools in the drug discovery process. This can be attributed to the fact that their basis for analysis is protein dynamics and entropic effects. This provides drug designers and researchers with a more accurate scenario in terms of protein-ligand binding (De Vivo et al. 2016). The drugs indinavir, ritonavir, saquinavir (Van Drie 2007), and tirofiban (Hartman et al. 1992), to name a few, have used CADD as an integral part of their development process, and many more are continuing to (Hillisch et al. 2015; Muegge et al. 2017). As mentioned earlier, protein dynamics is one of the key factors in deciding the type of ligands that are to be used in binding to a particular protein in structure-based drug designing. Keeping this in mind, a brief outline of the two most widely used computational methods in CADD, "*molecular docking*" and "*molecular dynamics simulations,*" is presented below.

3.4.1 Molecular Docking

Virtual Screening, a rational drug discovery approach, can be categorized into two methods: ligand-based and structure-based. The ligand-based method is employed when the ligand molecules are known, but the structure of the target protein is unknown; hence, methods like QSAR (Quantitative Structure Activity Relationship) and pharmacophore modeling are used. Whereas, in structure-based drug design, the structure of the target protein is known, and the molecular docking method is adopted (Kuntz et al. 1982). Structure-based virtual screening aims to identify new potential ligands for the specific protein target, and many such projects have been accomplished. The *molecular docking* approach is used to identify the interaction between a small molecule and protein, which in turn gives an insight into the nature of the small molecules in the binding site of the target protein (McConkey et al. 2002). During the process of docking, prior knowledge of the binding site adds to the docking efficiency. In the event that the binding site is unknown, the information can be obtained either by comparing the target protein with a class of proteins having similar structure and function or by using prediction tools and servers, such as GRID (Goodford 1985; Kastenholz et al. 2000), POCKET (Levitt and Banaszak 1992), and MetaPocket (Huang 2009). In some cases, docking is done without any knowledge of the binding site and is referred to as "blind docking." The docking process involves two inter-related steps: (1) sampling method: prediction of conformations of the ligand in the active site and (2) scoring schemes: evaluating the binding affinity via the scoring function (Meng et al. 2011). Both these methods are briefly explained below.

Sampling Algorithms

1. Matching algorithms—follow the principle of matching the shape and properties of the ligand to that of the active site (Brint and Willett 1987; Fischer et al. 1993; Norel et al. 1994).
2. Incremental construction methods—first breaks up the ligand into fragments at their rotatable bonds. The fragment of highest binding significance is first docked into the binding site followed by the addition of the remaining fragments, piece by piece (Des Jarlais et al. 1986; Leach and Kuntz 1992; Rarey et al. 1996).
3. Stochastic methods—work by modifying the conformation of the ligand or group of ligands, to identify the conformational space. The Monte Carlo (Goodsell et al. 1993; Hart and Read 1992) and Genetic algorithms (Jones et al. 1997; Oshiro et al. 1995) are examples of stochastic methods.

Scoring Functions

Scoring function estimates the binding affinity between the protein and ligand, thereby separating the correct orientations from incorrect ones. There are three types of scoring functions:

1. *Force-field based*: evaluating binding energy based on non-bonded interactions (electrostatic and Van der Waals).
2. *Empirical scoring functions*: Binding energy includes various energy components (ionic interactions, hydrophobic effects, hydrogen bonds, and binding entropy), and each contributes to the final score.
3. *Knowledge-based scoring function*: the distance between the ligand and protein is obtained using the statistical analysis of protein-ligand complexes.

Thus, molecular docking is a method that is used to predict the binding of a protein with potential ligands. Docking studies provide the best binding orientation of the ligand to the protein's binding pocket or active site. Docking can adopt two strategies: structure-based and fragment-based. Structure-based docking is more popular than fragment-based docking because the former uses the entire protein structure in the process, while the latter uses only the binding pocket region. Since protein flexibility and dynamics play an important role in protein-ligand binding, it is rather prudent to adopt the structure-based method. Furthermore, molecular docking methods can either cater for the various conformations that the protein structure might assume during the binding process (i.e., induced-fit mode) (Sherman et al. 2006) or use an existing ensemble of conformations (Erickson et al. 2004). This is also known as "*flexible docking,*" which is a better strategy to employ rather than "*rigid docking.*" The advantage of flexible docking is that the binding orientations with the least binding energies can be preferred rather than being given only one binding pose in rigid docking. A variation to the flexible docking mode is where most of the protein is frozen, and only specific binding site amino acid residues are left to "jiggle" and "wiggle" (Kuhn et al. 2016). Some of the most popular tools for docking are Dock (Allen et al. 2015), GOLD (Jones et al. 1997), AutoDock (Morris et al. 2009), Haddock (de Vries et al. 2010), and rDock (Ruiz-Carmona et al. 2014).

3.4.2 Molecular Dynamics Simulations

Molecular Dynamics (MD) (Cornell et al. 1995; Brooks et al. 1983; Weiner et al. 1984) is a computational method for studying the motion of atoms and molecules and their interactions based on Newton's physics. Though experimental techniques like X-ray crystallography and NMR provide information about the various conformations of protein structures, it is not sufficient in drug discovery, where the knowledge of all possible conformations and motions is necessary for proper ligand selection or specificity. This is where computational methods like MD step in to make up for the existing lacuna.

The very first step is to develop a computational model of the molecular system from experimental data (NMR, Crystallographic data, Homology modeling), followed by the estimation of forces on each of the system atoms. The energy parameters are made to fit quantum mechanical calculations and experimental data to mimic the behavior of real molecules in motion (Durrant and McCammon 2011). Together, these parameters describe the contribution of various atomic forces that influence MD and are referred to as "*force field.*" AMBER (Cornell et al. 1995), CHARMM (Weiner et al. 1984), and GROMOS (Christen et al. 2005) are some of the common force fields used in MD simulations. After the calculations of the forces acting on each of these system atoms, based on Newton's law of motion, the positions of these atoms are shifted. In the end, the simulation time is incremented by one or two quadrillionths of a second, and this process is iterated millions of times. This is the reason why MD simulations require to be run on supercomputers or clusters. Some of the popular MD software are AMBER (Case et al. 2005), CHARMM (Weiner et al. 1984), NAMD (Kalé et al. 1999; Phillips et al. 2005), and GROMACS (Pronk et al. 2013).

Since MD simulations cater for the flexibility and dynamics of both the protein and ligand, it is often used to predict the near accurate conformations of the ligand binding to the active site. Furthermore, in several cases, MD simulations have predicted novel binding sites in a protein structure. One of the success stories based on this is the discovery of a new binding pocket adjacent to the known active site of HIV integrase, which led to the designing of new inhibitors for it. Spearheaded by Merck & Co., this project led to the development of the drug Raltegravir (Summa et al. 2008).

In conclusion, when multiple conformations of both the ligand and protein are considered during docking calculations, apart from the chance predictions of new binding sites, the accuracy and efficacy of the interactions between the ligand and the amino acid residues may provide valuable insights about the kinetics of the process. Moreover, several studies have undertaken the evaluation of comparing and validating the MD data with experimental results (van Gunsteren et al. 2008), and a substantial number of such cases are seen to be highly comparable (LaConte et al. 2002; Markwick et al. 2010; Peter et al. 2003; Showalter and Bruschweiler 2007). In view of all these beneficial properties and the constant upgrade of computational technologies, contrary to the old notion, molecular dynamics along with other CADD techniques is becoming a standard feature in rational drug designing (Borhani and Shaw 2012; Durrant and McCammon 2011). For the benefit of the readers, a schematic, self-explanatory flow chart has been provided, which explains the general process of CADD (Fig. 3.3).

3.5 Case Study: Computational Screening of Potential Lead Molecules and Experimental Validation of Their Anti-Influenza Effect

This case study is a published work (Liu et al. 2016), which has been used as an example to demonstrate how CADD is an integral part of the drug designing pipeline.

Influenza viruses, of the virus family Orthomyxoviridae, are one of the most common causes of human respiratory infections. Influenza A virus eminently infects the mucosa of the upper respiratory tract and might induce respiratory diseases (Taubenberger and Morens 2008). Influenza viruses (A, B, C) are enveloped negative-strand RNA viruses, consisting of seven to eight gene segments (Lamb and Krug 1996), but all of them differ in host range and pathogenicity. All influenza A viruses contain eight genes encoding for 12 different proteins (Gao et al. 2012). The influenza A virus is classified into distinct subtypes based on the two proteins on the surface of the virus: hemagglutinin (H) and neuraminidase (NA) (Smith et al. 2009). Previous studies have reported the crucial role of NA for the replication and propagation of the influenza A virus; therefore, molecules inhibiting NA can act as a potential anti-influenza A drug (Michiels et al. 2013; Nayak and Jabbar 1989).

At present, there are three anti-influenza drugs: amantadine, oseltamivir, and zanamivir (Pica and Palese 2013), but owing to the emerging drug-resistant strains of the influenza A virus, it has become critical to develop a new anti-influenza A drug. This study demonstrates an *in silico* approach to identify potential anti-

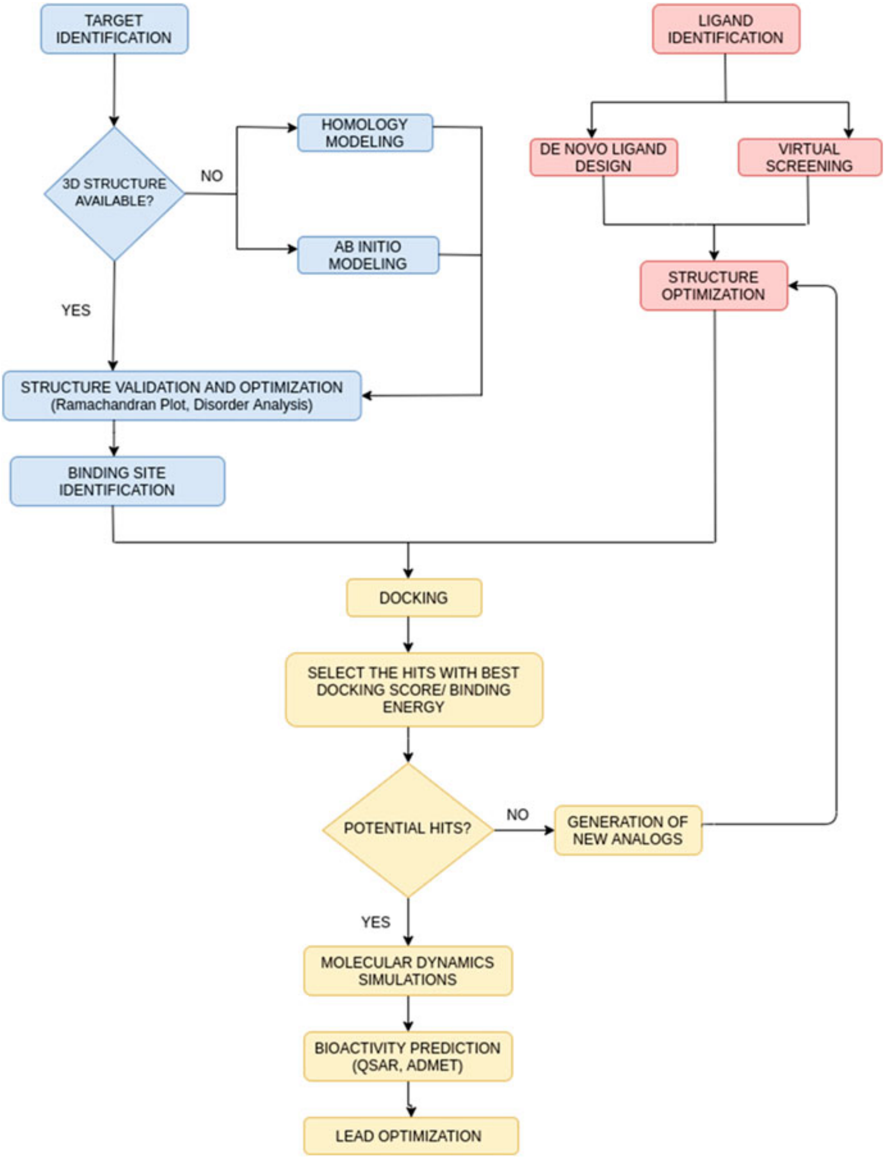

TARGET IDENTIFICATION
3D STRUCTURE AVAILABLE?
NO
HOMOLOGY MODELING
AB INITIO MODELING
YES
STRUCTURE VALIDATION AND OPTIMIZATION (Ramachandran Plot, Disorder Analysis)
BINDING SITE IDENTIFICATION
LIGAND IDENTIFICATION
DE NOVO LIGAND DESIGN
VIRTUAL SCREENING
STRUCTURE OPTIMIZATION
DOCKING
SELECT THE HITS WITH BEST DOCKING SCORE/ BINDING ENERGY
POTENTIAL HITS?
NO
GENERATION OF NEW ANALOGS
YES
MOLECULAR DYNAMICS SIMULATIONS
BIOACTIVITY PREDICTION (QSAR, ADMET)
LEAD OPTIMIZATION

Fig. 3.3 A schematic representation of a general Computer-Aided Drug Designing process

influenza molecules, followed by *in vitro* and *in vivo* experiments to validate the anti-influenza effects of the screened molecules (Liu et al. 2016).

In the current scenario, the most efficient approach to start with drug development is computational screening, *i.e.*, the *in silico* approach. In this work, the NA of A/PR/8/34 H1N1 has been taken as the target molecule. The NA structure of A/PR/8/34

H1N1 was computationally modeled using A/Brevig Mission/1/1918 H1N1 as a template, by the SWISS-MODEL Workspace. The sequence alignment, employing Clustal Omega, showed 93.25% identity for the two sequences, and the residues crucial for NA activity were found to be conserved. The structure comparison performed, using PyMOL, depicted close identity between the two structures, as well as indicated the structurally conserved site for NA activity.

Next, the identification of potential anti-influenza small molecules was done. Through literature survey, plants known to have anti-influenza effects were curated, and fifteen bioactive components were gathered for further process. From the ZINC database, the structures of small molecules were retrieved, and using the AutoDock software, molecular docking was performed with the modeled NA structure as the receptor. The binding energies calculated for the fifteen small molecules and zanamvir depicted that the binding energies of chlorogenic acid and quercetin were comparable with that of zanamvir. Further analysis of docking energies and calculation of the inhibition constant indicated the significant NA inhibition ability of chlorogenic acid and quercetin. Furthermore, the FAF-Drugs3 software was employed for the chemical informatics analysis of Zanamvir, chlorogenic acid, and quercetin. The results were promising as they indicated high drug-likeness of the two small molecules. The two small molecules with high binding potential were then further analyzed *in vitro* and *in vivo* to validate their NA inhibition ability.

3.6 Summary

It is evident that the knowledge of the three-dimensional structure and function of the protein is indispensable for the successful prediction and selection of target-specific drugs. This is where experimental techniques like X-ray crystallography, NMR, and Cryo-electron microscopy share the throne. However, experimental methods are usually very time consuming and cost intensive, and it is not certain that they will result in the desired output. Thus, computational methods were sought after as a means of reducing the time taken, especially in the case of virtual screening for specific and suitable small molecules. Furthermore, computational methods cater for and take into account the flexibility and dynamics of the target protein, as well as those of the binding pocket and ligand. Molecular docking provides a user-friendly and convenient yet reliable methodology for narrowing down to a set of ligands that could bind specifically to the target protein, through various orientations. MD simulations approach is gaining increasing popularity in drug designing, especially owing to its ability to predict new binding pockets and take into account the various movements in the protein molecule, even at the atomic level. Several successful drugs, as mentioned above, are the consequence of this combined approach in drug discovery, which is Computer Aided Drug Designing. As new pathogens are emerging and evolving with higher virulence, the need of the hour is to develop more robust computational tools to aid in the drug designing process.

Acknowledgements The authors gratefully acknowledge the facilities offered by the centre of excellence in structural biology and bio-computing and Department of Computational and Data Sciences. One of the authors (KS) thanks the Department of Biotechnology for a possible financial support in the form of a research grant.

Authors' Information KS is a Professor at the Department of Computational and Data Sciences (CDS), Indian Institute of Science (IISc), Bangalore. DM and NB are Project Assistants at CDS, IISc. AB has completed his PhD at IISc under KS and is currently a post-doctoral fellow at the National Centre for Biological Sciences (NCBS), Bangalore.

Competing Interests None.

References

Adami, C. (2004). Information theory in molecular biology. *Physics of Life Reviews, 1*, 3–22.

Adkar, B. V., Jana, B., & Bagchi, B. (2011). Role of water in the enzymatic catalysis: Study of ATP + AMP → 2ADP conversion by adenylate kinase. *The Journal of Physical Chemistry. A, 115*, 3691–3697.

Allen, W. J., Balius, T. E., Mukherjee, S., Brozell, S. R., Moustakas, D. T., Lang, P. T., et al. (2015). DOCK 6: Impact of new features and current docking performance. *Journal of Computational Chemistry, 36*, 1132–1156.

Anfinsen, C. B., & Haber, E. (1961). Studies on the reduction and re-formation of protein disulfide bonds. *The Journal of Biological Chemistry, 236*, 1361–1363.

Biswas, A., Shukla, A., Vijayan, R. S. K., Jeyakanthan, J., & Sekar, K. (2017a). Crystal structures of an archaeal Thymidylate kinase from Sulfolobus tokodaii provide insights into the role of a conserved active site arginine residue. *Journal of Structural Biology, 197*, 236–249.

Biswas, A., Shukla, A., Chaudhary, S. K., Santhosh, R., Jeyakanthan, J., & Sekar, K. (2017b). Structural studies of a hyperthermophilic thymidylate kinase enzyme reveal conformational substates along the reaction coordinate. *The FEBS Journal, 284*, 2527–2544.

Borhani, D. W., & Shaw, D. E. (2012). The future of molecular dynamics simulations in drug discovery. *Journal of Computer-Aided Molecular Design, 26*, 15–26.

Brink, C. B., Harvey, B. H., Bodenstein, J., Venter, D. P., & Oliver, D. W. (2004). Recent advances in drug action and therapeutics: Relevance of novel concepts in G-protein-coupled receptor and signal transduction pharmacology. *British Journal of Clinical Pharmacology, 57*, 373–387.

Brint, A. T., & Willett, P. (1987). Algorithms for the identification of three-dimensional maximal common substructures. *Journal of Chemical Information and Computer Sciences, 27*, 152–158.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., &Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry, 4*, 187–217.

Callaway, E. (2015). The revolution will not be crystallized: A new method sweeps through structural biology-move over X-ray crystallography. Cryo-electron microscopy is kicking up a storm by revealing the hidden machinery of the cell. *Nature, 525*, 172.

Case, D. A., Cheatham III, T. E., Darden, T., Gohlke, H., Luo, R., Merz Jr., K. M., et al. (2005). The Amber biomolecular simulation programs. *Journal of Computational Chemistry, 26*, 1668–1688.

Changeux, J. P., & Edelstein, S. (2011). Conformational selection or induced fit? 50 years of debate resolved. *F1000 Biol Rep, 3*, 1.

Chaudhary, S. K., Jeyakanthan, J., & Sekar, K. (2018). Structural and functional roles of dynamically correlated residues in Thymidylate kinase. *Acta Cryst D, 74*, 341–354.

Chothia, C., & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO Journal, 5*, 823–826.

Christen, M., Hünenberger, P. H., Bakowies, D., Baron, R., Bürgi, R., Geerke, D. P., et al. (2005). The GROMOS software for biomolecular simulation: GROMOS05. *Journal of Computational Chemistry, 26*, 1719–1751.

Corfield, A. (2017). Eukaryotic protein glycosylation: A primer for histochemists and cell biologists. *Histochemistry and Cell Biology, 147*, 119–147.

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., et al. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society, 117*, 5179–5197.

Coureux, P. D., Sweeney, H. L., & Houdusse, A. (2004). Three myosin V structures delineate essential features of chemo-mechanical transduction. *The EMBO Journal, 23*, 4527–4537.

Csermely, P., Palotai, R., & Nussinov, R. (2010). Induced fit, conformational selection and independent dynamic segments: An extended view of binding events. *Trends in Biochemical Sciences, 35*, 539–546.

De Vivo, M. (2011). Bridging quantum mechanics and structure-based drug design. *Frontiers in Bioscience, 16*, 1619–1633.

De Vivo, M., Masetti, M., Bottegoni, G., & Cavalli, A. (2016). Role of molecular dynamics and related methods in drug discovery. *Journal of Medicinal Chemistry, 59*, 4035–4061.

de Vries, S. J., van Dijk, M., & Bonvin, A. M. J. J. (2010). The HADDOCK web server for data-driven biomolecular docking. *Nature Protocols, 5*, 883–897.

Deocaris, C. C., Kaul, S. C., & Wadhwa, R. (2009). The versatile stress protein mortalin as a chaperone therapeutic agent. *Protein and Peptide Letters, 16*, 517–529.

Des Jarlais, R. L., Sheridan, R. P., Dixon, J. S., Kuntz, I. D., & Venkataraghavan, R. (1986). Docking flexible ligands to macromolecular receptors by molecular shape. *Journal of Medicinal Chemistry, 29*, 2149–2153.

Drews, J. (2000). Drug discovery: A historical perspective. *Science, 287*, 1960–1964.

Durrant, J. D., & McCammon, J. A. (2011). Molecular dynamics simulations and drug discovery. *BMC Biology, 9*, 71.

Eisenmesser, E. Z., Millet, O., Labeikovsky, W., Korzhnev, D. M., Wolf-Watz, M., Bosco, D. A., et al. (2005). Intrinsic dynamics of an enzyme underlies catalysis. *Nature, 438*, 117–121.

Erickson, J. A., Jalaie, M., Robertson, D. H., Lewis, R. A., & Vieth, M. (2004). Lessons in molecular recognition: The effects of ligand and protein flexibility on molecular docking accuracy. *Journal of Medicinal Chemistry, 47*, 45–55.

Fischer, D., Norel, R., Wolfson, H., & Nussinov, R. (1993). Surface motifs by a computer vision technique: Searches, detection, and implications for protein-ligand recognition. *Proteins, 16*, 278–292.

Fischer, E. (1894). Einfluss der Configuration auf die Wirkung der Enzyme. II. *Berichte der Deutschen Chemischen Gesellschaft, 27*, 3479–3483.

Frauenfelder, H., Sligar, S. G., & Wolynes, P. G. (1991). The energy landscapes and motions of proteins. *Science, 254*, 1598–1603.

Furst, M. J., Fiorentini, F., & Fraaije, M. W. (2019). Beyond active site residues: Overall structural dynamics control catalysis in flavin-containing and heme-containing monooxygenases. *Current Opinion in Structural Biology, 59*, 29–37.

Gao, Q., Chou, Y-Y., Doğanay, S., Vafabakhsh, R., Ha, T., & Palese, P. (2012). The influenza a virus PB2, PA, NP, and M segments play a pivotal role during genome packaging. *Journal of Virology, 86*, 7043–7051.

General, I. J., Liu, Y., Blackburn, M. E., Mao, W., Gierasch, L. M., & Bahar, I. (2014). ATPase subdomain IA is a mediator of Interdomain Allostery in Hsp 70 molecular chaperones. *PLoS Computational Biology, 10*, e1003624.

Geronikaki, A. (2019). Trends in enzyme inhibition and activation in drug design-part II. *Current Topics in Medicinal Chemistry, 19*, 317–318.

Gianni, S., Dogan, J., & Jemth, P. (2014). Distinguishing induced fit from conformational selection. *Biophysical Chemistry, 189*, 33–39.

Goodford, P. J. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of Medicinal Chemistry, 28*, 849–857.

Goodsell, D. S., Lauble, H., Stout, C. D., & Olson, A. J. (1993). Automated docking in crystallography: Analysis of the substrates of aconitase. *Proteins, 17*, 1–10.

Grunberg, R., Leckner, J., & Nilges, M. (2004). Complementarity of structure ensembles in protein-protein binding. *Structure, 12*, 2125–2136.

Guo, Z., Li, B., Cheng, L-T., Zhou, S., McCammon, J. A., & Che, J. (2015). Identification of protein−ligand binding sites by the level-SetVariational implicit-solvent approach. *Journal of Chemical Theory and Computation, 11*, 753–765.

Hammes, G. G., Chang, Y. C., & Oas, T. G. (2009). Conformational selection or induced fit: A flux description of reaction mechanism. *Proceedings of the National Academy of Sciences of the United States of America, 106*, 13737–13741.

Hart, T. N., & Read, R. J. (1992). A multiple-start Monte Carlo docking method. *Proteins, 13*, 206–222.

Hartman, G. D., Egbertson, M. S., Halczenko, W., Laswell, W. L., Duggan, M. E., Smith, R. L., et al. (1992). Non-peptide fibrinogen receptor antagonists. 1. Discovery and design of exosite inhibitors. *Journal of Medicinal Chemistry, 35*, 4640–4642.

Hasan, F., Shah, A. A., Javed, S., & Hameed, A. (2010). Enzymes used in detergents: Lipases. *African Journal of Biotechnology, 9*, 4836–4844.

Henrich, S., Salo-Ahen, O. M. H., Huang, B., Rippmann, F. F., Cruciani, G., & Wade, R. C. (2010). Computational approaches to identifying and characterizing protein binding sites for ligand design. *Journal of Molecular Recognition, 23*, 209–219.

Hensen, U., Meyer, T., Haas, J., Rex, R., Vriend, G., & Grubmüller, H. (2012). Exploring protein dynamics space: The dynasome as the missing link between protein structure and function. *PLoS One, 7*, 11.

Hillisch, A., Heinrich, N., & Wild, H. (2015). Computational chemistry in the pharmaceutical industry: From childhood to adolescence. *Chem Med Chem, 10*, 1958–1962.

Hol, W. G. J. (1986). Protein crystallography and computer graphics—Toward rational drug design. *Angewandte Chemie International Edition in English, 25*, 767–778.

Holmes, M. A., & Matthews, B. W. (1982). Structure of thermolysin refined at 1.6 Å resolution. *Journal of Molecular Biology, 160*, 623–639.

Hopkins, A. L., & Groom, C. R. (2002). The druggable genome. *Nature Reviews. Drug Discovery, 1*(9), 727–730. https://doi.org/10.1038/nrd892

Huang, B. (2009). MetaPocket: A meta approach to improve protein ligand binding site prediction. *OMICS, 13*, 325–330.

Imai, I. (1999). The haemoglobin enzyme. *Nature, 401*, 437–439.

Jana, B., Adkar, B. V., Biswas, R., & Bagchi, B. (2011). Dynamic coupling between the LID and NMP domain motions in the catalytic conversion of ATP and AMP to ADP by adenylate kinase. *The Journal of Chemical Physics, 134*, 035101.

Jones, G., Willett, P., Glen, R. C., Leach, A. R., & Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology, 267*, 727–748.

Jorgensen, W. L. (2004). The many roles of computation in drug discovery. *Science, 303*, 1813–1818.

Kalé, L., Skeel, R., Bhandarkar, M., Brunner, R., Gursoy, A., Krawetz, N., et al. (1999). NAMD2: Greater scalability for parallel molecular dynamics. *Journal of Computational Physics, 151*, 283–312.

Kastenholz, M. A., Pastor, M., Cruciani, G., Haaksma, E. E., & Fox, T. (2000). GRID/CPCA: A new computational tool to design selective ligands. *Journal of Medicinal Chemistry, 43*, 3033–3044.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., & Phillips, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature, 181*, 662–666.

Klabunde, T., & Hessler, G. (2002). Drug design strategies for targeting G-protein-coupled receptors. *Chembiochem, 3*, 928–944.

Klebe, G. (2015). The use of thermodynamic and kinetic data in drug discovery: Decisive insight or increasing the puzzlement? *Chem Med Chem, 10*, 229–231.

Koolman, J., & Roehm, K. H. (2005). *Color atlas of biochemistry* (2nd ed.). Stuttgart: Thieme.

Koshland, D. E. (1958). Application of a theory of enzyme specificity to protein synthesis. *Proceedings of the National Academy of Sciences, 44*, 98–104.

Kuhn, B., Guba, W., Hert, J., Banner, D., Bissantz, C., Ceccarelli, S., et al. (2016). A real-world perspective on molecular design. *Journal of Medicinal Chemistry, 59*, 4087–4102.

Kumar, A., Ellermann, M., & Sperandio, V. (2019). Taming the beast: Interplay between gut small molecules and enteric pathogens. *Infection and Immunity, 87*, e00131–e00119.

Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., & Ferrin, T. E. (1982). A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology, 161*, 269–288.

LaConte, L. E., Voelz, V., Nelson, W., Enz, M., & Thomas, D. D. (2002). Molecular dynamics simulation of site-directed spin labeling: Experimental validation in muscle fibers. *Biophysical Journal, 83*, 1854–1866.

Lamb, R. A., & Krug, R. M. (1996). Orthomyxoviridae: The viruses and their replication. In D. M. Knipe, P. M. Howley, & B. N. Fields (Eds.), *Fields virology*. Philadelphia: Lippincott-Raven Press.

Lamba, D. (2016). Wide-angle X-ray scattering (WAXS). In E. Drioli & L. Giorno (Eds.), *Encyclopedia of membranes*. Berlin: Springer.

Latorraca, N. R., Venkatakrishnan, A. J., & Dror, R. O. (2017). GPCR dynamics: Structures in motion. *Chemical Reviews, 117*, 139–155.

Leach, A. R., & Kuntz, I. D. (1992). Conformational analysis of flexible ligands in macromolecular receptor sites. *Journal of Computational Chemistry, 13*, 730–748.

Lee, Y., Basith, S., & Choi, S. (2018). Recent advances in structure-based drug design targeting class A G protein-coupled receptors utilizing crystal structures and computational simulations. *Journal of Medicinal Chemistry, 61*, 1–46.

Lee, Y., Lazim, R., Macalino, S. J. Y., & Choi, S. (2019). Importance of protein dynamics in the structure-based drug discovery of class a G protein-coupled receptors (GPCRs). *Current Opinion in Structural Biology, 55*, 147–153.

Levitt, D. G., & Banaszak, L. J. (1992). POCKET: A computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *Journal of Molecular Graphics, 10*, 229–234.

Linderstrom-Lang, K. U. (1952). *Proteins and enzymes: Lane medical lectures* (p. 1951). Oxford University Press: Stanford University Press.

Little, D. Y., & Chen, L. (2009). Identification of coevolving residues and coevolution potentials emphasizing structure, bond formation and catalytic coordination in protein evolution. *PLoS One, 4*, e4762.

Liu, Z., et al. (2016). Computational screen and experimental validation of anti-influenza effects of quercetin and chlorogenic acid from traditional Chinese medicine. *Scientific Reports, 6*, 19095.

Markwick, P. R. L., Cervantes, C. F., Abel, B. L., Komives, E. A., Blackledge, M., & McCammon, J. A. (2010). Enhanced conformational space sampling improves the prediction of chemical shifts in proteins. *Journal of the American Chemical Society, 132*, 1220–1221.

Marques, S. M., Daniel, L., Buryska, T., Prokop, Z., Brezovsky, J., & Damborsky, J. (2017). Enzyme tunnels and gates as relevant targets in drug design. *Medicinal Research Reviews, 37*, 1095–1139.

McConkey, B. J., Sobolev, V., & Edelman, M. (2002). The performance of current methods in ligand–protein docking. *Current Science, 83*, 845–856.

Meng, X-Y., Zhang, H-X., Mezei, M., & Cui, M. (2011). Molecular docking: A powerful approach for structure-based drug discovery. *Current Computer-Aided Drug Design, 7*, 146–157.

Michiels, B., Puyenbroeck, K. V., Verhoeven, V., Vermeire, E., & Coenen, S. (2013). The value of neuraminidase inhibitors for the prevention and treatment of seasonal influenza: A systematic review of systematic reviews. *PLoS One, 8*, e60348.

Monod, J., Wyman, J., & Changeux, J. P. (1965). On the nature of allosteric transitions: A plausible model. *Journal of Molecular Biology, 12*, 88–118.

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., et al. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry, 30*, 2785–2791.

Muegge, I., Bergner, A., & Kriegl, J. M. (2017). Computer-aided drug design at Boehringer Ingelheim. *Journal of Computer-Aided Molecular Design, 31*, 275–285.

Naderi, M., Lemoine, J. M., Govindaraj, R. G., Kana, O. Z., Feinstein, W. P., & Brylinski, M. (2019). Binding site matching in rational drug design: Algorithms and applications. *Briefings in Bioinformatics, 20*, 2167–2184.

Nayak, D. P., & Jabbar, M. A. (1989). Structural domains and organizational conformation involved in the sorting and transport of influenza virus transmembrane proteins. *Annual Review of Microbiology, 43*, 465–499.

Norel, R., Fischer, D., Wolfson, H. J., & Nussinov, R. (1994). Molecular surface recognition by a computer vision-based technique. *Protein Engineering, 7*, 39–46.

Nussinov, R., & Tsai, C.-J. (2015). Allostery without a conformational change? Revisiting the paradigm. *Current Opinion in Structural Biology, 30*, 17–24.

Okuhira, K., Shoda, T., Omura, R., Ohoka, N., Hattori, T., Shibata, N., et al. (2017). Targeted degradation of proteins localized in subcellular compartments by hybrid small molecules. *Molecular Pharmacology, 91*, 159–166.

Oshiro, C. M., Kuntz, I. D., & Dixon, J. S. (1995). Flexible ligand docking using a genetic algorithm. *Journal of Computer-Aided Molecular Design, 9*, 113–130.

Pan, A. C., Borhani, D. W., Dror, R. O., & Shaw, D. E. (2013). Molecular determinants of drug-receptor binding kinetics. *Drug Discovery Today, 18*, 667–673.

Peixoto, R. J. M., Alves, E. S., Wang, M., Ferreira, R. B. R., Granato, A., Han, J., et al. (2017). Repression of Salmonella host cell invasion by aromatic small molecules from the human fecal Metabolome. *Applied and Environmental Microbiology, 83*, e01148–e01117.

Perutz, M. (1960). Structure of hemoglobin. *Brookhaven Symposia in Biology*, *13*, 165–183.

Peter, C., Rueping, M., Wörner, H. J., Jaun, B., Seebach, D., & van Gunsteren, W. F. (2003). Molecular dynamics simulations of small peptides: Can one derive conformational preferences from ROESY spectra? *Chemistry, 9*, 5838–5849.

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., et al. (2005). Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry, 26*, 1781–1802.

Pica, N., & Palese, P. (2013). Toward a universal influenza virus vaccine: Prospects and challenges. *Annual Review of Medicine, 64*, 189–202.

Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., et al. (2013). GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics, 29*, 845–854.

Putri, J. F., Bhargava, P., Dhanjal, J. K., Yaguchi, T., Sundar, D., Kaul, S. C., et al. (2019). Mortaparib, a novel dual inhibitor of mortalin and PARP1, is a potential drug candidate for ovarian and cervical cancers. *Journal of Experimental & Clinical Cancer Research, 38*, 499.

Rarey, M., Kramer, B., Lengauer, T., & Klebe, G. (1996). A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology, 261*, 470–489.

Rees, D. C., Lewis, M., & Lipscomb, W. N. (1983). Refined crystal structure of carboxypeptidase a at 1.54 a resolution. *Journal of Molecular Biology, 168*, 367–387.

Rekka, E. A., Kourounakis, P. N., & Pantelidou, M. (2019). Xenobiotic Metabolising enzymes: Impact on pathologic conditions, drug interactions and drug design. *Current Topics in Medicinal Chemistry, 19*, 276–291.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering, Design and Selection, 12*, 85–94.

Ruiz-Carmona, S., Alvarez-Garcia, D., Foloppe, N., Garmendia-Doval, A. B., Juhos, S., Schmidtke, P., et al. (2014). rDock: A fast, versatile and open source program for docking ligands to proteins and nucleic acids. *PLoS Computational Biology, 10*, e1003571.

Sander, C., & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins, 9*, 56–68.

Schreyer, A., & Blundell, T. (2009). CREDO: A protein-ligand interaction database for drug discovery. *Chemical Biology & Drug Design, 73*, 157–167.

Sherman, W., Day, T., Jacobson, M. P., Friesner, R. A., & Farid, R. (2006). Novel procedure for modeling ligand/receptor induced fit effects. *Journal of Medicinal Chemistry, 49*, 534–553.

Showalter, S. A., & Bruschweiler, R. (2007). Validation of molecular dynamics simulations of biomolecules using NMR spin relaxation as benchmarks: Application to the AMBER99SB force field. *Journal of Chemical Theory and Computation, 3*, 961–975.

Silva, D-A., Bowman, G. R., Sosa-Peinado, A., & Huang, X. (2011). A role for both conformational selection and induced fit in ligand binding by the LAO protein. *PLoS Computational Biology, 7*, 26.

Singh, P., Abeysinghe, T., & Kohen, A. (2015). Linking protein motion to enzyme catalysis. *Molecules, 20*, 1192–1209.

Śledź, P., & Caflisch, A. (2018). Protein structure-based drug design: From docking to molecular dynamics. *Current Opinion in Structural Biology, 48*, 93–102.

Smith, G. J. D., Vijaykrishna, D., Bahl, J., Lycett, S. J., Worobey, M., Pybus, O. G., et al. (2009). Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza a epidemic. *Nature, 459*, 1122–1125.

Stank, A., Kokh, D. B., Fuller, J. C., & Wade, R. C. (2016). Protein binding pocket dynamics. *Accounts of Chemical Research, 49*, 809–815.

Stetz, G., & Verkhivker, G. M. (2017). Computational analysis of residue interaction networks and Coevolutionary relationships in the Hsp70 chaperones: A community-hopping model of allosteric regulation and communication. *PLoS Computational Biology, 13*, e1005299.

Summa, V., Petrocchi, A., Bonelli, F., Crescenzi, B., Donghi, M., Ferrara, M., et al. (2008). Discovery of raltegravir, a potent, selective orally bioavailable HIV-integrase inhibitor for the treatment of HIV-AIDS infection. *Journal of Medicinal Chemistry, 51*, 5843–5855.

Sweeney, H. L., & Holzbaur, E. L. F. (2018). Motor Proteins. *Cold Spring Harbor Perspectives in Biology, 10*, a0219.

Taubenberger, J. K., & Morens, D. M. (2008). The pathology of influenza virus infections. *Annual Review of Pathology, 3*, 499–522.

Van Drie, J. H. (2007). Computer-aided drug design: The next 20 years. *Journal of Computer-Aided Molecular Design, 21*, 591–601.

van Gunsteren, W. F., Dolenc, J., & Mark, A. E. (2008). Molecular simulation as an aid to experimentalists. *Current Opinion in Structural Biology, 18*, 149–153.

Varki, A. (2017). Biological roles of glycans. *Glycobiology, 27*, 3–49.

Vogt, A. D., & Di Cera, E. (2012). Conformational selection or induced fit? A critical appraisal of the kinetic mechanism. *Biochemistry, 51*, 5894–5902.

Wacker, D., Stevens, R. C., & Roth, B. L. (2017). How ligands illuminate GPCR molecular pharmacology. *Cell, 170*, 414–427.

Walther, D., Cohen, F. E., & Doniach, S. (2000). Reconstruction of low-resolution three-dimensional density maps from one-dimensional small-angle X-ray solution scattering data for biomolecules. *Journal of Applied Crystallography, 33*, 350–363.

Webb, E. C. (1992). *Enzyme nomenclature 1992: Recommendations of the nomenclature Committee of the International Union of biochemistry and molecular biology on the nomenclature and classification of enzymes*. San Diego: Academic. ISBN 978-0-12-227164-9.

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., et al. (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society, 106*, 765–784.

Wlodarski, T., & Zagrovic, B. (2009). Conformational selection and induced fit mechanism underlie specificity in noncovalent interactions with ubiquitin. *Proceedings of the National Academy of Sciences, 106*, 19346–19351.

Yari, M., Ghoshoon, M. B., Vakili, B., & Ghasemi, Y. (2017). Therapeutic enzymes: Applications and approaches to pharmacological improvement. *Current Pharmaceutical Biotechnology, 18*, 531–540.

Zhou, H. X. (2010). From induced fit to conformational selection: A continuum of binding mechanism controlled by the timescale of conformational transitions. *Biophysical Journal, 98*, 029.

Zhuravleva, A., & Gierasch, L. M. (2015). Substrate-binding domain conformational dynamics mediate Hsp70 allostery. *Proceedings of the National Academy of Sciences, 112*, E2865–E2873.

Chapter 4
Role of Advanced Computing in the Drug Discovery Process

Ajitha Mohan, Suparna Banerjee, and Kanagaraj Sekar

Computers are incredibly fast, accurate, and stupid; humans are incredibly slow, inaccurate, and brilliant; together they are powerful beyond imagination.

–Albert Einstein

Abstract The classical experimental approach to drug discovery is a tedious task for biological scientists as they are time-consuming and expensive. With the advent of advanced computing techniques such as artificial intelligence and high-performance computing, the problems of the traditional drug discovery approach can be circumvented. In particular, computational approaches help in analyzing and locating active binding sites and guide towards the selection of potential drug molecules that can effectively and specifically bind to these sites. Once lead molecules are identified, associated compounds can be chemically synthesized and tested. So the iterative time-consuming process of identifying potential drug molecules can be significantly reduced by implementing advanced computing techniques whereby it controls the spread of pandemic diseases. This chapter gives some of the popular high-performance techniques, automated statistical methods, and neural network algorithms for data mining in the drug discovery process along with their scope and application.

Ajitha Mohan and Suparna Banerjee contributed equally with all other contributors.

A. Mohan · S. Banerjee · K. Sekar (✉)
Department of Computational and Data Sciences, Indian Institute of Science, Bangalore, India
e-mail: sekar@iisc.ac.in

S. K. Singh (ed.), *Innovations and Implementations of Computer Aided Drug Discovery Strategies in Rational Drug Design*,
https://doi.org/10.1007/978-981-15-8936-2_4

4.1 Introduction

The field of drug design is a continuously developing area in which much progress has been done in recent years and has been fueled by the completion of the Human Genome Project and with the boom of genomic, proteomic, and structural information. The potential of designing drugs rationally using protein structures in the early 1980s was considered as an impractical approach by many structural biologists. Discovery of new lead drug targets is possible now due to enormous advances in high-throughput crystallography, for instance, automation at all stages, magnet and probe improvements in nuclear magnetic resonance (NMR), profound synchrotron radiation, and new progress in phase determination which in turn have lessened the timeline for structure determination. The availability of faster and relatively cost-effective computer clusters has improved the pace at which drug leads can be discovered and assessed in silico. The time committed to the drug design process may designate only a fraction of the total time towards generating end drug products for the market. Several years of research and development will be required to transform a potential drug into an effective drug through clinical trials which will be safe and tolerant for the human body. Based on the known structure of protein molecules, computer-aided drug designing is broadly classified into the structure and ligand-based drug designing (Fig. 4.1).

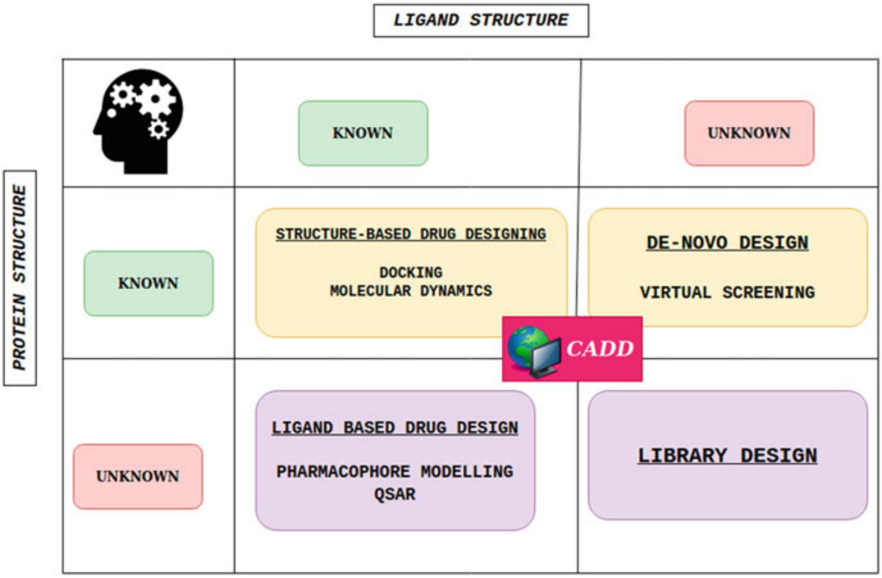


Fig. 4.1 Types of computer aided drug designing. This figure represents the types of Computer-Aided Drug Designing Methods. Structure-based and DE Novo methods can be applied if the three-dimensional structure of the protein is known. Ligand-based method can be applied if the ligand structure is known

4.1.1 Structure-Based Drug Designing Approach

Structure-based drug design is an iterative process and often continues through several cycles before an optimized lead enters into clinical trials phase1. The process starts with the cloning, followed by purification and eventually structure determination of the target protein by one of two principal methods: NMR or X-ray crystallography. Protein Data Bank is a repository of solved protein structures; for proteins that are difficult to isolate or crystallize, modeling methodology can be employed.

Comparative or Homology modeling is an approach that largely depends on sequence homology between the target protein and at least one known structure. This process involves the following steps: Template identification of protein having known homologous 3D structure(s); Alignment of sequences of a template and target proteins; Based on the alignment and 3D structure of the template, model generation for the target is done, followed by refinement and validation; To ascertain the rationality of modeled structures, few parameters have to keep under consideration like stereochemistry, energy profile, residue environment, and structure similarity (Huang et al. 2010).

Threading (Fold recognition) is a method that is used to model proteins that do not have homology with available protein structures. In the threading protocol, taking into consideration both protein surface area and the abode of residues interaction (Mishra 2009) similarity search is performed using the given amino acid sequence with the 3D structures in a database of known folds. From these folds, the construction of the structure of the query protein is done. For the prediction of unknown protein structure, if the aforementioned protocols fail, the ab initio method can be very instrumental although it is less convincing in terms of accuracy and identity (Huang et al. 1998). The ab initio method of modeling is also termed as de novo modeling or physics-based modeling. The basic aim is the prediction of native folds which starts with the primary amino acid sequence of the query protein that is searched for various conformations. After fold recognition and prediction, the model is assessed for verifying the quality of the predicted structure. Another important aspect of structure-based drug design is the determination of the active site of the ligand. By conducting cocrystallization studies, a good target site can be determined such that the target macromolecule is crystallized with an initial small molecule inhibitor. Active site verification by location mapping of a ligand in the crystal lattice is not possible for proteins that cannot be crystallized. To circumvent this problem, different types of binding site determination protocols have been designed wherein small molecular fragments are used as probes to explore the protein surface. One such example is FTMAP (Brenke et al. 2009) Spots, where small molecular fragments are clustered and expected to be the advantageous druggable sites.

After the identification of structure and target site, there are various ways both experimental and computer-aided to develop a potential lead based on the target structure. An instance of the experimental method is combinatorial chemistry along with high-throughput screening, in which millions of compounds are tested with biochemical assays. Structure-based design and combinatorial chemistry when integrated can guide the parallel synthesis of the resolved compound library (Antel 1999) The computer-aided methods can be categorized into three main classes: inspection, virtual screening, and de novo generation. In the first category, modification of known molecules that are reported to bind the target site is done to become inhibitors based on augmenting complementary interactions. Target sites may reside in enzymes such as substrates or cofactors or peptide forms in protein: protein or protein: nucleic acid interactions. As per the procedure of virtual screening, databases of identified compounds or fragments of compounds are docked into identified regions of the structure using computational algorithms. The identified compounds are ranked based on predicted steric and electrostatic interactions with the target site. Programs such as DOCK, SLIDE, FlexX, or FlexE and other dock compound databases and subsequently score them based on their interactions with the target site. In de novo generation, in silico positioning of small fragments of molecules, such as amino groups, carbonyl groups, benzene rings, etc. is done in the target site, followed by scoring and finally linking is performed. Subsequently from these linked fragments, synthesis of final compounds will be done. Examples of de novo lead generation programs are LUDI, GRID, MCSS, CONCERTS, SMoG, etc. (Anderson 2003).

After identification of a small molecule that has the potential to bind specifically to the target molecule, evaluation of the same has to be done before any further proceeding. Since the model of the target: ligand interaction is intrinsically an approximation, the rank assigned by the scoring function may not be suggestive of a true binding constant. It may so happen that the molecules that topped scoring in docking run may fail in vitro biochemical assays. Therefore, lead evaluation followed by lead optimization must be practiced for improved affinity. To become an orally bioavailable drug, “Lipinski’s Rule of 5” (Lipinski et al. 1997) must be satisfied which proposes that good leads generally have less than five hydrogen bond donors and less than ten hydrogen bond acceptors leads generally have less than five hydrogen bond donors and less than ten hydrogen bond acceptors, a molecular weight <500, and a calculated log of the partition coefficient (log P) <5. Veber and colleagues (Veber et al. 2002) opine that to increase the potential for oral bioavailability, the number of rotatable bonds should be <10. Various factors, such as cytotoxicity, chemical, and metabolic stability and the ease of synthesis, are also taken into consideration before proceeding with a particular candidate lead. For final evaluation best compounds or leads are brought into the wet lab for biochemical assays. Although even today it is still required to refine the process with an aim of more perfection, SBDD (Structure-based Drug Design) has become a fundamental part of most industrial drug discovery programs (Anderson 2003).

4.1.2 Ligand-Based Drug Design Approach

The ligand-based drug design approach is crucial when there is an absence of the three-dimensional structure of protein molecules which acts as a drug target. In this approach, lead compound optimization can be done by two methods such as Quantitative Structure-Activity Relationship study and pharmacophore modeling. This study correlates the activity of lead molecules based on the study of their molecular structures and drug properties adhering to the concept that the structure of the lead molecule contains some clue regarding their biological activity, that can be implemented for lead optimization. These approaches are carried out successfully with the help of prominent drug properties of a molecule such as Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) which plays a major role in the creation of pharmacophore and QSAR models. Initially, all such types of drug properties of a molecule will be collected from the available primary database. If data are scarce, then they can be retrieved from the literature survey. Based on the statistical analysis, the information which is not matching with the properties of a drug molecule is removed from the list. From the existing information, the standardization of Ki value was done initially, then 70% and 30% of that source data are used as training and test data to build the model respectively. Now, the collected list of information related to the structural and biochemical activity of the lead molecule will be implemented in the Quantitative study of structure and activity relationship. Finally, the performance of the model is evaluated with test data using cross-validation methods. In the Pharmacophore modeling method, superimposition algorithms are applied whereas, in the QSAR study, two types of renowned statistical approaches such as Multiple Linear Regression and Partial Least Square Regression were used eminently to build the model.

Pharmacophore Modelling

The International Union of Pure and Applied Chemistry defines the term "Pharmacophore" as an ensemble of electronic and steric features of a ligand molecule that is capable of binding to a target and alters its biological activity. This type of modeling method works on the base of the main concept that even if the two ligand is structurally varied, due to their ensemble steric and electronic features they will bind to the same biological targets. This modeling method predicts the novel ligand through the implementation of superimposition algorithms and the virtual screening process. This modeling process involves various significant steps, it initially screens the list of the known structurally diverse ligand. These data are used as training sets that contain both inactive and active compounds based on various features such as aromatic rings, hydrophobic centroids, anions, cations, and the number of hydrogen bond donors and acceptors. For the screened training sets the low energy conformations are identified for each molecule by several tools such as Discovery Studio and LigandScout etc. From this conformation, the best fit active compound is optimized by the superimposition of molecular frameworks. Finally, by including their ensemble features, all the raw data are transformed into the abstract annotation of molecular characteristics that contains the key information for the interaction of the biological molecule with that of a ligand. This method is mainly applied in the field of computational chemistry for developing the pharmacophore model (Prajapat et al. 2017).

The main role of advanced computing in generating the pharmacophore model is to automatically learn the rules from known compounds available at SCRATCH and design the new models by assembling the molecule based on the learned pharmacophore features. Thus the combination of modeling 3D-pharmacophore and Artificial intelligence results in the development of many academically free tools that are capable of generating 3D pharmacophores for virtual screening. Such type of generation takes place by including their hydration and thermodynamics properties.

QSAR

Multiple Linear Regression

Multiple Linear Regression approach is implemented in the study of the relationship between multiple sets of independent variables with the dependent variable. By fitting the linear line in the scatter plot, it predicts the linear relationship of dependent and independent values. The fitting of the line takes place on the following Eq. (4.1):

$$Y = b0 + b1x1 + b2x2 \tag{4.1}$$

where b0 is the y-intercept, b1 and b2 are regression coefficients, x1 and x2 are independent variables and Y is dependent query variable. This model is implemented to measure the quantitative relation between the training variables whereby the relativity of a query dependent variable can be predicted.

Partial Least Square Regression

Partial Least Square Regression works on the basic rules of principal component analysis as it fits the line in the scatter plot by projecting the observed and expected independent variables in new space with fewer dimensions. The models which are created by applying this method are also known as Bilinear factor models. This method is mostly applied in areas of computational biology, chemometrics, anthropology, and Neuroscience. At the time of model evaluation, it chooses its best model by identifying the least sum of squares between observed and predicted values. Specifically, Partial Least Square was applied at the condition when the number of features is higher than their observations. In this condition, multiple linear regression cannot be applied and this condition is called “Over Fitting”. In most cases, the validation process is done by applying LOO (Leave One Out) cross-validation method and its accuracy level is measured by the PRESS (Predicted Residual Error Sum of Squares) formula (4.2)

$$PRESS = \Sigma(X - \hat{X})^2 \tag{4.2}$$

where PRESS is the sum of the square of deviation of the observed mean value (X) from the predicted mean value ($\hat{X}$).

By the combo application of Multiple Linear Regression and Partial Least Square Methods, a QSAR model is designed for 33 lists of alpha1-adrenoreceptor antagonists retrieved from antipsychotic sertindole (Mehmood et al. 2012). This application disseminates the significance of properties of binding pockets such as polar interaction, molecular flexibility, and steric fit which is essential for drug-ligand interactions. These methods are considered and employed automatically using recent advanced computing techniques in the field of cheminformatics to solve 3D-QSAR problems. By applying such techniques the size and complexity of the massive source datasets can be reduced. Such that from the source of 240 million datasets from the PubChem database, the rules are learned by the advanced computing techniques and finally the new model can be built to predict the parameters related to pharmacological characteristics successfully.

4.2 Data Mining in Drug Discovery

In early periods of the drug discovery process, pattern identification from known data takes place by applying the Bayes Theorem and regression methods with limitations of training data sets. As the complexity and size of the input data increase tremendously, it reaches the stage that the manual extraction of data is not possible. But the advent of advanced computing techniques overcomes such limitations and initiates the automated process of data collection, manipulation, and analysis. Such an automated process is successfully proposed by the ensemble of statistics, machine learning, and database systems known as "Data Mining". Data Mining extracts the information from the large raw data by applying three approaches such as Artificial Intelligence, Machine Learning, and Deep Learning (Fig. 4.2). Information from Data Mining is applied successfully by all the researchers from various domains such as Predictive Analytics, Informatics, Business Intelligence, web mining, and the medical fields. In recent years, the drug discovery process is accelerated rapidly due to the application of the data mining process. Due to the unprecedented progress of biomedical records, there is a need to apply the data mining process in the biological field which gives safety hints to patients. Based on the biochemical records such as log (IC50) values and their pKi values, new models can be built by the following five indispensable steps in data mining processes such as Data preprocessing, Data splitting, Modelling, Evaluation, and Deployment.

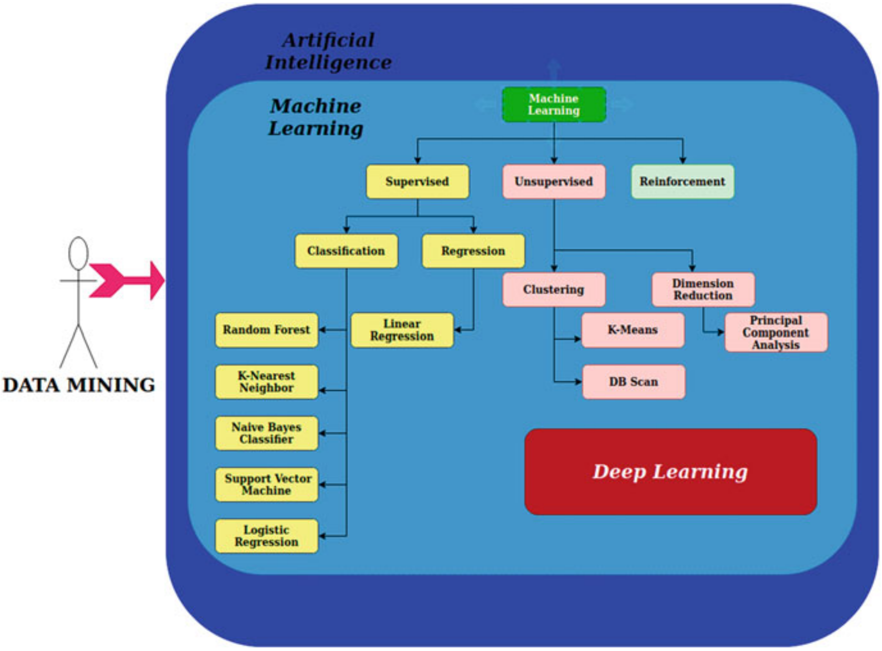


Fig. 4.2 Categories of data mining. This figure represents Data Mining that includes three components such as Artificial Intelligence, Machine Learning, and Deep Learning. Machine Learning is divided into three types such as supervised, unsupervised and reinforcement based on the availability of Actual outputs

In 2018, a novel model is developed using data mining process which predicts drug properties, in which the model is trained with 762 compounds with thirty-five physicochemical features such as binding energy to their target, exact mass, number of carbons, molecular surface area, polar surface area, number of hydrogen bond donors and acceptors, count of rotatable bonds, number of aromatic atoms and its polar properties, etc. Based on the approved status of drugs from DRUGBANK, 366, and 396 compounds are discerned into drug and non-drug molecules respectively from the 762 training data sets. As the first attempt in this model development process, data visualization is carried out by using the t-Distributed stochastic neighbor embedding method, as it is capable of analyzing the non-linear multi-dimensional data (Yosipof et al. 2018). This method works with the background concept that dissimilar and similar features are classified by the pairwise euclidean distance of far and nearby points from centroids respectively, thus it is capable of reducing the dimensionality of the training data. In specific, this work focuses on the disease related to antineoplastic agents, cardiovascular and nervous systems. Evaluation of this model is done and its calculated correlation rate is measured as 0.81. This model is implemented in the field of the pharmaceutical era as it saves time and cost of experimental drug design.

4.2.1 Artificial Neural Network

Artificial neural networks resemble the structure of interconnected neurons framework of the human brain. Latterly, it is highly used in chemoinformatics to design a drug molecule based on the knowledge of 2D and 3D known chemical descriptors of a compound. Such information can be retrieved from the database in the form of PDB, MDL (Molfile), SDF (Structure Data Format) format. Due to the elevation of descriptors data, these formats are simplified into line notation formats such as SMILE (simplified molecular-input line-entry system), WLN (Wiswesser line notation, ROSDAL and SYBYL using ASCII codes by applying morgan algorithm. In the computational-aided drug discovery process, MIF (Molecular Interaction Field) is considered as a prominent chemical descriptor as it plays an eminent role in binding with target molecules which can be predicted from the programs such as PRODRG and RDKit (Lo et al. 2019). Artificial neural networks are applied in 3D-QSAR study to search compounds with similar chemical descriptors and fingerprints by comparing their molecular interaction field (CoMIF), by the following steps (Fig. 4.3).

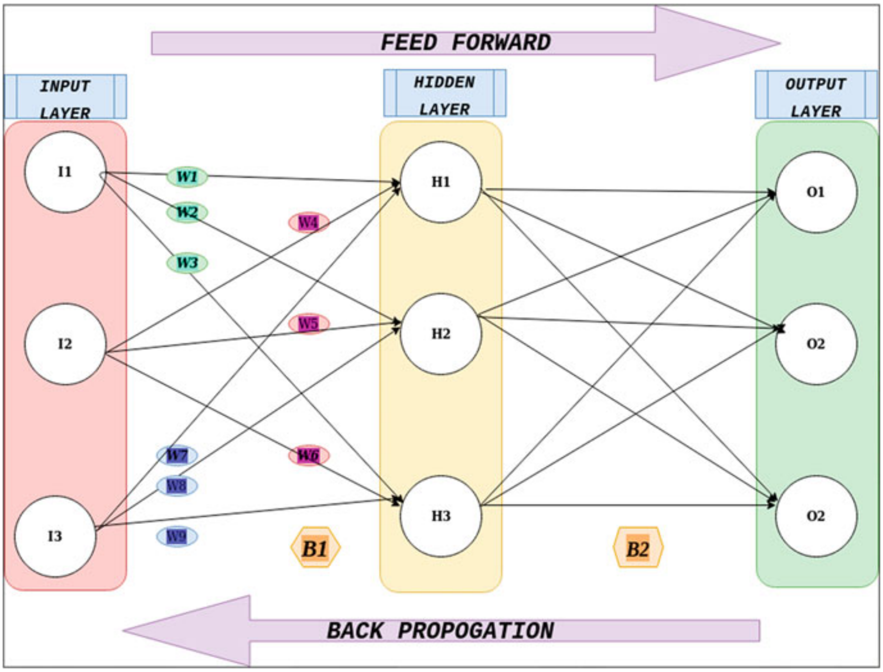


Fig. 4.3 The architecture of artificial neural networks. This figure represents artificial neural network which includes three layers such as Input (I), Hidden (H) and Outer layers (O). W and B indicate the value of weight and bias in the feed-forward process. The error value can be rectified by adjusting the value of weight in the backpropagation process

Initialization of Weights

Neural networks are similar to interconnected neurons of the human brain. It consists of three layers such as the input layer, hidden layer, and output layer and each layer consists of their corresponding nodes. If the input data are images then the input layer converts it to several pixel numerical data. Such types of artificial neural networks are applied to predict mystery data based on the final output of numerical data between 0 and 1, which indicates the probability of correct prediction. So initially the available data is fed into the input layer which contains nodes labeled as I1, I2, I3, these nodes are connected to the hidden layer by parameters known as weight (W) and bias (B). To get closer to the expected output, these weights and bias are initiated by random numbers and fed into the nodes of the hidden layer such as H1, H2, H3 by the following formula (4.3)

$$\mathrm{H}_i = \left(\sum_{i=1}^{n} \mathrm{I}_{\mathrm{i.}}\mathrm{W}_{\mathrm{i}} + \mathrm{B}_{\mathrm{i}}\right) \tag{4.3}$$

Multilayer Perceptrons

Multilayer Perceptrons are a type of artificial neural network in which the information flows only in the forward direction from the input node to the output node. The weighted sum of the inputs (H_i) with added bias (b) is explored in sigmoid or cost activation function (**φ**) as shown in (4.4) to get the desired output value between 0 and 1 in the forward direction and they are fed into the nodes of output layers (O1, O2, O3).

$$\text{Activation Function } (\varphi) = 1/\left(1 + \mathrm{e}^{-\mathrm{H}_i}\right) \tag{4.4}$$

If the output value is >0.5, mystery data is predicted to be the second category, if it is <0.5 it is predicted as the first category. This process is iterated automatically by algorithms for every training data sets which contain actual inputs excluding the actual outputs and from the nodes of the output layer the final outputs are predicted. Finally, the error of the network model can be predicted by the loss function which predicts the difference between the actual and predicted outputs as shown in formula (4.5).

$$\text{Loss function} = [\text{actual output} - \text{predicted output}]^2 \tag{4.5}$$

Backpropagation of Error

In this phase to reduce the error, weights and bias values are adjusted and passed through the neurons in a backward direction known as back-propagation. By adjusting the values several times until the final predicted output comes near the actual outputs, back-propagation is done repeatedly to attain higher accuracy. Finally, the minimum error value of a function is identified by the graphical method by plotting weight versus loss function which is known as Gradient descent. The correct weight is chosen based on the identification of the slope. A negative slope indicates weight to be decreased and a positive slope indicates the weight to be increased and this process comes to end when the zero slopes are identified. The weight which gives us zero slopes is considered as the correct weight and now it can be applied in the final model. Now the model is ready to predict the mystery data. By applying the above steps a new model is developed, by which the ADMET properties of drugs can be predicted for the query molecule. Such types of artificial intelligence techniques are proved as an eminent computing technique in the field of drug discovery in recent scenarios. Their success rate mainly depends on the characteristics of the input data such as the electronic medical report of patients which is to be fed into the neural networks. Thus the large input data should be compiled and annotated to learn the pattern automatically. Anyway to some extent it needs the human decision to direct the target of recurrent neural networks. One of the successful neural networks is ATOMWISE which learns the available structural interaction patterns of ligand and protein from huge data to predict new lead molecules and it performs better than the traditional drug discovery process.

4.2.2 Machine Learning

Machine Learning acts as one of the significant subgroups of Artificial Intelligence. Over the last decade, machine learning plays an eminent role in handling huge amounts of data and thus it is a boon to computer scientists. Especially, it accelerates the drug discovery process in a tremendous way. As the successful outcome of computer-aided drug discovery depends on various factors, for each issue different machine learning techniques are applied. Machine learning techniques are broadly categorized into two types such as supervised and unsupervised techniques which include various phases (Fig. 4.4). If the input and output data are available for the training of the model, then for prediction supervised machine learning techniques are applied. In the reverse case, if the training data contains only the input data then unsupervised techniques are applied. Thus in each issue, machine learning techniques have their advantages and disadvantages, but this can be overcome by the combination of machine learning techniques for model development.

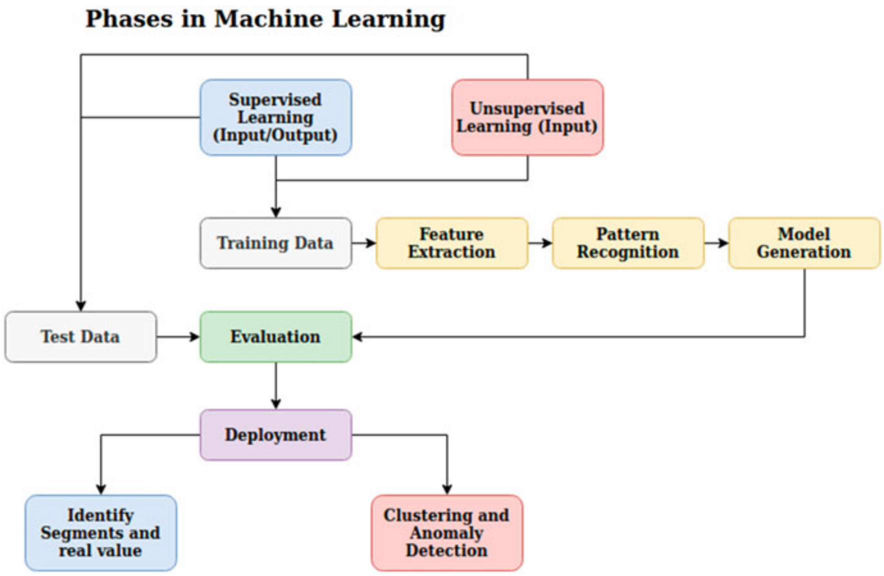

Phases in Machine Learning
Supervised Learning (Input/Output)
Unsupervised Learning (Input)
Training Data
Feature Extraction
Pattern Recognition
Model Generation
Test Data
Evaluation
Deployment
Identify Segments and real value
Clustering and Anomaly Detection

Fig. 4.4 Phases involved in machine learning techniques. This figure represents a supervised and unsupervised model that can be deployed to identify the real value and anomaly detection respectively

Supervised Machine Learning

In supervised machine learning techniques, the learning process depends on the pair of input and desired output data. It is broadly divided into two types such as Regression and Classification for quantifying and labeling problems respectively.

Regression

Regression is a type of supervised learning technique, which estimates the amount of dependency of the dependent variable on the independent variable. It is widely divided into linear and multiple linear regression based on the single and the multiple numbers of independent variables.

Linear and Multiple Linear Regression

Linear and Multiple Linear regression can be applied to predict continuous data using the linear function. It identifies the relationship between the dependent and independent variables. Such a relationship is predicted based on the dependency level of the dependent variable on the independent variable using a linear equation as shown in (4.6). Over the decade, it is highly applied in the ligand-based drug designing studies.

$$\mathrm{Y} = \mathrm{mx} + \mathrm{c} \tag{4.6}$$

where m is the slope. c is the y-intercept. x and Y are independent and dependent variables respectively. c is the y-intercept.

Classification

Classification is a type of supervised learning technique, which is applied to predict the discrete class labels based on their probability value. Predictions from this method can be assessed by the accuracy level which is not applied in regression analysis.

1. Logistic Regression

   Logistic regression is a type of supervised classification method in which the model is trained with the labeled information to predict the binary results using a sigmoid function. As the outcome of the dependent test variable is discrete and not continuous, the graph is plotted with a range of 0 to 1. The final prediction is successfully proposed based on the threshold value which is set to be a default value of 0.50. Recently in 2019, a novel model is developed using the biomedical records of 2098 patients who have undergone treatment for Hepatitis B and Hepatitis C virus in Kermanshah province to identify the factors associated with drug use transition injection (Najafi-Ghobadi et al. 2019). After the validation process, the average accuracy of the model is calculated as 91% and finally, it predicts some significant factors such as heroin, cocaine, and hallucinogens are responsible for drug use transition injection.
2. KNN

   The K-Nearest Neighbors (KNN) is a very significant method as it can develop a model using the available training data sets without limitation, where K indicates the number of nearest neighbor vectors that decide the categories of the classification. But the disadvantage of this method is due to updates, if the value of K changes, then the classification category also changes. Rather than this issue, it has more advantages such as learning from an instance-based training data set; thus it is time and cost-effective. Recently, it is proved that a combination of the k-nearest neighbor method along with the genetic algorithm improves their accuracy level and even their drawbacks can be amended by this method (Sarkate and Deorankar 2018). This combination strategy is successfully applied to classify the unknown novel type of drugs and their bias is rectified by the application of genetic algorithms. Thus, the K-Nearest Neighbor approach is highly recommended in the integrated form in the biomedical field.
3. Random Forest

   As the name indicates, random forest methods classify the input data based on the various number of decision trees and random majority votes. These decision trees are generated by applying bagging ensemble methods and its main advantages are it can be applied for both classification and regression problems.

Recently, these models are applied to predict the activity and inactivity of lead compounds against tumor cells based on the huge mutational status data from oncogenes (Lind and Anderson 2019). As this data from cancer patients indicates the response of each person for a drug candidate, this data is used as training data in random forest model generation. Due to its added advantages, novel random forest regression models are successfully applied in the field of precision medicine, where the log (IC50) values were used as training data set to optimize the inhibitors with a Pearson correlation coefficient value of 0.86. Among all supervised learning techniques, this method is proved as a top rank holder as it decides the final output by aggregating all the features of input training data sets.

4. Comprehensive Ensemble Methods

   As a recent trend, the novel predictive models are built by the combination of bagging (parallel training) and boosting (sequential training) ensemble methods to predict the activity of drugs in the study of the QSAR approach. As evidence of this comprehensive attempt in 2019, a novel ensemble method is proposed and it is easily accessible on the specific website (Kwon et al. 2019). By applying the second level-meta learning, this method breaks the caveats of single-subject modeling. The assessment process is carried out successfully by applying scaffold-based HIV data sets. This novel method proves that the accuracy level of comprehensive methods is very high when compared to that of other methods using ROC-AUC value.

5. Naive Bayes Classifier

   Naive Bayes classifier applies Bayes theorem to classify the data set by ignoring the dependency or correlation between the known input and output data. Initially, the Bayesian network model is developed using training data, and based on the conditional probabilities the classification of data takes place. Recently, it is proved that Bayes classifier models are deployed to estimate the activity of multidrug resistance reversal agents (MDRR) using 424 training data sets, and its accuracy is proved to be 82.2%. In 2017, by applying this classifier method, a novel model is proposed to predict the mutagenicity of drug substances. This model was rated by employing the k cross-fold validation, where the k value is 5 and the accuracy of the model is proved to be 89% and 74% for internal and external data (Zhang et al. 2017). This newly created model is leveraged to assess the risk condition based on the probability score derived from the list of mutagenic chemicals which plays a major role in the ligand-based drug discovery process.

   Similarly, in 2019, BANDIT is a new model which is proposed by applying the Bayesian network approach where the model is trained with the 2000+ integrated public known drug-target data (Madhukar et al. 2019). Then this model is deployed to predict the target of the ONC201-anticancer compound as their target remains uncertain still now. As one of the investigation processes, this model is used to predict latent kinase targets, and based on their outputs it is confirmed that the newly developed model is capable of guiding the experimental screens. The final results derived from the BANDIT model are compared with that of experimental results and confirmed to be similar. One of the outstanding

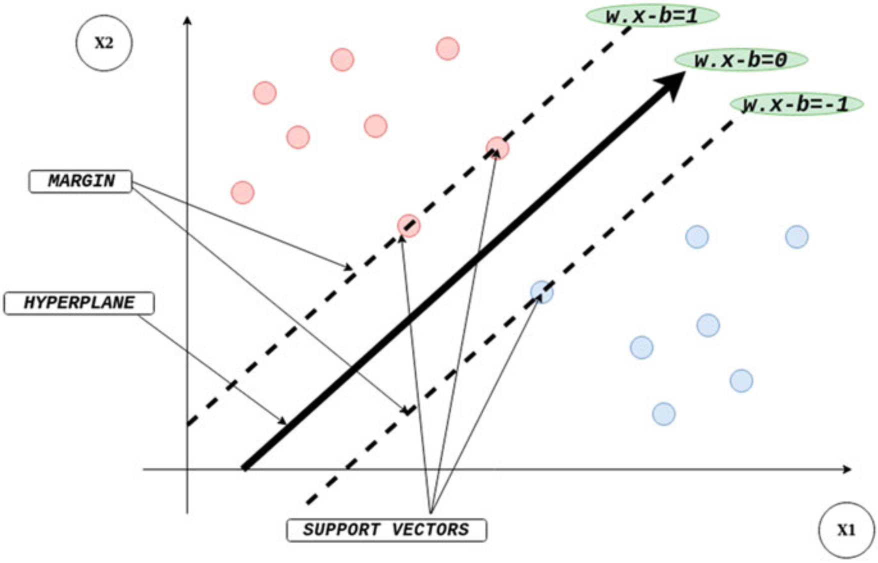


Fig. 4.5 Hyperplane of support vector machine. This figure represents a support vector machine that can be implemented to classify the non-linear data based on the support vectors which is separated by the Hyperplane

capabilities of this BANDIT model is it can augment even the upcoming diverse binding kinetics data sets by calculating their direct probabilities, whereby its predictive power can be increased. Thus in the medicinal field, it can be employed to identify the side effects of uncharacteristic off-targets.

6. Support Vector Machine

   A support vector machine is a type of supervised learning technique, in which the trained model can be applied to classify non-linear data with help of kernels. As this method is used as optimization of the logistic regression method, its application is very high over the last decade. It is also known as the Equal margin classifier. Classification is done by fixing the hyperplane to separate two distinct classes and this hyperplane is fixed based on points that are close to the opposing class which are known as support vectors (Fig. 4.5) and the remaining points in the training data sets are excluded.

A function that takes input vectors from its original space and transforms it into new vectors by their dot products in the feature space is called a kernel function or kernel tricks. Its main advantages are its reliability and cost-effectiveness, due to this reason this method is applied to classify the target data in recent periods. Usually, Non-linear support vector machines take three or four supporting vectors from the two classes. By applying the vectors in the mapping function, transformed new vectors can be identified. Then the bias unit is augmented with the newly formed vectors and applied in the linear equation. By applying the above kernel functions, the biological and chemical features of active compounds are predicted based on the available ADMET properties. It is also proved that meta-classifiers of the support vector machine approach overcome their limitations and speed up the drug discovery process (Heikamp and Bajorath 2014).

Unsupervised Learning

Principal Component Analysis

Principal Component Analysis is a linear transformation of data and its main application is the reduction of huge dimensions of data without loss of any significant features from training input data. It is a type of unsupervised learning techniques as it trains the model only with the training input data due to the absence of output data. The dimensional reduction is done by identifying highly correlated data and it considers that data into a single cluster by removing the consistency as the redundancy causes more bias. Final results are generally represented as a PCA plot which is a 2-D graph, which is plotted by analyzing the correlations of the input data. Based on the cluster in the graph, the highly correlated data can be identified, thus the number of clusters in the graph represents the number of correlated data in the training data sets. These clusters are formed based on the covariance value. If the covariance value is positive, then we can conclude that the data are positively correlated and vice versa. This covariance tells us only their associations and not their strength of the relationship. Based on this value, possible similar characteristics of training data are picked up and clustered into a group and thus many-dimensional data can be converted and plotted into a two-dimensional graph by the following steps.

1. Standardization of the Data

   Standardization of the data is the process of converting all the variables of the input data to a standard range as shown in formula (4.7)

$$\mathrm{Z} = (\mathrm{Variable} - \mathrm{mean})/\mathrm{standard\ deviation} \tag{4.7}$$

2. Computing the Covariance Matrix

   The covariance matrix is formed to identify the correlation between the different variables in the data set based on the number of dimensions without bias. It is a type of square matrix that contains variance values along their diagonal, covariance values at their off-diagonals.
3. Calculating Eigenvectors and Eigenvalues

   Eigenvectors and eigenvalues are calculated to compute the principal components based on the following Eq. (4.8) If the resultant vector formed by the product of a matrix and a vector, undergoes scaling function without any rotation, then that vector is known as eigenvector. The scaling factor which supports the eigenvector is known as an eigenvalue.

$$\mathrm{Ax} = \lambda \mathrm{x} \tag{4.8}$$

where A stands for a matrix. x stands for eigenvector and λ stands for eigenvalue.

4. Computing the Principal Components

   Principal components are the novel data sets which are formed after the dimensional reduction. Principal components are ranked in such a way that the first principal components contain the maximum number of variances compared to others as they have maximum eigenvectors and eigenvalues.

5. Reducing the Dimensions of the Data Set

   Replacement of initial data set with the newly formed principal component in a specific arrangement reduces the dimensionality of data. Thus, all the principal components are arranged in the order of high to low variance value. So that in case of more dimensional data, the last few principal components can be removed as it contains fewer variance data. Thus, there will not be any significant loss of data. By following the above five steps, an algorithm is designed and applied to decipher the multi-class mRNA expression data of control and PTSD (Post-traumatic Stress Disorder) from stressed mouse hearts (Taguchi et al. 2015)

K-Means Cluster

Clustering is the method of grouping the data sets into different clusters based on the possible similarity. This method is applied to develop a model with available input data as a training data set. As these training data sets do not have any output data, it belongs to the category of unsupervised learning. There are three types of clustering such as exclusive clustering, overlapping clustering, and Hierarchical clustering. K-means clustering comes under an exclusive clustering type, where the clustering is done very exclusively in such a way that data points of two different clusters do not overlap with each other. In contrast, overlapping will be there in overlapping clustering between the data points from two different clusters and the C-means clustering algorithm can be used to predict the overlapping clusters. Among the overlapping clusters, similar data points are again categorized hierarchically which is known as Hierarchical clustering. Over the last decade, the k-means clustering algorithm is applied in most of the cases, where k stands for the number of clusters or groups. It can be applied if the input data is continuous or numeric and it takes them as training data sets and based on their similarity it segments the whole given random data sets into clusters. Initially, for the segmentation process, it chooses two data points randomly from the given input data (X1={i1,i2..i10}, and X2={j1, j2. . .j10}) which is called as centroids ((i4,i7), (j4,j7)) Then it measures the Euclidean distance between the centroids and each data point from the data sets. Based on the range of distance, the data sets with minimum distance are grouped into the first cluster and the maximum distance is grouped into the second cluster. From the mean of these clustered data points, appropriate centroids are identified by the algorithm to form the new clusters. This process is iterated automatically by cluster algorithms using python or R modules.

1. Elbow Method

   The Elbow method is used to find the appropriate number of clusters for the given input data set. A graph is plotted with cluster numbers versus distortions derived from the iterated process. As the line graph looks like a human elbow, it is named as elbow method. The specific cluster number where the bend of elbow shape initiates is identified as the appropriate cluster number in the k-means cluster algorithm.
2. Silhouette Coefficient Method

   The Silhouette Coefficient method is applied to find the accuracy level of each cluster formation which can be done by applying in (4.9)

$$\text{Silhouette Coefficient} = \text{ba}/max(\text{a}, \text{b}) \tag{4.9}$$

where a is the distance between the centroids and data points and b is the distance between two centroids from two clusters. Finally, a cluster formation with a higher silhouette coefficient value indicates the higher accuracy level of the model. Based on the above steps a novel k-means model is proposed to cluster similar drug candidates based on their ATC (Anatomical Therapeutic Chemical) classification (Hameed et al. 2018) and the final model is evaluated successfully by comparing the clustered drug results (Thioridazine, Indomethacin, Chlorthalidone, and Metformin) with known clinical results.

DB-Scan

DB-scan is a type of unsupervised clustering method, where the clustering is done by the prototypes. These prototypes are derived from the squared error of the k-means clustering method. Their performance level is low, when it is applied individually due to its quadratic computational complexity. It’s better to combine with that of the k-means clustering method (Edla et al. 2012). Over the last period, a novel model is developed and validated successfully by the combo of DB-SCAN with the k-means method to cluster the gene expression data.

Reinforcement Learning

Reinforcement learning techniques apply dynamic programming to train the model by ignoring the labeled input data. But due to its complex functionality, it is less preferred individually when compared to that of other machine learning techniques. But with the combination of other methods, it shows significant results in the study of inhibitory activity against the Janus protein kinase-2 (Popova et al. 2018).

By applying both supervised and unsupervised methods, novel metaheuristic algorithms were designed to propose new drugs for various health issues such as to treat particular infectious diseases, cancer, cardiovascular disease, Rheumatoid Arthritis, and Parkinson's diseases, etc. By biomarker development, it abstains from the time consumption and expenditure of the traditional approach tremendously due to the implementation of quantum computers instead of supercomputers. All these projects are done successfully due to the huge availability of source data such as gene expression, proteome, and next-generation sequencing data.

4.3 Deep Learning

Deep learning is a subset of machine learning techniques, which uses the geometric conversion of multi neural layers for prediction from the known input data. The phrase "deep" is directly proportional to the number of layers in the neural networks. In these multilayers, the upper layer reads the highly significant data and the lower layer reads their edges. It is mainly applied in image processing, where each input pixel is convolved to a significant value. One of the main significance of this method is it automatically sorts the significant data from higher to lower level and it feeds that data into their corresponding layers. This method is classified broadly into two types as Convolution and Deep Belief Networks based on the presence and absence of a convolved process respectively.

Recently in 2016, by feeding the data from Pubchem Bioassays and STITCH database, a novel model is generated using a deep learning approach to predict the interaction between the compound and the protein. Evaluation is done by comparing the results of a deep learning approach with that of the outcome of classification and regression models of machine learning techniques and concluded that deep learning models have higher accuracy when compared to that of others. In 2019, a novel ensemble model of Convolution and Deep Belief Networks are generated with the merge of molecular fingerprints of binding molecules with that of protein primary sequence. All such types of data are retrieved from IUPHAR, KEGG, and DRUGBANK database. Even though the number of input data is about 32,000, they are fed into the multilevel layers and all the features of the data are deeply learned to predict the drug-target interactions (Lipinski et al. 2019).

A prominent DEEPSCREENING web server is also proposed in 2019, based on the deep learning approach to overcome the lack of tools to instantly perform virtual screening in CADD (Liu et al. 2019). Besides various open-source deep learning tools such as PACCMAN, INTERACT, PIMKL were proposed to predict drug sensitivity, decipher research publication, and phenotype prediction from omic data. Thus these methods act as robotic prototype drug designers by the collaboration of Biotech scientists with various pharmaceutical companies such as GlaxoSmithKline, Bayer Science, Cloud Pharmaceuticals, GNS Healthcare, Exscientia, etc. One of the main challenging tasks for implementing such novel computing techniques in the field of drug discovery is getting patent rights.

Thus, data science is an astrologer to the computational scientist as they predict and tell us the possible outcome of the upcoming experimental event in the drug discovery process in advance. Due to the advent of advanced parallelization computing techniques, High-Performance Computing acts as an empower for Data Mining Process.

4.4 High-Performance Computing in Drug Discovery

Advanced computing is an expansive term that is most commonly used to describe a specific type of high-end computer and the associated processing undertaken on it to solve a computationally-intensive problem. High-performance computing (HPC) is the most prevalently used advanced computing paradigm. HPC is the facility to process data and perform complex calculations at high speeds. Such problems are either too large for standard computers or would take too long (Ge et al. 2013). For instance, if HPC is not employed for a 10K chemical molecule repository, to carry out only tens of nanoseconds of MD simulations will require years of computer time. For an analogy, a modern laptop or desktop with a 3 GHz clock processor can perform around 3 billion calculations per second. While that is much faster than any human can achieve, it is insignificant in comparison to HPC solutions that can perform quadrillions of calculations per second. Supercomputers are the best-known example of HPC solutions. A supercomputer consists of thousands of computing nodes that coordinate to finish one or more tasks in parallel. Powerful supercomputers with state of the art sophisticated and complex algorithms can model real-world phenomenon. They are also effectively used to compute the binding energies between various small molecules and proteins. Few well-known HPC architectures are (1) Computer Clusters; a set of interconnected computers controlled by a centralized scheduler. (2) Grid Computing; a set of geographically distributed and logically organized (can be heterogeneous) computing resources. (3) Grid Computing; Dedicated parallel co-processor, used in computing data-parallel intensive segment. (4) FPGA (Field Programmable Gate Array); Integrated circuits containing an array of programmable logic blocks. (5) Cloud Computing; Pool of computation resources (e.g. processing, storage) offered by a third party, attainable on-demand, and ubiquitously over the Internet. (6) MIC (Many Integrated Core Architecture); Dedicated parallel coprocessor installable in common desktop computers, workstations, and servers.

In this section, we have focused our discussion on GPU and Cloud computing. GPUs are multi-core processors originally optimized for 3D rendering and image processing purposes. GPU devices are nowadays part of any desktop PC configurations and they can be programmed with general-purpose programming languages as well. These features make them easily accessible and cost-effective accelerator platform. Today’s CPUs have 8, 12, 16, or even 32 cores, while GPUs have up to 512 cores or more in a single chip. Today’s computing applications typically have data-intensive regions with scope of data computation parallelism. GPUs increase the speed of execution of parallel regions by spreading the calculation over hundreds of cores. This strategy delivers significant speedups for many applications.

Cloud computing delivers HPC in the form of a service similar to other forms of services that are already available in the cloud such as software as a service, platform as a service, and infrastructure as a service. HPC users derive benefit from the cloud in different aspects such as scalability, resources being available on-demand, fast, and inexpensive. On the other hand, moving HPC applications to the cloud have numerous challenges too. Few significant challenges are virtualization overhead in the cloud, multi-occupancy of resources, and network latency issues. Research is currently under progress to make HPC in the cloud a more realistic solution. Depending on the requirements, it may be much cheaper to use a public cloud's high-performance computing service rather than replacing the high-performance computing servers in the organization's data center, so the cloud providers are the ones buying lots of this high-end computation to keep up with the growing cloud demand.

4.5 GPU Computing

Graphics Processing Unit is used as a co-processor in GPU computing which accelerates the CPU's usage for scientific and engineering computing. Besides, it also accelerates the other programs running on the CPU by offloading some of the in-depth portions of the computational code thereby it reduces the time consumption. The user feels that the application is running faster because it's utilizing massively parallel processing power of the GPU to improve performance. A general-purpose CPU typically has multiple cores where multiple threads can be run, and a large cache for faster access to frequently used data, and also, sophisticated flow control techniques such as branch prediction, data and instruction pre-fetching, and out-of-order execution. Modern CPUs come with inbuilt floating-point ALUs which make them very useful for generic scientific and engineering computing tasks. In contrast, a GPU comprises a large number of execution units to process data in parallel. GPUs lack sophisticated control flow mechanisms similar to CPUs, however, they can run large numbers of threads, thereby providing large parallelism. If a program is split into many threads all doing the same computation on different data, GPU performance will be significantly faster than a CPU. On the other hand, if an application contains complex control flow, performance on CPU is going to be significantly faster than GPU. Programs that have leveraged the power of GPU computation have found applications in the field of Bioinformatics and Computational Biology such as sequence alignment, molecular dynamics, molecular docking, prediction and searching of molecular structures, spatial-temporal simulation, spectral analysis, cellular dynamics, genome-wide analysis, quantum chemistry, and Bayesian inference.

4.6 Parallel Programming Models

The availability of high-end computing hardware alone is not sufficient to improve computation performance. Compilers and APIs that exploit the parallel features of hardware should be available too. Parallel processing refers to the ability to tag the segment of the code that can be executed in parallel and subsequent execution across multiple processors or multiple cores. Broadly there are three types of parallel categories: (1) Shared memory systems, i.e, systems with multiple processing units attached to a common memory. (2) Distributed systems, i.e, systems comprising of multiple computing units, each having its processing unit and physical memory, that are connected with fast networks. (3) Graphic processor units (GPU), which are used as co-processors for solving computationally intensive problems.

4.6.1 OpenMP

OpenMP is a parallel programming paradigm that is best suited for writing parallel programs that are supposed to run on shared memory systems. It is not a programming language but an add-on to an existing language, usually FORTRAN or C/C++ made available through an application programming interface (API). API of OpenMP is a collection of (1) compiler directives, (2) run time callable functions, and (3) environment variables. Compiler directives of OpenMP instruct the compiler about the parallelism in the source code and provide instructions for generating the parallel code. Supporting functions permit programmers to control and utilize parallelism during the execution of a program. Environment variables permit the adoption of compiled programs to a particular parallel system. OpenMP program facilitates and coordinates various parallelization components, e.g. creating threads, distributing computation among threads, and synchronizing work among threads. OpenMP is significantly effective in parallelizing for-loops, where each thread is assigned to execute a single iteration of the loop. OpenMP is supported by open source community and major vendors, including AMD, IBM, Intel, Nvidia, Oracle.

4.6.2 Message Passing Interface (MPI)

Message Passing Interface (MPI) is a parallel programming model suitable for distributed memory architectures. Each processor has access to its memory and different processors are connected through high-speed communication interconnect. MPI provides a set of language-independent and platform-independent communication protocols for parallel computing, featuring point-to-point message passing as well as collective operations via user-specified processors. MPI standard prescribes facility for process creation and management, language bindings for C, C++ and

Fortran, point-to-point and collective communications, and group and communicator concepts. Processes are created by discrete processors/computing nodes executing different sections of the code. Each process gets its local variables and the memory space; the parallelism is achieved by establishing communications between processes by sending and receiving messages.

4.6.3 Compute Unified Device Architecture (CUDA)

Compute Unified Device Architecture (CUDA), developed by Nvidia is a programming model for facilitating general computing on its GPUs. CUDA comes to the advantage of programmers to accelerate the speed of compute-intensive applications by utilizing the computational features of GPUs for the parallelizable part of the computation. Software developers can access the CUDA platform through CUDA-accelerated libraries, compiler directives, and extensions to industry-standard programming languages including C, C++, and FORTRAN. CUDA makes of two abstractions, (1) Host—The CPU and its memory (host memory), (2) Device—The GPU and its memory (device memory) Basic steps in a GPU program are (1) Declaration and allocation of both the host and device memories, (2) Initialization of the host memory, (3) Transferring data from Host memory to device memory, (4) Executing GPU functions (aka kernels), (5) Transferring data back to the host memory.

4.6.4 Open Computing Language (OpenCL)

Open Computing Language (OpenCL) is an open-source standard for parallel programming for heterogeneous processors found in, servers, handset devices, personal computers, and embedded platforms. OpenCL describes constructs for writing programs that can execute across different platforms like central processing units (CPUs), graphics processing units (GPUs), and other types of processors or hardware accelerators. OpenCL standard consists of a set of APIs in C-like language to control a host processor and a variety of parallel devices and accelerators. A typical parallel application comprises of a C/C++ code for the host and a collection of kernels and special functions written in OpenCL for the accelerators. The parallelism is achieved at different levels, including SIMT (Single Instruction Multiple Threads), work-items, which are the smallest execution units, and work-groups in the order of increasing degree of coarse-grained parallelization level. Vendors such as Intel, AMD, NVIDIA, Altera Corp, Qualcomm, Samsung support OpenCL for their hardware.

4.6.5 Application of Parallel Programming Model and Tools in Binding Site Prediction and Docking

Knowledge of interactions of protein-ligand in the context of protein-ligand binding sites and ligand binding site residues is significant for understanding cellular mechanisms and is critical to understanding responses to drugs. These methods consider the knowledge about interaction energy and van der Waals (vdW) forces for binding site mapping. From a computational point of view, it represents a search problem where potential favorable binding sites need to be identified by scanning the entire protein surface. eFindSite (Feinstein and Brylinski 2016) is a program that employs OpenMP pragmas and dynamic workload balancing mechanism to launch parallel tasks for finding protein binding sites and residues. As reported earlier (Sánchez-Linares et al. 2012) BINDSURF leverages the power of CUDA based GPU parallel computing. BINDSURF divides the whole protein surface into independent regions called protein spots and then processes the divided spots in parallel. Such a procedure results in the screening of a large ligand database against the target protein over its whole surface simultaneously with docking simulations for each ligand being performed simultaneously using the massively parallel architecture of GPUs for all specified protein spots. As a result, new spots are found after analyzing the distribution of scoring function values over the entire protein surface. As per the study (Guerrero et al. 2011), the CUDA implementation of nonbonded interactions in parallel like electrostatics and van der Waals forces resulted in a speed of 260 times compared to the sequential version of the same algorithm. In another study (Guerrero et al. 2012) the parallelization effectiveness of the non-bonded electrostatic interactions kernel for Virtual Screening was benchmarked on three non-identical types of parallel architectures: a shared memory system, a distributed memory system, and a Graphics Processing Units (GPUs). Four combinations of implementations were implemented and tested based on MPI, OpenMP, Hybrid MPIOpenMP and CUDA programming models. The speed-up factor was significant to the sequential version for the aforementioned parallel modes: shared memory speed factor was 72× using OpenMP, for the MPI implementation and the Hybrid MPI-OpenMP implementation speed up factors were 60× and 229× respectively, and finally, speedup factor for CUDA implementation on the GPU architecture was 213× outperforming all the other implementations. AutoDock is a molecular modeling simulation program that is particularly effective for protein-ligand docking. AutoDock4.2.6 version has features implemented in OpenCL. The software utilizes the parallelism of its Lamarckian genetic algorithm (LGA) by processing ligand-receptor poses in parallel over multiple compute units. It targets platforms based on GPU as well as multi-core CPU accelerators. FRODRUG (Sánchez et al. 2014) is a method that has used spherical harmonic approximations method to increase the speed of the rotational part of the docking search by deploying in multi-core and GPU systems. Astex Diverse Set was used to benchmark the performance of the method. The speedup of GPU implementation compared to a single CPU core was 30 times eventually decreasing the docking time for a single ligand to only 50 ms. As per another study (Sánchez-Linares et al. 2011), the grid generation module of the program FlexScren was implemented using the CUDA language for the GPU architecture. For the ES and VDW grid calculations for proteins in the range of 1000 to 10,000 atoms, average speedups of up to 160 and 8 times respectively were obtained with high accuracy in double floating-point precision. Bristol university docking engine (BUDE) which is written in C++ and uses OpenMP and OpenCL is a general-purpose molecular docking program that utilizes GPU acceleration to perform (1) Virtual screening by docking of millions of small-molecule ligands, (2) Ligand binding site identification on protein surfaces, (3) Protein-protein docking in real space. MEGADOCK 4.0 is a structural bioinformatics software that implements FFT-grid-based protein-protein docking and exploits the advantage of the massively parallel CUDA architecture of NVIDIA GPUs and multiple computation nodes. MEGADOCK 4.0, is implemented using a combination of hybrid CUDA, MPI, and OpenMP parallelization.

4.7 Molecular Dynamics Simulation (MD)

MD is a computer simulation method that mimics the physical movements of atoms and molecules presenting the real environment and helps in providing detailed information on the fluctuations and conformational changes (Patodia et al. 2014) It plays a pertinent role in the theoretical study of the structure, dynamics, and thermodynamics of biological molecules and their complexes including the impact of solvent molecules. Because of the huge molecular system size generally, it is tedious to analyze such complex systems. Numerical methods in molecular dynamics simulation can be employed to circumvent such analytic intractability. During the simulation, the atoms are allowed to interact for a shorter period, which may help in computing their trajectory in and around the protein molecule thus providing intricate information about the individual motion of atoms as a function of time. Firstly, MD simulation assumes a given potential energy function i.e the energy function which allows us to calculate the force experienced by any atom given the positions of the other atoms and secondly, relies on Newton’s laws that tell us how those forces will affect the motions of the atoms. Using nuclear magnetic resonance (NMR), crystallographic, or homology-modeling data, an initial model of the system is obtained. The forces acting on each of the system atoms are estimated from force field parameters that comprise of bonded and non-bonded interaction terms. Bonded interactions include harmonic oscillator energy of bond lengths, bond angles, and sometimes improper dihedrals and torsional dihedral, while non-bonded interactions include Van der Waals and electrostatic interactions. Once the forces acting on individual atoms are obtained, a basic MD algorithm is followed: calculation of accelerations and velocities using classical Newton’s law of motion, updating the positions of every atom, and subsequently, calculation of forces applied on the investigated atom using inter-atomic potentials (Fig. 4.6).

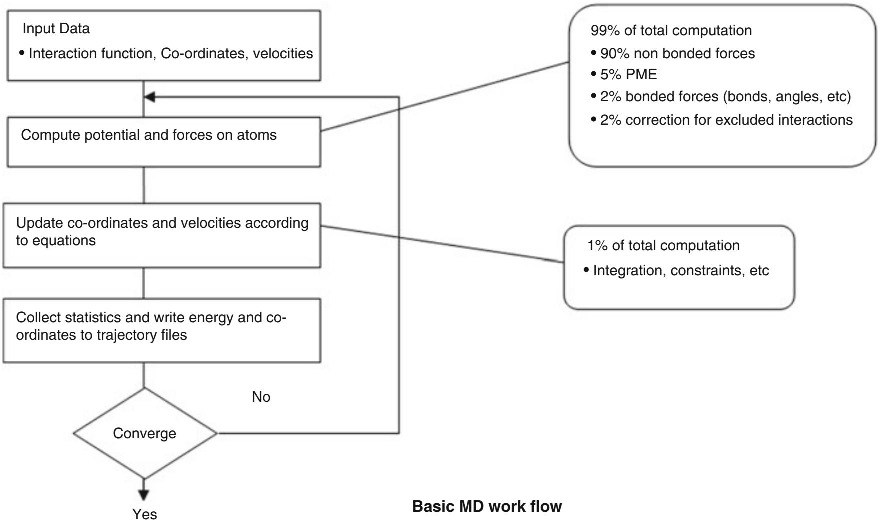

Input Data
• Interaction function, Co-ordinates, velocities
Compute potential and forces on atoms
Update co-ordinates and velocities according to equations
Collect statistics and write energy and co-ordinates to trajectory files
Converge
No
Yes
99% of total computation
• 90% non bonded forces
• 5% PME
• 2% bonded forces (bonds, angles, etc)
• 2% correction for excluded interactions
1% of total computation
• Integration, constraints, etc
Basic MD work flow

Fig. 4.6 Basic MD workflow. This figure represents the selection of initial conditions (positions, velocities), Selection of ensemble (NVE, NVT, NPT) Selection of target temperature, density/pressure. Performing simulation until equilibration is reached (property dependent), Performing production simulation run to collect thermodynamic averages, positions, velocities

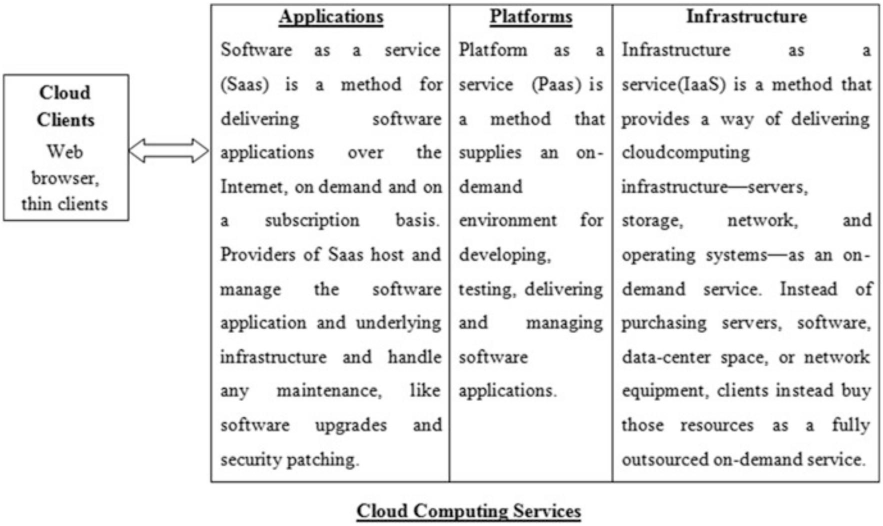
4.7.1 Applications of Advanced Computing in Molecular Dynamics Simulation (MD)

From the computational point of view, MD is significantly intensive due to (1) generation of millions to billions of time steps before converging, (2) a substantial amount of computation involved in each step, and dominated by nonbonded interactions. The use of the graphical processing unit (GPUs) and sophisticated algorithm-based optimization of energy calculation have significantly improved the performance of MD simulations. The technology involved in the present generation of computers takes the benefit of parallelism and accelerators to speed up the process. Modern simulation software packages, such as CHARMM, GROMACS, AMBER, and NAMD are compatible with the Message Passing Interface (MPI), which significantly facilitates the execution of complex tasks by software or program executing concurrently on multiple processors. As per the study (Ge et al. 2013) on three GPU based MD software AMBER, NAMD, and GROMACS, the speedup of GPU instance of GROMACS was about 3324 times than that of GROMACS PC instance. Significant speedups are reported for AMBER and NAMD as well. The latest version of AMBER running on a single workstation which contains single Titan-XP GPUs has achieved over 640 ns/day NVE benchmark. Consequent to an increase in the number of GPUs to 8, the accumulated throughput of over 5.1μs/day is possible.

4.8 Cloud Computing

Cloud computing is the facility to provide computing system resources, like data storage, computation power, various application functionality being made available on-demand, and without the need for management by the user. Cloud services are flexible wherein a user can limit the utilization of a service as they want at any given time, while the services are completely managed by the provider. The cloud allows customers to gain enhanced capabilities without investing upfront in new hardware or software. Instead, customers pay their cloud provider a subscription amount for only the resources they use. Users or Computing resources can be augmented on the fly. Some of the essential characteristics are (1) On-demand self-service, (2) Broad network access, (3) Resource pooling, (4) Rapid elasticity, and (5) Measured service. The array of available cloud computing services mostly fall into one of the following categories listed in Fig. 4.7. Each of the services provides varying degrees of control, flexibility, and management catering to the requirements of different categories of users.

Depending on how the cloud services are made available to users, Cloud computing models are classified into (1) Public cloud; this category of the model supports all users who want on a subscription basis to make use of a computing resource, such as hardware (OS, CPU, memory, storage) or software (application server, database), (2) Private cloud; this category of the model is used internally by an organization, (3) Hybrid cloud; an organization makes use of interconnected private and public cloud infrastructure that is distinct yet bound together,



Cloud Clients
Web browser, thin clients

Applications	Platforms	Infrastructure
Software as a service (Saas) is a method for delivering software applications over the Internet, on demand and on a subscription basis. Providers of Saas host and manage the software application and underlying infrastructure and handle any maintenance, like software upgrades and security patching.	Platform as a service (Paas) is a method that supplies an on-demand environment for developing, testing, delivering and managing software applications.	Infrastructure as a service(IaaS) is a method that provides a way of delivering cloudcomputing infrastructure—servers, storage, network, and operating systems—as an on-demand service. Instead of purchasing servers, software, data-center space, or network equipment, clients instead buy those resources as a fully outsourced on-demand service.

Cloud Computing Services

Fig. 4.7 Cloud computing services. This figure represents the cloud computing service model is divided into 3 broad categories (1) Applications (2) Platforms and (3) Infrastructure, each offering different levels of service abstraction for catering to the varying requirement of users

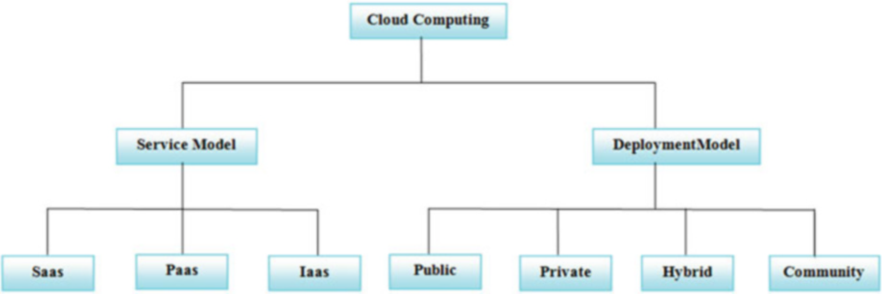


Fig. 4.8 Cloud computing architecture. This is based on the type of services it is offering as well as the environment where the user is located. It is divided into two models: (1) Service model which supports different types of services like Saas, Paas, and Iaas, whereas, (2) deployment model which includes Public, Private, Hybrid, and Community that supports the user depending on his access criteria

(4) Community cloud; this model supports multiple organizations sharing computing infrastructure and resources that are part of a community (Fig. 4.8).

4.8.1 Applications of Cloud Computing in Drug Discovery

The application of advanced molecular simulation techniques comes with the cost of additional advanced computational resources deployment and includes both hardware and software. Running advanced molecular simulation and analysis tasks in the Cloud eliminates the need for establishing in house advanced data centers or as well as access to high chargeable access to supercomputing resources. Consequently, it provides a cost-effective and practical solution for many modeling tasks for small and moderate size molecular dynamics simulations. As per the study (Guerrero et al. 2014), a comparison was done on the cost-effective performance of BINDSURF program in both HPC and cloud environment, processing 6000 different ligands. Each simulation had 5000 Monte Carlo steps. The results showed that the usage of local infrastructure should be significantly high, ranging between 50% and 100% so that local infrastructure sustainability is profitable; otherwise, cloud computing is a more cost-effective alternative. Cloud infrastructure is cheap upfront and expensive long term but provides flexibility, whereas setting local HPC infrastructure is expensive upfront but cheaper longer term, however, it is inflexible. Cost-benefit analysis has to be carried to decide from the two approaches.

The fact that the cloud computing platform provides access to HPC clusters through virtualization and abstractions of services is not abstracted enough. The end-users who do not have computing knowledge still need to understand and implement some details to configure an HPC cluster, install middleware and applications before making the system available for any scientific usage. A typical VMD/NAMD consists of (1) prepare simulation input data in VMD on a host computer; (2) locate an HPC cluster and book a slot; (3) transfer the prepared simulation input profile from the originating computer to the cloud-hosted HPC cluster and start the simulation; and (4) move the result back to the originating computer for analyzing in VMD.

VMD and NAMD have been integrated into a new software plugin for Amazon EC2 and this plugin also supports NAMD simulations on the Cluster Compute Instances. The plugin enables users to (1) quickly create an HPC cluster (2) submit a NAMD simulation from VMD to Amazon EC2 Plugin, and (3) transfer results from the Plug into the host computer running VMD for post-processing. Moreover, the Plug-in can integrate with the Interactive Molecular Dynamics (IMD) plug-in, which can make a standard simulation interactive and display the simulation in real-time. This implementation hides completely the details of any NAMD simulation deployment in the underlying HPC cloud.

AceCloud (Harvey and De Fabritiis 2015) is one solution that provides on-demand service for molecular dynamics simulation. It is designed to make possible secure execution of large group simulations on an external cloud computing service. The AceCloud client has been integrated into the ACEMD molecular dynamics simulation package. The client provides an interface that is easy to use and abstracts all aspects of interaction with cloud services. The user experiences that all simulations are running on their local machine, minimizing the time to learn the details associated with the usage of high-performance computing services. CovalentDock Cloud (Ouyang et al. 2013), is an algorithm to the model covalent binding. It is accessible directly online through a web server without any local installation and configuration. The application provides a user-friendly web interface to carry out covalent docking experiments and analysis online. The user has to enter the structures of both the ligand and the receptor or retrieve it from online databases with valid access id. AceCloud discovers the potential covalent binding patterns followed by carrying out the covalent docking experiments and eventually generates a visualization of the result for user analysis.

4.9 Conclusion

In the modern era, apart from reducing time and financial cost, advanced computing has proved their crucial role by identifying the various repurposing existing drugs which improves the quality of pharmaceutical research. Still now nearby 43 Pharmaceutical companies applied automated algorithms and parallelism techniques in the drug discovery process and showed the conquest outcomes. As it reduces the expenses of the Research and Development department of various Pharmaceutical companies, in the future they are planning to still improve the computing techniques whereby autonomous identification of promising drugs and novel pathways can be done with zero human interference. To extend the benefit of advanced computing they are also applied in the areas of personalized medicine based on the next-generation sequencing data. Thus in the future, the whole drug discovery process is expected to be in the hands of intelligence algorithms and high-performance techniques.

Acknowledgments The authors acknowledge the Department of Biotechnology, Govt. of India for providing the necessary computing facilities. Further, one of the authors (AM) thanks the Kothari Post-doctoral fellowship (PDF) funded by the University Grants Commission (UGC). The author (SB) thanks to the Council of Scientific and Industrial Research (CSIR) for Senior Research Fellowship.

References

Anderson, A. C. (2003). The process of structure-based drug design. *Chemistry and Biology, 10*(9), 787–797.

Antel, J. (1999). Integration of combinatorial chemistry and structure-based drug design. *Current Opinion in Drug Discovery and Development, 2*(3), 224–233.

Brenke, R., Kozakov, D., Chuang, G. Y., Beglov, D., Hall, D., Landon, M. R., et al. (2009). Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics, 25*(5), 621–627.

Edla, D. R., Jana, P. K., & Member, I. S. (2012) A prototype-based modified DBSCAN for gene clustering. *Procedia Technology*. 6(1), 485–492. https://doi.org/10.1016/j.protcy.2012.10.058

Feinstein, W., & Brylinski, M. (2016). Structure-based drug discovery accelerated by many-core devices. *Current Drug Targets, 17*(14), 1595–1609.

Ge, H., Wang, Y., Li, C., Chen, N., Xie, Y., Xu, M., et al. (2013). Molecular dynamics-based virtual screening: Accelerating the drug discovery process by high-performance computing. *Journal of Chemical Information and Modeling, 53*(10), 2757–2764.

Guerrero, G. D., Perez-S, H. E., Cecilia, J. M., & Garcia, J. M. (2012). Parallelization of virtual screening in drug discovery on massively parallel architectures. In *20th euromicro international conference on parallel, distributed and network-based processing* (pp. 588–595). Piscataway, NJ: IEEE.

Guerrero, G. D., Pérez-Sánchez, H., Wenzel, W., Cecilia, J. M., & García, J. M. (2011). Effective parallelization of non-bonded interactions kernel for virtual screening on GPUs. In M. P. Rocha, J. M. C. Rodríguez, F. Fdez-Riverola, & A. Valencia (Eds.), *5th International conference on practical applications of computational biology and bioinformatics (PACBB 2011) advances in intelligent and soft computing* (Vol. 93, pp. 63–69). Berlin, Germany: Springer.

Guerrero, G. D., Wallace, R. M., Vázquez Poletti, J. L., Cecilia, J. M., García, J. M., Mozos, D., et al. (2014). A performance/cost model for a CUDA drug discovery application on physical and public cloud infrastructures. *Concurrency and Computation: Practice and Experience, 26*(10), 1787–1798.

Hameed, P. N., Verspoor, K., Kusljic, S., & Halgamuge, S. (2018). A two-tiered unsupervised clustering approach for drug repositioning through heterogeneous data integration. *BMC Bioinformatics, 19*(1), 129. https://doi.org/10.1186/s12859-018-2123-4

Harvey, M. J., & De Fabritiis, G. (2015). AceCloud: Molecular dynamics simulations in the cloud. *Journal of Chemical Information and Modeling, 55*(5), 909–914.

Heikamp, K., & Bajorath, J. (2014). Support vector machines for drug discovery. *Expert Opinion on Drug Discovery, 9*(1), 93–104.

Huang, E. S., Koehl, P., Levitt, M., Pappu, R. V., & Ponder, J. W. (1998). Accuracy of side-chain prediction upon near-native protein backbones generated by ab initio folding methods. *Proteins: Structure, Function, and Bioinformatics, 33*(2), 204–217.

Huang, H. J., Yu, H. W., Chen, C. Y., Hsu, C. H., Chen, H. Y., Lee, K. J., et al. (2010). Current developments of computer-aided drug design. *Journal of the Taiwan Institute of Chemical Engineers, 41*(6), 623–635.

Kwon, S., Bae, H., Jo, J., & Yoon, S. (2019). Comprehensive ensemble in QSAR prediction for drug discovery. *BMC Bioinformatics, 20*(1), 521. https://doi.org/10.1186/s12859-019-3135-4

Lind, A. P., & Anderson, P. C. (2019). Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties. *PloS One, 14*(7), e0219774. https://doi.org/10.1371/journal.pone.0219774

Lipinski, C., Maltarollo, V., Oliveira, P., da Silva, A., & Honorio, K. (2019). Advances and perspectives in applying deep learning for drug design and discovery. *Frontiers in Robotics and AI, 6*, 108. https://doi.org/10.3389/frobt.2019.00108

Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews, 23*(1-3), 3–25.

Liu, Z., Du, J., Fang, J., Yin, Y., Xu, G., & Xie, L. (2019). DeepScreening: A deep learning-based screening web server for accelerating drug discovery. *Database, 2019*, baz104. https://doi.org/10.1093/database/baz104

Lo, Y. C., Ren, G., Honda, H., & Davis, K. L. (2019). Artificial intelligence-based drug design and discovery. In *Cheminformatics and its applications*. New York: IntechOpen. https://doi.org/10.5772/intechopen.89012

Madhukar, N. S., Khade, P. K., Huang, L., Gayvert, K., Galletti, G., Stogniew, M., et al. (2019). A Bayesian machine learning approach for drug target identification using diverse data types. *Nature Communications, 10*(1), 1–14.

Mehmood, T., Liland, K. H., Snipen, L., & Sabo, S. (2012). A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory Systems, 118*, 62–69.

Mishra, S. (2009). Function prediction of Rv0079, a hypothetical Mycobacterium tuberculosis DosR regulon protein. *Journal of Biomolecular Structure and Dynamics, 27*(3), 283–291.

Najafi-Ghobadi, S., Najafi-Ghobadi, K., Tapak, L., & Aghaei, A. (2019). Application of data mining techniques and logistic regression to model drug use transition to injection: A case study in drug use treatment centers in Kermanshah Province, Iran. *Substance Abuse Treatment, Prevention, and Policy, 14*(1), 55. https://doi.org/10.1186/s13011-019-0242-1

Ouyang, X., Zhou, S., Ge, Z., Li, R., & Kwoh, C. K. (2013). CovalentDock cloud: A web server for automated covalent docking. *Nucleic Acids Research, 41*(W1), 329–332.

Patodia, S., Bagaria, A., & Chopra, D. (2014). Molecular dynamics simulation of proteins: A brief overview. *Journal of Physical Chemistry and Biophysics, 4*(6), 1000166. https://doi.org/10.4172/2161-0398.1000166

Popova, M., Isayev, O., & Tropsh, A. (2018). Deep reinforcement learning for de novo drug design. *Science Advances, 4*(7), eaap7885. https://doi.org/10.1126/sciadv.aap7885

Prajapat, P., Agarwal, S., & Talesara, G. L. (2017). Significance of computer-aided drug design and 3D QSAR in modern drug discovery. *Journal of Medicinal Chemistry, 1*(1), 1.

Sánchez, S. G., Aportela, E. R., Garzón, J. I., Chacón, P., Montemayor, A. S., & Cabido, R. (2014). FRODRUG: A virtual screening GPU accelerated approach for drug discovery. In *2014 22nd Euromicro international conference on parallel, distributed, and network-based processing* (pp. 594–600). Piscataway, NJ: IEEE.

Sánchez-Linares, I., Pérez-Sánchez, H., Cecilia, J. M., & García, J. M. (2012). High-throughput parallel blind virtual screening using BINDSURF. *BMC Bioinformatics, 13*(Suppl 14), S13. https://doi.org/10.1186/1471-2105-13-S14-S13

Sánchez-Linares, I., Pérez-Sánchez, H., Guerrero, G. D., Cecilia, J. M., & García, J. M. (2011September). Accelerating multiple target drug screening on GPUs. In *Proceedings of the 9th international conference on computational methods in systems biology* (pp. 95–102). New York: ACM.

Sarkate, P. A., & Deorankar, A. V. (2018). Classification of chemical medicine or drug using K nearest neighbor (KNN) and genetic algorithm. *International Research Journal of Engineering and Technology, 5*(3), 833–834.

Taguchi, Y. H., Iwadate, M., & Umeyama, H. (2015). Principal component analysis-based unsupervised feature extraction applied to in silico drug discovery for posttraumatic stress disorder-mediated heart disease. *BMC Bioinformatics, 16*(1), 139. https://doi.org/10.1186/s12859-015-0574-4

Veber, D. F., Johnson, S. R., Cheng, H. Y., Smith, B. R., Ward, K. W., & Kopple, K. D. (2002). Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry, 45*(12), 2615–2623.

Yosipof, A., Guedes, R. C., & García-Sosa, A. T. (2018). Data mining and machine learning models for predicting drug-likeness and their disease or organ category. *Frontiers in Chemistry, 6*, 162. https://doi.org/10.3389/fchem.2018.00162

Zhang, H., Kang, Y. L., Zhu, Y. Y., Zhao, K. X., Liang, J. Y., Ding, L., et al. (2017). Novel naïve Bayes classification models for predicting the chemical Ames mutagenicity. *Toxicology in Vitro, 41*, 56–63.

Chapter 5
Protein Structure, Dynamics and Assembly: Implications for Drug Discovery

Arangasamy Yazhini, Sohini Chakraborti, and Narayanaswamy Srinivasan

Abstract Most of the therapeutic drugs available in the market today, are targeted against proteins. Drug molecules are designed to complement shape, size and electrostatic fingerprints of the functional site of a target protein so that they can bind to the protein and impede its molecular function. Details of functional site are derived from 3-D structure of the protein obtained either through experimental techniques or computational protein modeling and form the basis for structure-based drug design. Knowledge derived from homologous proteins facilitates this process by providing an understanding on common and unique features of the intended target with respect to its close and distant relatives. This helps to design a drug with high selectivity and affinity. Often inherent dynamic nature of proteins facilitates inter-protein interactions and aid them to perform major cellular activities as an assembled complex. With improved apprehension of structural biology, consideration of multi-protein machineries and their associated conformational dynamics is increasingly gaining importance in drug design and discovery. Susceptibility of protein-protein interactions in disease conditions is progressively being realized and this has attracted protein-protein interfaces as potential drug targets for therapeutic intervention in the last few decades. In this chapter, we have discussed the properties of protein structure, evolution, dynamics and protein complexes along with explanations on how each factor contributes to the design of an effective drug molecule that is safe and efficacious.

Keywords Protein structure · Evolution · Dynamics · Assembly · Protein-protein interactions · Drug design

A. Yazhini · S. Chakraborti · N. Srinivasan (✉)
Molecular Biophysics Unit, Indian Institute of Science, Bangalore, Karnataka, India
e-mail: ns@iisc.ac.in

S. K. Singh (ed.), *Innovations and Implementations of Computer Aided Drug Discovery Strategies in Rational Drug Design*,
https://doi.org/10.1007/978-981-15-8936-2_5

5.1 Introduction

Proteins are workforces of cells. All the efforts in DNA recombinant techniques, genome sequencing, structural genomics, etc., are eventually aiming for the holistic knowledge of how proteins in the organism function and how to manipulate them for human benefits. Proteins being the active rulers of cellular biomolecules, they use nucleic acids to synthesize themselves, produce energy from carbohydrates and guide lipids for energy storage. Therefore proteins play a dominant role in the cell and their design by the nature should be such that proteins fit with successful life stages of the organism during the course of divergent evolution (Pál et al. 2006). At the atomic level, proteins comprise of thousands of atoms. To decipher recognizable patterns, protein structures are described in hierarchical levels *viz.* primary, secondary, tertiary and quaternary. Further, proteins are intrinsically dynamic which is influenced by aqueous surroundings and molecular crowding in the cellular milieu. Proteins work independently and/or along with other biomolecules by forming complexes and large assemblies. Hence, function of the protein depends on the structure, dynamics and its assembly state in the cell.

Perturbation of the aforementioned factors could alter molecular recognition between a protein and its interacting partners and may cause disease. In many cases, undesirable modulation of protein function (which plays an important congenial role for the disease condition) can be treated by administration of suitable therapeutic agents (drugs) to alleviate the disease condition (Peng et al. 2019). Due to the superior druggable properties of proteins among the kinds of biomolecules (*viz.* lipids, nucleic acids, carbohydrates etc.), most of the therapeutic drugs available today are targeted against proteins. Receptors, enzymes, ion channels and transporters are the group of proteins that are predominantly targeted by drug molecules (Santos et al. 2016).

In this chapter, we describe the characteristics of protein structure, dynamics and molecular complexes formed by protein-protein interactions. Understanding of protein 3-D structure provides insights into characteristics of binding pocket of a drug target. Such information is essential in drug design and discovery. Here, we highlight a few successful examples of protein structure-based drug design (SBDD) approach. We discuss how knowledge derived from homologs contributes to the identification and characterization of target protein and for drug repurposing. The relationship between protein dynamics and function is further explained with implications in protein-drug interaction studies. Subsequently, the properties of protein complexes and their applications in drug discovery process are exemplified.

5.2 Protein Structure

Primary structure of a protein governs the folding and functional tertiary structure through formation of hydrogen bonds and other kinds of non-covalent interactions among amino acid residues (Anfinsen 1973). A globular 3-D structure comprises of stable secondary structures, namely α-helices and β-sheets that are characterized by series of hydrogen bonds. These regular secondary structures are connected by irregular turns and loops. In stable and compact three-dimensional (3-D) structure, individual elements of secondary structure are arranged with one another through tight packing of amino acid side chains. Depending on the composition of secondary structures and their topology, different proteins may have different globular 3-D structures (fold). Regardless of the nature of fold, water soluble proteins inevitably possess a hydrophobic core as a result of clustering apolar residues in the interior and a surface populated with polar residues.

An optimally folded 3-D structure holds functional residues at precise positions in the active site which is essential for the protein to carry out its dedicated biochemical functions. For example, alcohol dehydrogenase adopts GroES-like fold with 18 regions in α-helical conformation and 23 regions in β-strand conformation (Fig. 5.1) (Li et al. 1994). Substrate binding site, co-enzyme (NAD^+) binding residues are placed in appropriate positions in the 3-D structure of this protein to catalyze the conversion of variety of alcohols to respective aldehydes. Variations in its amino acid sequence could affect the 3-D structure and consequently influence substrate affinity and catalytic efficiency (Colby et al. 1998; Edenberg 2007). Hence, protein 3-D structure is a key determinant of protein function by providing suitable physicochemical environment for the active site and substrate binding residues.

In order to derive atomic details of protein 3-D structure, several experimental methods or computational methods are used. X-ray crystallography is a common

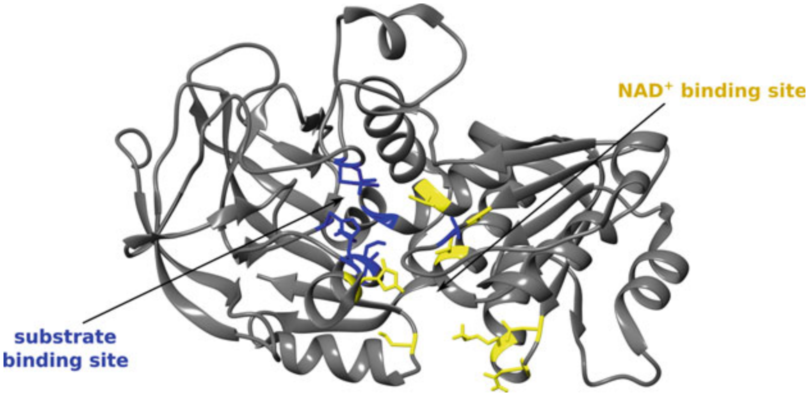


Fig. 5.1 Example of protein structure. Cartoon representation of alcohol dehydrogenase structure with annotation of critical residues (PDB code: 1ADG) (Li et al. 1994). Stick representation highlights substrate binding residues in blue, co-enzyme (NAD^+) binding residues in yellow

technique to determine 3-D structure of proteins (Blundell and Johnson 1976). This technique has significantly contributed to the availability of experimentally determined structures which are deposited in the protein structure repository (Protein Data Bank or PDB). Since mid-1980s, nuclear magnetic resonance (NMR) technique too is being employed to determine protein structure. It provides an ensemble of structures representing slightly different conformations of protein that are adopted under solution condition and hence helps us to study the dynamical nature of protein (Wuthrich 1995). Recently, cryo-electron microscopy (cryo-EM) has emerged as another technique to determine structure of proteins and biomolecular assemblies. Cryo-EM has enabled us to study the structural details of large protein-biomolecular complexes (Subramaniam et al. 2016; Cheng 2018).

Despite the development of these experimental methods, 3-D structure for innumerable proteins with known amino acid sequences available from exponentially growing data of next generation sequencing, is still unknown. Structure determination of such large number of proteins by experimental methods with an equal pace to sequence data is currently an unattainable task. In addition, some proteins may not be amenable to experimental techniques due to their incompatible nature such as intrinsic dynamics and size limitations for experimental conditions. To address this, computational methods have been developed to predict 3-D structure of proteins. Homology modeling or comparative modeling is a widely used method to generate 3-D structure of a protein. It uses structural information from homologous protein of known structure as a template and models protein 3-D structure (this approach has been explained in Sect. 5.3.2). Threading method based on a library of information about solvent accessibility, secondary structural state and neighbor contacts of known folds or *ab initio* modeling based on thermodynamics and potential energy landscape can also be employed to predict protein structure (Baker and Sali 2001). Recently, the application of artificial intelligence in protein structure prediction based on co-evolution and residue contact distance potential has resulted in a method called ‘alpha-fold’. It has the potential to build 3-D structure of proteins and to recognize plausible new protein folds (Senior et al. 2020). Thus, using experimental techniques or computational tools, knowledge on protein structures could be derived that is essential to understand the molecular basis of protein function.

In the following section, we have discussed the use of protein structure in drug design process for therapeutic treatment.

5.2.1 Use of Protein Structures in Rational Drug Design and Discovery

Function of a protein in cellular context largely depends on its interactions with other molecules, broadly referred to as ligands. These ligands can be macromolecules like other proteins, DNA, RNA and/or small endogenous molecules like neurotransmitters and ions or other organic molecules like carbohydrates. The feasibility and strength of such molecular recognitions mediated by binding of the partner molecules (i.e., protein and ligand) are regulated by complementarity in shape and electrostatic features of the interacting regions on the surface of each molecule (Sowdhamini et al. 1995; Voet et al. 2013). These features, i.e., shape and electrostatic fingerprints, are presented by the arrangement of the amino acid residues in the 3-D structure of proteins.

A drug molecule intended to bind to a protein target modulates the function of the target by altering its interactions with endogenous modulators. The drug molecules targeting the orthosteric sites are designed in such a way that they mimic the shape and electrostatic features of the endogenous modulators and can complement the protein binding site, thereby engaging the protein in interactions similar to that of the endogenous ligand. Understanding the features of the binding cavity helps the drug designers to decorate the chemical scaffold of the ligand molecule with relevant functional groups so that the interaction between the protein and the designed molecule is optimal. For example, if the protein binding site harbours a hydrogen bond donor (HBD), placement of a hydrogen bond acceptor (HBA) at a suitable position on the designed molecule, is likely to optimize the protein-ligand interactions. Similarly, when there is an aromatic amino acid residue in the binding cavity, rational introduction of aromatic/cationic substituents on the ligand is generally exploited to utilize the strength offered by π-π stacking or cation-π interactions, respectively. Likewise, the placement of polar and non-polar substituent on a ligand such that these groups are proximal to the polar and non-polar sub-pockets of a binding cavity respectively would also aid in optimizing protein-ligand interactions. It should be noted that such optimization of molecular interactions also considers that two interacting species should never come too close to each other which can result in steric clashes. Thus, if a sub-pocket of the protein binding site houses a residue with bulky side-chain, a small substituent of appropriate size and shape needs to be placed on the ligand's scaffold.

How favourable is the interaction between a protein and a ligand, is quantified in terms of free energy change (ΔG) associated with the binding event. A spontaneous binding would result when $\Delta G < 0$, indicating favourable accommodation of the ligand in the desired binding cavity of the protein. Better the complementarity in the shape and electrostatic features of the binding partners, stronger is the binding affinity and results in lower value of ΔG (Patrick 2013). It can thus be well comprehended that information (on shape, size and electrostatic features of the binding cavity) derived from 3-D structures of a protein play a very important role in designing the right ligand that fits favourably within the intended protein binding cavity and elucidate the desired therapeutic benefits. This is an interesting area and is commonly referred to as structure-guided/structure-based drug design or SBDD which is primarily driven by computational techniques (Batool et al. 2019) (Fig. 5.2). In the next section, we have provided a few examples of successful application of SBDD.

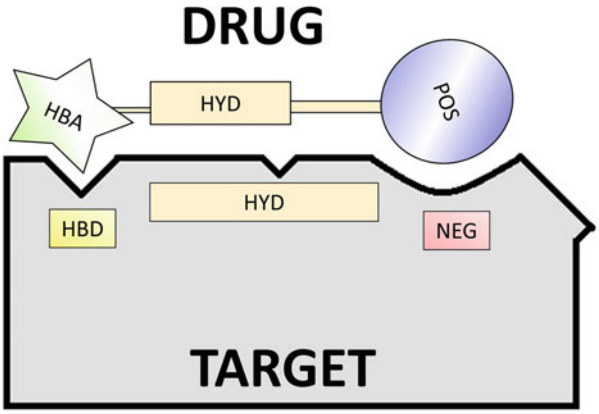


Fig. 5.2 Drug-target complementarity. This figure shows a schematic representation of complementarity in shape and chemical features of a target protein binding site (grey; bottom) with a bound drug (top) that enables a good fit. The HBA on the ligand complements the HBD region in the protein binding site. Similarly, the hydrophobic (HYD) and negatively charged (NEG) regions on the protein binding site are complemented by HYD and positively charged groups (POS) on the drug, respectively. This figure has been generated using Microsoft Powerpoint software

5.2.2 Overview of Drug Discovery Program and Few Successful Examples of SBDD

The conventional drug discovery process takes nearly 15 years from the time of target identification followed by target validation to its launching in the market after successful clinical trials and it requires huge investments (~$one million dollars) (Dimasi et al. 2010; DiMasi et al. 2016). With the advancement of computational technology, SBDD has now become an integral part of any drug discovery program which aims to reduce the overall timeline and investments required (Mohs and Greig 2017). In general, the first step in drug discovery programs involves the identification and validation of a target to ensure that modulation of function of the identified target would aid in addressing the disease condition. Once the target is validated, the hunt for suitable ligands which would be able to bind to the target of interest begins. SBDD principles play a vital role at this stage. The most popular technique that is applied at the initial stage is molecular docking of large libraries of compounds to identify the potential binders from a pool of non-binders (Abraham et al. 2010). Thus, the primary requirement for docking exercise is the availability of a reliable 3-D structure of the target protein with information on the binding site. Experimentally determined structures through X-ray crystallography (Petsko and Ringe 2010; Maveyraud and Mourey 2020), Nuclear Magnetic Resonance (Sugiki et al. 2018), Cryo-electron microscopy (Ceska et al. 2019) etc. serve as an important starting point for SBDD. In the absence of a reliable experimental structure, computationally generated models of the target protein are used (França 2015). In the recent times, artificial intelligence/machine learning based techniques are also gaining popularity in the identification and generation of potential hits and leads (Mak and Pichika 2019; Vamathevan et al. 2019). The potential binders predicted from computational studies (*in silico* hits) are then taken forward for experimental testing and the ones (experimental hits) which show promising results are subjected to multiple cycles of optimization to generate leads with improved pharmacokinetic and pharmacodynamic properties by taking inputs from synthetic chemists, biochemists, pharmacologists, formulation scientists, regulatory experts and many more group of scientists. A lead that passes all the criteria of preclinical phase is then considered as a drug candidate eligible for clinical trials in humans. With successful completion of clinical trials and approval from regulatory agencies, the drug can then make its entry into the market (McNamee et al. 2017; Mohs and Greig 2017). Thus, a successful SBDD is an iterative process that involves contributions from scientists of diverse expertise. Indeed, a typical drug discovery program in the modern era, definitely uses the principles of SBDD at some point of time during the entire cycle.

However, it is difficult to find plenty of well-documented scientific reports in the public domain which discusses the development of drug that is primarily centred around SBDD and have successfully made into the clinics. This, to a certain extent can be due to the limitations associated with intellectual property rights. Also, assimilating such data for a scientific report, would largely depend on the diminishing recall of the entire crew of scientists involved in the project over a long time span (~15 years or more) during the discovery, development and clinical studies of the particular drug. Moreover, it is not easy to deconvolute the factors attributing to the success of a drug discovery program and claim any one methodology (like SBDD only) to be the sole key to success unless it is predominantly obvious. This is so because, as it has already been mentioned, a successful drug discovery program is a result of combined efforts of a huge team of scientists with diverse expertise. For brevity, here, we have discussed only a few of the many well-known examples of successful SBDD programs.

The first successful case of SBDD dates back to the last decade of twentieth century reporting a series of Food and Drug Administration (FDA) approved Human immunodeficiency virus-1 (HIV-1) protease inhibitors (indinavir, saquinavir, etc.) (Roberts et al. 1990; Wlodawer and Vondrasek 1998; Ghosh et al. 2016). Other examples of successful story of SBDD involves the discovery of norfloxacin (antibiotic) (Rutenber and Stroud 1996), isoniazid (anti-tubercular) (Marrakchi et al. 2000), flurbiprofen (nonsteroidal anti-inflammatory agent) (Miller et al. 2015; Dadashpour et al. 2015), amprenavir (anti-HIV) (Wlodawer and Vondrasek 1998; Clark 2006), raltitrexed (anti-metabolite) (Anderson 2003). A major breakthrough in SBDD is the discovery and approval of imatinib (BCR-ABL kinase inhibitor) in the year 2001 for the treatment of chronic myelogenous leukaemia (Iqbal and Iqbal 2014). The history of discovery of imatinib has been a great example of learning for all drug designers and therefore, no doubt why it is still called a 'wonder drug'. The underlying science behind the discovery of imatinib showed that how information on conformational state (i.e., the inactive state which are less conserved among the different kinases as compared to the active state) of the target protein derived from its 3-D structure can be helpful to understand the selectivity of the investigating

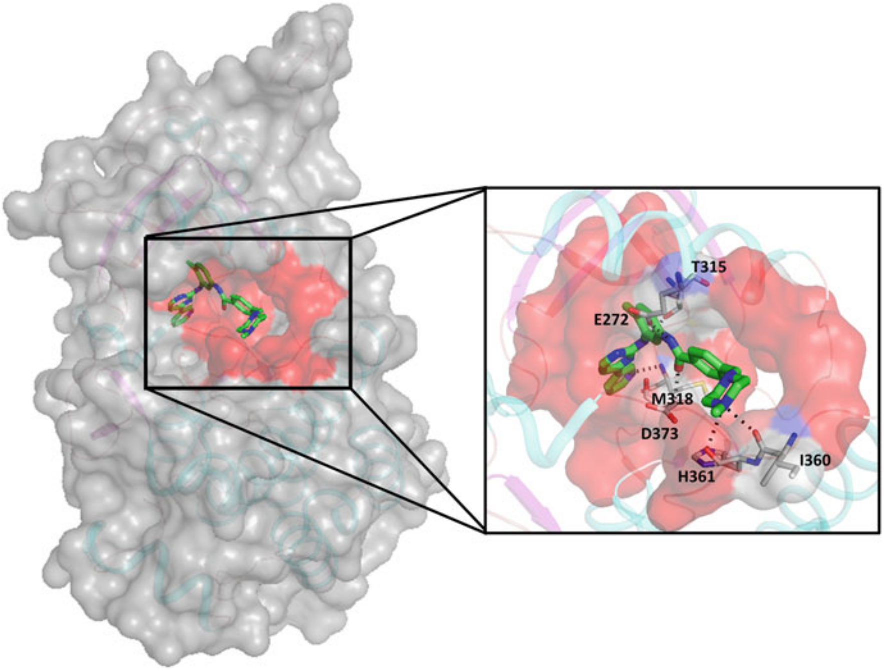


Fig. 5.3 Imatinib bound to inactive conformation of tyrosine kinase ABL complex. The protein tyrosine kinase ABL is shown as transparent surface representation (grey) and the ligand (imatinib) is shown as green (carbon atoms) stick (left panel). The binding site of imatinib is indicated in red colour. The protein backbone is depicted in thin cartoon representation where the helices are in cyan, beta sheets are in pink, and the loops are in orange. Enlarged view of binding site of imatinib (right panel). The protein residues which offer polar contacts (shown as black dashes) to imatinib are shown in stick representation (white carbon) and are labelled in black font. The image has been generated using the PDB entry 1IEP in PyMOL (The PyMOL Molecular Graphics System, Schrödinger, LLC)

molecule and therefore design a safer drug (Capdeville et al. 2002) (Fig. 5.3). There are many other examples of successful stories of SBDD as reviewed elsewhere (Batool et al. 2019). Notably, the success of a drug molecule does not solely lie in understanding the structure of its intended target. Knowledge of structures of other closely related proteins and evolutionarily aspects of structural features are also crucial in drug design as discussed in the subsequent sections.

5.3 Applications of Knowledge Derived from Homologous Proteins in Drug Design

Proteins that have diverged from common ancestor are referred as homologs. Homologous proteins retain much of structural and functional features. This feature of homologous proteins comes with both advantages and disadvantages in the

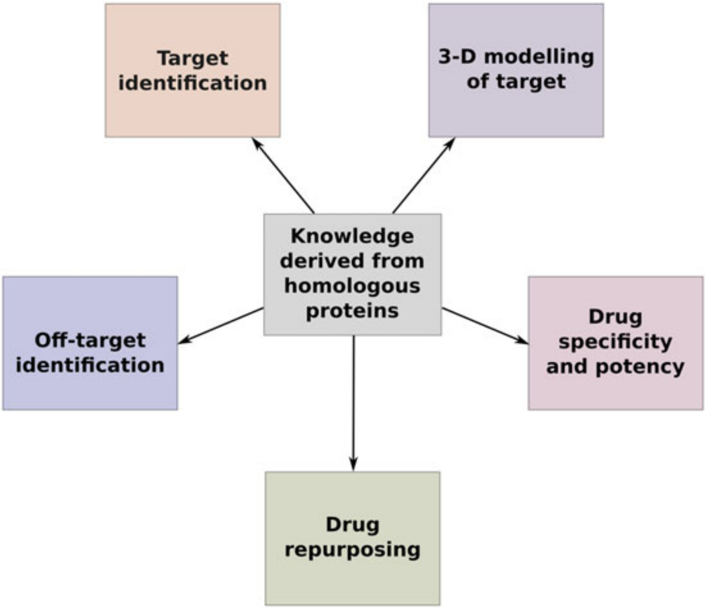

Target identification
3-D modelling of target
Knowledge derived from homologous proteins
Off-target identification
Drug specificity and potency
Drug repurposing

Fig. 5.4 Flow chart depicting applications of knowledge derived from homologous proteins in drug design

context of drug design and development. In the one hand, conservation of binding site properties of homologous proteins could end up in off-target binding of a drug causing adverse side effects. On the other hand, one can exploit the features of protein homology in repurposing drugs when the known target of a drug for a disease is homologous to a putative target in the context of a different disease. This section highlights the application of knowledge derived from homologs in SBDD, in understanding rationale behind off-target effects of a drug and in drug repurposing (Fig. 5.4).

5.3.1 Homology Detection in the Identification of Potential Drug Targets in Pathogens

Identification of homologs is an effective approach for drug target identification. Proteins that are essential for the organisms' survival is conserved among closely related species (Jordan et al. 2002). In general, for the treatment of infectious diseases, targeting proteins that are essential to the pathogen for its survival could be an effective way for managing therapeutic conditions. Homolog recognition helps in this process to identify proteins that are conserved among related pathogenic organisms which in turn, indicate their essentiality for the life-cycle of targeted pathogen. At the same time, the identified target should ideally have no homolog in the host system which further requires homolog recognition in host genome to confirm that the target is unique to only pathogen. Such approach referred as 'subtractive genomics' is commonly used for drug target identification of pathogenic infections (Rath et al. 2016; Sudha et al. 2019). Similarly, in case of newly detected pathogen, homolog search is used to understand whether the pathogen is similar to any known clinical pathogen. If such relationship could be detected, conserved proteins are further probed for identifying suitable drug targets in the newly emerged pathogen. For example, main protease of the SARS-CoV-2 (SARS-CoV-2 M^{pro}) has been identified as a potential drug target for the treatment of the ongoing pandemic, COVID-19. Based on the evolutionary relationship studies, it has been found that SARS-CoV-2 M^{pro} is closely related to main protease of SARS-CoV. The enzyme is known to be critical in maintaining viral life-cycle. Notably, active site residues of this protein is conserved among all closely related coronaviruses and does not have a homolog in human (Anand et al. 2003; Needle et al. 2015). Hence, targeting this protein is likely to be an effective strategy to kill the virus with least adverse effects to the human host. Earlier studies have already shown that main protease is a good drug target for treating SARS infection (Dai et al. 2020).

5.3.2 3-D Modeling of Drug Targets Using Structures of Homologous Proteins

Homology detection is immensely useful in preliminary stages of drug design. If a target protein does not have an experimentally determined 3-D structure, identification of homolog with known 3-D structure is helpful to obtain 3-D structure. With growing deposition of experimental protein structures in the PDB (Berman et al. 2002), sequence of drug target whose structure is unknown can be searched in the database using algorithms such as BLAST (Hu and Kurgan 2019). If homolog is detected with significant sequence similarity, comparative/homology modeling is widely employed to model 3-D structure of the target protein using structure(s) of the identified homolog(s) (Šali and Blundell 1993). A reliable 3-D model of target protein forms the basis for virtual screening and molecular docking (Muhammed and Aki-Yalcin 2019). Homology modeling provides a reliable 3-D structure of the target protein when it shares high sequence identity with homologous proteins (>60%) and use of multiple homologous structures have shown to improve the model quality of both structure and dynamics features of proteins (Yazhini and Srinivasan 2020). Whereas in case of poor sequence identity between target protein and homologous protein, i.e. below 30%, the atomistic details of the generated 3-D structure should be treated cautiously (Kryshtafovych and Fidelis 2009). Application of homology models in SBDD is particularly useful for proteins which are difficult to be solved by any experimental technique. For example, membrane proteins like G-protein coupled receptors (GPCRs) are difficult to be crystallized. Even though more than 800 GPCRs are encoded by human genome and these receptors form the leading drug targets, experimentally determined structures of GPCRs are available only for a small fraction of the total space of druggable human GPCR targets. Therefore, the GPCR community largely depend on homology models for SBDD strategies (Cavasotto and Palomba 2015).

5.3.3 Homology and Drug Promiscuity

For a molecule to be a successful drug, it should have high efficacy and low toxicity. Higher efficacy of a drug would result from its high binding affinity toward the intended protein target and lower toxicity which means that the drug binds weakly to off-targets. Now, one can have an obvious question that when a drug is designed to bind to a particular target, why would it bind to off-targets/undesired targets? The answer to this question again lies in understanding the 3-D structure of the protein. In the course of evolution, protein structures are conserved better than sequences (Chothia and Lesk 1986). Therefore, nature often repurposes protein folds leading to retention of global structures and sometimes even local structures. Hence, by virtue of similarity in local structures (shape and electrostatic features of binding sites), many proteins might be capable of offering reasonably favourable, but undesirable interactions to a given drug molecule (even though the drug molecule was not designed to bind to these undesired targets) and leading to promiscuous binding that results in adverse effects.

For example, protein kinases are a group of evolutionary related enzymes that phosphorylates serine/threonine/tyrosine residues of their substrate proteins to mediate multitude of signaling processes and hence become the second most attractive drug targets after GPCRs. However, the main challenge in targeting kinase is to achieve specificity. Since all typical serine-threonine kinases are known to adopt conserved bi-lobal structure with common motifs for ATP and substrate binding, the chances of cross-reactivity among different kinases for a given small molecule drug become the leading issue in most drug discovery programs aiming to target protein kinases (Shah et al. 2013). Hence, careful analysis of the binding site features is required to understand the commonalities (conserved features) and differences (diverse features) among binding sites with similar geometry and with similar chemical groups. In this line, a recent study aimed at the identification of off-targets by incorporating similarity search based on the 3-D target binding site against the PDB. Using this approach along with transcriptome profile, data of structure-activity relationships etc., isoforms of cyclin-dependent kinase and Glycogen Synthase Kinase 3 Beta were recognized as off-targets for protein kinase C theta inhibitor (A-1411735) which is used for chronic autoimmune and inflammatory diseases (Rao et al. 2019). Likewise, other tools for identifying similar drug molecule binding sites, such as PLIC (Anand et al. 2014) and ProBis (Konc et al. 2012), are also available.

5.3.4 Drug Specificity and Potency

The common features across similar binding sites in multiple proteins hint the importance of such features in executing the respective functions of the proteins which have been preserved during evolution. Engaging these conserved features in optimal interactions would help in gaining potency and thereby aid in improving the efficacy of the ligand. On the contrary, differences in certain features among the apparently similar looking binding sites across multiple proteins suggest that such features are not important from functional perspective of the proteins. However, such differences offer tremendous opportunities to a designer in developing safer ligands. Protein-ligand interactions which exploit the unique features of a binding site is expected to be more specific and hence safer than those which fail to do so. This is often the case with many new generations of drugs as compared to their predecessors (Neu 1996; French and Gazzola 2011; Baldoni et al. 2014; Barnhart and Shelton 2015; Chakraborty and Rhee 2015).

In this context, the recognition of unique and common features in a protein target upon comparing it with homologs in sequence as well as structure space, requires knowledge on evolutionarily history of the target protein. Let us again take the classic example of kinase inhibitors to understand this concept. The hinge region in kinases that connects the N-terminal lobe to the C-terminal lobe of the protein, are the structurally conserved region among all kinases. The adenine ring of ATP forms hydrogen bonds with the specific residues in the hinge region. This helps in placing the ATP in a suitable conformation that facilitates phosphorylation of substrate proteins. It has been observed that ATP mimetic agents which do not establish hydrogen bonds with those hinge residues fail to inhibit the enzyme either completely or partially. This signifies the importance of engaging functionally conserved residues in protein-inhibitor interactions (Arris et al. 2000; Xing et al. 2015). However, the gatekeeper residue just preceding the hinge residues are considerably variable among different groups of kinases and targeting the gatekeeper residue for protein-ligand interaction is one of the strategies to achieve kinase inhibitor selectivity (Huang et al. 2010). Further, specificity toward a particular isoform (Aurora kinase-A) within the Aurora kinase family has also been reported upon systematically targeting unique residue (Thr217) in the binding site sequence of the highly similar isoforms (Aurora kinase-A/B/C) (Bouloc et al. 2010; Bavetsias et al. 2013). Another interesting example in the area of achieving subtype specificity among closely related proteins is targeting the GPCRs. Studies have revealed that albeit there are universally conserved regions in GPCR which allow them to elicit a highly similar activation mechanism, there also exists “selectivity barcode” which facilitates uniqueness in molecular recognition. Such selectivity features include distinct pattern of amino acid residues which could be exploited in designing subtype specific modulators (Flock et al. 2017).

5.3.5 Homologous Proteins in Polypharmacology

Identification of similar binding pockets is useful in designing a drug that targets multiple related proteins involved in a given disease pathway without modulating the functions of undesired targets. Indeed, this approach is beneficial for the treatment of multigenic diseases such as heart diseases, type 2 diabetes, mental retardation, leukemia etc. (Anighoro et al. 2014). For example, sunitinib is a multi-targeted tyrosine kinase inhibitor that act upon PDGF receptor, VEGF receptors (1 and 2), fms like tyrosine kinase 3 and c-Kit. Such promiscuous nature of sunitinib is helpful in treating drug-resistant gastrointestinal stromal tumour and metastatic renal cell carcinoma (Faivre et al. 2007). This principle of targeting multiple proteins with a single drug is termed as 'polypharmacology' and it is being systematically exploited in drug repurposing/drug repositioning approaches (Jalencas and Mestres 2013; March-Vila et al. 2017). Drug repurposing has become an important branch of drug discovery which aims at identifying new use of an already existing drug. These approaches are less time-consuming and involve less investment of resources as compared to conventional drug discovery programs aiming to find a new chemical entity. The fundamental basis of most drug repurposing approaches revolves around the theory of neighbourhood behaviour where similar binding sites are expected to recognize similar molecules and vice-versa. This is especially relevant in the current scenario (at the time of writing this article, April 2020) when the world is facing a global health challenge of COVID-19 caused by a newly emerged coronavirus, SARS-CoV-2. Since conventional drug discovery programs aimed at identifying novel chemical entities are time-consuming, it is unlikely that any novel drug to treat COVID-19 will be available in the next few months. Therefore, the research community around the globe are putting efforts in identifying promising candidates from the repertoire of existing approved drugs which can be helpful in treating SARS-CoV-2 infection (Sanders et al. 2020). In the recent past, research from our group has helped in identifying potential drugs which could be repurposed against infectious diseases like malaria, tuberculosis, fungal infections etc. by exploiting the principles of protein evolutionary relationships (Ramakrishnan et al. 2015, 2017; Chakraborti et al. 2019a, 2019b). In continuity, we have recently identified potential anti-COVID-19 drugs using the principles of neighbourhood behaviour (Chakraborti et al. 2020). Albeit drug repurposing approaches contribute toward rapid identification of potential drugs, the greatest challenge of any drug repurposing program lies in exploiting the benefits of polypharmacology with minimum or no toxicity (Pushpakom et al. 2018).

In addition, it should be noted that under physiological conditions, the biological molecules are in dynamic state and undergo vibrational as well as large domain motions (Moroni et al. 2015). Thus, the degree of complementarity in shape and electrostatic features between a drug and its protein target(s) may vary with time. Knowledge derived solely from static structures as discussed so far, might not provide complete picture of molecular basis of a drug action. Hence, understanding the changes associated with protein-ligand complex structures with time, is central to any successful drug discovery program. In the following section, we introduce the concept of protein dynamics and emphasize on how dynamics play a role in drug design with examples of therapeutic drugs.

5.4 Protein Dynamics

Proteins are inherently dynamic that helps them to evolve and adapt to diverse cellular conditions (Tokuriki and Tawfik 2009). Protein function relies on the synergy between structure and dynamics. Therefore, a molecular level understanding of protein function requires the atomic details in both 3-D space as well as time. Protein dynamics can be viewed at multiple levels in timescale starting from fluid-like motions of atoms about its mean positions at picoseconds to sampling different conformational substates at milliseconds or seconds (Van Den Bedem and Fraser 2015). Large timescale dynamics drive proteins i) to be at equilibrium between different functional states, ii) to bind ligands, iii) to transport molecules, iv) to catalyze enzymatic reaction, v) to have allosteric regulations etc. In the following subsections, we have provided details of examples on how dynamics drive protein function, methods applied to study dynamics of protein and protein-ligand interactions, evolutionary aspect of protein dynamics and role of dynamics in drug design.

5.4.1 Role of Dynamics in Enzymatic Function

Specific and tightly regulated molecular recognition is central to all biomolecular interactions. It has been reported that binding of several therapeutic drugs involve dynamical motions of target proteins (Copeland 2011). Besides the classical Fisher's 'lock-and-key' and Koshland's 'induced-fit' models, a model based on 'conformational selection' by Frauenfelder is the widely accepted theory for molecular recognition (Frauenfelder et al. 1991). This model (and 'induced-fit' to some extent) considers inherent nature of protein dynamics which is observed to be involved in all binding events such as protein-small molecule, protein-protein, protein-DNA/RNA as well as RNA-ligand interactions (Kumar et al. 2008). Conformational selection process has been demonstrated in several enzymes including RNase A, adenylate kinase, aspartate transcarbamoylase and dihydrofolate reductase (Boehr et al. 2009). It explains the basis of molecular recognition as protein samples ensemble of conformations which may differ in their energy state and ligand selects a conformation that closely resembles the ligand-bound state (Frauenfelder et al. 1991).

Here, we have taken adenylate kinase as an example to explain how dynamics is linked to molecular function of the protein. This enzyme follows conformational selection for substrate binding (Kovermann et al. 2017). In addition, dynamical motions at picosecond to millisecond timescale drives the catalytic activity of this enzyme. It is known to catalyze reversible interconversion of adenine nucleotides and has three domains, namely core, NMP (nucleotide monophosphate) and lid. Collective motions involving opening and closing of lid domain in microseconds to milliseconds scale have been identified to be the rate-limiting step for the catalytic reaction (Fig. 5.5A). Fast motions such as backbone fluctuations by thermal energy in picoseconds to nanoseconds facilitate such motions of the lid domain. This large conformational motion is directly correlated with catalytic turnover rate of this enzyme (Henzler-Wildman et al. 2007). Hence, dynamics at all hierarchy in the timescale is essential and cognizance of such dynamical behaviour offers deeper understanding on the mechanistic basis of enzymatic function.

In the field of drug design, dynamics and associated functional substates are increasingly being recognized as important factors to design a highly selective drug molecule. One of the effective ways of selectively targeting a protein is to target its allosteric site, which is a secondary binding site that is capable of remotely altering the conformation of the orthosteric site. However, allosteric sites are not always evident. Computational techniques like molecular dynamics simulations as discussed later could be helpful in predicting allosteric binding sites on the protein surface. In the area of drug design, allosteric sites offer an advantage over the orthosteric sites as the former are evolutionarily less well conserved than the later and are most often unique. Therefore, achieving selectivity and hence better safety profile for a drug targeting the allosteric site is generally possible (Abdel-Magid 2015). For example, p38 MAP kinase is targeted for treating inflammatory diseases. It possesses a unique allosteric binding site whose accessibility to the therapeutic drug (BIRB 796) is governed by a large conformational change in the DFG motif of the kinase. Targeting this site has been shown to improve drug affinity (Pargellis et al. 2002). Likewise, as mentioned earlier, imatinib, a drug against ABL tyrosine kinase, is conformation specific drug molecule and binds only to the inactive conformation of this enzyme (Iqbal and Iqbal 2014). Therefore, knowledge of conformational dynamics of target protein lends a strategy to design conformation-specific drug molecule.

5.4.2 Role of Dynamics in Membrane Receptor Function

Like globular proteins, the functions of membrane proteins are also largely associated with their inherent dynamics. In this section, by taking few examples which are important drug targets, we have highlighted the importance of dynamics in designing drugs targeted against membrane receptors. Dynamics of membrane receptor proteins are particularly interesting to understand because these proteins generally undergo large conformational changes especially when they are involved in facilitating passage of molecules through the membrane channel.

N-methyl-D-aspartate (NMDA) receptors are glutamate-gated ion channels involved in brain development and synaptic plasticity by permeating Ca^{2+} ions across cellular membrane in neurons (Traynelis et al. 2010). NMDA receptors are

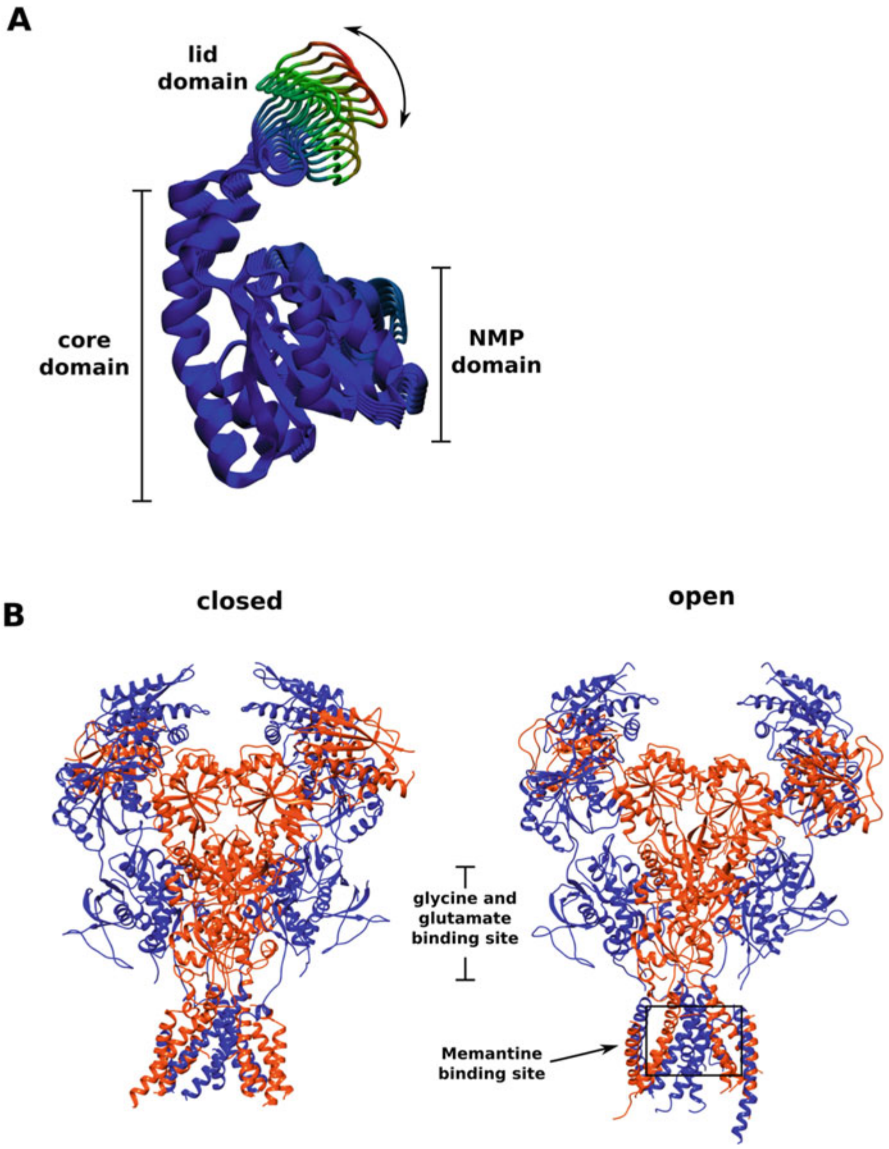


Fig. 5.5 Examples of protein dynamics. (**a**) Shown in cartoon representation is a large-scale motion of the lid domain of adenylate kinase. Snapshot of each conformation is derived from anisotropic network model based normal mode analysis using crystal structure of *E. coli* adenylate kinase (PDB code: 4AKE) (Müller et al. 1996; Atilgan et al. 2001). Color scale from blue to red indicates the magnitude of residue motions from small to large. (**b**) Cartoon representation of NMDA receptor in ‘closed’ (PDB code: 4TLL) (Lee et al. 2014) and ‘open’ conformation (PDB code: 6IRA) (Zhang et al. 2018). Subunits GluN1 and GluN2 are colored in blue and orange respectively. Memantine binding region is highlighted in a box. VMD (Humphrey et al. 1996) and Chimera (Pettersen et al. 2004) were used to generate Fig. (**a**) and (**b**), respectively

obligatory hetero-tetramers mainly comprises of two copies of GluN1 subunit that bind to glycine and two copies of GluN2 subunit that bind to L-glutamate. The receptor function is governed by two factors: membrane potential and glycine as well as L-glutamate binding. Receptor is inactive under membrane resting potential (-70 mV) by adopting 'closed' conformation. When membrane gets depolarized and ligands are bound, it acquires 'open' conformation in order to permeate Ca^{2+} ions into cytoplasm of the signal receiver neuron (post-synaptic neuron). For the next cycle of neurotransmission to happen, NMDA receptor reverts to 'closed' conformation. Hence, dynamic personality of NMDA receptor to switch between 'open' and 'closed' conformation is essential for neurotransmitter signaling. Indeed, application of cryo-EM techniques has unveiled a detailed knowledge of NMDA receptor dynamics and its regulation by proton and zinc concentration (Jalali-Yazdi et al. 2018). Duration of NMDA receptor in the 'open' conformation is directly related to cognitive skills such as learning and memory.

On the account of its importance in brain function, dysfunctional NMDA receptors are implicated in various brain diseases such as schizophrenia, Alzheimer's disease, Huntington's disease, Parkinson's disease and amyotrophic lateral sclerosis (Chen and Lipton 2006). For the therapeutic treatments, several antagonists have been designed against glutamate-binding and glycine-binding sites. However, such neuroprotective drugs have failed in clinical trials due to adverse effects by compromising the normal function of this receptor. Under neurological disease conditions, the activity of NMDA is elevated by retaining 'open' conformation for a longer period of time than it has to be. By taking the differences in the conformational states into account, memantine, a drug molecule to specifically target the 'open' conformation of the receptor as an uncompetitive antagonist, was designed (Fig. 5.5B). As a result, it has become successful therapeutic intervention and the drug blocks the 'open' conformation of the receptor under excessive glutamate condition and minimize the side-effects by not interfering with the receptor's basal level activity (Chen and Lipton 2006). Likewise, several allosteric modulators that consider dynamics states of NMDA receptor are used as therapeutic drugs for the treatment of many neurological diseases (Traynelis et al. 2010).

Along the same line, dynamics nature of GPCRs has been extensively exploited for the design of therapeutic drug molecules. GPCRs sample several conformations and undergo conformational transition between 'active' to 'inactive' states which is critical for signal transmission. The conformational transition involves a rearrangement of transmembrane helices (5–7) through inward and outward movements. It has also been shown that the intracellular component of the receptor is more dynamic than extracellular region however subtle changes in extracellular loops influence the binding kinetics for ligands (Latorraca et al. 2017). Interestingly, dynamics of these surface loops facilitates the formation of alternative binding sites for allosteric modulators (Dror et al. 2013). This observation emphasizes that the mechanism of ligand binding and signal transmission across membrane in GPCRs is facilitated by their dynamical motions.

Furthermore, in the ongoing life-threatening condition due to 'COVID-19 or SARS-CoV-2 infection', cryo-EM has provided details on the viral entry vis-à-vis how the viral spike protein interacts with human ACE2 receptor (Walls et al. 2020). Spike protein comprises of two subunits responsible for host receptor binding (S1) and membrane fusion (S2). S1 subunit interacts with ACE2 receptor in the trimeric form. For interaction between spike protein and ACE2 receptor to happen, receptor binding domain of S1 subunit has to undergo large scale motions from 'closed' to 'open' conformational state. Since 'open' conformation is essential for ACE2 receptor binding, an alteration in the conformational sampling between 'open' and 'closed' state appears to be correlated with the rate of virus transmissibility and disease severity (Walls et al. 2020). Hence, dynamics of spike protein plays a significant role in the viral entry to human cells.

Therefore, from the above examples it can be appreciated that conformation-specific drug design comes with the benefits of drug being highly selective to specific conformation that is related to disease condition. Consequently, it obviates adverse effects caused by perturbing basal level function of the target conferred by other conformational states or interactions with off-targets that are homologous to the intended drug target but largely differ by their characteristic conformations. Currently, there are several experimental and simulation methods available to study protein dynamics as mentioned in the following subsection.

5.4.3 Methods for Studying Protein Dynamics

Structures determined by X-ray crystallography are space and time averaged entities. However, analysis on multiple structures from independent X-ray crystallographic studies of the same protein provides structural basis for dynamics (Marino-Buslje et al. 2019). Also, ensemble structures from solution NMR technique provides details about large scale dynamics of proteins. Especially, Carr-Purcell-Meiboom-Gill relaxation dispersion, paramagnetic relaxation enhancement and native-state HD exchange NMR techniques are used to study higher energy conformations that are important to protein function (Baldwin and Kay 2009). Single molecule Forster Resonance Energy Transfer (smFRET) is also widely used to obtain information about protein dynamics (Kalinin et al. 2010). Recently, resolution revolution in cryo-EM have made this technique as a promising method to study large scale conformational dynamics of proteins and molecular assemblies (Cheng 2018). Albeit these experimental techniques give realistic insights into protein dynamics, they are time-consuming and require expensive resources. Alternatively, computational techniques such as molecular dynamic simulations are fast and inexpensive to obtain information of dynamics of biomolecules as explained below.

Molecular dynamics simulations allow us to study dynamics of proteins and protein-biomolecular interactions *in silico*. Using this method, flexibility associated with binding site residues of target protein and time-dependent variations in the interaction patterns between protein and ligand molecules in the aqueous medium can be studied. Since dynamics simulations account for protein flexibility and entropy effects, a combination of ligand docking and molecular dynamics helps to identify binding modes of ligand molecules (De Vivo et al. 2016). Generation of ensembles of target protein structure from dynamics simulations and docking of ligand molecules to each of the protein conformation obtained from simulations is helpful to accurately predict the binding energy of protein-ligand interactions. Along with molecular dynamics simulation, Monte Carlo simulations are also used to traverse through the conformational landscape of proteins (Cole et al. 2015). Hence, these simulation methods have also been employed in free-energy perturbations (FEP), thermodynamics integration (TI) and lamda-dynamics to study allosteric mechanisms and the role of water molecules in ligand binding (Wang et al. 2015). Recently, elastic network based normal mode analysis also widely used to study large scale motions of proteins that are relevant to their functional mechanisms. It comes with the advantages of analysing large proteins as well as membrane receptors with less demand on computational power and time which are in general the limitations of molecular dynamics simulations (Skjaerven et al. 2009; Zheng et al. 2017). Such approaches are employed during the process of lead optimizations in SBDD to design drug molecule with improved affinity for the intended target. Therefore, both experimental methods and computational simulation methods provide useful information to study dynamics associated with protein and aid to understand time-dependent interaction profiles between protein and drug molecule.

5.4.4 Evolution of Protein Dynamics in Target Identification

Dynamics being an inherent feature of a protein, not surprisingly, dynamics of homologous proteins with similar function is conserved. Like sequence and structure, the extent of conservation of dynamics features is higher among members of the same family than across members of difference families (Kalaivani et al. 2016; Narayanan et al. 2018). Such observation had led to a concept of ‘dynasome’ that states proteins with similar functions share common dynamic fingerprints (Hensen et al. 2012). Studies involving evolution of protein dynamics have become an emerging area and have shown that dynamics is manoeuvred in the course of protein evolution (Tokuriki and Tawfik 2009; Klinman and Kohen 2014). In conformity with this, differences in dynamics have been observed between two distantly related proteins. For example, hexokinase-1 from *H. sapiens* is a multidomain domain protein and its homolog in *E. coli* has only single domain (Vishwanath et al. 2018). Dynamics of these two homologous proteins are different as weaker correlations in residue motions are observed in single domain hexokinase-1 when compared to that of multi-domain hexokinase-1. Differences in the dynamics can explain the difference in substrate affinity of these two homologous enzymes which share identical substrate binding site. Hence, while target identification using information derived from distant homologs, it is important to be aware of the differences in the substrate affinity and associated dynamics between the homologs of interest. All these observations underscore that dynamics play an important role in biomolecular recognitions and analysis on dynamics is considered as an important aspect in the early phase of drug discovery (Amaro et al. 2018).

5.4.5 Role of Dynamics in Drug Design and its Therapeutic Response

Since dynamics helps a protein to respond to its environment and influence its interactions with different partners, it is important to consider protein dynamics in the process of drug design (Moroni et al. 2015). Sometimes, conformational changes of the protein uncover many treasures like cryptic sites which provide rare opportunities to target the otherwise undruggable protein (Beglov et al. 2018; Kuzmanic et al. 2020). The inherent dynamics of proteins often regulate allosteric signalling pathways. Understanding the molecular mechanism of such pathways have attracted considerable attention among the research community because of their potential benefits in the field of drug discovery and developments (Amamuddy et al. 2020). Although there have been striking advancements in predicting protein-ligand binding affinity, development of techniques to study protein-ligand dynamics is still an ongoing research. This is mainly limited by computational resources and the vastness as well as diversity of the chemical universe of ligands (Salsbury 2010). Nevertheless, in the recent times, there have been appreciable advancements which allow us to sample different conformations of protein-ligand complexes over a considerably long timescale using advanced molecular dynamics simulation techniques that captures atomistic details. This has opened avenues for studying even drug unbinding pathways which were otherwise computationally expensive and were beyond the resource capacity for many researchers in both academia as well as in industries (De Vivo et al. 2016; Hollingsworth and Dror 2018).

Another aspect of protein-drug molecule interactions is the stability. In general, greater the stability of the interactions between a drug and its target, better is the efficacy. However, a prolonged interaction is not desirable as that might lead to toxic effects due to drug accumulation. Thus, an optimum interaction between the drug and the target is needed so that the drug remains strongly bound to its intended target only during the timespan when the desired therapeutic benefit is expected. Once the therapeutic response is obtained, the drug should quickly leave the protein binding cavity to avoid any toxic outcomes. Hence, study of conformations and residence time of drug-target complexes help to gain understanding of various aspects of ligand binding and unbinding which in turn dictate the efficacy and safety of the drug (Schuetz et al. 2019; Gobbo et al. 2019).

As mentioned earlier, the dynamical nature of a protein helps it to interact with many other proteins or other biomolecules during its lifetime. Through the intermolecular interactions, protein forms permanent or transient complexes and performs its function. Studying such phenomena are increasingly becoming popular as these are the real machineries driving key cellular events. In the following section, we have discussed the importance of protein-protein complexes and their implications in drug discovery.

5.5 Protein-Protein Complexes

Proteins in general interact with other molecules to carry out their cellular functions. Although, proteins are likely to encounter random physical contacts with other biomolecules in cellular milieu, the frequency, duration and specificity of the interactions rely on functional significance. An observation that 85% of proteins in yeast genome have designated interaction with another protein(s) indicates the prevalence of specific interactions among proteins (Reid et al. 2010). Such interactions govern the formation of protein-protein complexes that mediate nearly all biological processes in the cell. A protein complex is an assembly of at least two protein molecules which behave as a single entity with an ability to play a specific role. The nature and number of proteins as well as their arrangement defines the quaternary structure of the protein complex. Polypeptide chains within a quaternary structure are usually permanently associated with each other during its functional lifetime in the cell.

Another functional form of a protein is in an assembly of several protein complexes together that work orchestrally. During the functional lifetime of a protein assembly, constituent proteins can associate and dissociate through transient interactions. Such sophisticated assemblies carry major cellular processes such as nuclear transport, electron transport system, splicing and protein translation (Pieters et al. 2016). Protein assemblies integrate several catalytic and structural activities with in-built regulatory mechanisms. Knowledge on the structure of protein complexes is indigent as compared to monomeric proteins. However, efforts have been undertaken for the characterization of complex structures. Rapid developments in experimental methodologies *viz.* cross-linking combined with mass spectrometry, synchrotron-based X-ray crystallography, serial femtosecond crystallography, methyl-transverse relaxation optimized spectroscopy NMR, cryo-EM have laid emerging paths for structural characterization of protein complexes. Availability of protein complex structures serve as a source to derive the properties of protein-protein interfaces. In several disease conditions, protein-protein interactions are perturbed and hence could serve as a potential drug target. Some of the advantages, recent achievements and challenges in this field are discussed below.

5.5.1 Protein-Protein Complexes in Drug Discovery

At this standpoint, studies on protein-protein complexes serve potential applications in therapeutics (Luo et al. 2016). For example, virus-like particles by self-assembly of bacteriophage MS2 coat protein has been successfully demonstrated to deliver chemotherapeutic drugs selectively to human hepatocellular carcinoma (Ashley et al. 2011). Protein design and engineering based on the knowledge of protein complexes has also become a promising avenue for successful design of biosensors (Gonçalves et al. 2014) and drug delivery system (Rohovie et al. 2017). In addition, targeting protein-protein interfaces (PPIs) is likely to lead to more efficient controlling of a disease condition with less side-effects as compared to targeting a binding pocket on the protein surface. Ribosome is a known example of a molecular assembly that is targeted by several drugs (macrolydes, fusidic acid, quinopristin and other antibiotics) to inhibit protein synthesis of pathogenic bacteria (Wilson 2011). Recently, COMMD3/8 complex was proposed to be a drug target for treating inflammatory diseases (Nakai et al. 2019). Mutations located at interface regions result in disease conditions due to the abruption of protein complex formation. For example, congenital fibrosis of extraocular muscles, a disorder of the nervous system that affects muscle movements around eyes, is caused by mutations in a tubulin gene (TUBB3). These mutations affect heterodimerization of tubulin that is essential for the formation of microtubules (Tischfield et al. 2010). Hence, inhibition of complex formation (Goncearenco et al. 2017) or inducing disassembly (Kim et al. 2012) of defective protein complexes have been investigated for disease treatment.

In order to target protein-protein interfaces, structural details especially for the interfacial region of protein-protein complex is essential to indicate the residues involved in the interactions and the 3-D geometry. If protein-protein complex structure is unavailable, computational approaches can be used to predict protein-protein interfaces. It has been well known that interface region is a surface patch which has modest sequence conservation. This feature is unique as compared to rest of the protein surface. Hence, details on residue composition, conservation profile and physicochemical properties of interface region could be used through tools such as Consurf (Ashkenazy et al. 2016), PSIVER (Murakami and Mizuguchi 2010) and ISIS (Ofran and Rost 2007) for interface prediction. Further, analogous to homology modeling of individual proteins, structural details and the orientation of protein partners at the interface could be inferred from 3-D structure of a homologous protein complex of known structure which shares significant sequence similarity with the protein complex of our interest. Furthermore, if the structure of two protein partners are available separately either by experiment or by computational modeling, protein-protein docking method could be employed to predict interfaces. In this context, HADDOCK, PIPER and RossetaDock tools have been developed to perform protein-protein docking (Pagadala et al. 2017). In all these prediction methods, protein partners are assumed to be rigid in which interface region does not undergo substantial conformational changes upon interactions.

Intrinsically disordered regions (IDRs) play a significant role in protein-protein interaction networks. IDRs do not have a stable structure on their own. When an IDR interacts with a protein partner, it undergoes a large conformational change and acquires stable conformation (Uversky 2019). IDRs are highly dynamic and unstructured in isolation because of which their structure determination using experiments are in general unsuccessful. As a result, identification of homologous protein complex of known structure to infer structural details of interfaces has become a difficult task. Another main challenge in the identification of interfaces involving IDRs is that the conformational changes of IDRs upon interaction remain unfeasible to predict and hence docking protocol cannot be applied even if the structure of protein partners are available in isolation. Furthermore, sequence of IDRs is typically not very well conserved (Brown et al. 2011). Together, prediction of protein-protein interaction sites mediated by IDRs remains a difficult problem. As a consequence, targeting interfaces with IDRs involved has thus far not been contemplated for therapeutic drug treatment.

Nonetheless, targeting PPIs of structured proteins by small molecules is again notoriously difficult task due to large and flat surface. The PPIs which hardly house any groove makes it challenging to design a small molecule inhibitor that can hold onto these flat surfaces and elucidate the desired inhibitory response. Further, lack of any endogenous ligands for PPIs leaves a drug designer clueless since no known starting scaffold is available to guide the design process. Also, the proprietary chemical libraries available for high throughput screening are unsuitable for targeting PPIs. As a result, PPIs are largely intractable at the moment. Nevertheless, research in the field of structural biology has shown some rays of hope. Despite the large surface involved in PPIs, only a small sub-set of residues are generally involved in key regulation of the protein-protein interactions. These residues are termed as ‘hotspots’ in PPIs. Targeting the hotspot residues with small molecules employing Fragment-Based Drug Discovery (FBDD) approaches have been shown to be an effective strategy in developing inhibitors against PPIs. This is an emerging area and new methodologies are explored by the community. An example of successful direct PPI inhibitor that reached the market is eltrombopag, a drug that is used to treat profound thrombocytopenia. In addition to this direct PPI inhibitor, allosteric PPI inhibitors are also known, such as the approved chemokine receptor modulators: maraviroc and plerixafor. Biased ligands which selectively modulate downstream cellular signalling mediated by certain protein-protein interactions over the others are also being explored as allosteric PPI modulators. For example, TRV027 is a therapeutic drug for AT1-receptor and acts as a biased agonist. It blocks G-protein activation by interacting with the angiotensin binding site in AT1-receptor while at the same time, it stimulates signals in the β-arrestin pathway. Through this control of biased signaling, the drug helps in the management of acute heart failures (Felker et al. 2015). Several other PPI inhibitors are currently undergoing clinical trials as reviewed elsewhere (Mabonga and Kappo 2019).

5.6 Conclusions

The biological outcome triggered from the interactions between two binding partners or multiple components in an assembly is a balanced interplay between several factors like sequence, structure, dynamics and molecular recognition as explained in this chapter. With increasing evidences of protein structure and dynamics influencing molecular recognition, drug design pipelines are now more focused on the analyses pertaining to these factors. Here, we have described principles of protein structure, dynamics and assembly and their contributions to molecular function. Selected examples discussed in this chapter indicate that perturbation of any these factors directly influences protein function or functional levels and can lead to disease conditions. Therapeutic interventions targeting the concerned protein could help in abrogating the disease condition. Modulators used as therapeutic drugs should have shape and chemical features which are complementarity to the binding pocket of the target protein. Such complementarities lead to optimal interactions between the target protein and the drug molecule. Protein-drug binding event can influence the dynamics of protein as well as can hinder assembly or disassembly of subunits in the protein complex. While the strength and stability of interactions between drug and target protein determines the potency, its potential to interact with unintended targets determines its toxicity profile. Therefore, knowledge on molecular details of 3-D structure, evolutionary relationships, dynamics, functional state and biomolecular interactions of target protein are essential for designing a safe and efficacious therapeutic agent.

Acknowledgements This work is supported by the Department of Biotechnology, Government of India in the form of IISc-DBT partnership programme, Mathematical Biology and FIST program, sponsored by the Department of Science and Technology. Support from UGC, India – Centre for Advanced Studies and Ministry of Human Resource Development, India, is gratefully acknowledged. SC acknowledges DST-INSPIRE fellowship program for funding her research work. NS is a J. C. Bose National Fellow.

References

Abdel-Magid, A. F. (2015). Allosteric modulators: An emerging concept in drug discovery. *ACS Medicinal Chemistry Letters, 6*(2), 104–107. https://doi.org/10.1021/ml5005365

Abraham, D. J., Spyrakis, F., Cozzini, P., & Kellogg, G. E. (2010). Docking and scoring in drug discovery. In D. J. Abraham (Ed.), *Burger's medicinal chemistry and drug discovery* (pp. 601–684). https://doi.org/10.1002/0471266949.bmc140

Amamuddy, O. S., Veldman, W., Manyumwa, C., Khairallah, A., Agajanian, S., Oluyemi, O., et al. (2020). Integrated computational approaches and tools for allosteric drug discovery. *International Journal of Molecular Sciences, 21*(3), 847. https://doi.org/10.3390/ijms21030847

Amaro, R. E., Baudry, J., Chodera, J., Demir, Ö., McCammon, J. A., Miao, Y., et al. (2018). Ensemble docking in drug discovery. *Biophysical Journal, 114*(10), 2271–2278. https://doi.org/10.1016/j.bpj.2018.02.038

Anand, K., Ziebuhr, J., Wadhwani, P., Mesters, J. R., & Hilgenfeld, R. (2003). Coronavirus main proteinase (3CLpro) structure: Basis for design of anti-SARS drugs. *Science, 300*(5626), 1763–1767. https://doi.org/10.1126/science.1085658

Anand, P., Nagarajan, D., Mukherjee, S., & Chandra, N. (2014). PLIC: Protein-ligand interaction clusters. *Database, 2014*(1). https://doi.org/10.1093/database/bau029

Anderson, A. C. (2003). The process of structure-based drug design. *Chemistry and Biology, 10*(9), 787–797. https://doi.org/10.1016/j.chembiol.2003.09.002

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science, 181*(4096), 223–230. https://doi.org/10.1126/science.181.4096.223

Anighoro, A., Bajorath, J., & Rastelli, G. (2014). Polypharmacology: Challenges and opportunities in drug discovery. *Journal of Medicinal Chemistry, 57*(19), 7874–7887. https://doi.org/10.1021/jm5006463

Arris, C. E., Boyle, F. T., Calvert, A. H., Curtin, N. J., Endicott, J. A., Garman, E. F., et al. (2000). Identification of novel purine and pyrimidine cyclin-dependent kinase inhibitors with distinct molecular interactions and tumor cell growth inhibition profiles. *Journal of Medicinal Chemistry, 43*(15), 2797–2804. https://doi.org/10.1021/jm990628o

Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., et al. (2016). ConSurf 2016: An improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Research, 44*(W1), W344–W350. https://doi.org/10.1093/nar/gkw408

Ashley, C. E., Carnes, E. C., Phillips, G. K., Durfee, P. N., Buley, M. D., Lino, C. A., et al. (2011). Cell-specific delivery of diverse cargos by bacteriophage MS2 virus-like particles. *ACS Nano, 5* (7), 5729–5745. https://doi.org/10.1021/nn201397z

Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., & Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical Journal, 80*(1), 505–515. https://doi.org/10.1016/S0006-3495(01)76033-X

Baker, D., & Sali, A. (2001). Protein structure prediction and structural genomics. *Science, 294* (5540), 93–96. https://doi.org/10.1126/science.1065659

Baldoni, D., Gutierrez, M., Timmer, W., & Dingemanse, J. (2014). Cadazolid, a novel antibiotic with potent activity against clostridium difficile: Safety, tolerability and pharmacokinetics in healthy subjects following single and multiple oral doses. *Journal of Antimicrobial Chemotherapy, 69*(3), 706–714. https://doi.org/10.1093/jac/dkt401

Baldwin, A. J., & Kay, L. E. (2009). NMR spectroscopy brings invisible protein states into focus. *Nature Chemical Biology, 5*(11), 808–814. https://doi.org/10.1038/nchembio.238

Barnhart, M., & Shelton, J. D. (2015). ARVs: The next generation. Going boldly together to new frontiers of HIV treatment. *Global Health Science and Practice, 3*(1), 1–11. https://doi.org/10.9745/GHSP-D-14-00243

Batool, M., Ahmad, B., & Choi, S. (2019). A structure-based drug discovery paradigm. *International Journal of Molecular Sciences, 20*(11), 2783. https://doi.org/10.3390/ijms20112783

Bavetsias, V., Faisal, A., Crumpler, S., Brown, N., Kosmopoulou, M., Joshi, A., et al. (2013). Aurora isoform selectivity: Design and synthesis of imidazo[4,5- B]pyridine derivatives as highly selective inhibitors of Aurora-A kinase in cells. *Journal of Medicinal Chemistry, 56* (22), 9122–9135. https://doi.org/10.1021/jm401115g

Beglov, D., Hall, D. R., Wakefield, A. E., Luo, L., Allen, K. N., Kozakov, D., et al. (2018). Exploring the structural origins of cryptic sites on proteins. *Proceedings of the National Academy of Sciences of the United States of America, 115*(15), E3416–E3425. https://doi.org/10.1073/pnas.1711490115

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., et al. (2002). The protein data bank. *Acta Crystallographica Section D: Biological Crystallography, 58*(6 I), 899–907. https://doi.org/10.1107/S0907444902003451

Blundell, T. L., & Johnson, L. N. (1976). *Protein crystallography* (Vol. 11, p. 331). Cambridge, MA: Academic Press.

Boehr, D. D., Nussinov, R., & Wright, P. E. (2009). The role of dynamic conformational ensembles in biomolecular recognition. *Nature Chemical Biology, 5*(11), 789–796. https://doi.org/10.1038/nchembio.232

Bouloc, N., Large, J. M., Kosmopoulou, M., Sun, C., Faisal, A., Matteucci, M., et al. (2010). Structure-based design of imidazo[1,2-a]pyrazine derivatives as selective inhibitors of Aurora-A kinase in cells. *Bioorganic and Medicinal Chemistry Letters, 20*(20), 5988–5993. https://doi.org/10.1016/j.bmcl.2010.08.091

Brown, C. J., Johnson, A. K., Dunker, A. K., & Daughdrill, G. W. (2011). Evolution and disorder. *Current Opinion in Structural Biology, 21*(3), 441–446. https://doi.org/10.1016/j.sbi.2011.02.005

Capdeville, R., Buchdunger, E., Zimmermann, J., & Matter, A. (2002). Glivec (ST1571, imatinib), a rationally developed, targeted anticancer drug. *Nature Reviews Drug Discovery, 1*(7), 493–502. https://doi.org/10.1038/nrd839

Cavasotto, C. N., & Palomba, D. (2015). Expanding the horizons of G protein-coupled receptor structure-based ligand discovery and optimization using homology models. *Chemical Communications, 51*(71), 13576–13594. https://doi.org/10.1039/c5cc05050b

Ceska, T., Chung, C. W., Cooke, R., Phillips, C., & Williams, P. A. (2019). Cryo-EM in drug discovery. *Biochemical Society Transactions, 47*(1), 281–293. https://doi.org/10.1042/BST20180267

Chakraborti, S., Ramakrishnan, G., & Srinivasan, N. (2019a). In Silico modeling of FDA-approved drugs for discovery of anticandida agents: A drug-repurposing approach. In K. Roy (Ed.), *In silico drug design* (pp. 463–526). Cambridge, MA: Academic Press. https://doi.org/10.1016/b978-0-12-816125-8.00016-x

Chakraborti, S., Ramakrishnan, G., & Srinivasan, N. (2019b). Repurposing drugs based on evolutionary relationships between targets of approved drugs and proteins of interest. In Q. Vanhaelen (Ed.), *Methods in molecular biology* (Vol. 1903, pp. 45–59). New York: Springer. https://doi.org/10.1007/978-1-4939-8955-3_3

Chakraborti, S., Bheemireddy, S., & Srinivasan, N. (2020). Repurposing drugs against the main protease of SARS-CoV-2: Mechanism-based insights supported by available laboratory and clinical data. *Molecular Omics, 16*(5), 474–491. https://doi.org/10.1039/d0mo00057d

Chakraborty, S., & Rhee, K. Y. (2015). Tuberculosis drug development: History and evolution of the mechanism-based paradigm. *Cold Spring Harbor Perspectives in Medicine, 5*(8), 1–11. https://doi.org/10.1101/cshperspect.a021147

Chen, H. S. V., & Lipton, S. A. (2006). The chemical biology of clinically tolerated NMDA receptor antagonists. *Journal of Neurochemistry, 97*(6), 1611–1626. https://doi.org/10.1111/j.1471-4159.2006.03991.x

Cheng, Y. (2018). Single-particle cryo-EM-how did it get here and where will it go. *Science, 361*(6405), 876–880. https://doi.org/10.1126/science.aat4346

Chothia, C., & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO Journal, 5*(4), 823–826. https://doi.org/10.1002/j.1460-2075.1986.tb04288.x

Clark, D. E. (2006). What has computer-aided molecular design ever done for drug discovery? *Expert Opinion on Drug Discovery, 1*(2), 103–110. https://doi.org/10.1517/17460441.1.2.103

Colby, T. D., Bahnson, B. J., Chin, J. K., Klinman, J. P., & Goldstein, B. M. (1998). Active site modifications in a double mutant of liver alcohol dehydrogenase: Structural studies of two enzyme – Ligand complexes. *Biochemistry, 37*(26), 9295–9304. https://doi.org/10.1021/bi973184b

Cole, D. J., Tirado-Rives, J., & Jorgensen, W. L. (2015). Molecular dynamics and Monte Carlo simulations for protein-ligand binding and inhibitor design. *Biochimica et Biophysica Acta – General Subjects, 1850*(5), 966–971. https://doi.org/10.1016/j.bbagen.2014.08.018

Copeland, R. A. (2011). Conformational adaptation in drug-target interactions and residence time. *Future Medicinal Chemistry, 3*(12), 1491–1501. https://doi.org/10.4155/fmc.11.112

Dadashpour, S., Kucukkilinc, T. T., Tan, O. U., Ozadali, K., Irannejad, H., & Emami, S. (2015). Design, synthesis and in vitro study of 5,6-diaryl-1,2,4-triazine-3-ylthioacetate derivatives as COX-2 and β-amyloid aggregation inhibitors. *Archiv Der Pharmazie, 348*(3), 179–187. https://doi.org/10.1002/ardp.201400400

Dai, W., Zhang, B., Su, H., Li, J., Zhao, Y., Xie, X., et al. (2020). Structure-based design of antiviral drug candidates targeting the SARS-CoV-2 main protease. *Science, 368*(6497), 1331–1335. https://doi.org/10.1126/science.abb4489

De Vivo, M., Masetti, M., Bottegoni, G., & Cavalli, A. (2016). Role of molecular dynamics and related methods in drug discovery. *Journal of Medicinal Chemistry, 59*(9), 4035–4061. https://doi.org/10.1021/acs.jmedchem.5b01684

Dimasi, J. A., Feldman, L., Seckler, A., & Wilson, A. (2010). Trends in risks associated with new drug development: Success rates for investigational drugs. *Clinical Pharmacology and Therapeutics, 87*(3), 272–277. https://doi.org/10.1038/clpt.2009.295

DiMasi, J. A., Grabowski, H. G., & Hansen, R. W. (2016). Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics, 47*, 20–33. https://doi.org/10.1016/j.jhealeco.2016.01.012

Dror, R. O., Green, H. F., Valant, C., Borhani, D. W., Valcourt, J. R., Pan, A. C., et al. (2013). Structural basis for modulation of a G-protein-coupled receptor by allosteric drugs. *Nature, 503* (7475), 295–299. https://doi.org/10.1038/nature12595

Edenberg, H. J. (2007). The genetics of alcohol metabolism: Role of alcohol dehydrogenase and aldehyde dehydrogenase variants. *Alcohol Research and Health, 30*(1), 5–13.

Faivre, S., Demetri, G., Sargent, W., & Raymond, E. (2007). Molecular basis for sunitinib efficacy and future clinical development. *Nature Reviews Drug Discovery, 6*(9), 734–745. https://doi.org/10.1038/nrd2380

Felker, G. M., Butler, J., Collins, S. P., et al. (2015). Heart failure therapeutics on the basis of a biased ligand of the angiotensin-2 type 1 receptor. Rationale and design of the BLAST-AHF study (Biased Ligand of the Angiotensin Receptor Study in Acute Heart Failure). *JACC: Heart Failure, 3*, 193–201. https://doi.org/10.1016/j.jchf.2014.09.008

Flock, T., Hauser, A. S., Lund, N., Gloriam, D. E., Balaji, S., & Babu, M. M. (2017). Selectivity determinants of GPCR-G-protein binding. *Nature, 545*(7654), 317–322. https://doi.org/10.1038/nature22070

França, T. C. C. (2015). Homology modeling: An important tool for the drug discovery. *Journal of Biomolecular Structure and Dynamics, 33*(8), 1780–1793. https://doi.org/10.1080/07391102.2014.971429

Frauenfelder, H., Sligar, S. G., & Wolynes, P. G. (1991). The energy landscapes and motions of proteins. *Science, 254*(5038), 1598–1603. https://doi.org/10.1126/science.1749933

French, J. A., & Gazzola, D. M. (2011). New generation antiepileptic drugs: What do they offer in terms of improved tolerability and safety? *Therapeutic Advances in Drug Safety, 2*(4), 141–158. https://doi.org/10.1177/2042098611411127

Ghosh, A. K., Osswald, H. L., & Prato, G. (2016). Recent progress in the development of HIV-1 protease inhibitors for the treatment of HIV/AIDS. *Journal of Medicinal Chemistry, 59*(11), 5172–5208. https://doi.org/10.1021/acs.jmedchem.5b01697

Gobbo, D., Piretti, V., Di Martino, R. M. C., Tripathi, S. K., Giabbai, B., Storici, P., et al. (2019). Investigating drug-target residence time in kinases through enhanced sampling simulations. *Journal of Chemical Theory and Computation, 15*(8), 4646–4659. https://doi.org/10.1021/acs.jctc.9b00104

Gonçalves, A. M., Pedro, A. Q., Santos, F. M., Martins, L. M., Maia, C. J., Queiroz, J. A., et al. (2014). Trends in protein-based biosensor assemblies for drug screening and pharmaceutical kinetic studies. *Molecules, 19*(8), 12461–12485. https://doi.org/10.3390/molecules190812461

Gonceareneo, A., Li, M., Simonetti, F. L., Shoemaker, B. A., & Panchenko, A. R. (2017). Exploring protein-protein interactions as drug targets for anti-cancer therapy with in silico workflows. In I. Lazar, M. Kontoyianni, & I. Lazar (Eds.), *Methods in molecular biology* (Vol. 1647, pp. 221–236). https://doi.org/10.1007/978-1-4939-7201-2_15

Hensen, U., Meyer, T., Haas, J., Rex, R., Vriend, G., & Grubmüller, H. (2012). Exploring protein dynamics space: The dynasome as the missing link between protein structure and function. *PLoS One, 7*(5). https://doi.org/10.1371/journal.pone.0033931

Henzler-Wildman, K. A., Lei, M., Thai, V., Kerns, S. J., Karplus, M., & Kern, D. (2007). A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature, 450*(7171), 913–916. https://doi.org/10.1038/nature06407

Hollingsworth, S. A., & Dror, R. O. (2018). Molecular dynamics simulation for all. *Neuron, 99*(6), 1129–1143. https://doi.org/10.1016/j.neuron.2018.08.011

Hu, G., & Kurgan, L. (2019). Sequence similarity searching. *Current Protocols in Protein Science, 95*(1), e71. https://doi.org/10.1002/cpps.71

Huang, D., Zhou, T., Lafleur, K., Nevado, C., & Caflisch, A. (2010). Kinase selectivity potential for inhibitors targeting the ATP binding site: A network analysis. *Bioinformatics, 26*(2), 198–204. https://doi.org/10.1093/bioinformatics/btp650

Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: Visual molecular dynamics. *Journal of Molecular Graphics, 14*(1), 33–38. https://doi.org/10.1016/0263-7855(96)00018-5

Iqbal, N., & Iqbal, N. (2014). Imatinib: A breakthrough of targeted therapy in cancer. *Chemotherapy Research and Practice, 2014*, 1–9. https://doi.org/10.1155/2014/357027

Jalali-Yazdi, F., Chowdhury, S., Yoshioka, C., & Gouaux, E. (2018). Mechanisms for zinc and proton inhibition of the GluN1/GluN2A NMDA receptor. *Cell, 175*(6), 1520–1532. https://doi.org/10.1016/j.cell.2018.10.043

Jalencas, X., & Mestres, J. (2013). Identification of similar binding sites to detect distant polypharmacology. *Molecular Informatics, 32*(11–12), 976–990. https://doi.org/10.1002/minf.201300082

Jordan, I., Rogozin, I. B., Wolf, Y. I., & Koonin, E. V. (2002). Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Research, 12*(6), 962–968. https://doi.org/10.1101/gr.87702

Kalaivani, R., de Brevern, A. G., & Srinivasan, N. (2016). Conservation of structural fluctuations in homologous protein kinases and its implications on functional sites. *Proteins, 84*(7), 957–978. https://doi.org/10.1002/prot.25044

Kalinin, S., Valeri, A., Antonik, M., Felekyan, S., & Seidel, C. A. M. (2010). Detection of structural dynamics by FRET: A photon distribution and fluorescence lifetime analysis of systems with multiple states. *Journal of Physical Chemistry B, 114*(23), 7983–7995. https://doi.org/10.1021/jp102156t

Kim, B., Eggel, A., Tarchevskaya, S. S., Vogel, M., Prinz, H., & Jardetzky, T. S. (2012). Accelerated disassembly of IgE-receptor complexes by a disruptive macromolecular inhibitor. *Nature, 491*(7425), 613–617. https://doi.org/10.1038/nature11546

Klinman, J. P., & Kohen, A. (2014). Evolutionary aspects of enzyme dynamics. *Journal of Biological Chemistry, 289*(44), 30205–30212. https://doi.org/10.1074/jbc.R114.565515

Konc, J., Česnik, T., Konc, J. T., Penca, M., & Janežič, D. (2012). ProBiS-database: Precalculated binding site similarities and local pairwise alignments of PDB structures. *Journal of Chemical Information and Modeling, 52*(2), 604–612. https://doi.org/10.1021/ci2005687

Kovermann, M., Grundström, C., Elisabeth Sauer-Eriksson, A., Sauer, U. H., & Wolf-Watz, M. (2017). Structural basis for ligand binding to an enzyme by a conformational selection pathway. *Proceedings of the National Academy of Sciences of the United States of America, 114*(24), 6298–6303. https://doi.org/10.1073/pnas.1700919114

Kryshtafovych, A., & Fidelis, K. (2009). Protein structure prediction and model quality assessment. *Drug Discovery Today, 14*(7–8), 386–393. https://doi.org/10.1016/j.drudis.2008.11.010

Kumar, S., Ma, B., Tsai, C.-J., Sinha, N., & Nussinov, R. (2008). Folding and binding cascades: Dynamic landscapes and population shifts. *Protein Science, 9*(1), 10–19. https://doi.org/10.1110/ps.9.1.10

Kuzmanic, A., Bowman, G. R., Juarez-Jimenez, J., Michel, J., & Gervasio, F. L. (2020). Investigating cryptic binding sites by molecular dynamics simulations. *ACS Applied Materials and Interfaces, 53*(3), 654–661. https://doi.org/10.1021/acs.accounts.9b00613

Latorraca, N. R., Venkatakrishnan, A. J., & Dror, R. O. (2017). GPCR dynamics: Structures in motion. *Chemical Reviews, 117*(1), 139–155. https://doi.org/10.1021/acs.chemrev.6b00177

Lee, C. H., Lü, W., Michel, J. C., Goehring, A., Du, J., Song, X., et al. (2014). NMDA receptor structures reveal subunit arrangement and pore architecture. *Nature, 511*(7508), 191–197. https://doi.org/10.1038/nature13548

Li, H., Hallows, W. H., Punzi, J. S., Goldstein, B. M., Marquez, V. E., Carrell, H. L., et al. (1994). Crystallographic studies of two alcohol dehydrogenase-bound analogues of Thiazole-4-carboxamide Adenine Dinucleotide (TAD), the active anabolite of the antitumor agent tiazofurin. *Biochemistry, 33*(1), 23–32. https://doi.org/10.1021/bi00167a004

Luo, Q., Hou, C., Bai, Y., Wang, R., & Liu, J. (2016). Protein assembly: Versatile approaches to construct highly ordered nanostructures. *Chemical Reviews, 116*(22), 13571–13632. https://doi.org/10.1021/acs.chemrev.6b00228

Mabonga, L., & Kappo, A. P. (2019). Protein-protein interaction modulators: Advances, successes and remaining challenges. *Biophysical Reviews, 11*(4), 559–581. https://doi.org/10.1007/s12551-019-00570-x

Mak, K. K., & Pichika, M. R. (2019). Artificial intelligence in drug development: Present status and future prospects. *Drug Discovery Today, 24*(3), 773–780. https://doi.org/10.1016/j.drudis.2018.11.014

March-Vila, E., Pinzi, L., Sturm, N., Tinivella, A., Engkvist, O., Chen, H., et al. (2017). On the integration of in silico drug design methods for drug repurposing. *Frontiers in Pharmacology, 8*, 298. https://doi.org/10.3389/fphar.2017.00298

Marino-Buslje, C., Monzon, A. M., Zea, D. J., Fornasari, M. S., & Parisi, G. (2019). On the dynamical incompleteness of the Protein Data Bank. *Briefings in Bioinformatics, 20*(1), 356–359. https://doi.org/10.1093/bib/bbx084

Marrakchi, H., Lanéelle, G., & Quémard, A. (2000). InhA, a target of the antituberculous drug isoniazid, is involved in a mycobacterial fatty acid elongation system, FAS-II. *Microbiology, 146*(2), 289–296. https://doi.org/10.1099/00221287-146-2-289

Maveyraud, L., & Mourey, L. (2020). Protein X-ray crystallography and drug discovery. *Molecules, 25*(5), 1030. https://doi.org/10.3390/molecules25051030

McNamee, L. M., Walsh, M. J., & Ledley, F. D. (2017). Timelines of translational science: From technology initiation to FDA approval. *PLoS One, 12*(5), e0177371. https://doi.org/10.1371/journal.pone.0177371

Miller, Z., Kim, K. S., Lee, D. M., Kasam, V., Baek, S. E., Lee, K. H., et al. (2015). Proteasome inhibitors with pyrazole scaffolds from structure-based virtual screening. *Journal of Medicinal Chemistry, 58*(4), 2036–2041. https://doi.org/10.1021/jm501344n

Mohs, R. C., & Greig, N. H. (2017). Drug discovery and development: Role of basic biological research. *Alzheimer's and Dementia: Translational Research and Clinical Interventions, 3*(4), 651–657. https://doi.org/10.1016/j.trci.2017.10.005

Moroni, E., Paladino, A., & Colombo, G. (2015). The dynamics of drug discovery. Current Topics in Medicinal Chemistry, 15(20), 2043–2055. https://doi.org/10.2174/1568026615666150519102950

Muhammed, M. T., & Aki-Yalcin, E. (2019). Homology modeling in drug discovery: Overview, current applications, and future perspectives. *Chemical Biology and Drug Design, 93*(1), 12–20. https://doi.org/10.1111/cbdd.13388

Müller, C. W., Schlauderer, G. J., Reinstein, J., & Schulz, G. E. (1996). Adenylate kinase motions during catalysis: An energetic counterweight balancing substrate binding. *Structure, 4*(2), 147–156. https://doi.org/10.1016/S0969-2126(96)00018-4

Murakami, Y., & Mizuguchi, K. (2010). Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites. *Bioinformatics, 26*(15), 1841–1848. https://doi.org/10.1093/bioinformatics/btq302

Nakai, A., Fujimoto, J., Miyata, H., Stumm, R., Narazaki, M., Schulz, S., et al. (2019). The COMMD3/8 complex determines GRK6 specificity for chemoattractant receptors. *Journal of Experimental Medicine, 216*(7), 1630–1647. https://doi.org/10.1084/jem.20181494

Narayanan, C., Bernard, D. N., Bafna, K., Gagné, D., Chennubhotla, C. S., Doucet, N., et al. (2018). Conservation of dynamics associated with biological function in an enzyme superfamily. *Structure, 26*(3), 426–436. https://doi.org/10.1016/j.str.2018.01.015

Needle, D., Lountos, G. T., & Waugh, D. S. (2015). Structures of the middle east respiratory syndrome coronavirus 3C-like protease reveal insights into substrate specificity. *Acta Crystallographica Section D: Biological Crystallography, 71*(5), 1102–1111. https://doi.org/10.1107/S1399004715003521

Neu, H. C. (1996). Safety of cefepime: A new extended-spectrum parenteral cephalosporin. *American Journal of Medicine, 100*(6), 68S–75S. https://doi.org/10.1016/s0002-9343(96)00110-6

Ofran, Y., & Rost, B. (2007). ISIS: Interaction sites identified from sequence. *Bioinformatics, 23*(2), e13–e16. https://doi.org/10.1093/bioinformatics/btl303

Pagadala, N. S., Syed, K., & Tuszynski, J. (2017). Software for molecular docking: A review. *Biophysical Reviews, 9*(2), 91–102. https://doi.org/10.1007/s12551-016-0247-1

Pál, C., Papp, B., & Lercher, M. J. (2006). An integrated view of protein evolution. *Nature Reviews Genetics, 7*(5), 337–348. https://doi.org/10.1038/nrg1838

Pargellis, C., Tong, L., Churchill, L., Cirillo, P. F., Gilmore, T., Graham, A. G., et al. (2002). Inhibition of p38 MAP kinase by utilizing a novel allosteric binding site. *Nature Structural Biology, 9*(4), 268–272. https://doi.org/10.1038/nsb770

Patrick, G. L. (2013). *An introduction to medicinal chemistry* (5th ed.). Oxford: Oxford University Press.

Peng, Y., Alexov, E., & Basu, S. (2019). Structural perspective on revealing and altering molecular functions of genetic variants linked with diseases. *International Journal of Molecular Sciences, 20*(3), 548. https://doi.org/10.3390/ijms20030548

Petsko, G. A., & Ringe, D. (2010). X-ray crystallography in the service of structure-based drug design. In C. H. Reynolds, D. Ringe, & M. J. M. Kenneth (Eds.), *Drug design* (pp. 17–29). Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511730412.004

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF Chimera – A visualization system for exploratory research and analysis. *Journal of Computational Chemistry, 25*(13), 1605–1612. https://doi.org/10.1002/jcc.20084

Pieters, B. J. G. E., Van Eldijk, M. B., Nolte, R. J. M., & Mecinović, J. (2016). Natural supramolecular protein assemblies. *Chemical Society Reviews, 45*(1), 24–39. https://doi.org/10.1039/c5cs00157a

Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., et al. (2018). Drug repurposing: Progress, challenges and recommendations. *Nature Reviews Drug Discovery, 18*(1), 41–58. https://doi.org/10.1038/nrd.2018.168

Ramakrishnan, G., Chandra, N. R., & Srinivasan, N. (2015). Recognizing drug targets using evolutionary information: Implications for repurposing FDA-approved drugs against Mycobacterium tuberculosis H37Rv. *Molecular BioSystems, 11*(12), 3316–3331. https://doi.org/10.1039/C5MB00476D

Ramakrishnan, G., Chandra, N., & Srinivasan, N. (2017). Exploring anti-malarial potential of FDA approved drugs: An in silico approach. *Malaria Journal, 16*(1). https://doi.org/10.1186/s12936-017-1937-2

Rao, M. S., Gupta, R., Liguori, M. J., Hu, M., Huang, X., Mantena, S. R., et al. (2019). Novel computational approach to predict off-target interactions for small molecules. *Frontiers in Big Data, 2*. https://doi.org/10.3389/fdata.2019.00025

Rath, S. N., Ray, M., Pattnaik, A., & Pradhan, S. K. (2016). Drug target identification and elucidation of natural inhibitors for Bordetella petrii: An in silico study. *Genomics & Informatics, 14*(4), 241. https://doi.org/10.5808/gi.2016.14.4.241

Reid, A. J., Ranea, J. A. G., & Orengo, C. A. (2010). Comparative evolutionary analysis of protein complexes in E. coli and yeast. *BMC Genomics, 11*(1), 79. https://doi.org/10.1186/1471-2164-11-79

Roberts, N. A., Martin, J. A., Kinchington, D., Broadhurst, A. V., Craig, J. C., Duncan, I. B., et al. (1990). Rational design of peptide-based HIV proteinase inhibitors. *Science, 248*(4953), 358–361. https://doi.org/10.1126/science.2183354

Rohovie, M. J., Nagasawa, M., & Swartz, J. R. (2017). Virus-like particles: Next-generation nanoparticles for targeted therapeutic delivery. *Bioengineering & Translational Medicine, 2*(1), 43–57. https://doi.org/10.1002/btm2.10049

Rutenber, E. E., & Stroud, R. M. (1996). Binding of the anticancer drug ZD1694 to E. coli thymidylate synthase: Assessing specificity and affinity. *Structure, 4*(11), 1317–1324. https://doi.org/10.1016/S0969-2126(96)00139-6

Šali, A., & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology, 234*(3), 779–815. https://doi.org/10.1006/jmbi.1993.1626

Salsbury, F. R. (2010). Molecular dynamics simulations of protein dynamics and their relevance to drug discovery. *Current Opinion in Pharmacology, 10*(6), 738–744. https://doi.org/10.1016/j.coph.2010.09.016

Sanders, J. M., Monogue, M. L., Jodlowski, T. Z., & Cutrell, J. B. (2020). Pharmacologic treatments for coronavirus disease 2019 (COVID-19): A review. *Journal of the American Medical Association, 323*(18), 1824–1836. https://doi.org/10.1001/jama.2020.6019

Santos, R., Ursu, O., Gaulton, A., Bento, A. P., Donadi, R. S., Bologa, C. G., et al. (2016). A comprehensive map of molecular drug targets. *Nature Reviews Drug Discovery, 16*(1), 19–34. https://doi.org/10.1038/nrd.2016.230

Schuetz, D. A., Bernetti, M., Bertazzo, M., Musil, D., Eggenweiler, H. M., Recanatini, M., et al. (2019). Predicting residence time and drug unbinding pathway through scaled molecular dynamics. *Journal of Chemical Information and Modeling, 59*(1), 535–549. https://doi.org/10.1021/acs.jcim.8b00614

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature, 577*(7792), 706–710. https://doi.org/10.1038/s41586-019-1923-7

Shah, D. R., Shah, R. R., & Morganroth, J. (2013). Tyrosine kinase inhibitors: Their on-target toxicities as potential indicators of efficacy. *Drug Safety, 36*(6), 413–426. https://doi.org/10.1007/s40264-013-0050-x

Skjaerven, L., Hollup, S. M., & Reuter, N. (2009). Normal mode analysis for proteins. *Journal of Molecular Structure: THEOCHEM, 898*(1–3), 42–48. https://doi.org/10.1016/j.theochem.2008.09.024

Sowdhamini, R., Srinivasan, N., Guruprasad, K., Rufino, S., Dhanaraj, V., Wood, S., et al. (1995). Protein three-dimensional structure and molecular recognition: A story of soft locks and keys. *Pharmaceutica Acta Helvetiae, 69*(4), 185–192. https://doi.org/10.1016/0031-6865(95)00002-Q

Subramaniam, S., Kühlbrandt, W., & Henderson, R. (2016). CryoEM at IUCrJ: A new era. *IUCrJ, 3*, 3–7. https://doi.org/10.1107/S2052252515023738

Sudha, R., Katiyar, A., Katiyar, P., Singh, H., & Prasad, P. (2019). Identification of potential drug targets and vaccine candidates in Clostridium botulinum using subtractive genomics approach. *Bioinformation, 15*(1), 18–25. https://doi.org/10.6026/97320630015025

Sugiki, T., Furuita, K., Fujiwara, T., & Kojima, C. (2018). Current NMR techniques for structure-based drug discovery. *Molecules, 23*(1), 148. https://doi.org/10.3390/molecules23010148

Tischfield, M. A., Baris, H. N., Wu, C., Rudolph, G., Van Maldergem, L., He, W., et al. (2010). Human TUBB3 mutations perturb microtubule dynamics, kinesin interactions, and axon guidance. *Cell, 140*(1), 74–87. https://doi.org/10.1016/j.cell.2009.12.011

Tokuriki, N., & Tawfik, D. S. (2009). Protein dynamism and evolvability. *Science, 324*(5924), 203–207. https://doi.org/10.1126/science.1169375

Traynelis, S. F., Wollmuth, L. P., McBain, C. J., Menniti, F. S., Vance, K. M., Ogden, K. K., et al. (2010). Glutamate receptor ion channels: Structure, regulation, and function. *Pharmacological Reviews, 62*(3), 405–496. https://doi.org/10.1124/pr.109.002451

Uversky, V. N. (2019). Intrinsically disordered proteins and their “Mysterious” (meta)physics. *Frontiers in Physics, 7*, 10. https://doi.org/10.3389/fphy.2019.00010

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., et al. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery, 18*(6), 463–477. https://doi.org/10.1038/s41573-019-0024-5

Van Den Bedem, H., & Fraser, J. S. (2015). Integrative, dynamic structural biology at atomic resolution – It's about time. *Nature Methods, 12*(4), 307–318. https://doi.org/10.1038/nmeth.3324

Vishwanath, S., de Brevern, A. G., & Srinivasan, N. (2018). Same but not alike: Structure, flexibility and energetics of domains in multi-domain proteins are influenced by the presence of other domains. *PLoS Computational Biology, 14*(2), e1006008. https://doi.org/10.1371/journal.pcbi.1006008

Voet, A., Berenger, F., & Zhang, K. Y. J. (2013). Electrostatic similarities between protein and small molecule ligands facilitate the design of protein-protein interaction inhibitors. *PLoS One, 8*(10), e75762. https://doi.org/10.1371/journal.pone.0075762

Walls, A. C., Park, Y. J., Tortorici, M. A., Wall, A., McGuire, A. T., & Veesler, D. (2020). Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell, 181*(2), 281–292. https://doi.org/10.1016/j.cell.2020.02.058

Wang, L., Wu, Y., Deng, Y., Kim, B., Pierce, L., Krilov, G., et al. (2015). Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *Journal of the American Chemical Society, 137* (7), 2695–2703. https://doi.org/10.1021/ja512751q

Wilson, D. N. (2011). On the specificity of antibiotics targeting the large ribosomal subunit. *Annals of the New York Academy of Sciences, 1241*(1), 1–16. https://doi.org/10.1111/j.1749-6632.2011.06192.x

Wlodawer, A., & Vondrasek, J. (1998). Inhibitors of HIV-1 protease: A major success of structure-assisted drug design. *Annual Review of Biophysics and Biomolecular Structure, 27*(1), 249–284. https://doi.org/10.1146/annurev.biophys.27.1.249

Wüthrich, K. (1995). *NMR in structural biology: A collection of papers by Kurt Wüthrich*. World Scientific Publishing Company.

Xing, L., Klug-Mcleod, J., Rai, B., & Lunney, E. A. (2015). Kinase hinge binding scaffolds and their hydrogen bond patterns. *Bioorganic and Medicinal Chemistry, 23*(19), 6520–6527. https://doi.org/10.1016/j.bmc.2015.08.006

Yazhini, A., & Srinivasan, N. (2020). How good are comparative models in the understanding of protein dynamics? *Proteins: Structure, Function and Bioinformatics, 88*(7), 874–888. https://doi.org/10.1002/prot.25879

Zhang, J. B., Chang, S., Xu, P., Miao, M., Wu, H., Zhang, Y., et al. (2018). Structural basis of the proton sensitivity of human GluN1-GluN2A NMDA receptors. *Cell Reports, 25*(13), 3582–3590. https://doi.org/10.1016/j.celrep.2018.11.071

Zheng, W., Wen, H., Iacobucci, G. J., & Popescu, G. K. (2017). Probing the structural dynamics of the NMDA receptor activation by coarse-grained modeling. *Biophysical Journal, 112*(12), 2589–2601. https://doi.org/10.1016/j.bpj.2017.04.043

Chapter 6
Recent Trends in Computer-Aided Drug Design

Seneha Santoshi and Puniti Mathur

Abstract The process of drug discovery begins with the identification of a potential target. Depending on the availability of data, various computational approaches and tools have been explored from time to time for target identification and lead design. In this chapter, two case studies have been discussed. The first one involves newer approaches for target identification based on subtractive genomics and comparative metabolomics in the pathogenic bacteria, *Pseudomonas aeruginosa,* followed by lead design. The availability of complete genome sequences of pathogenic bacteria has increased the possibility of identification of promising targets, while considering host-pathogen interactions and host toxicity simultaneously. Subtractive genomics involves comparison of whole genomes of the host, pathogen and symbiotic organisms to identify unique essential genes. Similarly, comparative metabolomics is performed by comparison of all the known metabolic pathways in the above three categories. The entire approach was designed to identify a potential target that plays an essential role in the pathogen's survival and constitutes a critical component in its metabolic pathway. The second case study describes various steps in identification of a potential lead compound against a target protein using molecular docking and molecular simulation methods. It elaborates on choosing a lesser known target protein of malaria, belonging to the pre-erythrocytic cycle of *Plasmodium falciparum*. Prediction of three dimensional structure of the target using comparative modelling, followed by detailed docking and simulation studies lead to the identification of a promising lead molecule. Wet laboratory studies are warranted on results of both the *in silico* case studies for further validation.

S. Santoshi · P. Mathur (✉)
Centre for Computational Biology and Bioinformatics, Amity Institute of Biotechnology, Amity University Uttar Pradesh, Noida, Uttar Pradesh, India
e-mail: pmathur@amity.edu

S. K. Singh (ed.), *Innovations and Implementations of Computer Aided Drug Discovery Strategies in Rational Drug Design*,
https://doi.org/10.1007/978-981-15-8936-2_6

6.1 Introduction

Drug design, discovery and development is a complex and arduous process. Drug design involves a series of steps to identify a compound for effective treatment of a disease. The drug discovery and development process starts with understanding the disease for which the drug is to be designed. Briefly, it constitutes of following steps: (a) Identification of a therapeutic target (b) Lead Discovery (c) Lead Optimization and candidate drug development (d) Pre-clinical trials (e) Clinical trials to assess the safety, efficacy and adverse effects (if any) of newly developed drugs.

In general, it takes around 12–15 years in order to bring a new drug to the market along with a very heavy cost of upto $2 billion. Any failure on the way, later stages in particular, shall lead to heavy losses in terms of both, money and time invested in the research. Therefore, there has been a continual need for incorporating inexpensive technologies that shorten the length of the drug discovery process. It is also important to get quick answers on safety, efficacy and toxicity of a new molecule, with increasing reliability and at a lower cost during the development process itself. One such technology is Computer aided drug design (CADD), which plays a significant role at all the stages of drug discovery process, from the initial stage to the last stage of clinical development. CADD methods and bioinformatics tools offer significant benefits including efficient selection which results in cost savings and reduction in time to bring the drug to the market. In-depth analysis of drug receptor interactions is facilitated, which could be useful for efficient designing of novel drugs and improvement of the existing drugs.

There are several key areas where CADD plays an important role in designing an effective drug (Humer 2005). These include target identification, sequence analysis, comparative structure modelling, virtual high throughput screening, drug receptor interactions/physicochemical modelling, drug optimization and ADMET property prediction. A variety of methods and tools are available and being used in the above-mentioned processes. In this chapter, various processes/tools used in CADD have been explained with the help of two case studies. In the first case study, a few *in silico* approaches of target identification, which is an important step in the drug discovery process, along with target validation and lead identification is discussed. The next case study explains target structure prediction along with lead discovery, optimization and refinement.

6.2 Computational Identification of Unique Therapeutic Drug Targets in Bacterial Pathogens and Designing Lead Molecules: A Case Study

The target identification stage is the first step in the drug discovery process (Terstappen and Reggiani 2001) and t forms the basis of any target based drug discovery process which further demands for years of dedicated research in the pharmaceutical industry. For any drug target to progress towards the next stage, it must fulfil certain criteria. First and foremost it should not be toxic to the host, which means homology between target and host low or nonexistent (Freiberg 2001), it should have a specific activity in the diseased state (Wang et al. 2004; Sanseau 2001) and the target should be essential for the growth and survival (Freiberg 2001; Roemer et al. 2003; Sassetti et al. 2003). Some of these selection criteria can be ascertained by using available bioinformatics resources like metabolic pathway databases such as KEGG (Kyoto encyclopedia of genes and genomes) (Wixon and Kell 2000) protein classification sets such as COGs (clusters of orthologous groups) (Tatusov et al. 1997) and databases of druggable (potentially useful as drug targets) proteins (Sanseau 2001; Hopkins and Groom 2002; Robertson 2005).

With the advancement in sequencing technologies and the availability of genome sequences of pathogens, a plethora of information has become readily available to the researchers for exploring various approaches towards identification of potent targets and hence aid the process of drug discovery. Subtractive genomics approach is one of the recently adopted strategies in which genome subtraction approach is used to find out a dataset of genes, after comparing the host and pathogen that are likely to be essential to the pathogen but absent in the host. Further using another approach called comparative metabolomics, results are verified, in which the subtraction dataset between the host and pathogen metabolome provides information for a set of enzymes that are absent in the host but likely to be essential to the pathogen which can serve as potential therapeutic targets.

In recent times there is an urgent need of identification of novel drug targets, in order to design new defense against antibiotic sensitive pathogens. In order to achieve this, strategies of drug design are progressively shifting from genetic to the genomic and metabolomic approaches. Based on an understanding of the related biological processes in bacterial pathogens and their hosts, comparative genomics and metabolomics provide new opportunities for finding optimal targets among previously unexplored cellular functions. In general, a target should provide adequate selectivity; yielding a drug which is specific or highly selective against the pathogen with respect to the host. Moreover, for the for growth and viability of the pathogen, the selected target should be essential either for its survival or at least under the condition of infection.

The search for potential based on genomic and metabolomics approaches is built on the assumption that the potential target must be a critical component in the metabolic pathway of a pathogen and must play an essential role in the pathogen's survival. At the same time, in order to rule out any possibilities of host toxicity, this target should not have any well-conserved homolog in the host and symbiotic organisms (if any). This would decrease the possibilities of unwanted cross-reactivity that might prove detrimental to the host. The above approach to target identification is termed as a subtractive approach because while comparing the two genomes and metabolomes under consideration, we use a subtraction dataset obtained after comparison, for further analysis. The focus is on the complement of the dataset of the pathogen that is essential for it but is not present in host.

In the present case study we have used the subtractive genomics and comparative metabolomics approaches in the pathogen, *Pseudomonas aeruginosa*, in order to identify a list of potential therapeutic targets. *P. aeruginosa*, a Gram-negative bacterium, is the embodiment of a highly opportunistic pathogen of humans. If the tissue defenses are compromised in some manner, then there is hardly any tissue this bacterium cannot infect. It is a common cause of respiratory system infections, dermatitis, urinary tract infections, gastrointestinal infections, soft tissue infections, bone and joint infections, bacteremia and a variety of systemic infections. It is particularly dangerous for cancer patients, patients with severe burns and in and AIDS patients who are immunosuppressed. *Pseudomonas aeruginosa* infection is a serious problem in hospitals where patients with cancer, cystic fibrosis, and burns are being treated. *Pseudomonas aeruginosa* can cause a variety of skin infections, both localized and diffuse. Breakdown of the integument which may result from burns, trauma or dermatitis; high moisture conditions such as those found in the ear of swimmers and the toe webs of athletes, hikers and combat troops, in the perineal region and under diapers of infants, and on the skin of whirlpool and hot tub users are the common predisposing factors (Stover et al. 2000). One of the biggest problems in treatment of bacterial infections is the presence of several multidrug efflux pumps from the major facilitator superfamily (MFS), multidrug and toxic compound extrusion (MATE) families, ATP-binding cassette (ABC) and small multi-drug resistance (SMR) which increased its intrinsic resistance to many efficient antibiotics. Owing to its resistance to antibiotics, *Pseudomonas aeruginosa* has become very infamously a dangerous and dreaded pathogen. (Mdluli et al. 2006). Thus, it has come up as a challenge to developing new antibacterial drugs against these kind of pathogens.

As a proof of concept, many of the genes identified by our approaches are also reported as essential by experimental methods. Furthermore, our approach successfully identified a number of promising protein targets for new antibiotic development.

6.2.1 Approaches for Drug Target Identification

There are many *in silico* approaches for finding drug targets in pathogenic bacteria (Bruccoleri et al. 1998) one of the research groups had developed a simple and computational tool that can determine concordances of putative gene products showing sets of proteins conserved across one set genomes, but are not present in another set of genomes, but the availability of this approach as an automated tool is limited. An automated tool, T-iDT developed by Singh et al. (2006) predict highly conserved genes, which are essential for pathogenic bacteria with no similarities with the host genes as potential drug targets. Geptop: A Gene Essentiality Prediction Tool based on orthology and phylogeny, offers gene essentiality annotations (Wei et al. 2013; Wen et al. 2019). This and other existing tools use only human genome sequence as a template for comparison against pathogens. However, comparison with the symbiotic organisms living within the human body cannot be ruled out for successful drug development. Fortunately the genome sequences of all these symbiotic bacteria are available and can be used as template for comparison with pathogen bacteria. Thus bioinformatics resources and tools are highly efficient in the identification of putative drug targets, but at the same time validation of these targets is again very essential. Traditionally the validation depends on laboratory-based work which helps in understanding the role of the gene or protein in the disease process.

To gain insights into the molecular mechanisms that underlie disease, bioinformatics plays a key role in the exploring and analyzing the genomic, transcriptomic, and proteomic data and to identify potential drug targets. Thus, genomics and proteomics technologies have created a paradigm shift in the drug discovery process. In this work we discuss the current state of the art for some of the bioinformatics approaches to identifying drug targets. It makes use of database of essential genes (DEG) (http://tubic.tju.edu.cn/deg/) and the comparative and subtractive genomic approaches to compare with the pathogen bacteria versus human as well as its symbiotic bacteria, including identifying new members of successful target classes and their functions, predicting disease relevant genes.

6.2.2 Identification of Unique Targets Based on Comparative Genomics

Protein sequences for *Homo sapiens* and *P. aeruginosa* (PA7) was downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/). The DEG database (Zhang et al. 2004) was downloaded from http://tubic.tju.edu.cn/deg/ and manually compiled to use as a stand-alone database for the BLAST program (Dutta et al. 2006). The *P. aeruginosa* genes were purged at 60% using CD-HIT to exclude paralogs from further analysis. The standalone BLAST executables including BLASTn, BLASTp and BLASTx were downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/) and installed locally. The set of essential genes in *P. aeruginosa* have been predicted based on homologous sequence search against DEG using BLASTn (E-value 10^{-8}). An interesting approach designated “differential genome display” has been proposed for the prediction of potential drug targets (Huynen et al. 1997; Huynen et al. 1998). The resultant set of essential genes of pathogenic bacteria was subjected to BLASTn (E-value, 10^{-8}) against complete genome of non-pathogenic bacteria (*Pseudomonas putida*; KT2440). Those genes are likely to be important for pathogenicity and, that are present in the genome of a pathogenic bacterium, but absent in the genome of a closely related free-living bacterium (non-pathogenic) and may be considered candidate drug targets. The subtracted essential genes of the pathogen were subjected to BLASTx (E-value, 10^{-5}) with complete human proteome to identify pathogen genes non-homologous in humans. The non-homologous entries were then subjected to BLASTx (E-value, 10^{-10}) against complete proteomes of four strains of symbiotic organisms: *Bacteroides thetaiotaomicron*, *Escherichia coli*, *Lactobacillus* *acidophilus* and *Lactobacillus johnsoni* for the identification of a total of 6 nonhomologous unique essential genes in the pathogen designated as potent therapeutic targets. The overall protocol used for identification of therapeutic targets based on comparative genomics is represented in Fig. 6.1 The identified genes were then classified into different groups based on gene names and subsequently matched against the list of essential genes in *P. aeruginosa* identified by mutagenesis (Jacobs et al. 2003). After thorough literature survey, out of the list of 6 unique essential genes only dapD was considered for virtual screening of lead molecules in *P. aeruginosa* (Table 6.1).

6.2.3 Subtractive Metabolomics Approach for Identification of Unique Enzymes in Unique Pathways (Naik et al. 2010)

In this approach the metabolic pathways present in the host were compared with the pathogen. We have used Kyoto Encyclopedia of Genes and Genomes to obtain the metabolic pathway information (Kanehisa et al. 2002). We have compared all the pathways of pathogen with host, 213 pathways of human being. An important question to be addressed while choosing potential drug targets is whether the biochemical pathway to be targeted is unique to bacteria. In case of *P. aeruginosa* the 20 pathways addressed in this study were all absent in the host *H. sapiens* and therefore unique to the pathogen *Pseudomonas aeruginosa*. The enzyme sequences in FASTA format from each metabolic pathway of *Pseudomonas aeruginosa* were taken and did BLASTp with the human proteome with threshold E-value of 10^{-5}. These filtered enzymes were then compared with proteome of non pathogen using BLASTp with an E-value of 10^{-8}. The enzymes which were found to be uniquely present in the pathogen but absent in non pathogen were further compared with proteomes of symbiotic organisms using BLASTp with an E-value of 10–10. All the enzymes which were found to be non homologous with symbiotic organisms were filtered out; they are unique enzymes and can be considered as therapeutic targets as listed in Table 6.2.

However, after thorough literature survey, only GspL and pilA were considered for virtual screening of lead molecules for *P. aeruginosa*. The overall protocol used for identification of therapeutic targets of *Pseudomonas aeruginosa* based on comparative metabolomics is represented in Fig. 6.2.

6.2.4 Sequence Analysis

Domain analysis (ProDom, Pfam) was carried out to check whether any domain present in the target protein could interfere in the action of the drug. The screening of

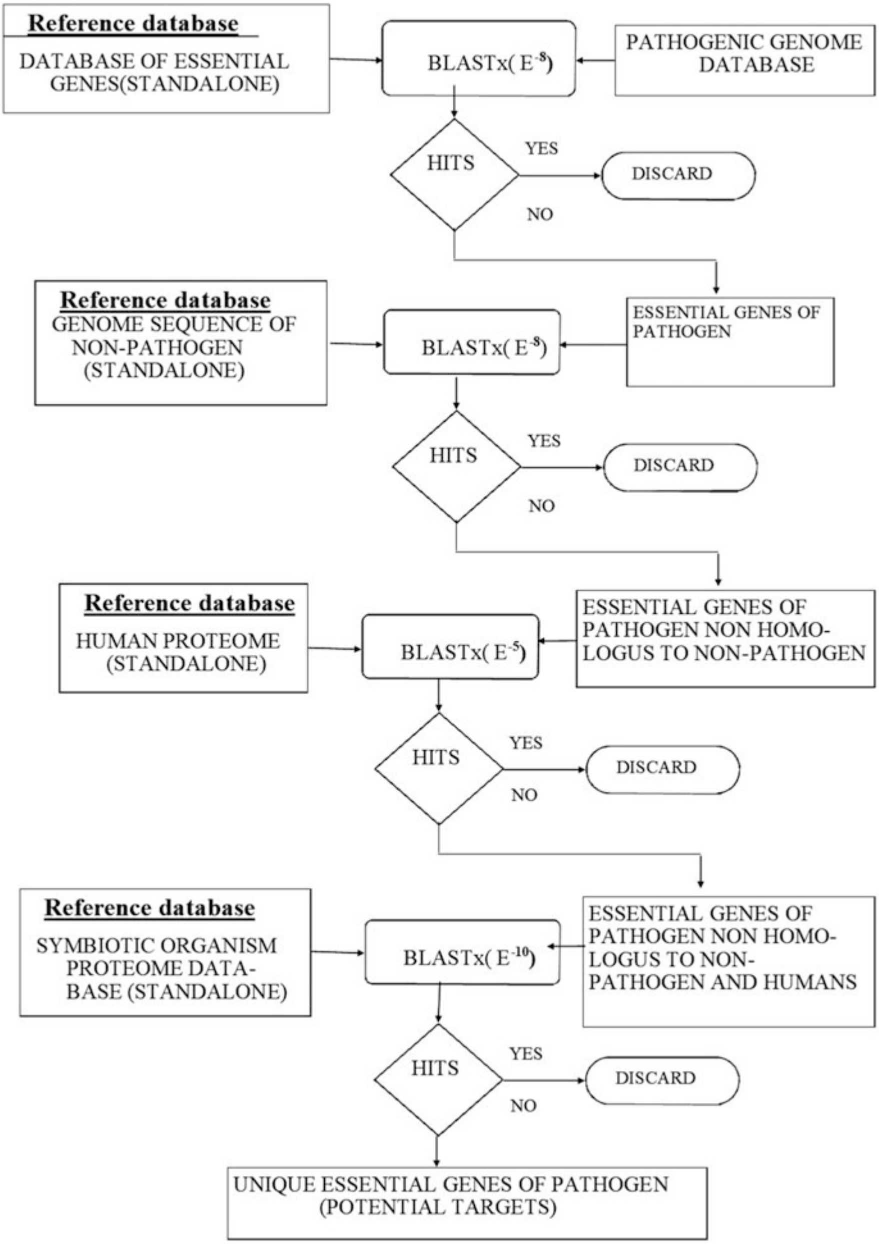

Reference database
DATABASE OF ESSENTIAL GENES(STANDALONE)
BLASTx(E^{-8})
PATHOGENIC GENOME DATABASE
HITS
YES
DISCARD
NO
Reference database
GENOME SEQUENCE OF NON-PATHOGEN (STANDALONE)
BLASTx(E^{-8})
ESSENTIAL GENES OF PATHOGEN
HITS
YES
DISCARD
NO
Reference database
HUMAN PROTEOME (STANDALONE)
BLASTx(E^{-5})
ESSENTIAL GENES OF PATHOGEN NON HOMO-LOGUS TO NON-PATHOGEN
HITS
YES
DISCARD
NO
Reference database
SYMBIOTIC ORGANISM PROTEOME DATA-BASE (STANDALONE)
BLASTx(E^{-10})
ESSENTIAL GENES OF PATHOGEN NON HOMO-LOGUS TO NON-PATHOGEN AND HUMANS
HITS
YES
DISCARD
NO
UNIQUE ESSENTIAL GENES OF PATHOGEN (POTENTIAL TARGETS)

Fig. 6.1 Workflow of the protocol of subtractive genomics for identification of potential targets which are essential essential genes/proteins of the pathogen which are not present in the non-pathogenic organism, human host and symbiotic organisms in human host (Naik et al. 2010)

Table 6.1 Drug targets identified in *P. aeruginosa* after four different levels of subtraction using comparative genomics (Naik et al. 2010)

No.	Gene Id	Gene	Protein name	Pathway
1	PSPA7_1473	dapD	Tetrahydrodipicolinate succinylase	Lysine biosynthesis (pap00300)
	PSPA7_1235	narH	nitrate reductase, beta subunit	Nitrogen metabolism (pap00910); Signal transduction (pap02020)
	PSPA7_1556	metE	5methyltetrahydropteroyltriglutamate homocysteine methyltransferase	Methionine metabolism (pap00271)
	PSPA7_3847	ccmF	cytochrome C-type biogenesis protein CcmF	Nitrogen metabolism (pap00910)
	PSPA7_5673	hfq	RNA-binding protein Hfq	Not available
	PSPA7_1508		putative inner membrane protein	Protein export (pap03060)

the database in this study proved that the domains present in selected targets of the pathogen, dapD, GspL and pilA are not ubiquitous in host. Finally, the proteins that have been selected (viz. dapD, GspL and pilA were further analyzed for structure availability. The structures of all the above-mentioned proteins were not available, therefore the structures were modelled using homology modelling method.

6.2.5 Homology Model Construction

The homology models of the proteins: dapD, GspL and pilA of *pseudomonas aeruginosa* were built using Prime (Prime version 1.5, Macromodel version 9.1, Schrodinger, LLC, New York, NY, 2005) accessible through the Maestro interface (Schrodinger, Inc.). To select the templates, a basic local alignment search tool (BLAST) search against the PDB was conducted using our protein sequence as a query. During the homology model building, Prime keeps the backbone rigid for the cases in which the backbone does not need to be reconstructed due to gaps in the alignment. Unfavorable steric contacts were screened and PRIME was used to remodel the structure using a rotamer library database. Explicit hydrogens were added to the protein and the protein model was energy minimized using the Macromodel (Prime version 1.5) force-field MMFFS. Energy minimization and relaxation of the loop regions was performed using 300 iterations in a simple minimization method. Again the steepest descent was carried out until the energy showed stability in the sequential repetition. Model evaluation was performed in PROCHECK v3.4.4 (Laskowski and MacArthur 1993) producing plots that were analyzed for the overall and residue-by-residue geometry. Ramachandran Plot (Ramachandran et al. 1963) provided by the program PROCHECK assured very good confidence for the predicted protein. There were only 0.3% residues in the disallowed region and 0.9% residues in generously allowed regions. Nevertheless,

Table 6.2 Twenty two unique potential targets in *Pseudomonas aeruginosa* (PA7) predicted from comparative metabolomics of *P. aeruginosa* with *H. sapiens*, *P. Putida*, and symbiotic organisms (Naik et al. 2010)

S No.	Gene id	locus	EC no.	Gene	Protein name
Amino Acid Metabolism					
1	PSPA7_0967	K01826	EC:5.3.3.10	hpcD	
Xenobiotics Biodegradation and Metabolism					
1,2-Dichloroethane degradation					
1	PSPA7_4708	K01560	EC:3.8.1.2	dehII	Putative haloacid dehalogenase
Gamma-Hexachlorocyclohexane degradation					
1	PSPA7_4709	K01561	EC:3.8.1.3	dehII	Putative haloacid dehalogenase
Benzoate degradation via hydroxylation					
1	PSPA7_0968	K01827	EC:5.3.3.11	hpcD	5-carboxymethyl-2-hydroxymuconate isomerase
Folding, Sorting and Degradation					
Type II secretion system					
1	PSPA7_2029	K02452		xcpP	General secretion pathway protein C
2	PSPA7_2033	K02457		gspH2	General secretion pathway protein H
3	PSPA7_1411	K02458		gspI3	General secretion pathway protein I
4	*PSPA7_2037*	*K02461*		*gspL*	*General secretion pathway protein L*
5	PSPA7_4870	K02282		CpaE	Pilus assembly protein
6	PSPA7_4872	K02279		cpaB	Flp pilus assembly protein
7	*PSPA7_4873*	*02651*		*PilA*	*Pilus assembly protein*
8	PSPA7_5780	K02663		PilN	Type IV pilus assembly protein
9	PSPA7_5779	K02664		pilO	Type IV pilus assembly protein
10	PSPA7_5195	K02675		pilY2	Type IV pilus assembly protein
11	PSPA7_5194	K02674		pilY1	Type IV pilus assembly protein
12	PSPA7_5193	K02673		pilX	Type IV pilus assembly protein
13	PSPA7_5192	K02672		pilW	Type IV pilus assembly protein
Type IV secretion system					
1	PSPA7_3697	K03195		VirB1	Conjugation TrbI family protein
2	PSPA7_3698	K03204		VirB9	Conjugal transfer protein VirB9
3	PSPA7_3699	03200		VirB5	Type IV secretion system protein VirB5
4	PSPA7_3703	K03199		VirB4	ATPase
5	PSPA7_3704	K03198		VirB3	Conjugal transfer protein type IV secretion

PROCHECK assured the reliability of the structure and the protein was subjected to VERIFY3D (Eisenberg et al. 1997) available from NIH MBI Laboratory Servers (Naik et al. 2010). These modelled structures were used for lead designing (Fig. 6.3).

In silico prediction of binding sites of these proteins was done using SiteMap (Schrodinger Inc.). SiteMap treat entire proteins to locate binding sites whose size, functionality, and extent of solvent exposure meet user specifications. SiteScore, the

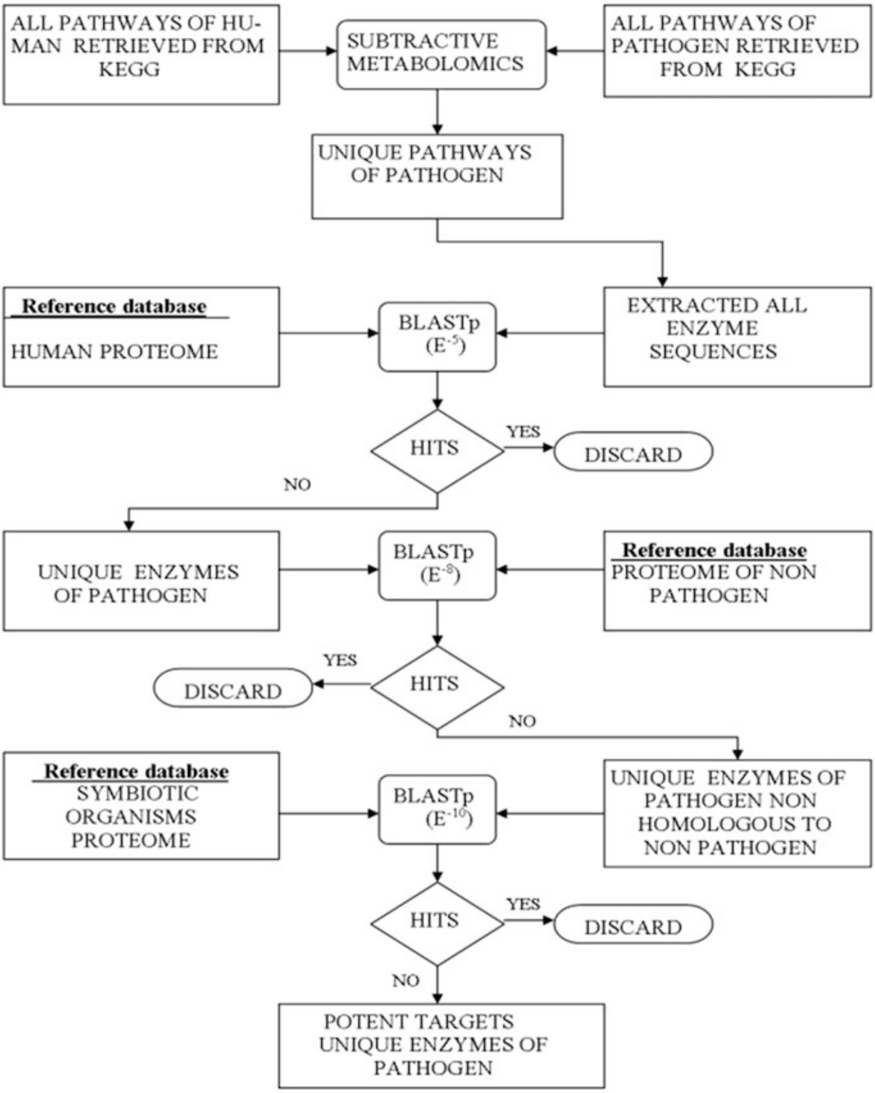

ALL PATHWAYS OF HUMAN RETRIEVED FROM KEGG
SUBTRACTIVE METABOLOMICS
ALL PATHWAYS OF PATHOGEN RETRIEVED FROM KEGG
UNIQUE PATHWAYS OF PATHOGEN
Reference database
HUMAN PROTEOME
BLASTp (E-5)
EXTRACTED ALL ENZYME SEQUENCES
HITS
YES
DISCARD
NO
UNIQUE ENZYMES OF PATHOGEN
BLASTp (E-8)
Reference database
PROTEOME OF NON PATHOGEN
YES
DISCARD
HITS
NO
Reference database
SYMBIOTIC ORGANISMS PROTEOME
BLASTp (E-10)
UNIQUE ENZYMES OF PATHOGEN NON HOMOLOGOUS TO NON PATHOGEN
HITS
YES
DISCARD
NO
POTENT TARGETS UNIQUE ENZYMES OF PATHOGEN

Fig. 6.2 The protocol for identification of unique enzyme targets using comparative metabolomics approach for *Pseudomonas aeruginosa.* These targets are unique to the pathogen while being absent in the non-pathogen, human host and symbiotic organisms (Naik et al. 2010)

scoring function used to assess a site's propensity for ligand binding, accurately ranks possible binding sites to eliminate those not likely to be pharmaceutically relevant. It identifies potential ligand binding sites by linking together "site points" that are suitably close to the protein surface and sufficiently well sheltered from the solvent. Given that similar terms dominate the site scoring function, this approach

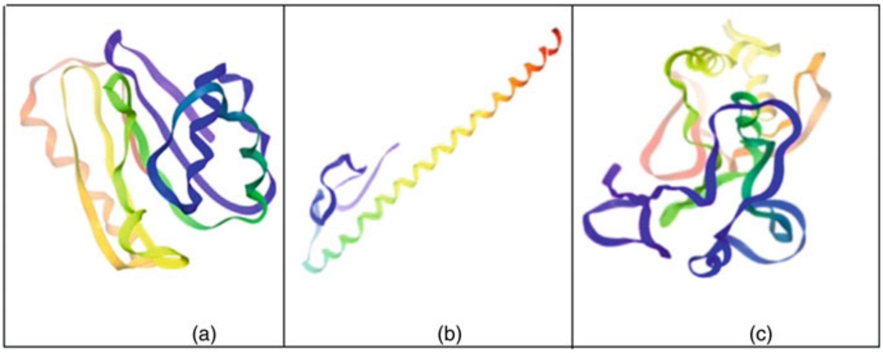


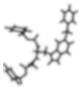
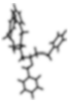
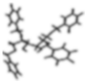
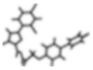
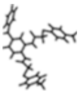
Fig. 6.3 Modeled three-dimensional structure of gspL (**a**), pilA (**b**) and dapD (**c**) of *P. aeruginosa* Ligand Binding Site Prediction (Naik et al. 2010)

ensures that the search focuses on regions of the protein most likely to produce tight protein-ligand or protein–protein binding. Subsites are merged into larger sites when they are sufficiently close and could be bridged in solvent-exposed regions by ligand atoms. SiteMap evaluates sites using a series of properties. The binding site with highest site score was taken for docking and virtual screening of the lead molecules.

6.2.6 Docking Studies

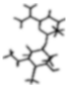
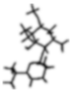
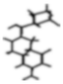
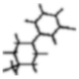
A library of 1,25,000 compounds taken from pubchem drug database (http://zinc.docking.org/vendor0/index_nfs.shtml) were compiled together to form a standalone library. These compounds were available in 3D-MOL2 file. These molecules were imported into maestro and finally prepared using ligPrep. LigPrep is a utility of Schrodinger suit that combines tools for generating 3D structures from 1D and 2D representation, searching for tautomers and steric isomers and perform a geometry minimization of ligands. The Schrodinger Glide program version 4.0 has been used for docking. Docking procedure consisted of three interrelated components; (a) identification of binding site, (b) a search algorithm to effectively sample the search space (the set of possible ligand positions and conformations on the protein surface) and (c) a scoring function. For each ligand in the virtual library, the pose with the lowest Glide (HTVS) score was refined using Glide (XP) docking. The best 20 ligands for the pathogen, shown in Tables 6.3, 6.4, and 6.5 respectively, chosen with Glide score proved their reliability in the Glide (XP). Since computational screenings always demand experimental testing in order to confirm the accurate drug molecule(s), the proposed LEAD molecules need to be optimized in further studies. The significance of this work is in providing a relatively inexpensive approach to screen compounds that are likely to inhibit the action of selected drug targets, dapD, GspL and pilA in *Pseudomonas aeruginosa*.(Naik et al. 2010).

Table 6.3 Database ID number, chemical structure, IUPAC names and Glide scores of top scored ligands docked with dapD of *Pseudomonas aeruginosa* using docking program Glide (Naik et al. 2010)

No.	Ligand	ZINC ID	Ligand name	Glide (HTVS) score	Glide (XP) score
1		ZINC04014946	phenylmethyl 2-[[2-(phenylmethoxycarbonylamino)-3-[5- (phenylmethoxy)-1H-indol-3-yl] propanoyl]amino]acetate	−7.01	−15.65
2		ZINC03994289	(2R,3S,4S,5R)-1,6-bis[(2,6-difluorophenyl)methoxy]-2,5- bis (phenylmethoxy)hexane-3,4-diol	−7.79	−15.19
3		ZINC03994290	(2R,3S,4S,5R)-1,6-bis [(2-fluorophenyl)methoxy]-2,5-bis[(4fluorophenyl) methoxy]hexane-3,4-diol	−8.73	−14.77
4		ZINC03146249	N-[[5-(2,4-dichlorophenyl)furan-2-yl]methylideneamino]-2-(4-phenylphenoxy)acetamide	−7.56	−13.56
5		ZINC02154335	N-[5-(benzoyl)-2-[[2-(4-methylphenyl)sulfanylacetyl] amino]phenyl]-2-(4-methylphenyl) sulfanylacetamide		−7.71 –13.50

The data presented here demonstrate that stepwise comparisons of data using simple biological criteria can be an effective way obtaining a set of experimentally manageable number of genes of interest. This process is an efficient way for enriching potential target genes, and for identifying those that are critical for normal cell function. The generation of a comprehensive essential gene list will allow an accelerated genetic dissection of traits such as metabolic flexibility and inherent drug resistance that render *P. aeruginosa* such a tenacious pathogen. Such a strategy will enable us to locate critical pathways and steps in pathogenesis; to target these steps by designing new drugs; and to inhibit the infectious agent of interest with new antimicrobial agents. We propose probable chemical compounds, which could be tested to devise drug molecules to retard the hazardous proliferation of *P. aeruginosa*. The scope of this work could be to use this data to do cost-effective experimental screening. The proposed potential chemical compounds could provide the prime lead for future experimental screening.

Table 6.4 Database ID number, chemical structure, IUPAC names and Glide scores of top scored ligands docked with GspL of *Pseudomonas aeruginosa* using docking program Glide (Naik et al. 2010)

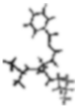
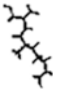
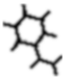
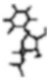
No.	Ligand	ZINC ID	Ligand Name	Glide (HTVS) score	Glide (XP) score
1		ZINC03831196	5-amino-2-(aminomethyl)-6-(4,6-diamino-2,3-dihydroxycyclohexyl)oxyoxane-3,4-diol	−7.45	−10.54
2		ZINC03830242	2-amino-3-[3-amino-6-(1-aminoethyl) oxan-2-yl]oxy-6-methoxy-5-(methylamino) cyclohexane-1,4-diol	−6.86	−7.83
3		ZINC03830241	2-amino-3-[3-amino-6-(1-aminoethyl)oxan-2-yl]oxy-6-methoxy-5-(methylamino)-cyclohexane-1,4-diol	−6.91	−7.49
4		ZINC03870170		−6.62	−7.48
5		ZINC02383344	1-(6-chloropyridin-2-yl)piperidin-4-amine	−6.86	−7.24

6.3 Modeling and Molecular Dynamics of *Plasmodium falciparum* SPECT Protein and Screening of Biogenic Compounds

Malaria is a major parasitic disease affecting a large population in tropical and subtropical countries and causing 1–2 million deaths every year. Non availability of an effective vaccine and emergence of multi-drug resistant strains of the malarial parasite *Plasmodium falciparum* continues to fuel search for newer and more effective ways of tackling the disease. Here, a target protein of the malarial parasite *P. falciparum* has been identified from the Tropical Disease Research (TDR) target database followed by structure-based drug design. TDR is a global research collaboration program for facilitating scientific efforts to combat diseases widely prevalent in the tropical countries. It is jointly sponsored by United Nations Development Programme (UNDP), United Nations Children's Education Fund (UNICEF), World Health Organization (WHO) and the World Bank.

Life cycle of *P. falciparum* is complex and multistaged; occuring within the female anopheles mosquito and human, with the former acting as vector and the latter as host. The parasite undergoes various stages of development namely, sporozoite, merozoite, trophozoite and gametocyte during its life cycle. The sporozoites are introduced into the dermis of the human host through the bite of the mosquito.

Table 6.5 Database ID number, chemical structure, IUPAC names and Glide scores of top scored ligands docked with pilA of *Pseudomonas aeruginosa* using docking program Glide (Naik et al. 2010)

No.	Ligand	ZINC ID	Ligand name	Glide (HTVS) score	Glide (XP) score
1		ZINC03871693		−8.90	−9.08
2		ZINC01888932	(2S)-2-[[(2S)-2,6-hydroxypropanoic acid	−5.46	−6.89
3		ZINC03786623		−5.77	−6.71
4		ZINC03870884	5-(5-methyl-2,4-dioxopyrimidin-1-yl)-2-(phosphonoxy methyl) oxolan-3-yl	−5.28	−6.24
5		ZINC03870179	5-(2,4-dioxo-1H-pyrimidin-5-yl)-3,4-dihydrogen phosphate	−4.61	−6.24

These sporozoites traverse through host cell barriers and enter blood by displaying various modes of motility such as gliding, host cell traversal and transmigration between cells (Victoria et al. 2009; Mali et al. 2008; Mota et al. 2001; Tavares et al. 2013). Once inside the liver, the sporozoites invade the hepatocytes and develop into exoerythrocytic forms (EEF) (Amino et al. 2008a; Yang and Boddey 2017; Sturm et al. 2006). The EEFs transform into merozoites which exit the liver and infect erythrocytes, leading to the clinical manifestation of malaria. Most of the efforts to find safe and effective drugs for malaria have been restricted to the blood stage of the parasite.

Recent studies have revealed a few drug targets in the pre erythrocytic or liver stage of the parasite (Amino et al. 2006; Yamauchi et al. 2007). Of particular interest is the phenomenon of host cell traversal by the parasite (Amino et al. 2008a; Derbyshire et al. 2011; Mazier et al. 2009; Vanderberg et al. 1990). Cell traversal involves entry of sporozoite into a host cell, transit through the host cell cytosol, and finally exit from the host cell plasma membrane (Mota et al. 2001; Mazier et al. 2009). The sporozoites enter and exit the hepatocytes through the formation of a pore in the plasma membrane, which may lead to necrosis. However, resealing of the wound and persistence of the cell has been observed in some cases (Mota et al. 2001; Amino et al. 2008a; Ishino et al. 2004; Formaglio et al. 2014). The sporozoites traverse the liver cells initially inside transient vacuoles which later lead to the formation of parasitophorous vacuoles (Frevert et al. 2005; Sibley 2004; Amino et al. 2008b). pH sensing and special proteins such as sporozoite microneme proteins essential for cell traversal are employed by sporozoites to exit these vacuoles and at the same time also avoid degradation by host lysosomes (Risco-Castillo et al. 2015).

SPECT (sporozoite microneme protein essential for cell traversal) and SPECT2 (also known as perforin like protein 1, PLP-1) are two secretory parasitic proteins which were initially thought to facilitate pore formation in the host cell membrane upon entry of the sporozoite (Mazier et al. 2009; Risco-Castillo et al. 2015). It was later reported that *P. berghei* and *P. yoelii* parasites which did not have these proteins could still readily enter hepatocytes. However these mutants cannot exit the vacuole or even the host cell (Amino et al. 2008b). Detailed studies in *Plasmodium berghei* reveal that both proteins, SPECT and SPECT2 have signal sequences and are located inside micronemes, secretory organelles present at the apical end of sporozoites. *P. berghei* SPECT2 (*Pb*SPECT2) is now known to have a direct role in pore formation (Kaiser et al. 2004; Ishino et al. 2005; Hamaoka and Ghosh 2014). *P. berghei* SPECT, a 25 kD protein, forms a four-helix bundle, with the rare feature of having all the helices in parallel or antiparallel alignment (Hamaoka and Ghosh 2014). However, an unfavorable packing of side chains in the protein leads to an unstable conformation. It is hypothesized that *Pb*SPECT could be triggered to undergo a conformational transition from soluble to membrane associated form. Homologs of *Pb*SPECT have been found only in the genomes of various Plasmodium species, including *P. falciparum* (Garg et al. 2013). These proteins have 92% sequence identity, suggesting that they share a common structure and mode of action. Three-dimensional structure prediction (Fran□a 2015) of *Pf*SPECT followed by docking studies using a portion of ZINC database (Sterling and Irwin 2015; Irwin et al. 2012) has been described below. Molecular dynamics studies on ligand-protein complexes have also been elaborated further.

6.3.1 Structure Modelling, Refinement and Validation

The 3D structure of the *Pf*SPECT protein (*Plasmodium falciparum* isolate 3D7, accession number XP_001350146) was predicted using knowledge-based homology modelling method. Template search was performed using BLAST and HHBlits against the SWISS-MODEL template library (SMTL). The target sequence was submitted to BLAST (Guex and Peitsch 1997) against the protein sequences contained in the template library of SWISS-MODEL. Seven templates were found against the target sequence. HHblits profile was built after which one iteration of HHblits was performed against NCBI's nr20(Remmert et al. 2011).Total ten templates were found after screening against all profiles of the SMTL. The quality of the templates was predicted using the target template alignment feature, which led to the selection of *Pb*SPECT (PDB ID: 4u5a) as the template for development of the 3D model. Promod2 was used to generate a model based on the target- template alignment process. Energy minimization of the modelled protein was performed using Macromodel (version 9.9, Schrodinger) and OPLS 2005 force field with

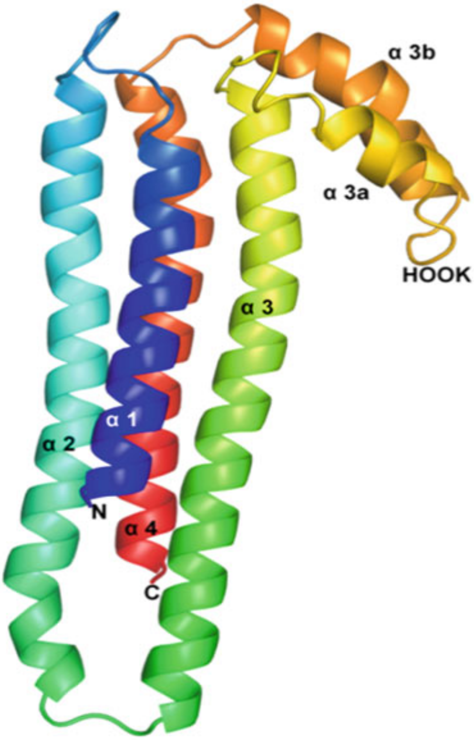


Fig. 6.4 Three-dimensional Structure of modelled *Pf*SPECT (Srivastava et al. 2017)

PRCG algorithm using 1000 steps of minimization and energy gradient of 0.001. The overall quality of the model was evaluated using SAVES server. The modeled protein was largely helical consisting of four α-helices, as shown in Fig. 6.4.

In order to refine the protein model further, Desmond 2.2 (D. E. Shaw Research 2009; Bowers et al. 2006) using OPLS 2005 force field was employed to perform a 50 nanoseconds molecular dynamics simulation. The protein was solvated in cubic box with a dimension of 20 Å using simple point charge water molecules, which were then replaced with 8 Na^+ counterions for electroneutrality. A total of 5000 frames were generated in the MD trajectories, out of which the last 2000 frames were used to generate structure of the *Pf*SPECT protein. The molecular dynamics simulation reduced the potential energy of *Pf*SPECT model from −50867.023 kJ/mol to −52162.293 kJ/mol, indicating an improved model. The rmsd plot (Fig. 6.5) shows that the protein conformation was largely stable. The model was further validated using PROCHECK (Laskowski and MacArthur 1993) and ERRAT (Colovos and Yeates 1993) PROCHECK results showed 91.4% of backbone angles were in allowed regions, 7.6% residues in additionally allowed regions and 1% residues in generously allowed regions (Srivastava et al. 2017).

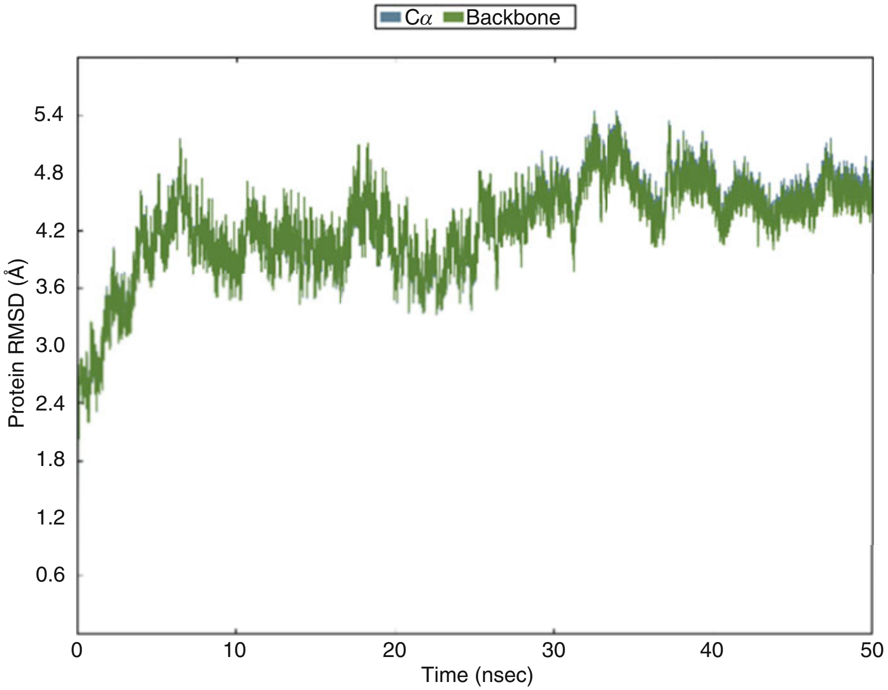


Fig. 6.5 50 ns molecular dynamics simulations run of *Pf*SPECT protein for refinement of structure showing RMSD of heavy atoms and back bone atoms (Srivastava et al. 2017)

6.3.2 Prediction of Binding Site

The binding site of PfSPECT was predicted using Sitemap, (version 3.6, Schrödinger). Five sites were predicted, the scores of which ranged from 0.637 to 1.171. A score of greater than 0.8 has been considered acceptable for selecting a binding site (Halgren 2009). Site1, with the best sitemap score 1.171, druggability score 1.240 and volume 378.6 (Å)^3 was selected for further interaction studies. The residues involved in the selected site were V11, S14, M15, V18, L19, T22, A24, S25, L26, V29, S30, H32, V33, I37, Y40, S41, I44, L48, L92, K93, L95, E96, N99, I102, K103, I106, I107, Y110, G111, N112, K113, N145, D148, K150, E151, L154, I158, N161, Y162, K164, F165 and L169. It was found out that the area covered by predicted binding site in *Pf*SPECT was similar to the reported binding pocket and cavity present in the already present structure of the *Pb*SPECT (Fig. 6.6). The position of some of the residues in the predicted binding site of *Pf*SPECT were common with those in *Pb*SPECT pocket and cavity, such as 99,102,164,169.

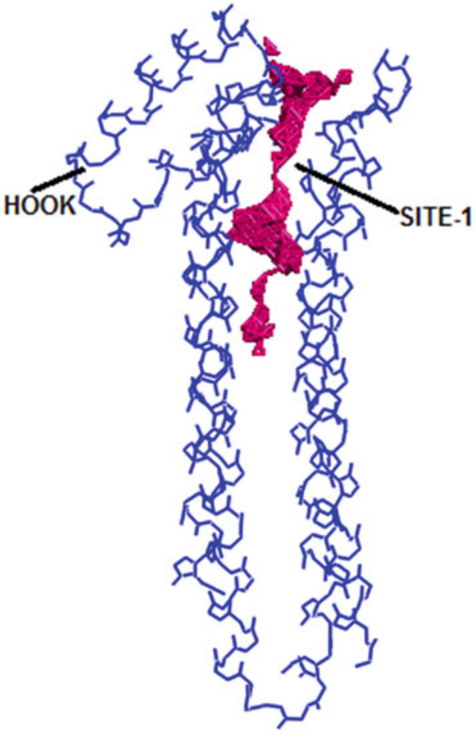


Fig. 6.6 Stereoview of the cavity of the predicted binding site in *Pf*SPECT depicted as sticks and lines (Srivastava et al. 2017)

6.3.3 Ligand Preparation

For virtual screening, compounds from ZINC database were used (Irwin et al. 2012). This library consisted of 276,784 molecules of biological origin. Zinc biogenic database (Zbc) constitutes primary and secondary metabolites and related entities. Therefore, there is an obvious advantage of increased possibility of obtaining a hit for a site, which could actually be the endogenous ligand/inhibitor for that site. The structures in the library were prepared for further analysis using LigPrep, version 3.5, Schrödinger. For each structure, a proper bond order was assigned and different tautomeric forms were produced. Stereoisomers were generated for each compound and were further used for carrying out protein interacting studies.

6.3.4 Molecular Docking

Protein ligand docking studies were carried out based on the MD simulated structure of *Pf*SPECT which was prepared using multi-step Schrödinger's protein preparation wizard before proceeding for docking calculations. Molecular docking calculation for all the compounds was performed using Glide (version 6.8, Schrödinger). Glide

Table 6.6 Glide energy, docking score and MMGBSA (ΔG_{bind}) score of selected ligands (Srivastava et al. 2017)

S. No.	ZINC ID	GLIDE energy	Docking score	ΔG_{bind}
1	**ZINC03851216**	**−36.078**	**−10.669**	**−90.092**
2	**ZINC0513454**	**−29.108**	**−10.544**	**−88.092**
3	ZINC77257422	**−25.172**	−9.285	−84.161
4	ZINC03851232	**−32.918**	−9.499	−83.553
5	ZINC03978926	**−32.84**	−9.669	−80.974
6	ZINC05124607	**−34.038**	−9.524	−80.195
7	ZINC00518290	**−31.536**	−9.968	−78.599
8	ZINC12496555	**−36.513**	−10.289	78.337
9	ZINC03847760	**−34.181**	−9.309	−78.007
10	ZINC15255743	**−22.767**	−9.373	−77.402

The bold values signify the two molecules with high binding affinity with the receptor

is a robust docking program that validates various poses of ligands in the binding pocket in a systematic manner. A box of size 10 Å × 10 Å × 10 Å was defined at the center of binding site and to occupy all the atoms of the docked poses one more enclosing box of 12 Å × 12 Å × 12 Å was also defined. The biogenic compounds of ZINC database were first subjected to filtering for drug-like properties using Qikprop, reactive, Lipinski's rule of five etc. and only the selected ones were used for high throughput virtual screening (HTVS). The screened compounds obtained as output of HTVS were further submitted for "Standard Precision" (SP) docking followed by "Extra Precision" (XP) algorithm of Glide Docking. On the basis of XP docking scores, 164 compounds were selected for further analysis. The binding affinity was calculated based on MMGBSA (Molecular Mechanics Generalised Born Surface Area) using Prime, (version 4.1, Schrödinger). First ten compounds from the above 164, with appreciable docking scores in the range of −10.669 to −9.285 and binding affinity, ΔG_{bind} from −90.092 to −77.402 kcal/mol. are shown in Table 6.6.

After analyzing the binding mode, two ligands, ZINC03851216 as ligand-1 and ZINC0513454 named as ligand-2 were selected for further calculations. Closer inspection of the mode of binding of *Pf*SPECT in its binding pocket showed hydrogen bonding (H-bond) patterns with both the poses of the docked ligands as shown in Fig. 6.7a–d. Only one H-bond was formed between (4-chlorophenyl)-(1-cyclohex-2-enyl)methanol (ligand-1) and Asn-99 at the binding site (Fig. 6.7a, b). For ligand 2, one side chain hydrogen bond with Asn-99 and a main chain H- bond with Met-15 stabilized interactions at the binding groove (Fig. 6.7c, d).

Per residue van der Waals (E_{vdw}) and electrostatic (E_{ele}) energy contribution was calculated within 12 A° of the docked ligands. The binding site amino acids showed significant contribution to the E_{vdw} and E_{ele} energy in both ligands 1 and 2. Specifically, appreciable E_{ele} energy contribution was made by Asn-99 in case of both ligands 1 and 2 and Met-15 for only ligand 2 (Fig. 6.8a, c). Phe 165, Tyr 162, Ile-158, Ile-102, Asn-99, Ile-44, Tyr-40, Leu-19, Met-15 contributed significantly to E_{vdw} in case of ligand 1(Fig. 6.8b) For ligand 2, two amino acids namely

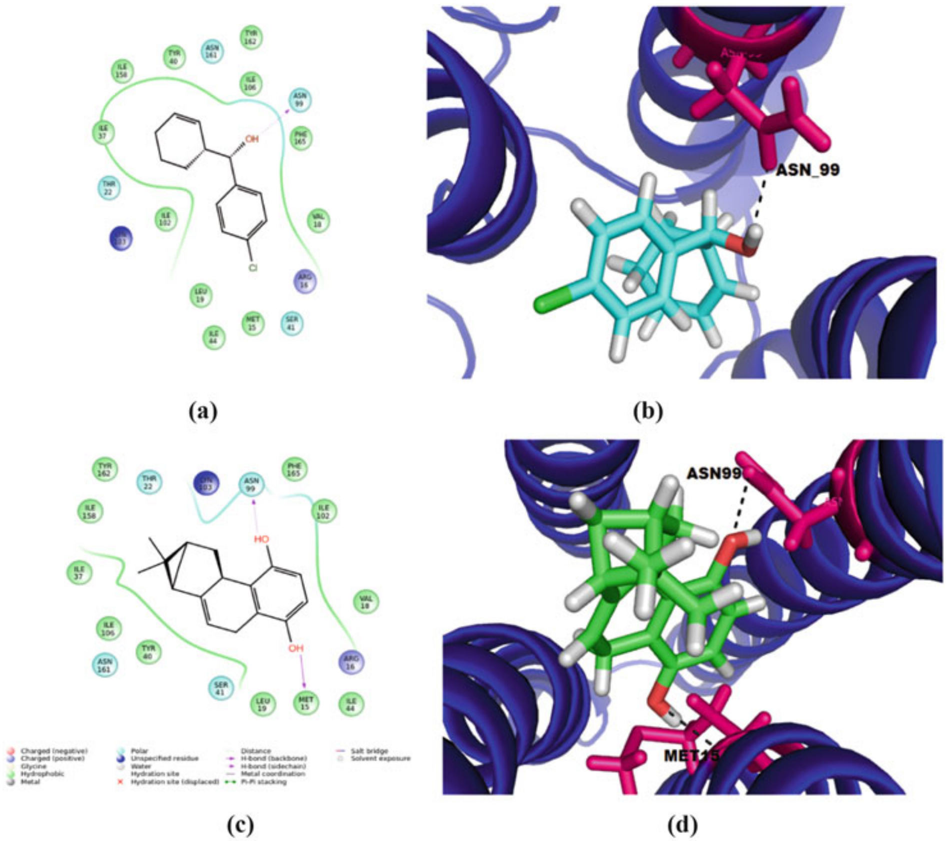


Fig. 6.7 (**a**) Ligand interaction of *Pf*SPECT showing hydrogen bond between ligand-1 and Asn-99. (**b**) 3-dimensional structure fitting of ligand-1 with the binding site of *Pf*SPECT protein showing hydrogen bond. (**c**) Ligand 2 showing main-chain (Met15) and side-chain H-bonds (Asn99) with *Pf*SPECT protein. (**d**) 3-dimensional view of H-bonding interaction with ligand-2 (Srivastava et al. 2017)

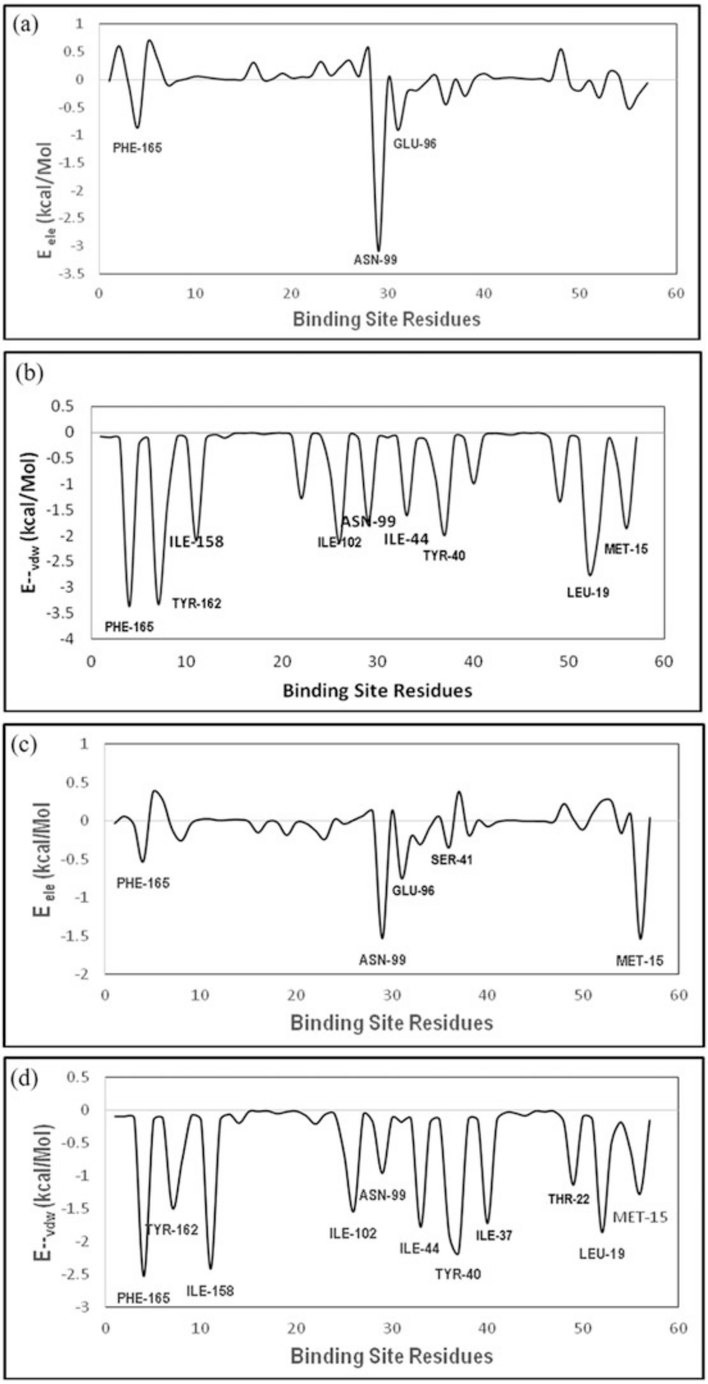


Fig. 6.8 Per residue energy interaction (**a**) Eele of ligand 1. (**b**) Evdw of ligand 1. (**c**) Eele of ligand 2. (**d**) Evdw of ligand 2 (Srivastava et al. 2017)

Ile-37 and Thr-22 showed additional contribution to the total van der waals energy (Fig. 6.8d).

6.3.5 Molecular Dynamics Simulation

In order to check the stability of the docked ligands and study the preferred binding mode and binding affinity of these two best ligands with *Pf*SPECT, the two docked complexes were subjected to molecular dynamics (MD) simulations using Desmond 2.2. The simulations were performed for a total duration of 50 nanoseconds. SPC (simple point charge) water model was used for the 10,067 water molecules positioned inside the ligand-receptor complex which was neutralized with sodium counter ions. The water molecules and heavy atom-hydrogen bonds were constrained using SHAKE algorithm (Ryckaert et al. 1977) while particle mesh Ewald method (Di Pierro et al. 2015) was used with Lennard Jones potential to treat electrostatic interactions. Periodic boundary conditions (PBC) were used (Cheatham et al. 1995).

Simulations were performed on the system using MD protocols of Maestro (Maestro, version 10.3). Full system minimization with restraints on solute was performed for a maximum of 2000 iterations using a hybrid of methods namely, steepest descent and the limited memory Broyden - Fletcher - Gold farb-Shanno (LBFGS), with a convergence threshold of 50.0 kcal/mol/Å^2. Similar minimization without any restraints was performed with a convergence threshold of 20.0 kcal/mol/Å^2. Simulations restraining non hydrogen solute atoms were performed in the NPT ensemble (constant number of atoms N, pressure P and temperature T) for 12 ps at a temperature of 10 K followed by a 24 ps simulation at 300 K and a final run of 24 ps (temperature 300 K) was performed without restraints to relax the system. The relaxed system was simulated for a period of 10,000 ps with a time step of 2 femtosecond (fs), using a Berendsen thermostat at 310 K and velocity resampling after every 1 ps. Trajectories were recorded after every 4.8 ps and a total of 10,000 frames were generated. Root mean square deviation (RMSD) and fluctuations in energy of the complex in each trajectory were analyzed during the simulation time. One of the important criteria to evaluate the conformational stability of the protein is the calculation of RMSD of Cα atoms from their initial coordinates as a function of time. RMSD measures average change in the displacement of a selection of atoms for a particular frame with respect to a reference frame and is calculated for all frames in the trajectory. Another important parameter namely, root mean square fluctuation or RMSF helps to determine stability of the protein chain during simulation. The backbone as well as side chains of each amino acid residue of *Pf*SPECT were monitored for fluctuations. The protein-ligand interactions were also assessed for evaluating the stability of the complex.

For the complex of ligand 1 and protein, the RMSD plot revealed that the complex was relatively stable throughout the simulation time as it found stability after 10 ns, shown in Fig. 6.9a. RMSD plot for the second complex with ligand

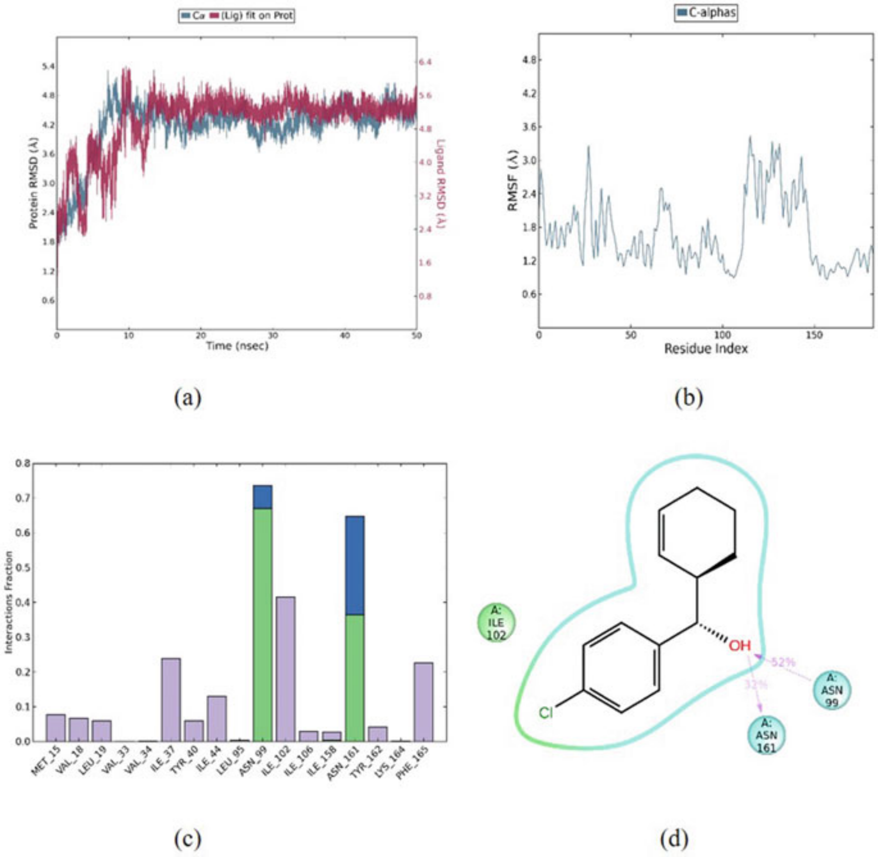
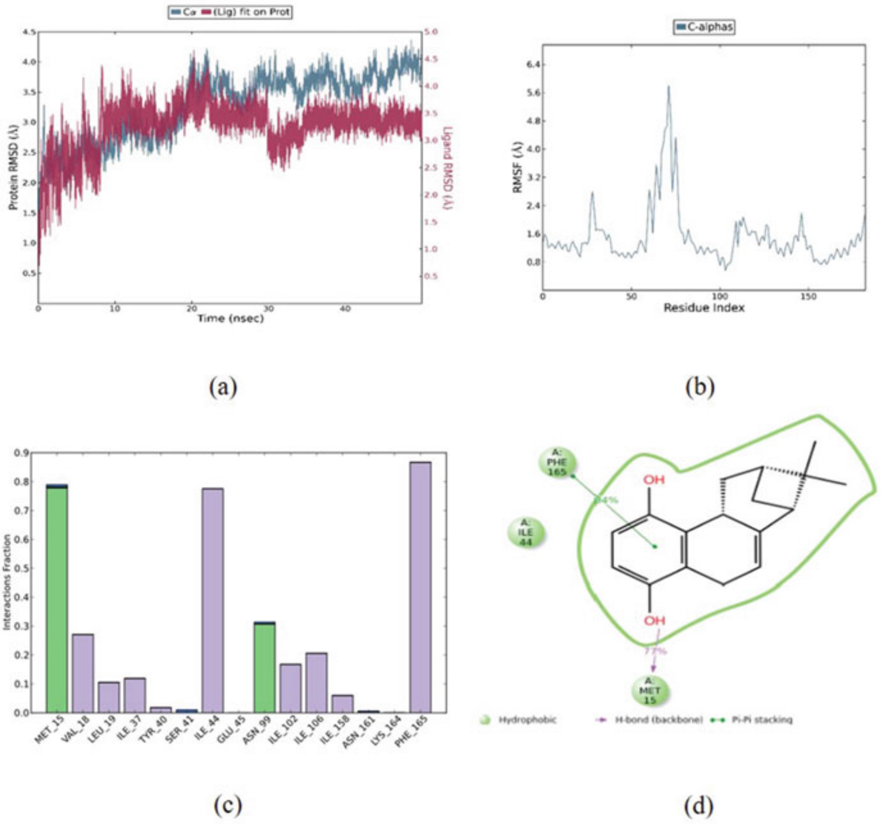


Fig. 6.9 Interactions in *Pf*SPECT-ligand 1 complex during 50 ns MD simulations run. (**a**) RMSD of Cα atoms of *Pf*SPECT alone and in complex form (**b**) *Pf*SPECT RMSF. (**c**) Protein-ligand interactions. (**d**) Protein- ligand complex showing 2 H bonds (Srivastava et al. 2017)

2 revealed that the complex stabilizes somewhat after 35 ns (Fig. 6.10a). This reveals that the first ligand forms a more stable complex and well within the binding pocket of *Pf*SPECT than the second ligand. Lower RMSF value was shown by the residues of the active site and alpha helical regions which indicates the stability of these regions (Figs. 6.9b and 6.10b). The trajectory data was analysed to obtain the ligand-protein binding interactions. The timeline representation shows that Asn-99 and Asn-161 were involved in protein-ligand contact over a large duration of the simulation time in case of PfSPECT-ligand 1 complex, while in case of PfSPECT-ligand 2 complex, three residues of the protein, namely, Met-15, Ile-44 and Phe-165 showed contact with ligand 2 (data not shown here). The stacked bar charts in Figs. 6.9c and 6.10c, represent the type of interactions between the protein and ligands and are normalized over the course of the trajectory. Figure 6.9c shows two hydrogen bonds (amino acids involved being Asn-99, Asn-165) and two



(a)

(b)

(c)

(d)

Fig. 6.10 Interactions of *Pf*SPECT and ligand 2 complex during MD simulations run. (**a**) RMSD of heavy atoms and back bone atoms. (**b**) Protein RMSF. (**c**) Protein-ligand interactions. (**d**) Protein-ligand complex showing various interactions (Srivastava et al. 2017)

hydrophobic bonds (amino acids involved being Ile-102 and Phe-165) between ligand 1 and *Pf*SPECT. Asn-99 forms a side chain H-bond for 52% of the simulation time, while Asn-161 forms a main chain H bond for 32% of the time (Fig. 6.9d). Ligand 2 forms a main-chain hydrogen bond with Met-15 and hydrophobic bonds with Ile-44 and Phe-165 (Fig. 6.10c). While ligand-2 forms a hydrogen bond with Met-15 for 77% of the time, it shows pi-pi stacking with Phe-165 for 84% of the total simulation time (Fig. 6.10d).

In summary, homology modeling and MD simulations have been used to derive the 3D structure of *Pf*SPECT. A C^{α} rmsd of 1.062 Å with the template protein *Pb*SPECT, Ramachandran plot showing 91.4% of backbone angles of *Pf*SPECT in the allowed regions and an ERRAT score of 85.9% further validate the quality of structure. Extensive docking studies are carried out and MMGBSA has been used to calculate binding affinity of ligands. ZINC03851216 (ligand 1) and ZINC0513454 (ligand 2) have been identified to demonstrate high affinity towards the binding site of *Pf*SPECT. Molecular dynamics simulations of these ligand-protein complexes reveal hydrogen bonds and pi-pi interactions which are conserved during major duration of the simulation time. This *in silico* study paves the way for medicinal chemists to design better drug-like compounds as inhibitors of *Pf*SPECT that eventually interfere with the process of host cell traversal in *P. falciparum* and contribute to malaria transmission-blocking strategies.

6.4 Conclusion

In the above two case studies we have successfully established subtractive genomics and subtractive metabolomics as an effective method for identification of unique essential targets in bacterial pathogens. Digging up genomic and proteomic data along with metabolic pathways, while taking care of the symbiotic organisms, is an innovative methodology for target identification which addresses the concern of unwillingly harming the gut microbes. Virtual screening of databases for selection of potential leads helps in exploring all the possible compounds in a very fast and cost-effective manner. This also brings to surface the potential of many compounds for drug repurposing. The case studies present a combination of tools and techniques like homology modeling, molecular docking and molecular dynamics simulations to gain an understanding of interactions between the target and ligand at the atomic level. These methods help to understand the interacting systems better which lead to more efficient designing of novel compounds, analogues and repurposed drugs. Other than the methods discussed above, a plethora of *in silico* tools and techniques are being constantly developed to contribute to the process of drug discovery, not only while designing leads but also in their experimental validation as well as clinical trials.

Certainly, with the help of CADD there has been a significant reduction in the drug attrition rates which has led to widespread adoption of *in silico* techniques of drug discovery in the pharmaceutical industry. It has increased the rate of drug target identification. In selection and screening of the compounds, efforts are focused on the early elimination of compounds that may cause several side effects or interaction with other molecules. Atomic level interaction details also pave a way for *de novo* drug development, target specific efficient drug development, drug repurposing and efficient analogue development. Due to increased computational power, a diverse range of *in silico* toxicity screening products have been developed. Owing to the efficacy and reliability of these models and to avoid later stage drug failures, pharmaceutical companies have increased the use of *in silico* ADME/Tox screening products. CADD plays and shall continue to play a major role in pharmaceutical drug discovery and development. It has also been envisioned that if implemented meticulously, it has the potential to surpass all the archaic techniques of drug development, the use of which will anyways become obsolete due to the increasing costs and time involved. Despite all the promises, there are still some constraints in the use of *in silico* methods for drug design and discovery. One of them is the risk of failure of a potentially safe and efficient drug candidate identified by utilizing *in silico* models and subsequently not performing the relevant *in vitro/in vivo* analysis; and the other being the lack of accurate and reliable experimental data for development of improved *in silico* models. The success of drug discovery and development depends on an efficient integration of computational technologies with the experience and accuracy of experimental methods.

References

Amino, R., Giovannini, D., Thiberge, S., Gueirard, P., & Boisson, B. (2008a). Host cell traversal is important for progression of the malaria parasite through the dermis to the liver. *Cell Host and Microbe, 3*(2), 88–96.

Amino, R., Giovannini, D., Thiberge, S., Gueirard, P., Boisson, B., Dubremetz, J. F., et al. (2008b). Host cell traversal is important for progression of the malaria parasite through the dermis to the liver. *Cell Host and Microbes, 3*, 88–96.

Amino, R., Thiberge, S., Martin, B., Celli, S., Shorte, S., Frischknecht, F., et al. (2006). Quantitative imaging of Plasmodium transmission from mosquito to mammal. *Nature Medicine, 12*, 220–224.

Bowers, K. J., Chow, E., Xu, H., Dror, R. O., Eastwood, M. P., Gregersen, B. A., et al. (2006). Scalable algorithms for molecular dynamics simulations on commodity clusters. In *Proceedings of the ACM/IEEE Conference on Supercomputing (SC06)*, Tampa, Florida.

Bruccoleri, R. E., Dougherty, T. J., & Davison, D. B. (1998). Concordance analysis of microbial genomes. *Nucleic Acids Research, 26*, 4482–4486.

Cheatham, T. E., Miller, J. H., Fox, T., Darden, P. A., & Kollman, P. A. (1995). Molecular dynamics simulations on solvated biomolecular systems: The particle Mesh ewald method leads to stable trajectories of DNA, RNA, and proteins. *Journal of the American Chemical Society, 117*(14), 4193–4194.

Colovos, C., & Yeates, T. O. (1993). Verification of protein structures: Patterns of non-bonded atomic interactions. *Protein Science, 2*(9), 1511–1519.

Desmond molecular dynamics system, version 2.2. New York: D. E. Shaw Research; 2009. *Maestro-desmond interoperability tools*, version 2.2, New York, NY: Schrödinger; 2009.

Derbyshire, E. R., Mota, M. M., & Clardy, J. (2011). The next opportunity in anti-malaria drug discovery: The liver stage. *PLoS Pathogens, 7*(9), e1002178.

Di Pierro, M., Elber, R., & Leimkuhler, B. (2015). A stochastic algorithm for the isobaric-isothermal ensemble with Ewald summations for all long range forces. *Journal of Chemical Theory and Computation, 11*(12), 5624–5637.

Dutta, A., Singh, S. K., Ghosh, P., Mukherjee, R., Mitter, S., & Bandyopadhyay, D. (2006). In silico identification of potential therapeutic targets in the human pathogen Helicobacter pylori. *In Silico Biology, 6*, 005.

Eisenberg, D., Luthy, R., & Bowie, J. U. (1997). VERIFY3D: Assessment of protein models with three-dimensional profiles. *Methods in Enzymology, 277*, 396–404.

Formaglio, P., Tavares, J., Menard, R., & Amino, R. (2014). Loss of host cell plasma membrane integrity following cell traversal by Plasmodium sporozoites in the skin. *Parasitology International, 63*(1), 237–244.

Fran⍰a, T. C. C. (2015). Homology modeling: An important tool for the drug discovery. *Journal of Biomolecular Structure and Dynamics, 33*(8), 1780e93.

Freiberg, C. (2001). Novel computational methods in anti–microbial target identification. *Drug Discovery Today, 6*, S72–S80.

Frevert, U., Engelmann, S., Zougbede, S., Stange, J., & Ng, B. (2005). Intravital observation of Plasmodium berghei sporozoite infection of the liver. *PLoS Biology, 3*(6), e192.

Garg, S., Agarwal, S., Kumar, S., Yazdani, S. S., Chitnis, C. E., & Singh, S. (2013). Calcium-dependent permeabilization of erythrocytes by a perforin-like protein during egress of malaria parasites. *Nature Communications, 4*, 1736.

Guex, N., & Peitsch, M. C. (1997). SWISS-MODEL and Swiss-Pdb Viewer: An environment for comparative protein modelling. *Electrophoresis, 18*(15), 2714–2723.

Halgren, T. A. (2009). Identifying and characterizing binding sites and assessing druggability. *Journal of Chemical Information and Modeling, 49*(2), 377–389.

Hamaoka, B. Y., & Ghosh, P. (2014). Structure of the essential plasmodium host cell traversal protein SPECT1. *PLoS One, 9*(12), e114685.

Hopkins, A. L., & Groom, C. R. (2002). The druggable genome. *Nature Reviews. Drug Discovery, 1*, 727–730.

Humer, F. (2005). *Innovation in the pharmaceutical industry—Future prospects*. Available: http://www.roche.com/fbh_zvg05_e.pdf

Huynen, M., Dandekar, T., & Bork, P. (1998). Differential genome analysis applied to the species–specific features of Helicobacter pylori. *FEBS Letters, 1998*(426), 1–5.

Huynen, M., Diaz-Lazcoz, Y., & Bork, P. (1997). Differential genome display. *Trends in Genetics, 13*, 389–390.

Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S., & Coleman, R. G. (2012). ZINC:A free tool to discover chemistry for biology. *Journal of Chemical Information and Modeling, 52*(7), 1757–1768.

Ishino, T., Chinzei, Y., & Yuda, M. (2005). A Plasmodium sporozoite protein with a membrane attack complex domain is required for breaching the liver sinusoidal cell layer prior to hepatocyte infection. *Cellular Microbiology, 7*(2), 199–208.

Ishino, T., Yano, K., Chinzei, Y., & Yuda, M. (2004). Cell-passage activity is required for the malarial parasite to cross the liver sinusoidal cell layer. *PLoS Biology, 2*(1), 77–84.

Jacobs, M. A., Alwood, A., Thaipisuttikul, I., Spencer, D., Haugen, E., Ernst, S., et al. (2003). Comprehensive transposon mutant library of Pseudomonas aeruginosa. *Proceedings of the National Academy of Sciences of the United States of America, 100*, 14339–14344.

Kaiser, K., Matuschewski, K., Camargo, N., Ross, J., & Kappe, S. H. I. (2004). Differential transcriptome profiling identifies Plasmodium genes encoding pre-erythrocytic stage-specific proteins. *Molecular Microbiology, 51*(5), 1365–2958.

Kanehisa, M., Goto, S., Kawashima, S., & Nakaya, A. (2002). The KEGG databases at genome net. *Nucleic Acids Research, 30*(1), 42–46.

Laskowski, M. W., & MacArthur, D. S. (1993). Moss, and Thornton. J.M. PROCHECK: A program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography, 26*, 283–291.

Mali, S., Steele, S., Slutsker, L., & Arguin, P. M. (2008). Malaria surveillance–United States, 2006. *MMWR Surveillance Summaries, 57*(5), 24–39.

Mazier, D., Rénia, L., & Snounou, G. (2009). A pre-emptive strike against malaria's stealthy hepatic forms. *Nature Reviews Drug Discovery, 8*, 854–864.

Mdluli, K. E., Witte, P. R., Kline, T., Barb, A. W., Erwin, A. L., Mansfield, B. E., et al. (2006). Molecular validation of LpxC as an antibacterial drug target in Pseudomonas aeruginosa. *Antimicrobial Agents and Chemotherapy, 2006*(50), 2178–2184.

Mota, M. M., Pradel, G., Vanderberg, J. P., Hafalla, J. C., & Frevert, U. (2001). Migration of Plasmodium sporozoites through cells before infection. *Science, 291*(5501), 141–144.

Naik, P. K., Santoshi, S., & Birmani, A. (2010). Computational prediction of potent therapeutic targets of Pseudomonas aeruginosa and in silico virtual screening for novel inhibitors. *Internet Electronic Journal of Molecular Design, 8*, 42–62.

Ramachandran, G. N., Ramakrishnan, C., & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology, 7*, 95–99.

Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2011). HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods, 9*, 173–175.

Risco-Castillo, V., Topçu, S., Marinach, C., Manzoni, G., Bigorgne, A. E., Briquet, S., et al. (2015). Malaria sporozoites traverse host cells within transient vacuoles. *Cell Host & Microbe, 18*, 593–603.

Robertson, J. G. (2005). Mechanistic basis of enzyme–targeted drugs. *Biochemistry, 44*, 8918.

Roemer, T., Jiang, B., Davison, J., Ketela, T., & Veillette, K. (2003). Large-scale essential gene identification in Candida albicans and applications to antifungal drug discovery. *Molecular Microbiology, 50*, 167–181.

Ryckaert, J.-P., Ciccotti, G., & Berendsen, H. J. C. (1977). Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *Journal of Computational Physics, 23*(3), 327–341.

Sanseau, P. (2001). Impact of human genome sequencing for in silico target discovery. *Drug Discovery Today, 6*, 316–323.

Sassetti, C. M., Boyd, D. H., & Rubin, E. J. (2003). Genes required for mycobacterial growth defined by high density mutagenesis. *Molecular Microbiology, 48*, 77–84.

Sibley, L. D. (2004). Intracellular Parasite Invasion Strategies. *Science, 304*(5668), 248–253.

Singh, N. K., Selvam, S. M., & Chakravarthy, P. (2006). T-iDT: Tool for identification of drug target in bacteria and validation by mycobacterium tuberculosis. *In Silico Biology, 6*(6), 485–493.

Srivastava, S., Santoshi, S., Malik, B. K., & Mathur, P. (2017). Molecular modeling and molecular dynamics studies of SPECT protein of Plasmodium falciparum and in silico screening of lead compounds. *International Journal of Pharmaceutical Sciences and Research, 8*(12), 5077–5087.

Sterling, T., & Irwin, J. J. (2015). ZINC-15-Ligand discovery for everyone. *Journal of Chemical Information and Modeling, 55*(11), 2324–2337.

Stover, C. K., Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warrener, P., Hickey, M. J., et al. (2000). Complete genome sequence of Pseudomonas aeruginosa PA01, an opportunistic pathogen. *Nature, 406*, 959–964.

Sturm, A., Amino, R., Van de Sand, C., Regen, T., Retzlaff, S., Rennenberg, A., et al. (2006). Manipulation of host hepatocytes by the malaria parasite for delivery into liver sinusoids. *Science, 313*(5791), 1287–1290.

Tatusov, R. L., Koonin, E. V., & Lipman, D. J. (1997). A genomic perspective on protein families. *Science, 278*, 631–637.

Tavares, J., Formaglio, P., Thiberge, S., Mordelet, E., & Van, N. (2013). Rooijen, Role of host cell traversal by the malaria sporozoite during liver infection. *The Journal of Experimental Medicine, 210*(5), 905–915.

Terstappen, G. C., & Reggiani, A. (2001). In silico research in drug discovery. *Trends in Pharmacological Sciences, 22*, 23–26.

Vanderberg, J. P., Serena, C., & Michael, J. S. (1990). Plasmodium sporozoite interactions with macrophages in vitro: A videomicroscopic analysis. *Journal of Eukaryotic Microbiology, 37*(6), 1550–7408.

Victoria, M., Granich, R., Gilks, F. C., Gunneberg, C., Hosseini, M., Were, W., et al. (2009). The global fight against HIV/AIDS, tuberculosis, and malaria: Current status and future perspectives. *American Journal of Clinical Pathology, 131*(6), 844–848.

Wang, S., Sim, T. B., Kim, Y. S., & Chang, Y. T. (2004). Tools for target identification and validation. *Current Opinion in Chemical Biology, 8*, 371–377.

Wei, W., Ning, L. W., Ye, Y. N., & Guo, F. B. (2013). Geptop: A gene essentiality prediction tool for sequenced bacterial genomes based on orthology and phylogeny. *PLoS One, 8*(8), e72343.

Wen, Q. F., Liu, S., Dong, C., Guo, H. X., Gao, Y. Z., & Guo, F. B. (2019). Geptop 2.0: An updated, more precise, and faster geptop server for identification of prokaryotic essential genes. *Frontiers in Microbiology, 10*, 1236.

Wixon, J., & Kell, D. (2000). The Kyoto encyclopedia of genes and genomes—KEGG. *Yeast, 17*, 48–55.

Yamauchi, L. M., Coppi, A., Snounou, G., & Sinnis, P. (2007). Plasmodium sporozoites trickle out of the injection site. *Cellular Microbiology, 9*, 1215–1222.

Yang, S. P., & Boddey, J. A. (2017). Molecular mechanisms of host cell traversal by malaria sporozoites. *Molecular BioSystems, 47*(2–3), 129–136.

Zhang, R., Ou, H. Y., & Zhang, C. T. (2004). DEG: A database of essential genes. *Nucleic Acids Research, 32*, D271–D272.

Chapter 7
Predicting Protein Folding and Protein Stability by Molecular Dynamics Simulations for Computational Drug Discovery

Ishwar Chandra, Chirasmita Nayak, and Sanjeev Kumar Singh

Abstract Biological function and properties depends on proteins three dimensional structure resolved through protein folding which is encoded in its polypeptide sequence. Protein acquires its native three dimensional structure by undergoing enormous conformational changes during protein folding. Predicting protein structure and its folding through computational methods give insights of the protein activity within the cell. Molecular dynamics simulations (MDS) utilize atomic interaction knowledge and are immensely effective in understanding of biomolecular structure-function relationship. MDS provides appropriate data about the dynamic properties of macromolecules to analyze the conformational ensemble over single structure of protein and its folding pathway. Advanced computational methods and techniques give details regarding pathway, intermediates and folding free energy landscapes that are both pragmatic and worthy. This chapter illustrates about computational protein structure prediction and different MDS methods to unveil protein folding. Significance of MDS studies on assessment of protein stability and computational drug discovery has been also summarized.

Keywords Protein prediction · Protein folding · Free energy landscape · MDS for protein folding · Protein stability · Drug discovery

Abbreviations

aMD	Accelerated molecular dynamics
CASP	Critical assessment of structure prediction
CG	Coarse grained
CVs	Collective variables
FEP	Free energy perturbation

I. Chandra · C. Nayak · S. K. Singh (✉)
Computer Aided Drug Design and Molecular Modeling Lab, Department of Bioinformatics, Alagappa University, Karaikudi, Tamil Nadu, India

S. K. Singh (ed.), *Innovations and Implementations of Computer Aided Drug Discovery Strategies in Rational Drug Design*,
https://doi.org/10.1007/978-981-15-8936-2_7

H-REMD	Hamiltonian REMD
MDS	Molecular dynamics simulation
MM-GBSA	Molecular mechanics, the generalized Born model and solvent accessibility
MM-PBSA	Molecular mechanics, Poisson–Boltzmann surface area
M-REMD	Multiplexed REMD
MSM	Markov state model
PMF	Potential of mean force
PSA	Path sampling approaches
RC	Reaction coordinate
REMD	Replica exchange molecular dynamics
R-REMD	Reservoir REMD
TPS	Transition path sampling
T-REMD	Temperature dependent REMD

7.1 Introduction

The major component of all the living systems ranging from simple organisms like bacteria, viruses, eukaryotes to complex organisms like plants and animals are proteins. It accounts for nearly fifty percent of the tissue dry weight in vertebrates. It is composed of carbon, oxygen, nitrogen, hydrogen and mostly sulfur. Proteins are also known as macromolecules, and are the polymers of structural units called *amino acids* that form the building blocks of proteins. According to the central dogma of biology the genetic information flows through DNA to RNA to protein (i.e. polypeptide chain) through the process of transcription followed by translation (Schneider-Poetsch and Yoshida 2018). Proteins are polymeric chains of amino acids; acquire its native conformation based on their amino acid sequences, folding pathway in the cellular environment (Jaenicke 1991). The process of attaining active three dimensional structure of protein from its amino acid sequence through protein folding is considered to be the secondary part of genetic code and has been the toughest challenges of molecular biology (Blanco and Blanco 2017; Szilágyi et al. 2007). Organisms and cells depend on proteins for their various biological processes like catalysis, signaling, mobility, ordering, pathogen clearing, shape recognition and stability. As proteins are polymeric chains of amino acids, the precise ordering of the amino acids determines the process that folds the polypeptide chain/protein to attain a specific, stable, and functional three dimensional structure which is termed protein folding (Khoury et al. 2014). Proteins can be divided into two distinctive molecular shape, *globular* and *fibrillary*. Globular proteins comprise the functionally active proteins where polypeptide chains fold into a smaller adjusted shape like an ovoid or spheroid with an irregular surface and are highly scattered in watery media. Enzymes, hormones, insulin, hemoglobin, immunoglobulin and antibodies are common examples of globular protein. In fibrillar proteins the polypeptide chains are

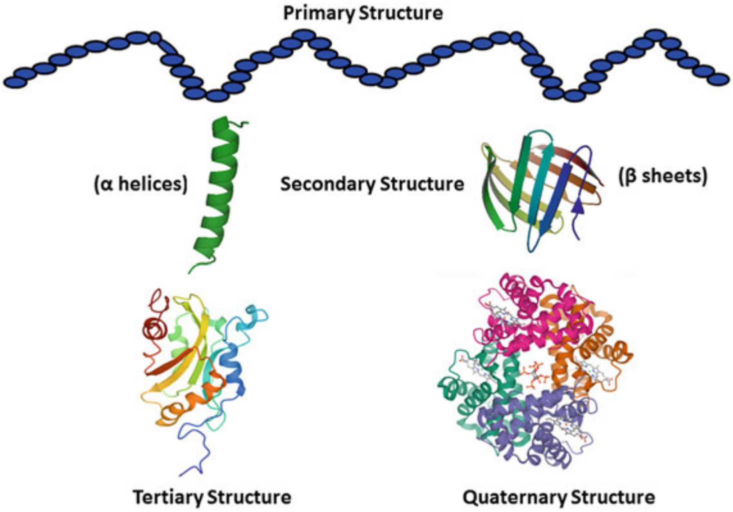


Fig. 7.1 Graphical representation of primary, secondary, tertiary and quaternary structure of proteins

stretched in long and narrow strands that are indispersible in water and establish structures of extraordinary physical strength, for example, the connective tissue strands, collagen, myosin, fibrin, keratin, actin, ellastin. Protein structure is organized into four levels by the biologists depending on its complexity that is, *primary structure*, *secondary structure*, *tertiary structure* and *quaternary structure*. The polypeptide chain in a linear form joined by peptide bonds forms the primary structure; whereas the arrangement of polypeptide chain in a repetitive and regular spatial form joined by hydrogen bonds constitutes secondary structure and is divided into α helices and β sheets. Once the polypeptide chains are held together by hydrogen bonds, ionic bonds, disulfide bridges and folds in globular shape brings forth the tertiary structure of protein, while more than one polypeptide chains are linked with each other, it is termed as quaternary structure of the protein (Blanco and Blanco 2017; Alberts et al. 2002). A schematic representation of the four levels of protein has been given in Fig. 7.1. Proteins become functional when it folds into its specific tertiary structure, though misfolding of proteins can cause various diseases like Alzheimer's, Parkinson's, Huntington's and different degenerative and neurodegenerative diseases (Chaudhuri and Paul 2006). Subsequently knowledge about the emergence of functional structure of protein from its primary amino acid arrangement is vital, as this information can help in understanding the enzymes function and the working ability of the immune system. The potential application of this science could be utilized in engineering novel enzymes, antibodies, hormones and biosensor for renewed biological usage. The effort to grasp the process that propels a protein into its distinctive, biologically functional and structural unit, from its amino acid sequence, is called the *protein-folding problem* (Hansmann 2003).

7.1.1 Protein Folding and Its Mechanism

The clear cut mechanism of protein folding is still unknown, and is an active field of multidisciplinary research which utilizes experimental and theoretical approaches to understand this complex process. Different theories on protein folding approaches have been reviewed by Yon (2001). Distinct folding models and pathways are recognized that characterizes intermediates in protein folding. Initially nucleation propagation model was proposed to know the folding of ribonuclease and was later superseded by nucleation condensation model where the nucleus is steadied through long range interactions. Hierarchical protein folding model has been also supported by many authors for hierarchical protein structure; where nucleation, formation of secondary structure, supersecondary structure, domains and finally active monomer formation occurs in a sequential stepwise way. Likewise framework model presumes the formation of secondary structures and role of short range interactions which directs the folding process. Another protein folding methodology is diffusion collision model postulated by Karplus and Weaver in which simultaneous nucleation occurs at different regions of polypeptide chain producing microstructures that diffuse combines and coalesce to generate substructures with native conformation. Hydrophobic impact is the key part of protein folding and stabilization in hydrophobic collapse model. In this model of protein folding, the secondary structure formation is preceded by collapse of long range hydrophobic interactions that gives rise to the stretches of secondary structure simultaneously with hydrophobic collapse. Protein folding has been normally represented as two state process comprising unfolded and native species that are significantly occupied and swings between unfolding and refolding state during equilibrium. The intermediates are normally unstable and scantly populated within equilibrium conditions. The intermediates are identified as "molten globule" with considerable secondary structures. Protein folding through lattice models and molecular dynamics based on energy landscape and the folding funnel has been proposed in theoretical studies. This model is based on the energy landscape terms that accommodate the thermodynamic and kinetic changes of the transformation of an ensemble of unfolded molecules to a prevalent native state (Yon 2001). Details related to protein unfolding and refolding through molecular simulation has been reviewed by Shea and Brooks (2001).

7.2 A Brief Outline of Protein Structure Prediction

Anfinsen through his classical experimentation in the 1970s showed that the necessary detail for protein folding is encrypted within its amino acid sequence and its arrangement can predict protein's three dimensional structure theoretically (Anfinsen 1973). As per the Levinthal's paradox protein folding takes enormously large number of chain conformation in the estimations of $(10^2)^L$, where L denotes the number of amino acid residues in the chain (typically, $L \sim 100$) and ten represents the microstates around the degrees of freedom of the torsion angles φ and ψ of the polypeptide chain. Thus the protein would achieve its properly folded native configuration through every possible chain conformations, in a time scale greater than the age of this Universe. But the time scale on which the protein acquires its natural fold is of the order of milliseconds, hence the protein needs to use some kind of folding pathway (Finkelstein and Garbuzynskiy 2016). Various multidisciplinary biological approaches like chemical biology, biophysics, evolutionary biology and mathematical assumptions have been used to decode the three dimensional structure of protein from the amino acid sequence which is often called as the second part of genetic code (Kolata 1986). The start of the Human Genome project outnumbered the sequenced genome compared to the experimentally determined protein structures which in turn brought the structural biologist and computational scientists together to determine the three dimensional protein structures that aids in understanding the mechanism of biological function of protein (National Research Council (US) Committee on Mapping and Sequencing the Human Genome 1988a, b). Also experimentally determining the protein's structure is exceptionally relentless and unpredictable considering the small changes in its sequence could profoundly affect its biophysical properties. Consequently incorporation of experimental and computational strategies is used to resolve the three dimensional protein structures (Schwede 2013). The protein folding from the amino acid sequence should fulfill two important criteria, first thermodynamic and the second kinetic. The thermodynamic criterion epitomizes that the sequence must have an exceptionally folded conformation and is stable under physiological conditions. The kinetic prerequisite is that the denatured polypeptide chain can fold in this conformation with feasible speed (Karplus and Šali 1995). The computational approaches can solve the Levinthal's paradox and can find the native folded protein among different conformations of polypeptide chain in biologically achievable time (Zwanzig et al. 1992). The underlying concept of computational protein folding is to obtain the least free-energy stable structure for an amino acid sequence, in accordance with the thermodynamic theory postulated by Anfinsen through probing the vast conformational space of the protein as defined by Levinthal (Gillet and Ghosh 2013).

7.2.1 Secondary Structure Prediction

Although the active structure of the protein can be determined by its amino acid sequence, advance studies expose the current challenges in the accurate prediction of tertiary structure form its sequence. In this scenario, prediction of protein secondary structure from sequence becomes an intermediate bridge covering the gap between the primary and tertiary structure (Zhang et al. 2018). Protein secondary structure traditionally categorized into two regular structure i.e. α-helix (H) and β-strand (E), and one irregular secondary structure type, the coil region (C). The DSSP database from there three state (Q3) structure proposed a detailed classification of the secondary structures by prolonging the three states into eight states (Q8): 310 helix (G), α-helix (H), π-helix (I), β-stand (E), bridge (B), turn (T), bend (S), and others (C) (Zhang et al. 2018; Kabsch and Sander 1983). In the last several decades, various advance tools, servers and software are developed to predict the protein secondary structure form its amino acid sequence such as Porter 5, GOR, PSIPRED, SOPMA, PredictProtein, J pred and many more. All the tools, servers and software use different algorithms, statistics and methods to predict protein secondary structure. Porter 5 is a fastest standalone state-of-the-art ab initio protein secondary structure prediction tool in 3 and 8 classes of proteins which shows 84% accuracy in prediction of Q3 and 73% accuracy on Q8 (Torrisi et al. 2018). GOR uses information theory and Bayesian statistics to predict the secondary structure and shows an accuracy of prediction Q3 of 73.5% (Sen et al. 2005). PSIPRED is a most widely used protein secondary structure prediction tool which uses two feed-forward neural networks obtained from PSI-BLAST and provides highly accurate secondary structure (McGuffin et al. 2000; Buchan and Jones 2019). Self OPtimised Prediction Method from multiple Alignments (SOPMA) predicts the protein secondary structure based on the nearest neighbor method and provides 69.5% accuracy for a three-state (Q3) description of the secondary structure (α-helix, β-sheet and coil) in a whole database (Geourjon and Deleage 1995). PredictProtein utilizes Profile-based neural network to predict the secondary structure and also provides other details of the protein such as solvent accessibility, globular regions, transmembrane helices, coiled-coil regions, structural switch regions, disulfide-bonds, sub-cellular localization and functional annotations (Rost et al. 2004). Jpred predicts secondary structure and solvent accessibility using Multiple Neural network assignment from PSI-BLAST and HMMER profiles. It provides a three-state (α-helix, β-strand and coil) prediction of secondary structure at an accuracy of 81.5% (Cole et al. 2008). Wang et al. (2016) present a machine learning method DeepCNF tool (Deep Convolutional Neural Fields) for protein secondary structure prediction which is integration of both conditional neural fields (CNF) 50 and deep convolutional neural networks (DCNN). DeepCNF obtain ~84% accuracy Q3, ~85% segment of overlap (SOV) and ~72% Q8 accuracy experimentally (Wang et al. 2016).

7.2.2 Tertiary Structure Prediction

Structure prediction techniques are mostly divided into homology modeling commonly called as template based modeling and ab initio or free modeling (Hatherley et al. 2016). Predicting the ultimate structure of protein from its sequence by exploiting the knowledge of the already resolved structure through template based methods have become popular methods for protein structure prediction. Although the ab initio methods of protein folding is also a way for exploration of secondary genetic code, intended to build the protein structure from the first principle of physics without relying on the already resolved structure, yet only ab initio methods are extremely scarce and challenging (Deng et al. 2018). Homology modeling is based on the principle of protein structure conservation which implies that the structure of a protein is much more conserved compared to its primary amino acid sequence (Vlachakis et al. 2017). Homology modeling or template based method predicts protein structure based upon template and target sequences where the target represents the sequences to be modeled and template pertains to the available known structures related to the target. The target sequence is aligned to the template structure; thereafter the structure is built by copying the aligned regions that favors the spatial constraint of the template and eventually build unaligned loop regions and addition of side chain atoms (Zhang 2008). Template based method also uses protein threading for fold recognition and backbone structure prediction through sequence similarity information and structural fitness information of the query sequence and the template structure (Khor et al. 2015). MODELLER is commonly used software for protein structure prediction for homology modeling while SWISS-MODEL, ModWeb, Phyre2, HHpred, Rosetta and I-TASSER are automated servers with their own computational sets and algorithms for protein structure prediction and comparative modeling. Some important commercial software packages for protein modeling includes DSModeler, Prime, LOOK, ICM, Sybyl and MOE (Nayeem et al. 2006). Phyre2 incorporates comparative and ab initio modeling whereas I-TASSER combines threading, fragment assembly and ab initio approaches for its template based modeling protocol (Hatherley et al. 2016). From the given example it is evident that integration of template based methods with ab initio methods makes protein structure prediction more efficient and relevant (Li et al. 2015a). A graphical framework for protein structure prediction from primary sequences is shown in Fig. 7.2. Since 1994 John Moult and others have biennially organized a competition for the prediction of protein structure in computational structural biology, called the Critical Assessment of Structure Prediction (CASP), which evaluates the updated advance technology in the modeling of protein structure (Moult et al. 1995). Members are given amino acid sequences of target proteins and submit the resolved three dimensional structures which are compared with the experimental structures by independent evaluators. Other than structure models, various facets of protein modeling like refinement of the modeled structure with the experimentally solved, exact estimation of whole structure and all residues, the capacity to improve models by utilizing an assortment of scanty information, modeling of protein structure oligomers and deriving the protein function by identification of its protein structure are also checked out in this contest (Kryshtafovych et al. 2019). Combining different types of computational methods and advancement of structural prediction techniques are required for ever growing structural genomics data that would give impetus to different areas of biology and would solve the biological puzzles related to protein structure and its function (Mills et al. 2015). An important method that plays a vital role in solving the Levinthal's paradox and in protein folding is molecular dynamics simulation (MDS) (Camilloni et al. 2008). Further we will be discussing on the theoretical concepts of protein folding through MDS.

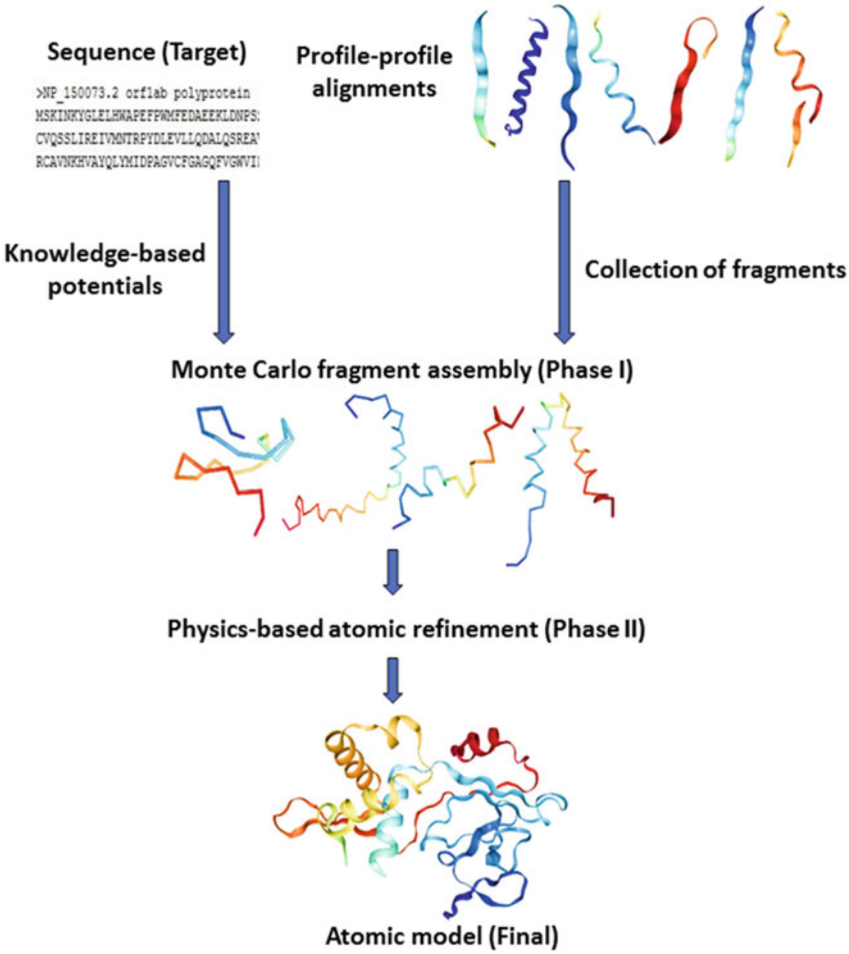

Sequence (Target)
>NP_150073.2 orf1ab polyprotein
MSKINKYGLELHWAPEFPWMFEDAEEKLDNPS
CVQSSLIREIVMNTRPYDLEVLLQDALQSREA
RCAVNKHVAYQLYMIDPAGVCFGAGQFVGWVI
Profile-profile alignments
Knowledge-based potentials
Collection of fragments
Monte Carlo fragment assembly (Phase I)
Physics-based atomic refinement (Phase II)
Atomic model (Final)

Fig. 7.2 A general outline of protein structure prediction from primary sequence

7.3 Computational Perspectives of Protein Folding

Proteins folds to their active native structure once they emerge from ribosome and remain continuously unfold and refold throughout their lifetime. The biological functions of protein determine by their proper folding and misfolding of the proteins cause a series of diseases. Therefore, it is important to analyze structural characterization and dynamic behavior of the proteins in order to understand their functional mechanism in theoretical and experimental studies (Duan et al. 2019; Miao and McCammon 2017). Computational molecular dynamics simulation (MDS) has become a popular tool to interpret experimental protein folding results, analyze the effects of small molecule on protein folding (Gershenson et al. 2020). However, many experimental and computational studies are limited to understand how small proteins fold but the folding of multi-domain proteins is less progressive. Due to high energy barriers, the conformational changes of large proteins require the time periods ranging from microsecond to millisecond or even longer (Markwick and McCammon 2011).

7.3.1 Molecular Dynamics Simulation and Its Application in Protein Folding

In this scenario, molecular dynamics (MD) simulation plays a significant role in the study of structural and dynamic properties of protein folding at the atomic level which offers exhaustive details on folding free energy landscapes, different intermediates and pathways protein undergoes while folding (Gershenson et al. 2020; Xu et al. 2018). Molecular dynamics simulations are techniques which uses theoretical and physics based approaches to study iterative calculations performed using classical mechanics (by using Newton's/Lagrange's equations of motions) to predict the movement of atoms or particles over time, thus enables in obtaining the coordinates and momenta of the particles along the folding and unfolding trajectories (Allen 2004; Scheraga et al. 2007). MDS uses the computational techniques and algorithms to calculate the molecular interactions (like bond lengths, bond angles, dihedral angles), non-bonded interaction (van der Waals and electrostatics) and force fields like AMBER, CHAARMM, OPLS, GROMOS to calculate the quantum mechanical calculations or experimental observables related to the conformational behavior of simulated proteins focusing on the parameters of backbone and side chain dihedral angles of the simulated protein (Allen 2004; Geng et al. 2019). MDS has been utilized from quite a while for studying protein's conformational dynamics for predicting ensemble structures of protein in the native state (McCammon et al. 1977; Gaalswyk et al. 2018). It is highly effective in examining the intricacies of protein folding and its impact on protein motions during catalysis and ligand binding. Both these protein dynamics are inherently thermodynamic; protein conformational changes, and coupled fluctuations, are the two types of dynamics that are most compliable to investigate by molecular dynamics simulations, and have become specifically significant for pharmacological invention with the advancement of broadly emerging allostery concepts. According to the classical view of generalized allostery, virtually all proteins are allosteric to some extent; by virtue of either coupling of conformational changes or due to long-range communication between parts of proteins or protein complexes, and this allostery can be utilized in the drug-discovery process (Salsbury 2011). Details about MDS and its applications in protein dynamics can be found in above referenced articles of this section, as well as in (Hollingsworth and Dror 2018; Patodia et al. 2014; Childers and Daggett 2017) and equally through a simple web search.

Additionally, increased availability of large-scale supercomputing resources and methodological advances provides a possible shortcut to reach long time scale folding simulation of large proteins at atomic level (Duan et al. 2019). However, all researchers have not been able to access the supercomputing facilities and still there is a time scale gap between molecular dynamics simulation and experimental observation. So the problem can be resolved by thorough sampling of protein conformational space, which includes different methods like umbrella sampling (Torrie and Valleau 1977), multicanonical algorithms (Berg and Neuhaus 1992), simulated tempering (Marinari and Parisi 1992), transition path sampling (Bolhuis et al. 2002), targeted molecular dynamics (Schlitter et al. 1994; Ma et al. 2000), replica exchange molecular dynamics (REMD) also known as parallel tempering methods (Hukushima and Nemoto 1996; Hansmann 1997; Sugita and Okamoto 1999; Garcia and Sanbonmatsu 2002), and accelerated molecular dynamics (aMD) (Hamelberg et al. 2004) have been developed which can effectively extend the timescale in the simulation.

7.3.2 Molecular Dynamics Simulation Complementing the Experiments

Experimentally determining the atomic details of in vivo protein folding which is fast and efficient is difficult to study, while MDS can resolve the spatial and temporal resolution through local fluctuations around equilibrium conformations to huge conformal variations, which helps in reviewing the proteins at expedient time scale (Childers and Daggett 2017). Still modeling of protein folding on biological timescales (ranging in microsecond to second) is challenging due to rigorous force fields, adequate sampling and robust data analysis within stipulated timescales. Additionally protein folding through MDS requires enormous computational resources and efforts with advanced hardware, software and sampling techniques in order to reach the millisecond timescales simulations (Lane et al. 2013). The trend to use long MDS with advanced sampling methods and technologies has increased. For example Voelz et al. in 2010 reported 1.5 ms simulations of numerous folding trajectories from the unfolded state of the 39 residue protein NTL9 (1–39) by using implicit solvent model and distributed computing network on Folding@home (Voelz et al. 2010). Later, Bowman et al. adopted explicit solvent model to reveal the folding dynamics of lambda repressor in millisecond time span (Lane et al. 2013). Villin headpiece simulation by Duan and Kollman for 1μs in 1998 and 1 ms simulation of bovine pancreatic trypsin inhibitor (BPTI) are remarkable benchmark of long simulations (Duan and Kollman 1998; Shaw et al. 2010). Exclusively designed supercomputer for protein simulation viz. ANTON was used for the folding prediction of BPTI protein for 1 ms encased within 100μs folding prediction in a single trajectory (Shaw et al. 2010; Schuler and Hofmann 2013). Though the generation of longer MDS in feasible time is expected to increase in future, the accessible timescales presently limits the study to fast folders with folding time less than a millisecond (Gelman and Gruebele 2014; Daggett 2006). Efficiency of the MDS can be increased through the reduction in number of atoms of the protein like using coarse grained models over atomistic models and decreasing the collision within the simulated system by using implicit solvent model over explicit solvent model (Rudzinski 2019).

7.3.3 Free Energy Landscapes and Protein Folding

The reversible folding of protein molecules into its three-dimensional native state as shown by Anfinsen, opines that these structures correspond to the global minimum of a rocky funnel like energy landscape (Onuchic et al. 1997). The hierarchical folding theory postulated by Baldwin and Rose states, the initiation of the protein folds start with the formation of local structural components, viz. α-helices and β-strands (Baldwin and Rose 1999a, b). Mutual interaction within these secondary structure components, gives rise to the folded protein. Genesis of local structural components lowers the protein entropy, like the side chains of helical residues are heavily restrained by the remaining helix. Later favorable short-range interactions, comprising hydrogen bonding and desolvation of backbone polar groups balances the loss of entropy. This is contemplated as the elementary trait of proteins, and MDS mimic this property to simulate protein folding (Burkoff et al. 2012). Exploring the free energy landscapes (FELs) aids in the recognition of the way the proteins fold and function (Maisuradze et al. 2010). Energy-landscape theory has significantly advanced the understanding of protein folding kinetics, protein structure prediction and protein design. Funnel landscapes depicts protein folding and binding and determines its kinetics from protein topology. Landscape-optimized energy functions derived from bioinformatics input are used for correctly predicting low-resolution protein structures and to design novel proteins spontaneously (Wolynes 2005). FEL is the foundation pillar within protein folding research. Its most integral factor is such the protein folding is energetically biased to the globally funneled minimum for the native state (Chong and Ham 2019). Classical MD are sufficient for acquiring native state dynamics, but not enough to shun from the native basin and overcome the energy barrier, hence are limited in their aptness to characterize the energy landscape besides the global minimum. The breakthrough across such barriers under normal temperature and pressures rarely occurs; hence, the MDS methods used in the study of protein folding strives to bypass this restriction in various ways that trims the size of the barriers or increment the likelihood of barrier crossing. Protein folding biased MDS is explicitly utilized for determining the folding simulations of large proteins like adenylate kinase, GFP, TIM barrels, dihydrofolate reductase, a DNA polymerase, and serpins. These specialized MDS have successfully produced folding pathways and intermediates that concur experimentally and give verifiable theories on which intermediate states are probably going to be populated while folding (Gershenson et al. 2020). Further we will be discussing various MDS methods useful in protein folding.

7.4 MDS Methods in Protein Folding

This section outlines different MDS methods used and applied extensively in protein folding studies.

7.4.1 Replica Exchange Molecular Dynamics (REMD)

REMD is a robust method to examine the free energy surface and different thermodynamic properties of biomolecules initiated by Okamato and his colleagues on pentapeptide Met-enkephalin, wherein MDS and Monte Carlo algorithm are combined to overcome the high energy barriers and adequately sample the conformational space of proteins (Sugita and Okamoto 1999). REMD is a generalized ensemble of system containing several copies (replicas) of the same system simulated parallelly using MDS at various temperatures or at identical temperature though applying diverse Hamiltonians. Swapping between neighboring copies (replicas) is intermittently endeavored with a likelihood given by the Metropolis rule. In this manner REMD competently overcomes high energy barriers quickly and decently samples the conformational space, which provides the analysis of free energy landscape of protein folds. Besides the parallel aspect of REMD makes it reasonable to distribute the job over exceptionally parallel computing clusters that is frequently accessible these days (Qi et al. 2018). Though REMD effectively samples the energy landscapes by removing the limitation of kinetic traps and energetic barriers existing at lower temperatures it is not suitable for establishing protein folding pathways and kinetics due to the variation within moving states (Beck et al. 2007). Several variants of the conventional temperature dependent REMD (T-REMD), like reservoir REMD (R-REMD), Hamiltonian REMD (H-REMD), multiplexed REMD (M-REMD), pH REMD and free energy perturbation (FEP) combined λ-REMD has been used to improve the conformational sampling of MDS (Bernardi et al. 2015).

7.4.2 Coarse Grained Models for MDS

Coarse graining methodology reduces the complexity of a system by considering groups of atoms/molecules as single pseudo particles (Saunders and Voth 2013). In coarse graining the complexity of a system is resolved by clustering atoms into new bunches with simplified coarse grained beads or pseudo atoms. CG models are used with the purpose to cross the edge of accessible temporal and spatial scales of biomolecular systems. CG beads adequately reduces the overall flexibility i.e. the absolute degrees of freedom present in the system linked with one another under improved computational potential (Merchant and Madura 2011). Simulations of longer time scales or potentially larger systems could be proficiently reproduced by CG model. The model inducts varying level of reduction in the polypeptide chain where the protein primary chain could be communicated by every single heavy atoms or by linking two atoms per residue, although the side chain is substituted by just a couple of joined molecules. This unlocks the plausibility of multiscale modeling, derived by combining the computational speed of CG models with the highly precise classical all-atom MDS. It has been effectively applied for resolving protein folding mechanism in accordance with generalized protein like model or simulations of actual proteins. This is the finest available technique for the challenging de novo modeling and is crucial for advanced comparative modeling. Other tasks carried out includes predicting protein structure, complex dynamic process modeling, protein interaction with proteins and peptides, besides modeling of membrane proteins (Kmiecik et al. 2016).

7.4.3 Accelerated Molecular Dynamics (aMD)

In aMD enhancement of sampling is done by adding a non-negative potential boost to shrink the energy barrier and quickens the changes or transitions within low energy states. In this way it raises the low-energy regions on the potential energy landscape, and lowers the energy barriers, thereby accelerating the interchange of low-energy conformational states unaccessible in traditional MDS. One of the advantages of aMD is that the canonical average of an observable can be obtained which enables in determining the thermodynamic and other equilibrium properties unambiguously. Moreover, prior knowledge about the underlying potential energy surface is not required in aMD (Bucher et al. 2011). aMD has been successfully used to study slow scale time scale dynamics of protein like ubiquitin, HIV protease, H-Ras and IKBA as well as for enhancing the configurational sampling in ab initio simulations (Bucher et al. 2011). Additionally, simulation of four fast folding proteins including chignolin, Trp-cage, villin headpiece and WW domain by aMD took less time for folding compared to conventional MDS, and simultaneously the folded protein conformation were in the range of 0.2 to 2.1 Å of the native NMR or X-ray crystal structures (Miao et al. 2015). Similarly Duan et al. have used aMD and traditional MDS for studying the folding process of eight helical proteins using explicit solvent model at different temperatures under AMBER14SB force field and found aMD to be more better and effective technique for protein folding (Duan et al. 2019). Likewise protein folding simulations of five proteins in implicit solvent model at room temperature through aMD was conducted by Zong-Chao et al. with conformational resolution nearer to the native structures, whereas the protein folding of the same structures failed in normal MDS (Li et al. 2015b). Thus the above examples fortify that the aMD could be used for studying protein folding simulations.

7.4.4 Umbrella Sampling

Umbrella sampling, as an enhanced sampling method allows the sampling by accelerating the conformational dynamics along reaction coordinates thus we can estimate the change in free energy in each window (Xu et al. 2017). Afterwards, the weighted histogram analysis method or umbrella integration methods are applied to combine each window by utilizing a restraint bias quadratic or harmonic formed potential. At this stage, the bias potential can be adjusted to bring about an even dispersion between the end states, and then this entire range can be spread over in single window (Kästner 2011; Bowman and Lindert 2018). Umbrella sampling can be utilized to investigate the differences in the free energy of
protein folding windows and allows the system to explore the region which may not be touched at the time of conventional MDS due to presence of high energy barriers (Bowman and Lindert 2018; Ito et al. 2018). Umbrella sampling computed potential of mean force (PMF) by generating binding affinity and energy barriers which showed more or less similarities with experimental date (You et al.2017). These PMF seem to influenced by many practical aspects such as method used to generate the initial dissociation pathway, collective variables (CVs) used in reaction coordinate (RC), and how CV restrained the conformational space of the sampling (You et al. 2017; Demuynck et al. 2017).

7.4.5 Markov State Model (MSM) for Protein Folding

MSM is another valuable MDS tool broadly utilized for protein and peptide folding for quantitatively studying the conformational dynamics of macromolecules. In this model, the biomolecular dynamics are modeled as a stochastic network of transitions between metastable conformational states based on discrete master equation approach called Markov state model (MSM) (Sirur et al. 2016; Weber and Pande 2011). Initially a protein's configuration space is partitioned into a set of kinetically distinct states and parallel sampling methods are applied to obtain MSMs. When the time period where changes between the states stay memoryless is settled, a MSM transition network propels dynamics to longer time spans required to plot the folding procedure (Weber and Pande 2011). Memoryless state here means, the system's evolution with the increment of time depends on the properties in current time and not on its past history i.e. if a system transitions from X to Y the system does not remember how it entered X (Swope et al. 2004; Husic and Pande 2018). Shortly, MSM could portray the long time behavior of protein folding in a solvated protein system through a number of relatively short time MDS. Recently protein NTL9 and the four-helix bundle of λ-repressor were simulated in millisecond timescale that ensures the potential of MSMs in simulating slowly folding protein system (Weber and Pande 2011).

7.4.6 Path Sampling Approaches for Protein Folding

Path sampling approaches (PSA) of protein folding computationally focuses the functional transitions in place of stable states without inducting bias in the results. PSA considerably increases MDS application in rare events like protein conformational changes, unfolding and unbinding which is difficult to dispose through conventional MDS (Chong et al. 2017). Transition path sampling (TPS) method under PSA is numerical methods that adequately apply Monte Carlo sampling on the ensemble of transition paths, which is solely biased towards the transition without disturbing the path and the potential itself. TPS spotlights on shorter MD trajectories that cross the barriers, thereby rapidly expedites the rare events sampling (Swenson et al. 2019). Different types of PSA in combination with variety of MDS techniques has been used to determine the conformation transitions of rare events like open and closed states of HIV reverse transcriptase, binding of p53 peptide in α helical conformation with MDM2 are reviewed by Chong et al. (2017).

7.5 Assessment of Protein Stability Through MDS

Alteration in protein structures and functions offers to create and design proteins with greater application and improved stability over their natives by the use of protein engineering. Strengthening the protein stability while subjecting to severe conditions is largely practiced in commercial and research sectors for its beneficial impact on swift chemical reactions, enhanced dissolvability of substrates, better resistance towards microbial deterioration and simpler stockpiling and treatment of proteins, will help the biotechnological advancements in entirety or part (Zhang and Lazim 2017). One of the principle focuses in protein engineering is enhancing protein stability. MDS enabled detailed analysis of the complex dynamics of biological macromolecules helps in gaining the theoretical knowledge of conformational changes in protein, its folding, unfolding mechanism, and its stability (Karplus 1987). These exciting applications accessible by MDS are useful in protein engineering because it gives insight on the achievability of the conformation of the native and mutated protein by giving information involving critical interactions, for example hydrophobic, Van der Waals, electrostatic and hydrogen bonding interactions that could direct the stability of the protein in study. Mutants are rationally designed to boost protein stability by improving stabilizing interactions or by decreasing destabilizing features. These elements of MDS remove the odds of disturbing the tertiary structure of the native protein when mutagenesis is performed experimentally; making MDS an appealing innovation to be used by experimental scientists engaged in protein research that increases their efficiency by reducing the tasks at hand and expenses (Zhang and Lazim 2017; Pikkemaat et al. 2002). MDS was used to determine the stability of apomyoglobin and the effect of some mutation on the general conformation of the protein in urea solution by Zhang and Lazim by analyzing variations in RMSD, native contacts and solvent accessible surface area whereas RMSF, correlation matrix, principal component analysis, and hydrogen bond analysis was used to study the destabilization of apomyoglobin variants (Zhang and Lazim 2017). Similarly, MDS was used for identifying flexible regions in haloalkane dehalogenase proteins that can serve as a target for stability improvement by inserting a disulfide bond (Pikkemaat et al. 2002). The effect of temperature on the stability of α subunit of tryptophan synthase from hyperthermophilic, mesophilic, and psychrophilic protein homologs was determined by comparative MDS (Khan et al. 2016). Cases of modulation of protein stability and functions through conformational changes and interaction as well as the impact on protein folding and unfolding due to conformational transition has been used for designing proteins by using MDS has been discussed by Childers and Daggett (Childers and Daggett 2017).

7.6 Enhancing Drug Discovery by MDS

MDS strategies are regularly utilized in computational drug discovery process. Its major influence is in specifically recognizing the structural flexibility and entropic impacts that permit the thermodynamics and energy correlation for drug-target detection and binding. The availability of enhanced algorithms, improved forcefield calculations and superior hardware framework augments its utilization in the drug discovery field. Impartial or unbiased MDS these days permit examining of ligand-target binding, evaluating drug target affinity and drug dwelling time within the target pocket as drug efficacy criteria (De Vivo et al. 2016; Salmaso and Moro 2018). Other topic examined through MDS includes allosteric regulation and the role water particles play in ligand binding and optimization (Vettoretti et al. 2016; Prabhu and Singh 2019; Nayak et al. 2019; Jung et al. 2018). MDS are excellent tools for identifying cryptic and allosteric binding sites on protein which are many times not evident from the crystal structure. MDS played an important role in finding the drug target for HIV integrase and later the drug raltegravir was approved by FDA as first HIV integrase inhibitor. It has been effectively used to generate numerous receptor conformations sample to find inhibitors of FKBP, T. brucei RNA editing ligase 1, T. brucei GalE, T. brucei FPPS, and Mycobacterium tuberculosis dTDP-6-deoxy-Llyxo-4-hexulose (Durrant and McCammon 2011). MDS have been utilized broadly in exploring the pathogenic systems of infections brought about by protein misfolding, in virtual screening, and in investigating drug resistance mechanisms because of mutations of the target. The applications of MDS in novel drug discovery, including the pathogenic mechanisms of amyloidosis diseases comprising neurodegenerative disorders like Alzheimer disease, Parkinson disease, Gerstmann-Straussler-Scheinker (GSS) disease, Creutzfeldt-Jakob disease (CJD) and bovine spongiform encephalopathy (BSE) and the affect of mutations on protein structure is immensely helpful. The combined use of crystal structures and the conformational clusters obtained through MDS in ensemble docking have much better performance

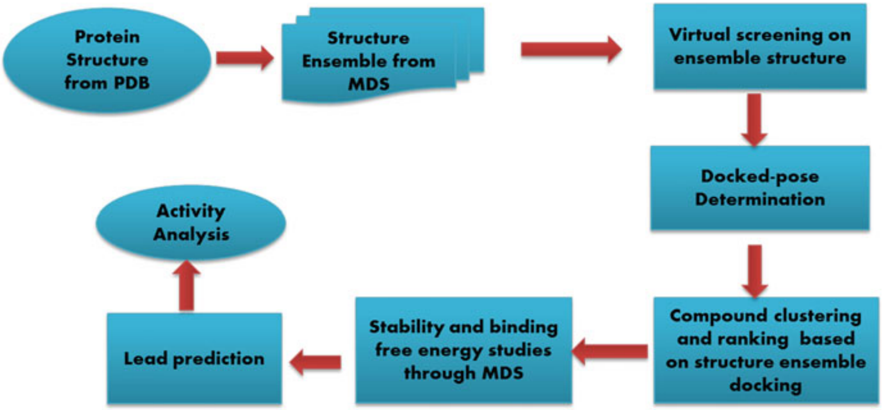

Protein Structure from PDB
Structure Ensemble from MDS
Virtual screening on ensemble structure
Docked-pose Determination
Compound clustering and ranking based on structure ensemble docking
Stability and binding free energy studies through MDS
Lead prediction
Activity Analysis

Fig. 7.3 Integrating MDS with virtual screening for computational drug discovery

than the use of either crystal structures or conformational clusters only through MDS (Liu et al. 2018). MDS have been applied for conformational refinement and the ranking of candidates by combining binding free energy calculations through MM-PBSA and MM-GBSA calculations (Panwar and Singh 2018; Suryanarayanan and Singh 2015; Reddy et al. 2013; Selvaraj et al. 2016; Shukla et al. 2019; Dhasmana et al. 2018; Selvaraj and Singh 2014; Aarthy et al. 2018; Patidar et al. 2019; Gupta et al. 2017; Reddy and Singh 2014; Panwar et al. 2019; Tripathi and Singh 2014) which could be further utilized for lead optimization studies and elucidation of drug resistance mechanism (Selvaraj et al. 2014; Nayak et al. 2019; Shafreen et al. 2013; Vijayalakshmi et al. 2013; Tripathi et al. 2012). A scheme for implementing MDS in drug discovery process has been illustrated in Fig. 7.3.

7.7 Conclusions

The information related to the prediction of protein folding from the primary polypeptide sequences through protein prediction and molecular dynamics simulation tools is covered in this chapter. Difficulty in studying the atomic details of protein folding only through in vivo techniques is obvious. This is where the MDS techniques have complemented well in solving some of the protein folding problem computationally and has been discussed. Here we have described about some of the specialized MDS methods like umbrella sampling, replica exchange molecular dynamics, coarse grained MDS, accelerated molecular dynamics, markov state model and path sampling approaches that are extensively applied to understand the protein folding mechanism. Folding pattern of globular proteins by using single MDS is renowned whereas the complicated systems include multidomain protein, the simulation that perturbs the protein protein interaction to mimic as such in living cell is challenging, since it requires state-of-the-art computational resources to undertake such simulations. Nevertheless the MDS field has expanded immensely and its importance in structural biology and protein folding could not be relegated. Further research and advances in algorithmic and computational functions will bring more depth in the protein folding area through MDS. This will help in assessing the mechanism of several protein folding disorders like Alzheimer, Parkinson and could aid in their therapies through protein engineering and design (Childers and Daggett 2017). MDS significance in protein stability studies and drug discovery has been also purported in the chapter which highlights its versatility in biological and medicinal research.

Acknowledgements The authors are pleased to acknowledge the research facility provided by Alagappa University. The authors also thank RUSA—Phase 2.0 (grant sanctioned vide letter No. F. 24-51/2014—U, Policy (TNMulti-Gen), Dept. of Edn. Govt. of India, Dt. 09.10.2018), Department of Biotechnology (DBT) Ministry of Science and Technology, New Delhi, Grant/ Award (No. BT/PR8138/BID/7/458/2013, dated 23rd May, 2013), DST-PURSE 2ndPhase Programme Order No. SR/PURSE Phase 2/38 (G dated 21.02.2017 and FIST (SR/FST/LSI - 667/2016), MHRD for the financial support.

Conflict of Interest The authors declare no conflict of interest.

References

Aarthy, M., Kumar, D., Giri, R., & Singh, S. K. (2018). E7 oncoprotein of human papillomavirus: Structural dynamics and inhibitor screening study. *Gene, 658*, 159–177. https://doi.org/10.1016/j.gene.2018.03.026

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). *Molecular biology of the cell* (4th ed.). New York: Garland Science. The shape and structure of proteins. Available from: https://www.ncbi.nlm.nih.gov/books/NBK26830/ Accessed 13 Feb 2020

Allen, M. P. (2004). *Introduction to molecular dynamics simulation*. Retrieved February 13, 2020, from https://udel.edu/~arthij/MD.pdf

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science, 181*, 223–230. https://doi.org/10.1126/science.181.4096.223

Baldwin, R. L., & Rose, G. D. (1999a). Is protein folding hierarchic? I. Local structure and peptide folding. *Trends in Biochemical Sciences, 24*, 26–33. https://doi.org/10.1016/s0968-0004(98)01346-2

Baldwin, R. L., & Rose, G. D. (1999b). Is protein folding hierarchic? II. Folding intermediates and transition states. *Trends in Biochemical Sciences, 24*, 77–83. https://doi.org/10.1016/s0968-0004(98)01345-0

Beck, D. A. C., White, G. W. N., & Daggett, V. (2007). Exploring the energy landscape of protein folding using replica-exchange and conventional molecular dynamics simulations. *Journal of Structural Biology, 157*, 514–523. https://doi.org/10.1016/j.jsb.2006.10.002

Berg, B. A., & Neuhaus, T. (1992). Multicanonical ensemble: A new approach to simulate first-order phase transitions. *Physical Review Letters, 68*, 9–12. https://doi.org/10.1103/PhysRevLett.68.9

Bernardi, R. C., Melo, M. C. R., & Schulten, K. (2015). Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochimica et Biophysica Acta, 1850*, 872–877. https://doi.org/10.1016/j.bbagen.2014.10.019

Blanco, A., & Blanco, G. (2017). Proteins. In *Medical biochemistry* (pp. 21–71). Cambridge, MA: Academic.

Bolhuis, P. G., Chandler, D., Dellago, C., & Geissler, P. L. (2002). Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annual Review of Physical Chemistry, 53*, 291–318. https://doi.org/10.1146/annurev.physchem.53.082301.113146

Bowman, J. D., & Lindert, S. (2018). Molecular dynamics and umbrella sampling simulations elucidate differences in troponin C isoform and mutant hydrophobic patch exposure. *The Journal of Physical Chemistry. B, 122*, 7874–7883. https://doi.org/10.1021/acs.jpcb.8b05435

Buchan, D. W., & Jones, D. T. (2019). The PSIPRED protein analysis workbench: 20 years on. *Nucleic Acids Research, 47*, W402–W407. https://doi.org/10.1093/nar/gkz297

Bucher, D., Pierce, L. C. T., McCammon, J. A., & Markwick, P. R. L. (2011). On the use of accelerated molecular dynamics to enhance configurational sampling in Ab initio simulations. *Journal of Chemical Theory and Computation, 7*, 890–897. https://doi.org/10.1021/ct100605v

Burkoff, N. S., Várnai, C., Wells, S. A., & Wild, D. L. (2012). Exploring the energy landscapes of protein folding simulations with Bayesian computation. *Biophysical Journal, 102*, 878–886. https://doi.org/10.1016/j.bpj.2011.12.053

Camilloni, C., Sutto, L., Provasi, D., Tiana, G., & Broglia, R. A. (2008). Early events in protein folding: Is there something more than hydrophobic burst? *Protein Science, 17*, 1424–1433. https://doi.org/10.1110/ps.035105.108

Chaudhuri, T. K., & Paul, S. (2006). Protein-misfolding diseases and chaperone-based therapeutic approaches. *The FEBS Journal, 273*, 1331–1349. https://doi.org/10.1111/j.1742-4658.2006.05181.x

Childers, M. C., & Daggett, V. (2017). Insights from molecular dynamics simulations for computational protein design. *Molecular Systems Design and Engineering, 2*, 9–33. https://doi.org/10.1039/C6ME00083E

Chong, L. T., Saglam, A. S., & Zuckerman, D. M. (2017). Path-sampling strategies for simulating rare events in biomolecular systems. *Current Opinion in Structural Biology, 43*, 88–94. https://doi.org/10.1016/j.sbi.2016.11.019

Chong, S. H., & Ham, S. (2019). Folding Free Energy Landscape of Ordered and Intrinsically Disordered Proteins. *Scientific Reports, 9*, 1–9. https://doi.org/10.1038/s41598-019-50825-6

Cole, C., Barber, J. D., & Barton, G. J. (2008). The Jpred 3 secondary structure prediction server. *Nucleic Acids Research, 36*, W197–W201. https://doi.org/10.1093/nar/gkn238

Daggett, V. (2006). Protein folding-simulation. *Chemical Reviews, 106*, 1898–1916. https://doi.org/10.1021/cr0404242

De Vivo, M., Masetti, M., Bottegoni, G., & Cavalli, A. (2016). Role of molecular dynamics and related methods in drug discovery. *Journal of Medicinal Chemistry, 59*, 4035–4061. https://doi.org/10.1021/acs.jmedchem.5b01684

Demuynck, R., Rogge, S. M. J., Vanduyfhuys, L., Wieme, J., Waroquier, M., & Van Speybroeck, V. (2017). Efficient construction of free energy profiles of breathing metal-organic frameworks using advanced molecular dynamics simulations. *Journal of Chemical Theory and Computation, 13*, 5861–5873. https://doi.org/10.1021/acs.jctc.7b01014

Deng, H., Jia, Y., & Zhang, Y. (2018). Protein structure prediction. *International Journal of Modern Physics B, 32*, 1840009. https://doi.org/10.1142/S021797921840009X

Dhasmana, D., Singh, A., Shukla, R., Tripathi, T., & Garg, N. (2018). Targeting nucleotide binding domain of multidrug resistance-associated protein-1 (MRP1) for the reversal of multi drug resistance in cancer. *Scientific Reports, 8*, 11973. https://doi.org/10.1038/s41598-018-30420-x

Duan, L., Guo, X., Cong, Y., Feng, G., Li, Y., & Zhang, J. Z. (2019). Accelerated molecular dynamics simulation for helical proteins folding in explicit water. *Frontiers in Chemistry, 7*, 540. https://doi.org/10.3389/fchem.2019.00540

Duan, Y., & Kollman, P. A. (1998). Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science, 282*, 740–744. https://doi.org/10.1126/science.282.5389.740

Durrant, J. D., & McCammon, J. A. (2011). Molecular dynamics simulations and drug discovery. *BMC Biology, 9*, 71. https://doi.org/10.1186/1741-7007-9-71

Finkelstein, A. V., & Garbuzynskiy, S. O. (2016). Solution of Levinthal's paradox is possible at the level of the formation and assembly of protein secondary structures. *Biophysics, 61*, 1–5. https://doi.org/10.1134/S0006350916010085

Gaalswyk, K., Muniyat, M. I., & MacCallum, J. L. (2018). The emerging role of physical modeling in the future of structure determination. *Current Opinion in Structural Biology, 49*, 145–153. https://doi.org/10.1016/j.sbi.2018.03.005

Garcia, A. E., & Sanbonmatsu, K. Y. (2002). α-helical stabilization by side chain shielding of backbone hydrogen bonds. *Proceedings of the National Academy of Sciences of the United States of America, 99*, 2782–2787. https://doi.org/10.1073/pnas.042496899

Gelman, H., & Gruebele, M. (2014). Fast protein folding kinetics. *Quarterly Reviews of Biophysics, 47*, 95–142. https://doi.org/10.1017/S003358351400002X

Geng, H., Chen, F., Ye, J., & Jiang, F. (2019). Applications of molecular dynamics simulation in structure prediction of peptides and proteins. *Computational and Structural Biotechnology Journal, 17*, 1162–1170. https://doi.org/10.1016/j.csbj.2019.07.010

Geourjon, C., & Deleage, G. (1995). SOPMA: Significant improvement in protein secondary structure prediction by consensus prediction from alignments and joint prediction. *Computer Applications in the Biosciences, 11*, 681–684. https://doi.org/10.1093/bioinformatics/11.6.681

Gershenson, A., Gosavi, S., Faccioli, P., & Wintrode, P. L. (2020). Successes and challenges in simulating the folding of large proteins. *The Journal of Biological Chemistry, 295*, 15–33. https://doi.org/10.1074/jbc.REV119.006794

Gillet, J. N., & Ghosh, I. (2013). Concepts on the protein folding problem. *Journal of Biomolecular Structure & Dynamics, 31*, 1020–1023. https://doi.org/10.1080/07391102.2012.748546

Gupta, S., Suryanarayanan, V., Yadav, S., Singh, S. K., & Saxena, J. K. (2017). Delineating the role of ionic interactions in structural and functional integrity of B. malayi Guanylate kinase. *International Journal of Biological Macromolecules, 98*, 357–365. https://doi.org/10.1016/j.ijbiomac.2017.01.098

Hamelberg, D., Mongan, J., & McCammon, J. A. (2004). Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *The Journal of Chemical Physics, 120*, 11919–11929. https://doi.org/10.1063/1.1755656

Hansmann, U. H. E. (1997). Parallel tempering algorithm for conformational studies of biological molecules. *Chemical Physics Letters, 281*, 140–150. https://doi.org/10.1016/S0009-2614(97)01198-6

Hansmann, U. H. E. (2003). Protein folding in silico: An overview. *Computing in Science & Engineering, 5*, 64–69. https://doi.org/10.1109/MCISE.2003.1166554

Hatherley, R., Brown, D. K., Glenister, M., & Tastan Bishop, Ö. (2016). PRIMO: An interactive homology modeling pipeline. *PLoS One, 11*, e0166698. https://doi.org/10.1371/journal.pone.0166698

Hollingsworth, S. A., & Dror, R. O. (2018). Molecular dynamics simulation for all. *Neuron, 99*, 1129–1143. https://doi.org/10.1016/j.neuron.2018.08.011

Hukushima, K., & Nemoto, K. (1996). Exchange Monte Carlo method and application to spin glass simulations. *Journal of the Physical Society of Japan, 65*, 1604–1608. https://doi.org/10.1143/JPSJ.65.1604

Husic, B. E., & Pande, V. S. (2018). Markov state models: From an art to a science. *Journal of the American Chemical Society, 140*, 2386–2396. https://doi.org/10.1021/jacs.7b12191

Ito, S., Wang, Y., Okamoto, Y., & Irle, S. (2018). Quantum chemical replica-exchange umbrella sampling molecular dynamics simulations reveal the formation mechanism of iron phthalocyanine from iron and phthalonitrile. *The Journal of Chemical Physics, 149*, 072332. https://doi.org/10.1063/1.5026956

Jaenicke, R. (1991). Protein folding: Local structures, domains, subunits, and assemblies. *Biochemistry, 30*, 3147–3161. https://doi.org/10.1021/bi00227a001

Jung, S. W., Kim, M., Ramsey, S., Kurtzman, T., & Cho, A. E. (2018). Water pharmacophore: Designing ligands using molecular dynamics simulations with water. *Scientific Reports, 8*, 10400. https://doi.org/10.1038/s41598-018-28546-z

Kabsch, W., & Sander, C. (1983). DSSP: Definition of secondary structure of proteins given a set of 3D coordinates. *Biopolymers, 22*, 2577–2637. https://doi.org/10.1002/bip.360221211

Karplus, M. (1987). Molecular dynamics simulations of proteins. *Physics Today, 40*, 68–72. https://doi.org/10.1063/1.881116

Karplus, M., & Šali, A. (1995). Theoretical studies of protein folding and unfolding. *Current Opinion in Structural Biology, 5*, 58–73. https://doi.org/10.1016/0959-440X(95)80010-X

Kästner, J. (2011). Umbrella sampling. *Wiley Interdisciplinary Reviews: Computational Molecular Science, 1*, 932–942. https://doi.org/10.1002/wcms.66

Khan, S., Farooq, U., & Kurnikova, M. (2016). Exploring protein stability by comparative molecular dynamics simulations of homologous hyperthermophilic, mesophilic, and psychrophilic proteins. *Journal of Chemical Information and Modeling, 56*, 2129–2139. https://doi.org/10.1021/acs.jcim.6b00305

Khor, B. Y., Tye, G. J., Lim, T. S., & Choong, Y. S. (2015). General overview on structure prediction of twilight-zone proteins. *Theoretical Biology & Medical Modelling, 12*, 15. https://doi.org/10.1186/s12976-015-0014-1

Khoury, G. A., Smadbeck, J., Kieslich, C. A., & Floudas, C. A. (2014). Protein folding and de novo protein design for biotechnological applications. *Trends in Biotechnology, 32*, 99–109. https://doi.org/10.1016/j.tibtech.2013.10.008

Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A. E., & Kolinski, A. (2016). Coarse-grained protein models and their applications. *Chemical Reviews, 116*, 7898–7936. https://doi.org/10.1021/acs.chemrev.6b00163

Kolata, G. (1986). Trying to crack the second half of the genetic code. *Science, 233*, 1037–1039. https://doi.org/10.1126/science.3738524

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., & Moult, J. (2019). Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins: Structure, Function, and Bioinformatics, 87*, 1011–1020. https://doi.org/10.1002/prot.25823

Lane, T. J., Shukla, D., Beauchamp, K. A., & Pande, V. S. (2013). To milliseconds and beyond: Challenges in the simulation of protein folding. *Current Opinion in Structural Biology, 23*, 58–65. https://doi.org/10.1016/j.sbi.2012.11.002

Li, J., Adhikari, B., & Cheng, J. (2015a). An improved integration of template-based and template-free protein structure modeling methods and its assessment in CASP11. *Protein and Peptide Letters, 22*, 586–593. https://doi.org/10.2174/0929866522666150520145717

Li, Z. C., Duan, L. L., Feng, G. Q., & Zhang, Q. G. (2015b). All-atom direct folding simulation for proteins using the accelerated molecular dynamics in implicit solvent model. *Chinese Physics Letters, 32*, 118701. https://doi.org/10.1088/0256-307X/32/11/118701

Liu, X., Shi, D., Zhou, S., Liu, H., Liu, H., & Yao, X. (2018). Molecular dynamics simulations and novel drug discovery. *Expert Opinion on Drug Discovery, 13*, 23–37. https://doi.org/10.1080/17460441.2018.1403419

Ma, J., Sigler, P. B., Xu, Z., & Karplus, M. (2000). A dynamic model for the allosteric mechanism of GroEL. *Journal of Molecular Biology, 302*, 303–313. https://doi.org/10.1006/jmbi.2000.4014

Maisuradze, G. G., Liwo, A., & Scheraga, H. A. (2010). Relation between Free Energy Landscapes of Proteins and Dynamics. *Journal of Chemical Theory and Computation, 6*, 583–595. https://doi.org/10.1021/ct9005745

Marinari, E., & Parisi, G. (1992). Simulated tempering: A New Monte Carlo Scheme. *EPL, 19*, 451–458. https://doi.org/10.1209/0295-5075/19/6/002

Markwick, P. R. L., & McCammon, J. A. (2011). Studying functional dynamics in bio-molecules using accelerated molecular dynamics. *Physical Chemistry Chemical Physics, 13*, 20053–20065. https://doi.org/10.1039/c1cp22100k

McCammon, J. A., Gelin, B. R., & Karplus, M. (1977). Dynamics of folded proteins. *Nature, 267*, 585–590. https://doi.org/10.1038/267585a0

McGuffin, L. J., Bryson, K., & Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics, 16*, 404–405. https://doi.org/10.1093/bioinformatics/16.4.404

Merchant, B. A., & Madura, J. D. (2011). A review of coarse-grained molecular dynamics techniques to access extended spatial and temporal scales in biomolecular simulations. *Annual Reports in Computational Chemistry, 7*, 67–87. https://doi.org/10.1016/B978-0-444-53835-2.00003-1

Miao, Y., Feixas, F., Eun, C., & McCammon, J. A. (2015). Accelerated molecular dynamics simulations of protein folding. *Journal of Computational Chemistry, 36*, 1536–1549. https://doi.org/10.1002/jcc.23964

Miao, Y., & McCammon, J. A. (2017). Gaussian accelerated molecular dynamics: Theory, implementation, and applications. *Annual Reports in Computational Chemistry, 13*, 231–278. https://doi.org/10.1016/bs.arcc.2017.06.005

Mills, C. L., Beuning, P. J., & Ondrechen, M. J. (2015). Biochemical functional predictions for protein structures of unknown or uncertain function. *Computational and Structural Biotechnology Journal, 13*, 182–191. https://doi.org/10.1016/j.csbj.2015.02.003

Moult, J., Pedersen, J. T., Judson, R., & Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Genetics, 23*, ii–iv. https://doi.org/10.1002/prot.340230303

National Research Council (US) Committee on Mapping and Sequencing the Human Genome. (1988a). Implications for medicine and science. In *Mapping and sequencing the human genome*. Washington, DC: National Academies Press. https://www.ncbi.nlm.nih.gov/books/NBK218245/. Accessed 13 Feb 2020

National Research Council (US) Committee on Mapping and Sequencing the Human Genome. (1988b). Introduction. In *Mapping and sequencing the human genome*. Washington, DC: National Academies Press. https://www.ncbi.nlm.nih.gov/books/NBK218247/. Accessed 13 Feb 2020

Nayak, C., Chandra, I., & Singh, S. K. (2019). An in silico pharmacological approach toward the discovery of potent inhibitors to combat drug resistance HIV-1 protease variants. *Journal of Cellular Biochemistry, 120*, 9063–9081. https://doi.org/10.1002/jcb.28181

Nayeem, A., Sitkoff, D., & Krystek, J. S. (2006). A comparative study of available software for high-accuracy homology modeling: From sequence alignments to structural models. *Protein Science, 15*, 808–824. https://doi.org/10.1110/ps.051892906

Onuchic, J. N., Luthey-Schulten, Z., & Wolynes, P. G. (1997). Theory of protein folding: The energy landscape perspective. *Annual Review of Physical Chemistry, 48*, 545–600. https://doi.org/10.1146/annurev.physchem.48.1.545

Panwar, U., & Singh, S. K. (2018). Structure-based virtual screening toward the discovery of novel inhibitors for impeding the protein-protein interaction between HIV-1 integrase and human lens epithelium-derived growth factor (LEDGF/p75). *Journal of Biomolecular Structure and Dynamics, 36*, 3199–3217. https://doi.org/10.1080/07391102.2017.1384400

Panwar, U., Chandra, I., Selvaraj, C., & Singh, S. K. (2019). Current computational approaches for the development of Anti-HIV inhibitors: An overview. *Current Pharmaceutical Design, 25*, 3390–3405. https://doi.org/10.2174/1381612825666190911160244

Patidar, K., Panwar, U., Vuree, S., Sweta, J., Sandhu, M. K., Nayarisseri, A., et al. (2019). An in silico approach to identify high affinity small molecule targeting m-TOR inhibitors for the clinical treatment of breast cancer. *Asian Pacific Journal of Cancer Prevention, 20*, 1229–1241. https://doi.org/10.31557/APJCP.2019.20.4.1229

Patodia, S., Bagaria, A., & Chopra, D. (2014). Molecular dynamics simulation of proteins: A brief overview. *Journal of Physical Chemistry & Biophysics, 4*, 166. https://doi.org/10.4172/2161-0398.1000166

Pikkemaat, M. G., Linssen, A. B., Berendsen, H. J., & Janssen, D. B. (2002). Molecular dynamics simulations as a tool for improving protein stability. *Protein Engineering, 15*, 185–192. https://doi.org/10.1093/protein/15.3.18

Prabhu, S. V., & Singh, S. K. (2019). Energetically optimized pharmacophore modeling to identify dual negative allosteric modulators against group I mGluRs in neurodegenerative diseases. *Journal of Biomolecular Structure & Dynamics*, 1–12. https://doi.org/10.1080/07391102.2019.1640794

Qi, R., Wei, G., Ma, B., & Nussinov, R. (2018). Replica exchange molecular dynamics: A practical application protocol with solutions to common problems and a peptide aggregation and self-assembly example. *Methods in Molecular Biology, 1777*, 101–119. https://doi.org/10.1007/978-1-4939-7811-3_5

Reddy, K. K., & Singh, S. K. (2014). Combined ligand and structure-based approaches on HIV-1 integrase strand transfer inhibitors. *Chemico-Biological Interactions, 218*, 71–81. https://doi.org/10.1016/j.cbi.2014.04.011

Reddy, K. K., Singh, S. K., Tripathi, S. K., & Selvaraj, C. (2013). Identification of potential HIV-1 integrase strand transfer inhibitors: in silico virtual screening and QM/MM docking studies. *SAR and QSAR in Environmental Research, 24*, 581–595. https://doi.org/10.1080/1062936X.2013.772919

Rost, B., Yachdav, G., & Liu, J. (2004). The predictprotein server. *Nucleic Acids Research, 32*, W321–W326. https://doi.org/10.1093/nar/gkh377

Rudzinski, J. F. (2019). Recent progress towards chemically-specific coarse-grained simulation models with consistent dynamical properties. *Computation, 7*, 42. https://doi.org/10.3390/computation7030042

Salmaso, V., & Moro, S. (2018). Bridging molecular docking to molecular dynamics in exploring ligand-protein recognition process: An overview. *Frontiers in Pharmacology, 9*, 923. https://doi.org/10.3389/fphar.2018.00923

Salsbury Jr., F. R. (2011). Molecular dynamics simulations of protein dynamics and their relevance to drug discovery. *Current Opinion in Pharmacology, 10*, 738–744. https://doi.org/10.1016/j.coph.2010.09.016.Molecular

Saunders, M. G., & Voth, G. A. (2013). Coarse-graining methods for computational biology. *Annual Review of Biophysics, 42*, 73–93. https://doi.org/10.1146/annurev-biophys-083012-130348

Scheraga, H. A., Khalili, M., & Liwo, A. (2007). Protein-folding dynamics: Overview of molecular simulation techniques. *Annual Review of Physical Chemistry, 58*, 57–83. https://doi.org/10.1146/annurev.physchem.58.032806.104614

Schlitter, J., Engels, M., & Krüger, P. (1994). Targeted molecular dynamics: A new approach for searching pathways of conformational transitions. *Journal of Molecular Graphics, 12*, 84–89. https://doi.org/10.1016/0263-7855(94)80072-3

Schneider-Poetsch, T., & Yoshida, M. (2018). Along the central dogma—Controlling gene expression with small molecules. *Annual Review of Biochemistry, 87*, 391–420. https://doi.org/10.1146/annurev-biochem-060614-033923

Schuler, B., & Hofmann, H. (2013). Single-molecule spectroscopy of protein folding dynamics—Expanding scope and timescales. *Current Opinion in Structural Biology, 23*, 36–47. https://doi.org/10.1016/j.sbi.2012.10.008

Schwede, T. (2013). Protein modeling: What happened to the "protein structure gap"? *Structure, 21*, 1531–1540. https://doi.org/10.1016/j.str.2013.08.007

Selvaraj, C., Krishnasamy, G., Jagtap, S. S., Patel, S. K., Dhiman, S. S., & Kim, et al. (2016). Structural insights into the binding mode of d-sorbitol with sorbitol dehydrogenase using QM-polarized ligand docking and molecular dynamics simulations. *Biochemical Engineering Journal, 114*, 244–256. https://doi.org/10.1016/j.bej.2016.07.008

Selvaraj, C., & Singh, S. K. (2014). Validation of potential inhibitors for SrtA against Bacillus anthracis by combined approach of ligand-based and molecular dynamics simulation. *Journal of*

Biomolecular Structure & Dynamics, 32, 1333–1349. https://doi.org/10.1080/07391102.2013.818577

Selvaraj, C., Sivakamavalli, J., Vaseeharan, B., Singh, P., & Singh, S. K. (2014). Structural elucidation of SrtA enzyme in Enterococcus faecalis: an emphasis on screening of potential inhibitors against the biofilm formation. *Molecular BioSystems, 10*, 1775–1789. https://doi.org/10.1039/C3MB70613C

Sen, T. Z., Jernigan, R. L., Garnier, J., & Kloczkowski, A. (2005). GOR V server for protein secondary structure prediction. *Bioinformatics, 21*, 2787–2788. https://doi.org/10.1093/bioinformatics/bti408

Shafreen, R. M. B., Selvaraj, C., Singh, S. K., & Pandian, S. K. (2013). Exploration of fluoroquinolone resistance in Streptococcus pyogenes: comparative structure analysis of wild-type and mutant DNA gyrase. *Journal of Molecular Recognition, 26*, 276–285. https://doi.org/10.1002/jmr.2270

Shaw, D. E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R. O., Eastwood, M. P., et al. (2010). Atomic-level characterization of the structural dynamics of proteins. *Science, 330*, 341–346. https://doi.org/10.1126/science.1187409

Shea, J. E., & Brooks 3rd, C. L. (2001). From folding theories to folding proteins: A review and assessment of simulation studies of protein folding and unfolding. *Annual Review of Physical Chemistry, 52*, 499–535. https://doi.org/10.1146/annurev.physchem.52.1.499

Shukla, R., Shukla, H., & Tripathi, T. (2019). Structural and energetic understanding of novel natural inhibitors of Mycobacterium tuberculosis malate synthase. *Journal of Cellular Biochemistry, 120*, 2469–2482. https://doi.org/10.1002/jcb.27538

Sirur, A., De Sancho, D., & Best, R. B. (2016). Markov state models of protein misfolding. *The Journal of Chemical Physics, 144*, 075101. https://doi.org/10.1063/1.4941579

Sugita, Y., & Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters, 314*, 141–151. https://doi.org/10.1016/S0009-2614(99)01123-9

Suryanarayanan, V., & Singh, S. K. (2015). Assessment of dual inhibition property of newly discovered inhibitors against PCAF and GCN5 through in silico screening, molecular dynamics simulation and DFT approach. *Journal of Receptors and Signal Transduction, 35*, 370–380. https://doi.org/10.3109/10799893.2014.956756

Swenson, D. W. H., Prinz, J. H., Noe, F., Chodera, J. D., & Bolhuis, P. G. (2019). OpenPathSampling: A python framework for path sampling simulations. 1. Basics. *Journal of Chemical Theory and Computation, 15*, 813–836. https://doi.org/10.1021/acs.jctc.8b00626

Swope, W. C., Pitera, J. W., & Suits, F. (2004). Describing protein folding kinetics by molecular dynamics simulations. 1. Theory. *The Journal of Physical Chemistry. B, 108*, 6571–6581. https://doi.org/10.1021/jp037421y

Szilágyi, A., Kardos, J., Osváth, S., Barna, L., & Zavodszky, P. (2007). Protein folding. In *Handbook of neurochemistry and molecular neurobiology: Neural protein metabolism and function* (pp. 303–343). Springer. https://doi.org/10.1007/978-0-387-30379-6_10

Torrie, G. M., & Valleau, J. P. (1977). Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics, 23*, 187–199. https://doi.org/10.1016/0021-9991(77)90121-8

Torrisi, M., Kaleel, M., & Pollastri, G. (2018). Porter 5: Fast, state-of-the-art ab initio prediction of protein secondary structure in 3 and 8 classes. *bioRxiv*, 289033. https://doi.org/10.1101/289033

Tripathi, S. K., Selvaraj, C., Singh, S. K., & Reddy, K. K. (2012). Molecular docking, QPLD, and ADME prediction studies on HIV-1 integrase leads. *Medicinal Chemistry Research, 21*, 4239–4251. https://doi.org/10.1007/s00044-011-9940-6

Tripathi, S. K., & Singh, S. K. (2014). Insights into the structural basis of 3, 5-diaminoindazoles as CDK2 inhibitors: Prediction of binding modes and potency by QM–MM interaction, MESP and MD simulation. *Molecular BioSystems, 10*, 2189–2201. https://doi.org/10.1039/c4mb00077c

Vettoretti, G., Moroni, E., Sattin, S., Tao, J., Agard, D. A., Bernardi, A., et al. (2016). Molecular dynamics simulations reveal the mechanisms of allosteric activation of Hsp90 by designed ligands. *Scientific Reports, 6*, 23830. https://doi.org/10.1038/srep23830

Vijayalakshmi, P., Selvaraj, C., Singh, S. K., Nisha, J., Saipriya, K., & Daisy, P. (2013). Exploration of the binding of DNA binding ligands to Staphylococcal DNA through QM/MM docking and molecular dynamics simulation. *Journal of Biomolecular Structure and Dynamics, 31*, 561–571. https://doi.org/10.1080/07391102.2012.706080

Vlachakis, D., Armaos, A., & Kossida, S. (2017). Advanced protein alignments based on sequence, structure and hydropathy profiles; the paradigm of the viral polymerase enzyme. *Mathematics in Computer Science, 11*, 197–208. https://doi.org/10.1007/s11786-016-0287-8

Voelz, V. A., Bowman, G. R., Beauchamp, K., & Pande, V. S. (2010). Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39). *Journal of the American Chemical Society, 132*, 1526–1528. https://doi.org/10.1021/ja9090353

Wang, S., Peng, J., Ma, J., & Xu, J. (2016). Protein secondary structure prediction using deep convolutional neural fields. *Scientific Reports, 6*, 1–11. https://doi.org/10.1038/srep18962

Weber, J. K., & Pande, V. S. (2011). Characterization and rapid sampling of protein folding Markov state model topologies. *Journal of Chemical Theory and Computation, 7*, 3405–3411. https://doi.org/10.1021/ct2004484

Wolynes, P. G. (2005). Energy landscapes and solved protein–folding problems. *Philosophical Transactions of the Royal Society A - Mathematical Physical and Engineering Sciences, 363*, 453–467. https://doi.org/10.1098/rsta.2004.1502

Xu, M., Zhu, T., & Zhang, J. Z. H. (2018). A force balanced fragmentation method for ab initio molecular dynamic simulation of protein. *Frontiers in Chemistry, 6*, 189. https://doi.org/10.3389/fchem.2018.00189

Xu, Y., Cheng, S., Sussman, J. L., Silman, I., & Jiang, H. (2017). Computational studies on acetylcholinesterases. *Molecules, 22*, 1324. https://doi.org/10.3390/molecules22081324

Yon, J. M. (2001). Protein folding: A perspective for biology, medicine and biotechnology. *Brazilian Journal of Medical and Biological Research, 34*, 419–435. https://doi.org/10.1590/s0100-879x2001000400001

You, W., Tang, Z., & Chang, C. A. (2017). Evaluating the accuracy of the umbrella sampling plots with different dissociation paths, conformational changes, and structure preparation. *bioRxiv*, 169532. https://doi.org/10.1101/169532

Zhang, B., Li, J., & Lü, Q. (2018). Prediction of 8-state protein secondary structures by a novel deep learning architecture. *BMC Bioinformatics, 19*, 293. https://doi.org/10.1186/s12859-018-2280-5

Zhang, D., & Lazim, R. (2017). Application of conventional molecular dynamics simulation in evaluating the stability of apomyoglobin in urea solution. *Scientific Reports, 7*, 44651. https://doi.org/10.1038/srep44651

Zhang, Y. (2008). Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology, 18*, 342–348. https://doi.org/10.1016/j.sbi.2008.02.004

Zwanzig, R., Szabo, A., & Bagchi, B. (1992). Levinthal's paradox. *Proceedings of the National Academy of Sciences of the United States of America, 89*, 20–22. https://doi.org/10.1073/pnas.89.1.20

Chapter 8
Magnitude and Advancements of CADD in Identifying Therapeutic Intervention against Flaviviruses

Murali Aarthy, Umesh Panwar, and Sanjeev Kumar Singh

Abstract Flaviviridae is an enveloped viruses composed of positive sense single stranded RNA genome. Flaviviruses causes a major outbreak around the globe through the allegation of life threatening diseases with strident risk to death. Every year, people around the globe reports for the victim of various diseases like Dengue fever, Yellow fever, Encephalitis, Microcephaly and other neurological implications with the dominance of flaviviruses. Flaviviruses are composed of various therapeutic targets which is the important source of human diseases. The diseases caused by these viruses are referred to as the life threatening disease since, no drug is available or reported till date. Hence, there develops an emergency need for the identification of small molecules or vaccines to treat various diseases caused by flaviviruses. During the course of time, the computer aided drug designing strategies plays a major role in the identification of small molecules in short period of time which in turn develops the cost-effective techniques to scrutinize the compounds from the large set of databases or groups. This chapter highlights the importance of structure based design and other *in silico* strategies implemented in the identification of small molecule inhibitors for the intervention of therapeutic targets of flaviviruses in concern with the various infections to the human health.

8.1 Introduction

Flaviviruses constitute a major human health concern over the world causing several diseases with countless ailments leading to lifelong impairment and even death (Therkelsen et al. 2018). These viruses belongs to the family flaviviridae which is the positive stranded RNA viruses and this family is divided into three different generas namely the flaviviruses, pestiviruses and hepatitisvirus. Further, the genera flavivirus is sub-divided into tick-borne viruses, mosquito-borne viruses and

M. Aarthy · U. Panwar · S. K. Singh (✉)
Computer Aided Drug Design and Molecular Modelling Lab, Department of Bioinformatics, Alagappa University, Karaikudi, Tamil Nadu, India

S. K. Singh (ed.), *Innovations and Implementations of Computer Aided Drug Discovery Strategies in Rational Drug Design*,
https://doi.org/10.1007/978-981-15-8936-2_8

arthropod borne viruses (Hasan et al. 2018). The word Flaviviridae has been derived from the Latin word Flavus which means yellow emerged due to the jaundice induced by Yellow fever virusThe viruses belonging to this family is lipid enveloped and icosahedral that infects the mammals (Lindenbach and Rice 2003; Huang et al. 2014).

Among the three different genera of flavivirus, the tick-borne viruses are monophyletic group consisting of single "serocomplex" despite distinct differences in the disease caused by the respective viruses whereas the mosquito borne is diverse from the tick-borne comprising of the viruses namely the Zika Virus (ZIKV), Dengue Virus (DENV), Japanese Encephalitis Virus (JEV), West Nile Virus (WNV) and Yellow Fever Virus (YFV) among many others (Pettersson and Fiz-Palacios 2014). These viruses has transmitted the viral content to human beings through mosquitoes where the humans become the hosts. The mosquitoes that causes transmission and the main source for the infection is the *Aedes aegypti* and the *Aedes Albopictus* in the case of yellow fever, dengue and Zika. Whereas in the case of West nile virus, the *culex pipiens* mosquitoes plays a major role (Gould and Solomon 2008; Mazeaud et al. 2018). These viruses cause diverse clinical indications and impediments such as hemorrhagic fever with plasma leakage and encephalitis leading to death. Generally, the flaviviruses are zoonotic whereas survival and replication depends mostly on the non-human animal vectors except for the Dengue virus that transmits in human (Petersen and Marfin 2005).

The lifecycle of flavivirus is completely dependent on the cytoplasmic fate of genomic viral RNA whose replication occurs in the cytoplasm where no DNA generates immediately (Mazeaud et al. 2018). It includes the attachment of the virion to the host cell surface followed by internalization and further the transfer of the viral RNA genome into the cytoplasm where the translation of the viral protein and replication of the genomic RNA and maturation happened resulting in the release of the progeny viruses from the cell. The genome of the flavivirus is packed with the multiple copies of protein containing capsid with the host-derived lipid bilayer surrounded by 180 copies of both the envelope and membrane glycoproteins (Kaufmann and Rossmann 2011).

The flavivirus contains the single open reading frame that is translated into polyprotein precursor which is subsequently glycosylated by cellular glycotransferases and cleaved by the combination of the viral and host proteases to release the three structural and seven non-strucutral proteins (Medigeshi 2011). The structural proteins constitute the virion, core or capsid and the envelope protein whereas the non-strucural proteins forms the replicase complex catalyzing the RNA accumulation with modified cytoplasmic membranes (Murray et al. 2008). The Non-structural proteins are NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5. The NS1 possess its important role in the viral replication whereas NS2A, NS4A and NS4B represents its role in the RNA accumulation. The NS3 protein is the viral protease which also possess helicase and nucleoside triphosphatase activities required for the replication. Finally, the NS5 proteins constitute the viral RNA-dependent RNA polymerases (Westaway et al. 1997; Murray et al. 2008). The structural and non-structural proteins are major source for the development of various diseases and causing the global threat through flaviviruses. Vaccines has been developed for the Yellow fever virus, Japanese Encephalitis virus and Tick borne Encephalitis among the other flaviviruses. But vaccines of other viruses has not been identified which possess continuous challenges. Hence, in the nonexistence of the vaccines, the need for specific drugs increases but till date no small molecule inhibitors or the peptidic inhibitors are identified. Here comes, the role of computer aided drug design to develop specific inhibitors in a short period of time keeping in the mind the challenges and drawbacks of the experimental evidences. *In silico* strategies helps in identifying or discovering small molecule inhibitors from the large sets of databases which can be proceeded further for the validation of the activity to inhibit the virus through experimental evidence. Hence, it was decided to depict the importance of computer aided drug designing aspects in identifying the small molecule inhibitors that could inhibit the viral mechanism theoretically. The theoretical studies identified and reported behaves as a major base to proceed further for the *in vitro* and *in vivo* studies. This chapter highlights the molecular modelling strategies and other computer aided drug design methods implemented for the identification of the small molecule inhibitors for the various proteins of flaviviruses in concern with West Nile Virus, Yellow Fever Virus, Dengue Virus and Zika virus.

8.2 Geographical Distribution of Flaviviruses

8.2.1 Dengue

Dengue virus spreads through the mosquito species Aedes *aegypti* and *Aedes albopictus that* transmits the dengue fever throughout the tropics. Around 1779–1780, the dengue like diseases has been recordedin Asia, Africa, Pasific, Americas and the caribbean (Daep et al. 2014). The Dengue hemorrhagic fever was initially identified in 1950s around Thailand and Philippines during the epidemic of dengue. The Dengue fever was first recorded in the year 1992 in the Medical Encyclopedia of Chinese and bas been registered as the pandemic over the world after the break out of world war II in Southeast Asia (Gubler and Clark 1995; Alshammari et al. 2018).

The etiology of the virus and the mosquito transmission were recognized in twentieth century and millions of the population around the world are at potential risk areas of dengue transmission. It was reported that the diseases has been registered as the endemic over 100 countries statin the prevalence of the fever followed by the illness. Dengue virus can cause dengue fever, to dengue hemorrhagic fever or dengue shock syndrome (Daep et al. 2014). The infections caused by DENV are self-limiting and ranges from asymptomatic to relatively mild with homogenousailmentultimatelyleading to complete recovery. An average of 5& of cases having symptomatic illness develops to severe ailment whose indications were fever, acute abdominal pain, vomiting and myalgia. These symptoms further develops to hypertension, decreased peripheral perfusion, myocarditis and tachycardia. Additionally, post infection 7 days, a hemorrhage develops (Malavige et al. 2004).

Researchers identified the virus with various serotypes namely DEN-1, DEN-2, DEN-3, and DEN-4) stating that they are closely related but possess variations in genetic information. These serotypes developed from the same ancestor but progressed separately with diverse antigenicity and degrees of illnesses ultimately leading to the dengue hemorrhagic fever (Kawaguchi et al. 2003). The epidemiology of the dengue virus is unveiled with the co-circulation of the different serotypes at impulsive prevalence at different times. Researchers throughout the world tried developing epidemiological and mathematical models to predict the epidemics of DENV but the transmission mode of the DENV made it difficult for the model prediction (Racloz et al. 2012). The difficulty in vaccine prediction arises due to the randommode of DENV epidemics and transmission.

Millions of people are affected by the dengue infection based on the different serotypes causing dengue hemorrhagic fever or dengue shock syndrome. In general, the infections caused naturally promotes strong neutralizing antibodies and develops immunity that lasts long for the serotype that is homologous. The vaccine development for the dengue infections is complicated due to the presence of distinct serotypes (de Silva and Harris 2018). The approach of developing the LAVs holds to be promising for the dengue infection due to the probability of monovalent vaccine that induces cross-relative immunity enhancing the replication and diseases developed by the serotypes of DENV. The vaccines that has been identified till date are based on the tetravalent formulations to induce simultaneous protective responses to the available four different types of serotypes (Murphy and Whitehead 2011).

The flaviviruses when gets transmitted through the bites of mosquito into the human body interacts with the dendritic cells present in the dermis and epidermis layers of the skin. When the DENV enters into the human body especially to these cells, the circulating dendritic cells gets decreased and unresponsive altering the host-immune response. It is clearly understood that the flavivirus use the dendritic cells as the gateways to successfully infect their human host (Hober et al. 1996). Followed by the dendritic cells, monucytes are the natural hosts of DENV and are implicated in the pathogenesis of dengue fever and dengue haemorrhagic fever. The infection caused by the dengue virus quickens the distinction of monocytes into macrophages and expedites cellular movement into central nervous system where the chemokines, inflammatory cytokines and viral proteins are produced. The infected CD 16+ monocytes in the human hostproduce cytokines and chemokines like TNF alpha, IL-1beta and CCL3 which is strongly involved in the blood brain barrier integrity loss and Developing central nervous sytem disorder (Halstead 1988; Azeredo et al. 2010).

8.2.2 ZIKA

The identification of zika virus is observed first in the forests of Uganda in the year 1947 from an infected sentinel rhesus monkey. Later, the infection has been transmitted to the humans through the bites of several mosquito species like *Aedes Aegypti, Aedes africanus, Aedes albopictus and Aedes hensilli* causing mild and febrile syndrome which is clinically inapparent (Dick and Haddow 1952; Gorshkov et al. 2019). Later, the infections has also been observed in 2007 around Yap island and a huge outbreak of ZIKV has been reported in French Polynesia resulting in more than 25,000 cases during 2013. Most of the patients affected with ZIKV infection exhibits mild symptoms whereas Guillain-Barre Syndrome has been recorded in one patient after a week the infection has been observed.

Guillain-Barre Syndrome is an auto-immune disease that affects the peripheral nervous system which leads to weekness in mild illness and paralysis resulting in death in the case of severe illness (Musso et al. 2014). These infections exploded and it is related with the microcephaly and other inherited abnormalities in newborns and fetuses of the pregnant women infected with the ZIKV (Olagnier et al. 2016). The other modes of transmission reported for the ZIKV infection is the Sexual transmission. It was observed with the case which has been reported in 2008 in the person who has returned home from Senegal with hematospermia and other symptoms of ZIKV infection. Lately, wife of the traveler who is actively involved in intercourse with the traveler and has not travelled internationally also developed symptoms. But in 2016, americas has experienced with the active transmission of the infection from the travelers who has traveled to North America and Europe confirms the sexual mode of transmission (Gregory et al. 2017).

The ZIKV epidemiology has been registered with the neurological complications in newborns and adults around Caribbean and Central/South Americas during 2014 in which 10% of the cases has been recorded for fetal death with irresistible impact on the societies of infected personalities. In the year 2016, the epidemic of ZIKV has reached peak with the reports of large number of cases in United States and however, in 2017 and 2018, the decline of ZIKV infection has been documented (Khaiboullina et al. 2019). Researchers has stated that the keratinocytes, fibroblasts and dendritic cells that is present in the human skin with immature dendritic cells is mostly lenient and responsible for the ZIKV isolate and the epidemic in French Polynesia. The mosquitoes that transmits the infection to the host deposits the virus in dermis and epidermis during the blood meal. The infection present in the skin resulted in existence of high RNA copy and intensification in the production of ZIKV elements representing the viral duplication in diseased cells (Hamel et al. 2015). Further, it is understood rom the reports that ZIKV forms single serotype with different ZIKV lineages like African and Asian. In 2016, Wang has completed the analysis of the genetic and phylogenetic characteristics along with the structural modelling on the complete open reading frames of ZIKV. The reports stated clearly that the human outbreak evolved from the Asian lineage and all the strains identified were related to the strains observed from the French Polynesia in 2013 (Wang et al. 2016). Till date, there are no specific antivirals or the vaccines that has been out for the clinical use. Development of the prophylactic immunization is the greatest approach to fight the disease. Numerous vaccines has entered the clinical trials but not released and approved. The preventive measures in the control and eradication of the infection is the control of mosquito vectors, prevention against mosquito bites and the safety measures during the intercourse and pregnancy (Garg et al. 2018). It is reported that the vaccines tested for other flaviviruses like Japanese encephalitis and yellow fever cannot be used for Zika even though it falls under the same category of flaviviruses because of the contraindications observed in pregnant womens (Kudchodkar et al. 2018).

8.2.3 West Nile

In the year 1937, West Nile virus has been first isolated in the districts of Northern Ugands from patient who is febrile. During the early 1950s and 1960s, the ecology and epidemiology of the West Nile Virus has been distinguished during numerous outbreaks in the Mediterranean. Periodic outbreaks and cases of west nile virus infection has been reported from various countries like Europe, Africa, Romania and India. These viruses are mostly distributed in the regions of Africa, South Asia and the Middle East which has stated endemic conditions (Johnston and Conly 2000; Sejvar 2003). Till 1999, the spread of west nile virus has not been reported in Americas but a serotype that is closely related to west nile namely the Kunjin is found in Australia and Southeast Asia (38). Based on the variations in antigen observed in Envelope protein and the presence of N-glycosylation site different subtype has been distinguished (Jia et al. 1999).

West Nile virus has been proposed to obtain nine lineages whereas mostly the outbreaks of human have been attributed to lineage 1 and 2. These lineages are globally spread and the clades are distinct which comprises of strains isolated from various parts of the world like Europe, Americas and Africa with the implication of neurological complications. The different clades are available represented as Kunjin virus restricted to oceania but the neurological disease is not reported often. The lineages 2 was exclusively reported only in 2004 which was isolated from humans and bird in hungary, Greece and Italy. It is also considered to be less pathogenic than other lineages until it cause severe diseases in south Africa and causing encephalitis among birds (Fall et al. 2017). In Czech Republic, the Rabensburg virs has been isolated which is referred to as the Lineage 3 whereas the Lineage 4 is reported in Russia. In India also, the lineages has also been observed and is similar to the lineage 1 with different clade. In spain, the sixth lineage has been observed whereas the Koutango virs represented as the lIneage 7 which is a distinct lineage of west nile virus. These strains were observed from the ticks, rodents (Lvov et al. 2004; Fall et al. 2014).

The west nile virus enters to the human host through receptor-mediated endocytosis after the attachment of the virus to the cell surface. Various molecules present in huans are signified to be receptors like DC-sign, mannose receptor and several glycosaminoglycans (Colpitts et al. 2012). Maturation of the virus happens with the endosmoes present during the internalization from the cell surface whereas the pH drops during early endosome from neutral to slight acidic and at the late endosome become more acidic (Modis et al. 2004). After the capsid dissociation, the RNA genome replicates and initiation of virus assembly is progresses. The polyproteins of the virus gets translated and gets processed on the intracellular membranes which results in expressing viral proteins. The viral RNA gets replicated by the cellular proteins into multiple copies which can be used in the production of new virions. In the endoplasmic reticulum, the structural proteins gather in the membrane that gets connected with the nucleocapsid and flourish into cytoplasm through the Golgi network. The surface of the cell present in an exocytic vesicle is the place where the virus travels and gets matured since, the enzymes cleave the precursor membrane that results in the entry of the matured virus from the surface of cell (Rice 1996; Modis et al. 2004).

From the emergence of the west nile virus spread over the world in 1990, the significant investigation on the development of vaccine has been procured. But till date, no vaccines has been identified for the effective use of humans. Various technological platforms has been developed along with the use of various licensed vaccines from the other flaviviruses like attenuated and chemically inactivated strains. It was clearly understood from the studies that, the protection from the west nile virus can be achieved only with the large variety of immunization techniques and the other protective immune response are the antibodies against the E-protein. The vaccines developed based on the envelope protein of the west nile virus are protective against the genetic lineages 1 and 2. The reasons for the lack of the human vaccine include various challenges like safety concerns, problems in the difficulty of designing clinical aspects or may be the economic considerations. The other possibility than developing vaccines is the development of immune responses to west nile virus throught he support of different immunization technologies (Botha et al. 2008; Ulbert 2019).

8.2.4 Yellow Fever Virus

Yellow fever virus, one of the major human threat causing virus is identified at the late nineteenth century and to be specific on December, 2016. Yellow fever virus first originated from the Africas then in western hemisphere record has shown the epidemic of the diseases in the Yucatan around 1648. In the late 1800s, a Cuban scientist named Carlos Finlay determined that mosquitoes is the main source of the disease (Gardner and Ryman 2010). This virus causes considerable mortality and morbidity with the density of numerous hosts which is fecilitated by the transmission through the mosquito vector speices Aedes Aegypti (Sall et al. 2010). Yellow fever virus is first isolated in 1927 from the Ghanaian patient and the strain identified from the patient is widely used still. The yellow fever virus impacted the economy of the Americas more than that of Africas (Chippaux and Chippaux 2018).

Africa has taken several safety measures with the immunization in emergency need and persuaded the human threat and severity of the epidemics but still larger populations were remained at risk in more than 30 countries. The massive outbreak has registered in Ethiopia during 1959–1962 and followed by the endemic in Kenya which covers mostly the West Africa (Lepiniec et al. 1994). The yellow fever epidemics has discouraged the Napolean Bonaparte from accomplishing the defeat of United States of America (Marr and Cathey 2013). It was also observed that the construction of the Panama Canal has also been delayed because of the epidemic of Yellow fever (Chippaux and Chippaux 2018).

The presence and epidemic condition of the yellow fever virus made it a global health threat whereas the people travelling from Europe and North America to various parts of the world has the increased percentage of spreading the diseases through their expose to the virus (Johansson et al. 2010). Understanding the transmission mode of the yellow fever infections is critical which is very essential for the development of vaccine. Virus transmission always require a competent mosquito feed on an infected human and survives with an incubation period where the replication of virus is observed and disseminates to the salivary glands and gets transmitted to other hosts (Hindle 1930). Understanding the incubation period of yellow fever is the important strategy for the detection, prevention and control of the outbreaks. The intrinsic incubation period is the important factor in the differential analysis of illness in people around the endemic areas as well as the travelers form the endemic areas. Researchers stated that analysis of the outbreak with the mathematical model provides a quantitative status for making public health decisions which is aimed to control the disease spread (Johansson et al. 2010).

In 1930s, group of researchers have developed the live-attenuated vaccine strain designed with the 17D attenuated for the viscerotropic disease in mammals but it remained immunogenic. The vaccine used today clinically was derived from the 17D strain and the researcher Max Theiler has been awarded the Nobel Prize for life saving research around 1951 (Chippaux and Chippaux 2018). At present there are several substrains of 17D is available for production which has been manufactured in several countries like Brazil, Russia, France, Senegal, China and USA. Millions of people are receiving vaccine and routine immunization strategy has been followed for the prevention of these infections. Scientific and Health Community conducts the EYE (Eliminating Yellow Fever Epidemics) plan which is successful in the control of infection (Collins and Barrett 2017). Even though vaccines has been in clinical use for the eradication of yellow fever but till date, there are no antivirals available to treat the infections caused by the Yellow fever.

8.2.5 Role of Structural and Non-Structural Proteins of Flaviviruses

Structural Proteins

Flaviviruses are positive sense RNA enveloped viruses which contains proteins which is associated with membrane that has organized in the icosahedral matrices at the surface of the virion differentiating the immature and mature representation of the virion (Blazevic et al. 2016). The flaviviruses encompasses of nucleocapsid which is poised of capsid protein which is surrounded by the lipid bilayer derived from the host membranes. Further the envelope of flaviviruses is sprinkled with the membrane and envelope glycoprotein. The virus is observed with the diameter of 40 nm with the projections in surface of 5–10 nm (Gardner and Ryman 2010). Glycoprotein is the major component of the surface in the virion and possess most of the biological activity that includes assembly of the virion, cell-surface receptor binding and the activity of the fusion at pH along with immunogenicity (Chambers et al. 1990).

The virions with immature forms contains trimers of the precursor membrane – envelope heterodimers and gets assembled by the budding into the endoplasmic reticulum. The maturation of the virus happens during the transis of the immature particles which is triggered by the pH change of trans-Golgi network (TGN) (Yu et al. 2008; Blazevic et al. 2016). The structural changes of rearrangement of precursor membrane and envelope protein to the herringbone like lattice are the characteristics of mature virions which results in the exposure of cyptic cleavage of protease site in precursor Membrane protein that is further cleaved by the golgi network furin protease to obtain precursor and membrane protein. The fragment from the precursor protein attaches with the enveloep protein at the surface of the virus which contains acidic pH and later falls down when the other cells happens to attain neutral pH upon the proclamation from the cell by exocytosis (Li et al. 2008).

Envelope protein shows its stability in mature conformation and also the functions of virus entry that attaches to the cell gets mediated by the envelope protein. These proteins possess anchor that clings to double membrane which helps in the processing of polyprotein that occurs at the endoplasmic reticulum (Blazevic et al. 2016). The assembly of the virus is a very vital stage in the infection cycle of the virus but the mechanism is not understood clearly. The capsid protein which is the important protein that gets translated plays an important role when it interacts with the genome of viral RNA that packs during the assembly process (Tan et al. 2020). Preceding to encapsidation and genome replication, the dimers of capsid are stored on the lipid droplets. The ability of the capsid protein is to interact with the droplets which is very much essential for the production of viral particles. These proteins possess the ability to enter into nucleus which causes problems on the biogenesis of ribosome and transcriptome of the host (Sotcheff and Routh 2020).

Non Structural Proteins

The Non Structural protein (NS1) is a 48 kDa glycoprotein is not present in the particles of the virus. But it is an essential gene present within the infected cells and functions as the cofactor for replication of the RNA and gets colocalized with the replication of double-stranded RNA. This NS1 gets synthesized within the infected cell as the soluble monomer and gets dimerized after the post translational modifications in the endoplasmic reticulum with consequent transportation to the cell surface and release into the extracellular (Avirutnan et al. 2007). The characteristics of the NS1 is a hexamer and accumulates within the serum in huge amounts. It also contributes highly to the flavivirus pathogenesis remains unclear that facilitates the formation of immune complex, extracellular matrix proteins, elicit auto-antibodies react with platelet and directly enhance the infection (Avirutnan et al. 2007).

The Non-structural NS2A/B protein is the 22 kDa protein which is hydrophobic in nature in which the N-terminus gets generated in the lumen of endoplasmic reticulum through the host protease whereas the C terminus is generated in the cytoplasm through the viral protease. This structural protein depicts its importance for the replication of the virus and viral pathogenesis and it also functions in the synthesis of viral RNA. Each flavivirus non-structural protein NS2A interacts with the various human targets and possess its main role in the viral assembly (Falgout and Markoff 1995; Xie et al. 2013).

The polyprotein gets cleaved by the proteases of the host cell which includes signalases and furin along with the serine proteases encoded by the virus namely NS2B-NS3 protease. This protease is established to be the valuable target of therapeutic interest which is 375 kDa (Noble et al. 2012). The N-terminal domain of the NS3 protease adopts a chymotrypsin like fold with beta barrels composed of beta strands along with that a catalytic triad is also located at the cleft between two beta barrels. In general, the NS3 protease requires an additional stretch of amino acids from the cystolic domain of NS2B which is important for the catalytic activity which is represented as the two-component protease (Erbel et al. 2006; Gupta et al. 2015).

The Non-structural protein NS4 is initiated by the proteases and the signal peptidases of the host which produces NS4A and NS4B. The cleavage of the non-structural protein 4A/B is essential for the signaling pathway of IFN (68). Also, it indicates that this non strucutral protein is very much essential for the infection of the virus which interacts with the non-strucutral protein 1. It is also capable of inducing the rearrangement or bend of the membrane (Yu et al. 2017; Chen et al. 2017).

The nonstructural protein NS5 is the protein which is largely encoded in the genome of the flavivirus and is conserved among other proteins of flavivirus. The main objective of this protein is to depict its role in replication and itpossess its dominance in drug development. The highlights of the nonstructural 5 protein is that it contains two domains which is structural and functional domains namely the RNA methyltransferaase domain in the N-terminal and the RNA dependent RNA polymerase domain in the C-terminal (Valente and Moraes 2019). The MTase domain of the protein is involved in the formation of cappinig in mRNA of the virus and to be specific it methylates the mRNA cap in different positions which is the necessary component for the replication of the virus. Researchers stated that this protein displays an important function with the RdRp doaim which is treated as an important target for the drug discovery (Dong et al. 2014; Brecher et al. 2015). The RdRp domain develops the synthesis of RNA through *de novo* which is present in the negative polar-sense which is essential for the synthesis of the RNA of positive sense. Along with this NS5 also behaves as the antagonist in the host interferon response (Valente and Moraes 2019).

8.3 Computational Strategies for Drug Discovery

The rapid development of the technology and the techniques have benefited the life of humans in significant manner which helps in managing the time and cost of production. In drug development and production, researchers faces difficulty to find the exact therapeutic intervention for targeting various diseases which in turn targets the protein. The recent trends in the drug development and discovery helps in understanding the mechanism of the diseases in turn understanding the characteristics of the target and identification of lead for treating the disease (Katsila et al. 2016). Significant approach of the pharmaceutical industry in the drug discovery strategies. Bioinformatics plays an important role in the drug discovery which helps reveal the mechanism of genes from the genomic data. This also predicts the possible target proteins for the design and discovery of drugs (Lin et al. 2020). Over three decades in the drug discovery, computer aided drug discovery plays a major role in the pharmaceutical industry through development of therapeutically important small molecules in a short period of time. Structure and Ligand based drug discovery are the methods classified from the computer aided drug design (Sliwoski et al. 2014). Based on the biological problems, bimolecular simulations has been incorporated in the development of models with multiscale. The quantum and molecular mechanics were combined to study and analyze the electronic properties along with the study of chemical reaction & mechanism (Fan et al. 2017).

8.3.1 Bioinformatics Approaches Involved Indrug Design

Structure based drug design is the technique involves the available models of the target either identified or developed by the X-ray diffraction, homology modelling and the nuclear magnetic resonance. This is more efficient calculation method for the early identification of the novel potent drug candidate from large set of molecules to the particular target. This methods helps in saving experimental expenses through the significant reduction of the large number of compounds to smaller size (Panwar

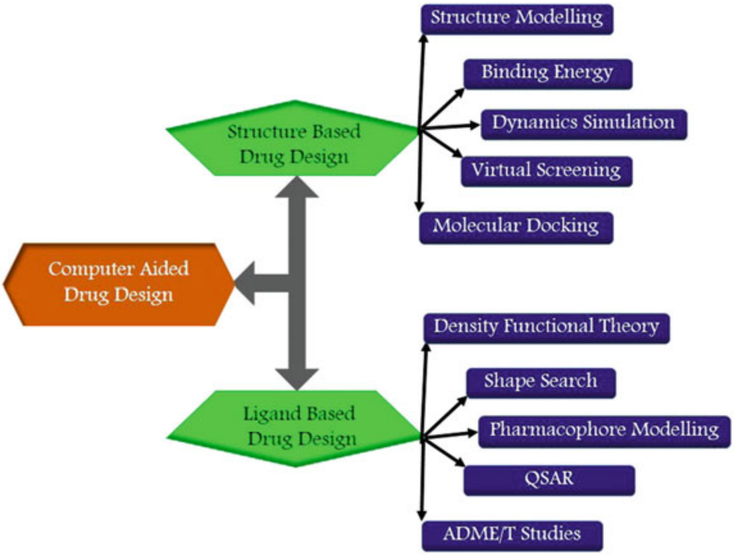


Fig. 8.1 Workflow of the Computer Aided Drug Design

and Singh 2018; Lin et al. 2020). Structure based high throughput is one of the important technique implemented for the identification of potent inhibitors from the pool of large small molecules (Sharda et al. 2017; Sharma et al. 2018). The other technique which possess the same advantage as of the screening strategies is the Molecular docking. This helps in identifying the potent inhibitors that can bind to the receptor helping in understanding the mechanism of inhibition (Tripathi et al. 2012; Pradiba et al. 2018). The Molecular dynamics simulation is used extensively for calculating the stability of three dimensional structure and also helps in accessing the ligand bound conformation (Grover et al. 2012; Shanmuganathan et al. 2018). But Ligand based drug design follows the different criteria where it uses the knowledge and information of the available/reported ligands for the identification or discovery of new ligands for targeting the therapeutic targets. This method involves various methods like quantitative structure activity relationship of the compounds for predicting its binding ability, pharmacophore modelling to analyse the characteristics of the molecule and pharmacophore based screening to screen the molecules from larger datasets based on the characteristics of the available molecules (Singh et al. 2007; Dessalew et al. 2007; Reddy et al. 2014). The types of computer aided drug design and its sub sections are represented in the Fig. 8.1.

Apart from these methods, the *de novo* drug design and fragment based drug discovery methods has appeared to be the next promising approach for the development of optimized ligand which helps in decreasing the erosion rates in drug design. The success in the clinical aspects has been obtained from the fragment based de novo design. Even though drug discovery has achieved success in identification of efficient drugs, but also gets affected by various factors like multiscale optimization methods which should be concentrated beyond the molecular levels (Lin et al. 2020). Recently, the machine learning approach which has been developed into the deep learning approach is the booming era helps in the analysis of the biological properties that has to be extracted from the enormous amount of databases (Varpa et al. 2008).

8.4 Drug Discovery on Flaviviruses

8.4.1 Homology Modelling and Sequence Alignment on Flaviviral Proteins

The NS3 proteins from the Dengue virus (DENV), Japanese Encephalitis virus (JEV), West Nile virus (WNV), Murray Valley Encephalitis virus (MVEV) and the Usutu virus (USV) has been used for the development of Sequence alignment by the research group of Jitendra and Vinay 2011 which revealed that the murray valley encephalitis shared 61%, 86%, 88% and 81% with Dengue, Japanese encephalitis, Usutu and West Nile respectively. The NS3 of MVEV of PDB ID: 2WV9 is the template for other flavivirus used in the study which helps in the development of three dimensional structure (Jitendra and Vinay 2011). The crystal structure of the ZIKV NS5 protein is not available which provoked to develop modelled structure with the help of blast search. The crystal structure of four different flavivirus proteins are used as the template for ZIKV NS5 modelled structure through Modeller suite version 9.1 separating the domains into RDRp and MTase. The structure is further validated with the MolProbity and Ramachandran plot analysis stating that 97.2% residues are in favoured region (Ramharack and Soliman 2018). The dengue possess four different serotypes and the NS1, NS2B/NS3 and the NS5 proteins of the different serotypes has been carried out for the sequence alignments and it has been observed that the numerous hubs of conserved regions are available among the proteins. The similarity between the proteins are represented with the 73%, 79% and 77% respectively among each proteins (ul Qamar et al. 2019).

8.4.2 Structure Based Drug Design for the Identification of Potent Small Molecules

Entry inhibitors categorized as the peptide and small molecules that targets the hydrophobic pocket of the Envelope protein has been identified with the experiments of docking and other algorithms. It has been reported by Nicholson et al. stating that two peptide inhibitors namely 1OAN1 and DN59 could inhibit the envelope protein of dengue virus that could prevents the disease outcome of dengue hemorrhagic fever (Nicholson et al. 2011). Also, it is clearly evident from the reports of Yang et al. representing that the tetracycline derivatives interacts with the envelope protein in the hydrophobic pocket and prevents conformational rearrangements and viral fusion (Yang et al. 2007). Compound NITD448, reported for the inhibition of fusion in envelope protein through its binding to the active pocket which is hydrophobic in nature (Poh et al. 2009).

The envelope protein of the Dengue protein possess the X-ray diffracted structure with the small molecule binding pocket with the bounded n-octyl-beta-D-glucoside. This identification and development of the crystal structure helps in the discovery potential inhibitors better than the available small molecule (Botting and Kuhn 2012). Aarthy et al. utilized the crystal structure for the prediction of novel potent inhibitors based on the ligand binding pocket with the small molecule n-octyl-beta-D-glucoside through the Schrodinger and identified best five compounds (Aarthy and Singh 2018). Recently, it has been reported that the Epigallocatechingallate possses the ability to inhibit the entry of ZIKV into the host cells but the mechanism of action is not demonstrated clearly. Sharma et al. has investigated mechanism of EGCG when it is bounded into the binding site of ZIKV envelope protein which provided better insights about the interactions of the molecule (Sharma et al. 2017).

Non-structural (NS) protein 1 has been reported as the therapeutic target for the drug discovery which is involved mostly in the replication of viral RNA. Raza et al. has obtained the crystallographic structure of the non-structural protein of the zika virus and reported inhibitors from the literature. GOLD and Autodock/Vina are the docking tools which raza has used for the identification of potent inhibitors to treat the zika virus which in turn inhibits the viral replication (Raza et al. 2019). Ul Qamar et al. has obtained a total of 2200 flavonoids from the group of antiviral medicinal plants against the dengue nonstructural 1 protein. The author has utilized the Molecular Operating Environment platform for providing a valuable evidence for the development of drug which would be beneficial through the implementation of the computer aided drug designing (ul Qamar et al. 2014).

The crystal structure of the NS3 reveals that a fold is present with the chemotrypsin consisting of beta barrels at the N-terminal with helicase and RNA tri-phosphatase at the C terminal. The NS3 helicase protein had a significant role in the replication process of the viral genome. Kumar *et al.* has assessed the modulatory effects of the derivatives belonging to bis-coumarin against the NTPase of NS3 helicase of Zika through *in vitro* studies. In order to carry out the *in vitro,* the computational studies through molecular docking has represented as the base. A unique activity is present in the NS3 helicase when compared to the other proteins of flavivirus is that it contains two different binding which is very important for the inhibition of the replication process. One binding site contains the ATP whereas the other binding site is meant the position of RNA (Kumar et al. 2020a, 2020b). The major function of the NS3 helicase is the unwinding of the double stranded RNA during its replication which provokes the researchers to think it as a promising drug target. The NTPase site of the helicase is implemented for the identification of potent inhibitors to treat the infection with Zika from the ZINC library and Helicase focused library. These studies helps in reducing the time taken for proceeding with the *in vitro* studies. The best hit compounds obtained from the study is represented in the figure (Kumar et al. 2019). Also, the author tries to identify the inhibitory action of EGCG in both the NTPase site and the RNA binding site which helps in better insights of the mechanism (Kumar et al. 2020b).

The hydrophobic region of the Non-structural 3 (NS3) protein gets shielded by hydrophilic part of the NS2B which develops the chimera. The functional disturbance of the NS2B/NS3 progresses into the inhibition of the replication and infection of the virus (ul Qamar et al. 2019). The inhibition site of the protein has been revealed and a catalytic triad is present at the amino terminal domain where the binding pocket of the substrate is present within the NS3 protein. Based on the seed molecule present in the diffracted structure was the query structure used for generation of the ligands (Borowski et al. 2002). With the ligands generated, the best conformers of the top 10 ligands were selected with the base of maximum binding affinity. The best hit compounds were utilized for the purpose of binding with the protein through the Autodock Vina. This approach uses the Lamarckian Genetic algorithm method (Jitendra and Vinay 2011). Experimental evidences are reported stating that the flavonoids are potent inhibitors for the NS2B/NS3 protease but the interactions between the inhibitors at the molecular and the atomic levels are not explored. The inhibitors reported were obtained from the literature to examine the interaction through the docking studies with the Schrodinger. The best hits among the flavonoids obtained based on the maximum binding energy shows strong interaction with the catalytic triad of NS2B-NS3 which plays a vivacious role in the inhibition of the viral propagation (Yadav et al. 2020). With the help of the structure based drug design, a potent molecule which has been already reported to use during pregnancy has been identified to be the potent inhibitor inhibiting the viral replication of NS2B-NS3 protease of Zika virus. With the facts observed through *in silico* docking studies, the compound has been tested *in vitro* which reduces the time of selecting the small molecules over the large set of molecules (Kumar et al. 2018).

Methyltransferase domain of the Nonstructural protein 5 is convoluted in the establishment of mRNA capping. These proteins are mostly conserved and amongst the other viral proteins. Coutard et al. investigated the functional and structural properties of the MTase domain of the Zika and stated that the methylation of RNA capping occurs at the seventh position of the N-terminal and the 2’-O position nucleotide which yields a capping structure. The author reported that based on the biochemical and structural analyses, the similarity with the DENV MTase domain (Lim et al. 2015; Coutard et al. 2017). Ramharack et al. proposed two novel compounds for inhibiting MTase and RdRp with promising interactions and physicochemical properties. The strategy of *in silico* studies serve as a beneficial technique to enhance the drug discovery and develops as cost effective resources. The compounds listed from the author has represented in the figure (Ramharack and Soliman 2018). Ul Qamar et al. has reported three inhibitors namely Canthin-6-1-9-O-beta-glucopyraside, Kushenol W and Kushenol K from the phytochemical databses for the inhibiton of NS3 pretoin of DENV along with is NS3, NS1 and NS5 proteins. The author stated that this study will be a platform for the development and optimization of the identified compounds as a potent inhibitor which can be proceded further (ul Qamar et al. 2019).

8.4.3 Evolutionary Studies Based on the Computational Technique

Flaviviruses family comprise of large number of viruses and it is an important strategy to identify the genetic relationship between the viruses of the flavivirus family. Kuno et al. has carried out a phylogenetic analysis to understand the relationship between other viruses and reported that from the ancestor, non-vector and vector borne clusters has been evolved whereas the tickborne and the mosquito borne has emerged from the latter (Kuno et al. 1998). Moureau et al. tried to generate the diversed phylogenetic datas and based on that they have determined the sequence of the genome and phylogeny of the 14 flaviviruses and stated that various species of Culex mosquitoes has been involved in the evolution (Moureau et al. 2015). *In silico* studies on the NS2B/NS3 protease proteins are being developed to decipher the transmission mechanics and evolution of the virus. Various models with networking are proposed to understand the viral inheritance pattern and phylogeny (Fathima and Murugaboopathi 2019).

8.4.4 Ligand Based Studies on the Flaviviral Proteins

The known inhibitors of DENV NS3 helicase from the literature report were used for the development of pharmacophore through the Ligand Scout software. The pattern matching based on the alignment is used to create pharmacophore. The pharmacophore is used to align through the shared feature pharmacophore or merged feature pharmacophore which represents the common interaction features or the extended pharmacophore respectively. Further these pharmacophore model is used for screening potent small molecule from the databases (Halim et al. 2017). The complex of DENV NS2B/NS3 protease with the substrate benzoyl-norleucine (P4)-lysine (P3)-arginine (P2)-arginine (P1)-aldehyde (Bz-Nle-Lys-Arg-Arg-H) is used for the development of compex based pharmacophore. The pharmacophore tool implemented in the MOE was employed for the development of pharmacophore. The test set of 20 known inhibitors of dengue protease was used for the validation of identified pharmacophore which is further carried out for the screening large number of inhibitors from ZINC database (ul Qamar et al. 2016). The density functional theory studies for the best identified compounds for the inhibiton of NS2B/NS3 zika proteins with large scoring in the docking methods performed by Balajee et al. these studies were successfully employed as the structural for the screening large set of databases for the compounds (Balajee et al. 2016).

8.4.5 Free Energy Calculation

The identified small molecules for the inhibition of the zika virus NS1 protein is the initial concept for the *in silico* strategies. But the evaluation of the energy of the binding is the major aspect in which illustrates the strength of inhibition. Hence, the affinity of the ligands before and after the simulation can be predicted with the MM/GBSA and the MM/PBSA methods implemented in various soft wares. Raza et al. stated that the three compounds Sanggenon-O, AC1M2ZZJ and Deoxycalyxin-A showed potential for the ZIKV NS1 in the bonding pattern and the binding free energy analysis (Raza et al. 2019). The free energy calculation has been carried out with the identified compounds from the docking methods by Balajee et al. and provided significance towards the binding affinity. These results observed also states that the binding energy of soalvation is an important factor that helps in providing the force for ligand binding (Balajee et al. 2016).

8.4.6 Conformational Analysis on the Flaviviral Proteins

The identification of small molecule inhibitors through computer aided drug design is not sufficient for the report of therapeutic intervention against flaviviruses. The dominance of the small molecules in protein helps in providing better insights about the mechanism. These dominance can also develops rearrangement of the conformations (McCammon et al. 1977). The conformational changes can be better analysed with the help of molecular dynamics simulation which has been developed in the late 1970s (Davidson et al. 2020). Insights into the molecular motions which is absurd can be appropriate to drug discovery. The binding mechanism between the ligand and target protein is analysed with the conformational analysis through various softwares available through open server like Gromacs, Amber and Desmond. Researchers throughout the world has worked on various flaviviral proteins to depict the conformational changes and analysis of the binding site along with the mechanism.

Fusion of the membrane proteins of the flavivirus is triggered by the endosomal pH during virus endocytosis. These strategies involves in the protonational changes happened due to the conserved histidine residues of the envelope protein available in the surface of the virus resulting in fusion between endosome and bilayers of the virus. The behaviors of the envelope protein of Dengue studied through simulation of the whole systems including protein, solvent and ions states that the ionization states gets updated cyclically. The simulation results states the other residues which is conserved also plays a major role in the rearrangement helps in in the understanding of the structural changes leading to fusion process (Fuzo and Degrève 2014). The Dengue envelope protein with the 23 structurally different compounds are identified for the inhibition among which 2 inhibitors docked with the envelope protein revealed the characterization of the protein. These studies paves way for the future endeavors which highlights the possibility of developing new inhibitors with efficient binding affinity.

The NS1, NS2B/NS3 proteases and NS5 of the Dengue virus of four serotypes with the docked complexes with the phytochemicals has been carried for the assessment of the binding behaviors through the simulation strategies (84). Raza et al. has carried out *in silico* docking strategies which has identified best three compounds which are possessed to be non-toxic to human and can inhibit the zika virus NS1 protein. The significant interactions of the small molecule with various residues of the protein were analyzed with the molecular dynamics simulation (Raza et al. 2019). Santos et al. has formed complexes of the NS2B-NS3 protease and the most promising ligand has been submitted to carry out simulations studies for the conformational analysis. The best identified compound chlorcyclizine validated through the docking and simulation studies has been carried out further for in vitro antiviral studies among the large set of compounds reducing the time (Santos et al. 2019).

Davidson et al. has performed the molecular dynamics which is unbiased with the microseconds and the simulations that is based on the umbrella sampling of the ZIKV NS3 helicase to investigate the dependence of RNA structural conformations. The open and closed conformations of the single stranded RNA bound to the NS3 helicase is studied extensively. It was clearly identified from the results that the loop is conjectured to the feasible region in the protein for targeted inhibition with the small molecule (Davidson et al. 2020). Badshah et al. has performed molecular dynamics simulation for the best complex composed of ZIKV NS3 and 1,4-benzothiazine derivatives which depicts the rigid and flexible domains of the protein. These conformational rearrangements observed through the simulation is the dominance of the small molecules. It is also observed that these domains possess a major role in the replication process which can be targeted for the drug design strategies (Badshah et al. 2019). NS3 of protease dengue virus is targeted by the research group of Almeida helps in fighting against the dengue through the multiple template homology modelling followed by the simulation studies. The reports states that the model comprises of the flexible link in glycine and the proteolytic domain. The simulation studies sampled the binding site on the proteins with the conformational landscape before the binding of the ligand. Almeida et al. proposed the multiple binding modes in the proteins for the inhibitor through the binding plots and the interaction analysis (de Almeida et al. 2013).

The Modelled Zika NS5 protein structure by the Ramharack and Solliman has been carried out for the conformational changes in the structure before binding it with the small molecules for the inhibition through AMBER package simulation methods (Ramharack and Soliman 2018). Molecular interactions between selected compounds and the non-structural enzyme 5 RNA-dependent RNA-polymerase were obtained from molecular docking. The selected compounds exhibited high docking score, binding affinity and suitable protein-ligand interactions. Molecular dynamics simulations were subsequently used to get better idea of the interaction between the selected compounds and the binding pocket of the targeted non-structural-5 RNA dependent RNA polymerase of zika virus (Ahmad et al.

2020). Chuang et al. carried out the together Guassian and normal molecular dynamics simulations to investigate the structural conformations of the Zika NS5 protein with the two components namely the S-adenosyl-L-homocysteine and the m7GTP which is the RNA analogue. The simulation revealed that the some of the substructures and the residues are responsible for the selectivity of the new Zika virus drugs (Chuang et al. 2018).

8.5 Conclusion

Flaviviridae is the major family of viruses which constitute more than 80 types and differentiated into three generas namely the flavivirus, hepatis virus and pestivirus. The flavivirus is further grouped into tick-borne viruses, mosquito-borne viruses and arthropod borne viruses. These viruses has caused a global health pandemic throughout the world over 1800s. Till date no vaccines has been identified for the virus present within the family. Less number of vaccines has been identified for only some viruses which has not shown strong inhibitory activity of the viruses. Due to the time consumed and the expenditure to be involved in the identification of small molecule inhibitors, the researchers around the world faced strong challenges. Herein, the role of computer aided drug design has played a major role in scrutinizing the number of molecules to be tested or analyzed for the inhibitory activity of the flaviviruses. Various research groups around the world are working extensively in in silico strategies and experimental techniques o identify potent inhibitors for treating the diseases caused by these flaviviruses. The use of computational studies with the implementation of various techniques has been put together in this chapter which helps in overcoming the challenges.

Acknowledgements MA, UP and SKSgratefully acknowledge the research facility provided by Alagappa University. SKS thanks RUSA - Phase 2.0 (grant sanctioned vide letter No. F. 24-51/ 2014 - U, Policy (TNMulti-Gen), Dept. of Edn. Govt. of India, Dt. 09.10.2018), Department of Biotechnology (DBT) Ministry of Science and Technology, New Delhi, Grant/Award (No. BT/PR8138/BID/7/458/2013, dated 23rd May, 2013), DST-PURSE 2ndPhase Programme Order No. SR/PURSE Phase 2/38 (G dated 21.02.2017 and FIST (SR/FST/LSI - 667/2016), MHRD for the financial support. MA thanks UGC for providing RGNF Fellowship award no. F1-17/22016-17/RGNF-2015-17-SC-TAM-18759. UP thanks ICMR for provding SRF fellowship with ISRM/11/(19)/2017. Dated 09.08.2018.

Conflict of Interest The authors declare no conflict of interest.

References

Aarthy, M., & Singh, S. K. (2018). Discovery of potent inhibitors for the inhibition of dengue envelope protein: An in silico approach. *Current Topics in Medicinal Chemistry, 18*(18), 1585–1602.

Ahmad, N., Rehman, A. U., Badshah, S. L., Ullah, A., Mohammad, A., & Khan, K. (2020). Molecular dynamics simulation of zika virus NS5 RNA dependent RNA polymerase with selected novel non-nucleoside inhibitors. *Journal of Molecular Structure, 1203*, 127428.

Alshammari, S. A., Alamri, Y. S., Rabhan, F. S., Alabdullah, A. A., Alsanie, N. A., Almarshad, F. A., et al. (2018). Overview of dengue and Zika virus similarity, what can we learn from the Saudi experience with dengue fever? *International Journal of Health Sciences, 12*(1), 77.

Avirutnan, P., Zhang, L., Punyadee, N., Manuyakorn, A., Puttikhunt, C., Kasinrerk, W., et al. (2007). Secreted NS1 of dengue virus attaches to the surface of cells via interactions with heparan sulfate and chondroitin sulfate E. *PLoS Pathogens, 3*(11), e183.

Azeredo, E. L., Neves-Souza, P. C., Alvarenga, A. R., Reis, S. R., Torrentes-Carvalho, A., Zagne, S. M. O., et al. (2010). Differential regulation of toll-like receptor-2, toll-like receptor-4, CD16 and human leucocyte antigen-DR on peripheral blood monocytes during mild and severe dengue fever. *Immunology, 130*(2), 202–216.

Badshah, S. L., Ahmad, N., Rehman, A. U., Khan, K., Ullah, A., Alsayari, A., et al. (2019). Molecular docking and simulation of Zika virus NS3 helicase. *BMC chemistry, 13*(1), 1–8.

Balajee, R., Srinivasadesikan, V., Sakthivadivel, M., & Gunasekaran, P. (2016). In silico screening, alanine mutation, and DFT approaches for identification of NS2B/NS3 protease inhibitors. *Biochemistry Research International, 2016,* 7264080.

Blazevic, J., Rouha, H., Bradt, V., Heinz, F. X., & Stiasny, K. (2016). Membrane anchors of the structural flavivirus proteins and their role in virus assembly. *Journal of Virology, 90*(14), 6365–6378.

Borowski, P., Niebuhr, A., Schmitz, H., Hosmane, R. S., Bretner, M., Siwecka, M. A., et al. (2002). NTPase/helicase of Flaviviridae: Inhibitors and inhibition of the enzyme. *Acta Biochimica Polonica, 49*(3), 597–614.

Botha, E. M., Markotter, W., Wolfaardt, M., Paweska, J. T., Swanepoel, R., Palacios, G., et al. (2008). Genetic determinants of virulence in pathogenic lineage 2 West Nile virus strains. *Emerging Infectious Diseases, 14*(2), 222.

Botting, C., & Kuhn, R. J. (2012). Novel approaches to flavivirus drug discovery. *Expert Opinion on Drug Discovery, 7*(5), 417–428.

Brecher, M., Chen, H., Liu, B., Banavali, N. K., Jones, S. A., Zhang, J., et al. (2015). Novel broad spectrum inhibitors targeting the flavivirus methyltransferase. *PLoS One, 10*(6), e0130062.

Chambers, T. J., McCourt, D. W., & Rice, C. M. (1990). Production of yellow fever virus proteins in infected cells: Identification of discrete polyprotein species and analysis of cleavage kinetics using region-specific polyclonal antisera. *Virology, 177*(1), 159–174.

Chen, S., Wu, Z., Wang, M., & Cheng, A. (2017). Innate immune evasion mediated by flaviviridae non-structural proteins. *Viruses, 9*(10), 291.

Chippaux, J. P., & Chippaux, A. (2018). Yellow fever in Africa and the Americas: A historical and epidemiological perspective. *Journal of Venomous Animals and Toxins Including Tropical Diseases, 24*(1), 1–14.

Chuang, C. H., Chiou, S. J., Cheng, T. L., & Wang, Y. T. (2018). A molecular dynamics simulation study decodes the Zika virus NS5 methyltransferase bound to SAH and RNA analogue. *Scientific Reports, 8*(1), 1–9.

Collins, N. D., & Barrett, A. D. (2017). Live attenuated yellow fever 17D vaccine: A legacy vaccine still controlling outbreaks in modern day. *Current Infectious Disease Reports, 19*(3), 14.

Colpitts, T. M., Conway, M. J., Montgomery, R. R., & Fikrig, E. (2012). West Nile virus: Biology, transmission, and human infection. *Clinical Microbiology Reviews, 25*(4), 635–648.

Coutard, B., Barral, K., Lichière, J., Selisko, B., Martin, B., Aouadi, W., et al. (2017). Zika virus methyltransferase: Structure and functions for drug design perspectives. *Journal of Virology, 91* (5). https://doi.org/10.1128/JVI.02202-16

Daep, C. A., Muñoz-Jordán, J. L., & Eugenin, E. A. (2014). Flaviviruses, an expanding threat in public health: Focus on dengue, West Nile, and Japanese encephalitis virus. *Journal of Neurovirology, 20*(6), 539–560.

Davidson, R. B., Hendrix, J., Geiss, B. J., & McCullagh, M. (2020). RNA-dependent structures of the RNA-binding loop in the flavivirus NS3 helicase. *The Journal of Physical Chemistry B, 124* (12), 2371–2381.

de Almeida, H., Bastos, I. M., Ribeiro, B. M., Maigret, B., & Santana, J. M. (2013). New binding site conformations of the dengue virus NS3 protease accessed by molecular dynamics simulation. *PLoS One, 8*(8), e72402.

de Silva, A. M., & Harris, E. (2018). Which dengue vaccine approach is the most promising, and should we be conerned about enhanced disease after vaccination? The path to a dengue vaccine: Learning from human natural dengue infection studies and vaccine trials. *Cold Spring Harbor Perspectives in Biology, 10*(6), a029371.

Dessalew, N., Bharatam, P. V., & Singh, S. K. (2007). 3D-QSAR CoMFA study on Aminothiazole derivatives as Cyclin-dependent kinase 2 inhibitors. *QSAR & Combinatorial Science, 26*(1), 85–91.

Dick, G. W. A., & Haddow, A. J. (1952). Uganda S virus: A hitherto unrecorded virus isolated from mosquitoes in Uganda..(I). Isolation and pathogenicity. *Transactions of the Royal Society of Tropical Medicine and Hygiene, 46*(6), 600–618.

Dong, H., Fink, K., Züst, R., Lim, S. P., Qin, C. F., & Shi, P. Y. (2014). Flavivirus RNA methylation. *Journal of General Virology, 95*(4), 763–778.

Erbel, P., Schiering, N., D'Arcy, A., Renatus, M., Kroemer, M., Lim, S. P., et al. (2006). Structural basis for the activation of flaviviral NS3 proteases from dengue and West Nile virus. *Nature Structural & Molecular Biology, 13*(4), 372–373.

Falgout, B., & Markoff, L. (1995). Evidence that flavivirus NS1-NS2A cleavage is mediated by a membrane-bound host protease in the endoplasmic reticulum. *Journal of Virology, 69*(11), 7232–7243.

Fall, G., Di Paola, N., Faye, M., Dia, M., de Melo Freire, C. C., Loucoubar, C., et al. (2017). Biological and phylogenetic characteristics of west African lineages of West Nile virus. *PLoS Neglected Tropical Diseases, 11*(11), e0006078.

Fall, G., Diallo, M., Loucoubar, C., & Faye, O. (2014). Vector competence of Culex neavei and Culex quinquefasciatus (Diptera: Culicidae) from Senegal for lineages 1, 2, Koutango and a putative new lineage of West Nile virus. *The American Journal of Tropical Medicine and Hygiene, 90*(4), 747–754.

Fan, J., Lin, L., & Wang, C. K. (2017). Excited state properties of non-doped thermally activated delayed fluorescence emitters with aggregation-induced emission: A QM/MM study. *Journal of Materials Chemistry C, 5*(33), 8390–8399.

Fathima, A. J., & Murugaboopathi, G. (2019). Computer aided drug design for finding a therapeutics for dengue virus targets. *International Journal of Innovative Technology and Exploring Engineering, 9*, 766–771.

Fuzo, C. A., & Degrève, L. (2014). The pH dependence of flavivirus envelope protein structure: Insights from molecular dynamics simulations. *Journal of Biomolecular Structure and Dynamics, 32*(10), 1563–1574.

Gardner, C. L., & Ryman, K. D. (2010). Yellow fever: A reemerging threat. *Clinics in Laboratory Medicine, 30*(1), 237–260.

Garg, H., Mehmetoglu-Gurbuz, T., & Joshi, A. (2018). Recent advances in Zika virus vaccines. *Viruses, 10*(11), 631.

Gorshkov, K., Shiryaev, S. A., Fertel, S., Lin, Y. W., Huang, C. T., Pinto, A., et al. (2019). Zika virus: Origins, pathological action, and treatment strategies. *Frontiers in Microbiology, 9*, 3252.

Gould, E. A., & Solomon, T. (2008). Pathogenic flaviviruses. *The Lancet., 371*(9611), 500–509.

Gregory, C. J., Oduyebo, T., Brault, A. C., Brooks, J. T., Chung, K. W., Hills, S., et al. (2017). Modes of transmission of Zika virus. *The Journal of Infectious Diseases, 216*(suppl_10), S875–S883.

Grover, A., Katiyar, S. P., Singh, S. K., Dubey, V. K., & Sundar, D. (2012). A leishmaniasis study: structure-based screening and molecular dynamics mechanistic analysis for discovering potent inhibitors of spermidine synthase. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 1824*(12), 1476–1483.

Gubler, D. J., & Clark, G. G. (1995). Dengue/dengue hemorrhagic fever: The emergence of a global health problem. *Emerging Infectious Diseases, 1*(2), 55.

Gupta, G., Lim, L., & Song, J. (2015). NMR and MD studies reveal that the isolated dengue NS3 protease is an intrinsically disordered chymotrypsin fold which absolutely requests NS2B for correct folding and functional dynamics. *PLoS One, 10*(8), e0134823.

Halim, S. A., Khan, S., Khan, A., Wadood, A., Mabood, F., Hussain, J., et al. (2017). Targeting dengue virus NS-3 helicase by ligand based pharmacophore modeling and structure based virtual screening. *Frontiers in Chemistry, 5*, 88.

Halstead, S. B. (1988). Pathogenesis of dengue: Challenges to molecular biology. *Science, 239* (4839), 476–481.

Hamel, R., Dejarnac, O., Wichit, S., Ekchariyawat, P., Neyret, A., Luplertlop, N., et al. (2015). Biology of Zika virus infection in human skin cells. *Journal of Virology, 89*(17), 8880–8896.

Hasan, S. S., Sevvana, M., Kuhn, R. J., & Rossmann, M. G. (2018). Structural biology of Zika virus and other flaviviruses. *Nature Structural & Molecular Biology, 25*(1), 13–20.

Hindle, E. (1930). The transmission of yellow fever. *Lancet, 216* (5590), 835–842.

Hober, D., Shen, L., Benyoucef, S., De Groote, D., Deubel, V., & Wattré, P. (1996). Enhanced TNFα production by monocytic-like cells exposed to dengue virus antigens. *Immunology Letters, 53*(2–3), 115–120.

Huang, Y. J. S., Higgs, S., Horne, K. M., & Vanlandingham, D. L. (2014). Flavivirus-mosquito interactions. *Viruses, 6*(11), 4703–4730.

Jia, X. Y., Briese, T., Jordan, I., Rambaut, A., Chi, H. C., Mackenzie, J. S., et al. (1999). Genetic analysis of West Nile New York 1999 encephalitis virus. *The Lancet, 354*(9194), 1971–1972.

Jitendra, S., & Vinay, R. (2011). Structure based drug designing of a novel antiflaviviral inhibitor for nonstructural 3 protein. *Bioinformation, 6*(2), 57.

Johansson, M. A., Arana-Vizcarrondo, N., Biggerstaff, B. J., & Staples, J. E. (2010). Incubation periods of yellow fever virus. *The American Journal of Tropical Medicine and Hygiene, 83*(1), 183–188.

Johnston, B. L., & Conly, J. M. (2000). West Nile virus-where did it come from and where might it go? *Canadian Journal of Infectious Diseases, 11*, 175–178.

Katsila, T., Spyroulias, G. A., Patrinos, G. P., & Matsoukas, M. T. (2016). Computational approaches in target identification and drug discovery. *Computational and Structural Biotechnology Journal, 14*, 177–184.

Kaufmann, B., & Rossmann, M. G. (2011). Molecular mechanisms involved in the early steps of flavivirus cell entry. *Microbes and Infection, 13*(1), 1–9.

Kawaguchi, I., Sasaki, A., & Boots, M. (2003). Why are dengue virus serotypes so distantly related? Enhancement and limiting serotype similarity between dengue virus strains. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 270*(1530), 2241–2247.

Khaiboullina, S., Ribeiro, F. M., Uppal, T., Martynova, E., Rizvanov, A., & Verma, S. C. (2019). Zika virus transmission through blood tissue barriers. *Frontiers in Microbiology, 10*, 1465.

Kudchodkar, S. B., Choi, H., Reuschel, E. L., Esquivel, R., Kwon, J. J. A., Jeong, M., et al. (2018). Rapid response to an emerging infectious disease–lessons learned from development of a synthetic DNA vaccine targeting Zika virus. *Microbes and Infection, 20*(11–12), 676–684.

Kumar, A., Liang, B., Aarthy, M., Singh, S. K., Garg, N., Mysorekar, I. U., et al. (2018). Hydroxychloroquine inhibits Zika virus NS2B-NS3 protease. *ACS Omega, 3*(12), 18132–18141.

Kumar, D., Aarthy, M., Kumar, P., Singh, S. K., Uversky, V. N., & Giri, R. (2019). Targeting the NTPase site of Zika virus NS3 helicase for inhibitor discovery. *Journal of Biomolecular Structure and Dynamics, 38*, 1–11.

Kumar, D., Kaur, N., Giri, R., & Singh, N. (2020a). A biscoumarin scaffold as an efficient anti-Zika virus lead with NS3-helicase inhibitory potential: In vitro and in silico investigations. *New Journal of Chemistry, 44*(5), 1872–1880.

Kumar, D., Sharma, N., Aarthy, M., Singh, S. K., & Giri, R. (2020b). Mechanistic insights into Zika virus NS3 helicase inhibition by Epigallocatechin-3-gallate. *ACS Omega, 5*(19), 11217–11226.

Kuno, G., Chang, G. J. J., Tsuchiya, K. R., Karabatsos, N., & Cropp, C. B. (1998). Phylogeny of the genus Flavivirus. *Journal of Virology, 72*(1), 73–83.

Lepiniec, L., Dalgarno, L., Huong, V. T. Q., Monath, T. P., Digoutte, J. P., & Deubel, V. (1994). Geographic distribution and evolution of yellow fever viruses based on direct sequencing of genomic cDNA fragments. *Journal of General Virology, 75*(2), 417–423.

Li, L., Lok, S. M., Yu, I. M., Zhang, Y., Kuhn, R. J., Chen, J., et al. (2008). The flavivirus precursor membrane-envelope protein complex: Structure and maturation. *Science, 319*(5871), 1830–1834.

Lim, S. P., Noble, C. G., & Shi, P. Y. (2015). The dengue virus NS5 protein as a target for drug discovery. *Antiviral Research, 119*, 57–67.

Lin, X., Li, X., & Lin, X. (2020). A review on applications of computational methods in drug screening and design. *Molecules, 25*(6), 1375.

Lindenbach, B. D., & Rice, C. M. (2003). Molecular biology of flaviviruses. *Advances in Virus Research, 59*, 23–62.

Lvov, D. K., Butenko, A. M., Gromashevsky, V. L., Kovtunov, A. I., Prilipov, A. G., Kinney, R., et al. (2004). West Nile virus and other zoonotic viruses in Russia: Examples of emerging-reemerging situations. In *Emergence and control of zoonotic viral encephalitides* (pp. 85–96). Vienna: Springer.

Malavige, G. N., Fernando, S., Fernando, D. J., & Seneviratne, S. L. (2004). Dengue viral infections. *Postgraduate Medical Journal, 80*(948), 588–601.

Marr, J. S., & Cathey, J. T. (2013). The 1802 saint-Domingue yellow fever epidemic and the Louisiana purchase. *Journal of Public Health Management and Practice, 19*(1), 77–82.

Mazeaud, C., Freppel, W., & Chatel-Chaix, L. (2018). The multiples fates of the Flavivirus RNA genome during pathogenesis. *Frontiers in Genetics, 9*, 595.

McCammon, J. A., Gelin, B. R., & Karplus, M. (1977). Dynamics of folded proteins. *Nature, 267* (5612), 585–590.

Medigeshi, G. R. (2011). Mosquito-borne flaviviruses: Overview of viral life-cycle and host–virus interactions. *Future Virology, 6*(9), 1075–1089.

Modis, Y., Ogata, S., Clements, D., & Harrison, S. C. (2004). Structure of the dengue virus envelope protein after membrane fusion. *Nature, 427*(6972), 313–319.

Moureau, G., Cook, S., Lemey, P., Nougairede, A., Forrester, N. L., Khasnatinov, M., et al. (2015). New insights into flavivirus evolution, taxonomy and biogeographic history, extended by analysis of canonical and alternative coding sequences. *PLoS One, 10*(2), e0117849.

Murphy, B. R., & Whitehead, S. S. (2011). Immune response to dengue virus and prospects for a vaccine. *Annual Review of Immunology, 29*, 587–619.

Murray, C. L., Jones, C. T., & Rice, C. M. (2008). Architects of assembly: Roles of Flaviviridae non-structural proteins in virion morphogenesis. *Nature Reviews Microbiology, 6*(9), 699–708.

Musso, D., Nhan, T., Robin, E., Roche, C., Bierlaire, D., Zisou, K., et al. (2014). Potential for Zika virus transmission through blood transfusion demonstrated during an outbreak in French Polynesia, November 2013 to February 2014. *Eurosurveillance, 19*(14), 20761.

Nicholson, C. O., Costin, J. M., Rowe, D. K., Lin, L., Jenwitheesuk, E., Samudrala, R., et al. (2011). Viral entry inhibitors block dengue antibody-dependent enhancement in vitro. *Antiviral Research, 89*(1), 71–74.

Noble, C. G., Seh, C. C., Chao, A. T., & Shi, P. Y. (2012). Ligand-bound structures of the dengue virus protease reveal the active conformation. *Journal of Virology, 86*(1), 438–446.

Olagnier, D., Muscolini, M., Coyne, C. B., Diamond, M. S., & Hiscott, J. (2016). Mechanisms of Zika virus infection and neuropathogenesis. *DNA and Cell Biology, 35*(8), 367–372.

Panwar, U., & Singh, S. K. (2018). Structure-based virtual screening toward the discovery of novel inhibitors for impeding the protein-protein interaction between HIV-1 integrase and human lens epithelium-derived growth factor (LEDGF/p75). *Journal of Biomolecular Structure and Dynamics, 36*(12), 3199–3217.

Petersen, L. R., & Marfin, A. A. (2005). Shifting epidemiology of Flaviviridae. *Journal of Travel Medicine, 12*(suppl_1), s3–s11.

Pettersson, J. H. O., & Fiz-Palacios, O. (2014). Dating the origin of the genus Flavivirus in the light of Beringian biogeography. *Journal of General Virology, 95*(9), 1969–1982.

Poh, M. K., Yip, A., Zhang, S., Priestle, J. P., Ma, N. L., Smit, J. M., et al. (2009). A small molecule fusion inhibitor of dengue virus. *Antiviral Research, 84*(3), 260–266.

Pradiba, D., Aarthy, M., Shunmugapriya, V., Singh, S. K., & Vasanthi, M. (2018). Structural insights into the binding mode of flavonols with the active site of matrix metalloproteinase-9 through molecular docking and molecular dynamic simulations studies. *Journal of Biomolecular Structure and Dynamics, 36*(14), 3718–3739.

Racloz, V., Ramsey, R., Tong, S., & Hu, W. (2012). Surveillance of dengue fever virus: A review of epidemiological models and early warning systems. *PLoS Neglected Tropical Diseases, 6*(5), e1648.

Ramharack, P., & Soliman, M. E. (2018). Zika virus NS5 protein potential inhibitors: An enhanced in silico approach in drug discovery. *Journal of Biomolecular Structure and Dynamics, 36*(5), 1118–1133.

Raza, S., Abbas, G., & Azam, S. S. (2019). Screening pipeline for Flavivirus based inhibitors for Zika virus NS1. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. https://doi.org/10.1109/TCBB.2019.2911081

Reddy, K. K., & Singh, S. K. (2014). Combined ligand and structure-based approaches on HIV-1 integrase strand transfer inhibitors. *Chemico-Biological Interactions, 218*, 71–81.

Rice, C. M. (1996). Flaviviridae: The viruses and their replication. In B. N. Fields, D. M. Knipe, & P. M. Howley (Eds.), *Fields virology* (pp. 931–959). Philadelphia/London: Wolters Kluwer Health/Lippincott Williams & Wilkins.

Sall, A. A., Faye, O., Diallo, M., Firth, C., Kitchen, A., & Holmes, E. C. (2010). Yellow fever virus exhibits slower evolutionary dynamics than dengue virus. *Journal of Virology, 84*(2), 765–772.

Santos, F. R., Nunes, D. A., Lima, W. G., Davyt, D., Santos, L. L., Taranto, A. G., et al. (2019). Identification of Zika virus NS2B-NS3 protease inhibitors by structure-based virtual screening and drug repurposing approaches. *Journal of Chemical Information and Modeling, 60*(2), 731–737.

Sejvar, J. J. (2003). West Nile virus: An historical overview. *Ochsner Journal, 5*(3), 6–10.

Shanmuganathan, B., Suryanarayanan, V., Sathya, S., Narenkumar, M., Singh, S. K., Ruckmani, K., et al. (2018). Anti-amyloidogenic and anti-apoptotic effect of α-bisabolol against Aβ induced neurotoxicity in PC12 cells. *European Journal of Medicinal Chemistry, 143*, 1196–1207.

Sharda, S., Sarmandal, P., Cherukommu, S., Dindhoria, K., Yadav, M., Bandaru, S., et al. (2017). A virtual screening approach for the identification of high affinity small molecules targeting BCR-ABL1 inhibitors for the treatment of chronic myeloid leukemia. *Current Topics in Medicinal Chemistry, 17*(26), 2989–2996.

Sharma, N., Murali, A., Singh, S. K., & Giri, R. (2017). Epigallocatechin gallate, an active green tea compound inhibits the Zika virus entry into host cells via binding the envelope protein. *International Journal of Biological Macromolecules, 104*, 1046–1054.

Sharma, K., Patidar, K., Ali, M. A., Patil, P., Goud, H., Hussain, T., et al. (2018). Structure-based virtual screening for the identification of high affinity compounds as potent VEGFR2 inhibitors for the treatment of renal cell carcinoma. *Current Topics in Medicinal Chemistry, 18*(25), 2174–2185.

Singh, S. K., Dessalew, N., & Bharatam, P. V. (2007). 3D-QSAR CoMFA study on oxindole derivatives as cyclin dependent kinase 1 (CDK1) and cyclin dependent kinase 2 (CDK2) inhibitors. *Medicinal Chemistry, 3*(1), 75–84.

Sliwoski, G., Kothiwale, S., Meiler, J., & Lowe, E. W. (2014). Computational methods in drug discovery. *Pharmacological Reviews, 66*(1), 334–395.

Sotcheff, S., & Routh, A. (2020). Understanding Flavivirus capsid protein functions: The tip of the iceberg. *Pathogens, 9*(1), 42.

Tan, T. Y., Fibriansah, G., & Lok, S. M. (2020). Capsid protein is central to the birth of flavivirus particles. *PLoS Pathogens, 16*(5), e1008542.

Therkelsen, M. D., Klose, T., Vago, F., Jiang, W., Rossmann, M. G., & Kuhn, R. J. (2018). Flaviviruses have imperfect icosahedral symmetry. *Proceedings of the National Academy of Sciences, 115*(45), 11608–11612.

Tripathi, S. K., Selvaraj, C., Singh, S. K., & Reddy, K. K. (2012). Molecular docking, QPLD, and ADME prediction studies on HIV-1 integrase leads. *Medicinal Chemistry Research, 21*(12), 4239–4251.

ul Qamar, M. T., Kiran, S., Ashfaq, U. A., Javed, M. R., Anwar, F., Ali, M. A., et al. (2016). Discovery of novel dengue NS2B/NS3 protease inhibitors using pharmacophore modeling and molecular docking based virtual screening of the zinc database. *International Journal of Pharmacology, 12*(6), 621–632.

ul Qamar, M. T., Maryam, A., Muneer, I., Xing, F., Ashfaq, U. A., Khan, F. A., et al. (2019). Computational screening of medicinal plant phytochemicals to discover potent pan-serotype inhibitors against dengue virus. *Scientific Reports, 9*(1), 1–16.

ul Qamar, M. T., Mumtaz, A., Rabbia Naseem, A. A., Fatima, T., Jabbar, T., Ahmad, Z., et al. (2014). Molecular docking based screening of plant flavonoids as dengue NS1 inhibitors. *Bioinformation, 10*(7), 460.

Ulbert, S. (2019). West Nile virus vaccines–current situation and future directions. *Human Vaccines & Immunotherapeutics, 15*(10), 2337–2342.

Valente, A. P., & Moraes, A. H. (2019). Zika virus proteins at an atomic scale: How does structural biology help us to understand and develop vaccines and drugs against Zika virus infection? *Journal of Venomous Animals and Toxins Including Tropical Diseases, 25,* e20190013.

Varpa, K., Iltanen, K., & Juhola, M. (2008). Machine learning method for knowledge discovery experimented with otoneurological data. *Computer Methods and Programs in Biomedicine, 91* (2), 154–164.

Wang, L., Valderramos, S. G., Wu, A., Ouyang, S., Li, C., Brasil, P., et al. (2016). From mosquitos to humans: Genetic evolution of Zika virus. *Cell Host & Microbe, 19*(5), 561–565.

Westaway, E. G., Mackenzie, J. M., Kenney, M. T., Jones, M. K., & Khromykh, A. A. (1997). Ultrastructure of Kunjin virus-infected cells: Colocalization of NS1 and NS3 with double-stranded RNA, and of NS2B with NS3, in virus-induced membrane structures. *Journal of Virology, 71*(9), 6650–6661.

Xie, X., Gayen, S., Kang, C., Yuan, Z., & Shi, P. Y. (2013). Membrane topology and function of dengue virus NS2A protein. *Journal of Virology, 87*(8), 4609–4622.

Yadav, R., Selvaraj, C., Aarthy, M., Kumar, P., Kumar, A., Singh, S. K., et al. (2020). Investigating into the molecular interactions of flavonoids targeting NS2B-NS3 protease from ZIKA virus through in-silico approaches. *Journal of Biomolecular Structure and Dynamics*, 1–13.

Yang, J. M., Chen, Y. F., Tu, Y. Y., Yen, K. R., & Yang, Y. L. (2007). Combinatorial computational approaches to identify tetracycline derivatives as flavivirus inhibitors. *PLoS One, 2*(5), e428.

Yu, I. M., Zhang, W., Holdaway, H. A., Li, L., Kostyuchenko, V. A., Chipman, P. R., et al. (2008). Structure of the immature dengue virus at low pH primes proteolytic maturation. *Science, 319* (5871), 1834–1837.

Yu, L., Takeda, K., & Gao, Y. (2017). Characterization of virus-specific vesicles assembled by West Nile virus non-structural proteins. *Virology, 506*, 130–140.

Chapter 9
Elucidating Protein-Ligand Interactions Using High Throughput Biophysical Techniques

Nipanshu Agarwal, Vivek Chetry, and Krishna Mohan Poluri

Abstract Proteins are large, complex molecules that functionally regulates almost all cellular and biochemical processes. As proteins are important component in cell physiology, their interaction with small molecules that modulates the function are clinically significant. Nearly, all essential biomolecular processes are highly sensitive and selective involving molecular recognition and binding of ligands/macromolecules to proteins. Hence, techniques that can reveal detailed information like binding energetics, kinetics, stoichiometry, thermodynamics, structural changes and conformational dynamics are of great importance. Current arsenal of techniques that enable characterization of these interactions have been well established and progressing towards advancement at a faster pace. The current chapter details four major biophysical techniques namely Nuclear Magnetic Resonance spectroscopy (NMR), Surface Plasmon Resonance (SPR), Isothermal Titration Calorimetry (ITC) and Fluorescence Spectroscopy that are vastly used in characterizing the thermodynamics and kinetics of protein-ligand interaction.

Keywords Protein-ligand interaction · Nuclear magnetic resonance spectroscopy · Surface Plasmon resonance · Isothermal titration calorimetry · Fluorescence spectroscopy

Nipanshu Agarwal and Vivek Chetry have contributed equally to this work.

N. Agarwal · V. Chetry
Department of Biotechnology, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India

K. M. Poluri (✉)
Department of Biotechnology, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India

Centre for Nanotechnology, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India
e-mail: krishfbt@iitr.ac.in

S. K. Singh (ed.), *Innovations and Implementations of Computer Aided Drug Discovery Strategies in Rational Drug Design*,
https://doi.org/10.1007/978-981-15-8936-2_9

9.1 Introduction

Biomolecules in cellular milieu interact with various types of components and bring about certain physiological outcomes. The partners that are often involved in such reactions are generally classified broadly into macromolecules, ions, small molecules, etc. Interaction of proteins with their counter-interactive molecules essentially defines the functions of the particular protein (Frederick et al. 2007). Such interactions are necessary, for biological processes, and act as a pre-requisite ground, to shape the binding site(s) for further intuitive ligand interaction (Gossert and Jahnke 2016; Meyer and Peters 2003). For instance, the signal transduction requires the flow of synaptic impulse mediated by the ions (Dolphin and Lee 2020); whereas cellular metabolism pathways employ exuberant use of cooperativity (Liu et al. 2019; Tang and Yengo 2018), allostery (Zhang and Nussinov 2019), conformation modulatory switches (Liu et al. 2019), and trans-acting molecules (Guo and Zhou 2016; Meyer and Peters 2003). Owing to ubiquitous nature of protein interactions with other molecules, characterization of such interactions is important to explore the underlying mechanisms, generating therapeutic proteins using protein engineering methods and nano-biotechnology approaches (Gulati and M Poluri 2016; Gulati and Poluri 2019; Poluri and Gulati 2016; Poluri and Gulati 2017; Sarkar et al. 2018; Sarkar et al. 2020).

Apart from macromolecular assemblies the small molecules, namely, ligands (entities with low molecular weight) and ions, often remain unsung for their discrete role in molecular response pathways. These small molecules/ligands can invariably alter the dynamics of the system components by changing the intra/inter molecular interactions (Chandel et al. 2018; Liu et al. 2019). The altered dynamics could be understood by the basic principles of thermodynamics, wherein the bond interaction and change in state could be derived in the form of free energy. Protein-ligand interactions are central in mediating enzyme catalysis, signal transduction, receptor binding, and bio-cognition (Chandel et al. 2018). Hence, to facilitate the quantitative and qualitative description of molecular communication, certain techniques based on principles of physics are put to use.

Protein-ligand interaction is a complex phenomenon that involves domain relocation, conformation switch, functional group re-arrangement, allostery and ionic interactions etc. Undoubtedly, to understand these molecular interactions, native conditions of system are mimicked *in vitro* by employing suitable buffer systems. Often the ions from buffer salts bind at the active site/ probable site of ligand interaction thereby inhibiting the interactions perpetuating through salt bridges. Salts like Na_2SO_4, NaCl, NaSCN can impart a profound effect on protein stability, and henceforth a careful choice of buffer system is a pre-requisite (Damodaran and Kinsella 1981). Prior characterization of ion-protein binding energetics can prove beneficial in galvanizing partaking of ion(s) in the interaction. For instance, high salt concentrations can lead to an undesirable pulse width increase in NMR spectroscopy experiments. On a similar note, certain fluorescent dyes are incompatible with buffer components; e.g. Alexa fluor cannot be used in Tris buffer due to presence of amine-moiety. Hence, the sole discretion for the choice of buffer system with particular ionic strength depends on the technique, scope of study, required precision and the reagents involved as per the experimental design.

From biological perspective, protein-ligand interactions are indispensable. Hence, efforts were put forth to understand the static behavior of the molecules. Several techniques like X-ray Crystallography, Mass Spectrometry, Isothermal Titration Calorimetry (ITC), Nuclear Magnetic Resonance (NMR) Spectroscopy, and Differential Scanning Calorimetry (DSC), Absorption spectroscopy, Circular Dichroism spectroscopy (CD), and Fluorescence spectroscopy etc., were developed. These analytical techniques illustrate the proportional input from each component in a reaction as well as their rates of reaction and thermodynamic parameters. Previously, X-ray Crystallography study of protein bound to ligand was considered the ultimate. With the advent of advanced technologies and techniques, the interest of the researchers shifted towards studying the dynamic behavior of the molecules. Presently, to understand disease progression, to either design bio-mimetic antagonists or agonists through Structure Based Drug Discovery approach (SBDD) (Bocquet et al. 2018; Śledź and Caflisch 2018), information from structural biophysics and computational biology is often amalgamated with the chemo-genomics approaches (Gulati and Poluri 2019; Jones and Bunnage 2017; Kubinyi 2006; Sarkar et al. 2018). Structural biophysics techniques impart a heightened upper edge by predicting the site of interaction (up to atomic resolution in certain cases), conformational transitions, thermodynamic parameters, binding kinetics and stoichiometry.

Henceforth, in this chapter we attempt to rationally portray the recent instances where the techniques are implemented to extract high throughput results and advancements that led to the drastic improvements in biophysical techniques like (a) Nuclear Magnetic Resonance spectroscopy (NMR), (b) Surface Plasmon Resonance (SPR), (c) Isothermal Titration Calorimetry and, (d) Fluorescence spectroscopy for elucidating the protein-ligand interactions. Interestingly, each technique described herein could be used to investigate the interaction by keeping both protein and ligand as static while probing the interaction of the other counterpart.

9.2 Nuclear Magnetic Resonance Spectroscopy

Nuclear magnetic Resonance (NMR) spectroscopy relies on the intrinsic magnetic property of the nucleus present in the atoms of a given molecule. Any perturbation bought about, whether due to interaction with the external molecule or change in vicinal environment can be detected by use of standard/specialized radio frequency pulse sequence. Unequivocal applicability for characterization of all the interactional permutations and combinations of proteins, nucleic acids, carbohydrates, lipids, small molecules and ions with precision attained due to selective determination at atomic level, makes NMR spectroscopy an outstanding technique at disposal.

9.2.1 Theoretical Aspects of NMR Spectroscopy and Beyond

With the intricate attention to the Electron cloud model and Bohr' Model of atom, subatomic particles traverse a prejudiced trajectory around the nucleus (Garrett 1962). The central nuclei have existent neutron and positron constituting a net positive charge with existent mass. This mass and charge at the central core result in inherent magnetic properties and hence, a net magnetic dipole denoted by magnetic moment (μ), parallel to spin axis is generated (Garrett 1962; Keeler 2011). NMR utilizes the magnetic properties of nuclei to extract the information from surrounding atoms in near vicinity (Agarwal et al. 2018). According to the principles of quantum physics, a nucleus with net spin of *I* can orient in different possible combinations that are restricted to 2*I* + 1 ways. In case, of nucleus with spin = ½, there are only two probable orientations which are equal in energy in the absence of magnetic field, whereas under the influence of applied magnetic field these two states are equal and opposite in magnitude that either reinforces the spin or oppose the spin. This splitting of the spins in the two energy states is known as *Zeeman effect* (Bray et al. 1991; Keeler 2011). The magnetic moment (μ) being a quantized vector simultaneously existing in two states, could not align exactly to the magnetic axis. Hence, by virtue of force due to magnetic axis (B_0), the magnetic moment processes at a particular angle to the angular velocity with a specific angular frequency (ω_0) known as *Larmor frequency* (Keeler 2011). In simplified terms the *Larmor frequency* is dependent on gyromagnetic ratio (γ) of a particular nucleus and the applied magnetic field (B_0), as follows:

$$\omega_0 = \gamma B_0$$

When a sample is exposed to radiation in the form of radiofrequency (MHz), the nucleus absorbs energy and a significant change in angle of precession is observed. This absorption of energy happens only in a condition where the frequency of radiation matches (resonates) with the precession frequency. Hence, NMR phenomenon relies on the nuclear spin and absorption of electromagnetic radiation from the radiofrequency region by the nuclei of atoms. NMR chemical shifts depend on local neighborhood of nuclei and are strongly affected by surrounding electronic configuration. Such a phenomenon allows us to distinguish different nuclei in a given NMR spectrum that appears at specific resonance frequencies (Agarwal et al. 2018; Legchenko 2013). Further removal of this external radiation relaxes the nuclei back to the original state and release of energy is detected as free induction decay (FID). This FID, thus obtained is Fourier transformed (FT) to get a fingerprint spectrum of the analyte (Keeler 2011) (Fig. 9.1).

The characteristic spectrum of an analyte is represented as peaks against a particular chemical shift value. The chemical shift value is nevertheless the measure of frequency at which a nucleus resonates, converted to a universal scale, for the ease of uniformity in convention (Agarwal et al. 2018; Legchenko 2013). Thus, any change in the chemical shift after addition of the interacting partner primarily

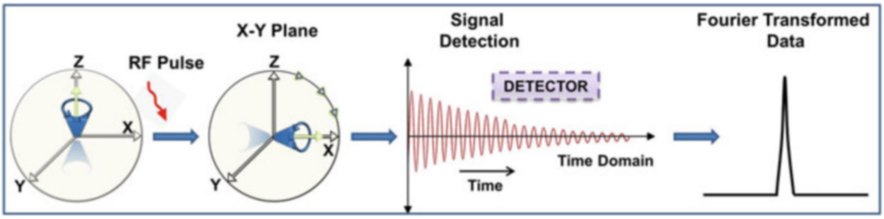


Fig. 9.1 Schematic showing the signal generation and detection in a typical NMR spectroscopy experiment

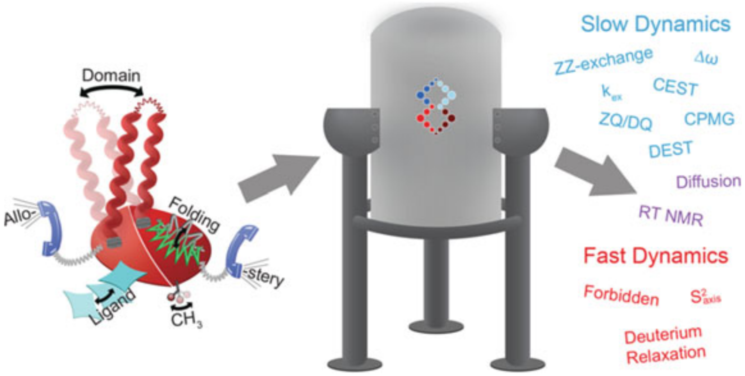


Fig. 9.2 Illustrative depiction of the biochemical phenomenon that can be investigated using NMR spectroscopy experiments. [Adapted with permission from Boswell et al. (Boswell and Latham 2018), Copyright (2018) American Chemical Society]

denotes the interaction and represented as chemical shift perturbation (CSP). Through all the NMR experiments like 1D, 2D and multidimensional experiments, in concept the binding of protein to ligand is discerned by comparing spectrum in free and bound state (Cala et al. 2014). Several experimental approaches have been devised to understand protein-ligand interaction through NMR spectroscopy. Recently, an innovative method of In-cell NMR was utilized to screen potential drug molecules in human cells *in vitro* (Dansereau et al. 2019; Yamaoki et al. 2020).

Major aim of the NMR experiments is to detect the binding site, mechanism, kinetics, and energetics; and if at all there are any changes in the structure upon binding. These parameters can be ascertained by keeping both proteins and ligands, under consideration at constant concentration while adding other interacting counterpart sequentially. The changes are observed in the form of magnetic tensors like chemical shifts, scalar couplings, Nuclear Overhauser Effects (NOEs), paramagnetic interactions or dipolar couplings and intensity change (Billeter et al. 2008; Cala et al. 2014; Fielding 2003; Kay 2011) (Fig. 9.2). The chemical shifts thus obtained are then compared to get the CSPs which are suggestive of interaction. The ligand binding site could be galvanized using the information of interacting residues and it is easier to differentiate a specific binding from a non- specific binding. But, the

Table 9.1 List of NMR techniques employed to probe protein-ligand interactions. The observed parameter(s), and the process of evaluation for each technique is also described

NMR technique	Observed entity	Observed parameter	Procedure for analysis
Chemical shift perturbations	Protein/ ligand	Chemical shifts	Net chemical shift is calculated using reference spectrum
Relaxation	Protein/ ligand	T_1/R_1 or T_2/R_2	Relaxation of the molecule is monitored over time and then relaxation parameter (T_1/ T_2) is inferred using exponential decay fit
Exchange spectroscopy	Protein/ ligand	^{1}H- NOEs	Proton intensity ratio of spatially coupled nucleus is calculated
Saturation transfer difference NMR	Protein/ ligand	^{1}H/^{13}C/^{15}N chemical shifts	$I_{STD} = I_0$-I_{SAT}, is calculated to account for binding I_0: Unsaturated spectrum, I_{SAT}: Saturated spectrum
Water-LOGSY	Ligand	^{1}H chemical shifts	Obtained spectrum is a measure of average weighted population in terms of bound and unbound states Positive signals denote interacting protons and negative signals arise from non-interacting protons
Translational diffusion measurements	Protein/ ligand	Rate of molecular tumbling	Differential tumbling/diffusion coefficient of molecules due to interaction is correlated to the molecular weight

major concern while using these methods lies in the limitation due to constraints related to (a) size/stability/solubility/isotopic labeling/large quantity of proteins, (b) time consuming longer set of 2D and 3D experiments, (c) Backbone Resonance Assignment (Fielding 2007). As per the reports in recent past (Becker et al. 2018; Jabar et al. 2017; Pellecchia et al. 2002), large proteins have been analyzed using NMR SOLVE (Structurally Oriented Library Valency Engineering) technique (Kuzuyama et al. 2000; Pellecchia et al. 2002), suggesting least limitation due to size of protein. Few NMR based experimental techniques that are routinely used to elucidate the protein-ligand interactions are summarized in Table 9.1 and described in the following sections.

9.2.2 *Chemical-Shift Perturbations (CSPs)*

This is a preliminary method to quantitate the binding of ligand. For this spectrum of free protein is used to decipher the change in the chemical shift values after addition of ligand in sequentially increasing order. Once the protein- ligand complex is formed the initial local electron density is affected. The chemical shifts can be studied by overlaying the spectra of a series, where the displaced peaks are suggestive of interaction (Becker et al. 2018; Furukawa et al. 2016). The overall appearance of peak is dependent on the life-time of the complex formation. Reliance on protein labeling and anomalies in the form of spectral line broadening are major setbacks for this method. Further, line shape analysis could be applied to a series of NMR titration spectra, to conclude precise kinetic parameters and fit binding model using software tools like NMRkin, TITAN, Lineshapekin, IDAP, NmrLineGuru etc. (Feng et al. 2019; Waudby et al. 2016).

In the case of fast exchange, the interaction observed is weak binding and often appears as a single peak due to short lived protein-ligand complex. Contrastingly, in case of slow exchange, the protein-ligand complex is stable for longer period and could be identified as separate entity in the overlaid spectrum (Gossert and Jahnke 2016; Liu et al. 2016; Waudby et al. 2016). CSP methods are commonly used to derive binding kinetics from a series of 2D ^{1}H-^{13}C/^{15}N HSQC/TROSY NMR experiments as was demonstrated in case of protein- glycosamine interaction (Joseph et al. 2015; Poluri et al. 2013). Recently, Yu et al. used chemical shift perturbations for refining the protein-ligand complex. The theoretical chemical shifts were compared to empirically calculated CSPs. Also, a computational approach was designed using quantum mechanics to study protein-ligand complex on the basis of existing datasets. The method hence is called as ^{1}H Empirical Chemical Shift perturbation (HECSP) method (Yu et al. 2017). Similarly, Balazs et al. used CSPs derived from proton magnetic resonance to study the three dimensional conformations adopted by a linear ligand in solution (Balazs et al. 2019). NMR data analysis requires intricate understanding of principles and theory associated with NMR spectroscopy. Thus, a sense of intimidation and reluctance among non-experts prevails; recently, Mureddu et al. published a report discussing the use of NMR software packages for data analysis. The report provided a comprehensive account on the data analysis capabilities of softwares like NMR View, Computer Assisted Resonance Assignment (CARA), CcpNmr V3 and CcpNmr V2. The authors could simulate the peak movement using CSP under slow and fast exchange regime (Mureddu and Vuister 2019). In this past decade ^{19}F based NMR chemical shift methodology has gained popularity for ligand screening and competitive binding assays have been improvised (Abboud et al. 2017; Dalvit and Knapp 2017; Dalvit and Vulpetti 2018; Gao et al. 2017; Gee et al. 2018; Usui et al. 2017).

9.2.3 Relaxation Derived Conformational Dynamics

Likewise, relaxation of nucleus after the irradiation with the radio-frequency pulse, each amino acid can be considered to relax back to its initial position. This restoration of equilibrium in the form of relaxation can be recorded as time dependent fluctuations affecting overall protein conformation as a function of diminishing intensity (Becker et al. 2018). The 3D flexibility attained by the protein by virtue of inherent molecular motions affects protein function, thermodynamic stability, folding and ligand binding. Altered dynamics of the conformational ensemble in the form of restricted motions and enhanced flexibility can provide insights into the binding event. Relaxation dynamics could enable study of proteins at different time scales over the range of slow to fast motion regime. There are three types of relaxations that are generally used to study motions related to Protein Ligand Interactions, (a) Longitudinal/Spin-Lattice relaxation (T_1/R_1), (b) Transverse/Spin-Spin relaxation (T_2/R_2), and (c) Nuclear Overhauser Effect (het-NOE) (Kay 2011). These relaxation parameters denote the transfer of energy to the neighboring spin/ lattice of the element. Transverse relaxation rates are sensitive to the motions in a relatively slow time frame of micro to milliseconds, while het- NOEs are mostly sensitive to high frequency motions in protein backbone (Kay 2011).

Relaxation experiments like T_1/R_1 provide quantitative involvement of the proton in interaction, and are considered as an essential parameter to elucidate the protein-ligand binding (Yang et al. 2016). Lafutidine, a commonly used drug for the treatment of gastresophageal reflux disease (GERD) was tested for interaction with human serum albumin (HSA) by Yang et al. Authors also report usage of other NMR methods (STD-NMR and Water- LOGSY, discussed later in this chapter) to validate this binding (Yang et al. 2016). Importantly, the relaxation time depends on the dynamics of molecule, and can be correlated to NOE values. Most of the times, the ligand molecules/small molecules are partially or fully insoluble in water, hence they are dissolved in DMSO leading to a minute amount of DMSO in final protein-ligand solution. Interestingly, addition of DMSO leads to change in the viscosity of the final solution and conformational change of the protein that finally leads to the change in the dynamics of the protein side chain (Chakraborty et al. 2012). Thus, is it is important to consider the change bought about in relaxation dynamics and correlate with the rotational diffusion coefficient (τ_c) (Wallerstein and Akke 2019). In so far, conformational dynamics was thought to be informative in enzyme catalysis only, but a recent work by Liu et al. describes the use of relaxation parameters to study ligand binding. The study focuses on the differential binding of Tiam1 PDZ domain and an engineered QM PDZ domain (QM, quadruple mutant) toward a small peptide. The relaxation parameters were shown to contribute significantly towards understanding of thermodynamics and binding specificity (Liu et al. 2019; Liu et al. 2016).

9.2.4 Exchange Spectroscopy (EXSY)

In a Nuclear Overhauser Effect Spectroscopy (NOESY) experiment recorded with small ligand in presence of protein, the correlation time is increased due to binding and relatively slow tumbling of bound molecules relative to unbound molecules, a strong negative NOESY peaks is detected (Post 2003). Otherwise, in absence of protein, the small molecules experience unhindered motions and have very weak positive values. This method has been successfully used to screen the binding partner from a compound library with multiple lead compounds at FragmenTech NMR screening Centre (Aguirre et al. 2015; Barelier and Krimm 2011). In order to account for competitive binding for a single binding pocket the inter ligand NOEs were considered and termed as INPHARMA (Inter- Ligand NOEs for

PHARmacophore Mapping) (Bartoschek et al. 2010; Orts et al. 2009; Sánchez-Pedregal et al. 2005). Whereas, in case if two binding pockets are existing on a single molecule, the NOEs are termed as ILOEs (Inter-Ligand NOEs) (Li et al. 1999; Rega et al. 2011). These two methods are regularly used for Fragment-Based screening and structural Investigation of protein-ligand complexes by Isabelle Krimm and coworkers (Aguirre et al. 2015; Ahmed-Belkacem et al. 2016).

ZZ- exchange is another form of longitudinal relaxation based on multi- dimensional correlation spectroscopy that can prominently characterize the internal motions in a molecule that arise because of inter-conversion and auto- relaxation between conformational states. The time frame under observation is milli-second to sub-second within slow exchange regime. ZZ- exchange method has its own unique features like enhanced sensitivity towards slower molecular motions relative to R_2 dispersion, because of perturbation of longitudinal magnetization for extended duration during experimental relaxation delays due to slow internal motions. Moreover, this method does not require large induced chemical shift perturbations resulting from ligand binding, as the longitudinal relaxation remains unaffected by chemical shifts. In this method, the sample is made such that the half of population remains unbound, hence, the longitudinal relaxation profiles of free form, complex form and two exchange peaks (free- bound and bound- free) are analyzed, simultaneously. 2D ZZ-exchange experiments show correlation between ground state (on diagonal) and excited state (off- diagonal). This spectrum when acquired using different mixing times result in diminished cross peaks due to relaxation phenomenon. Using this time dependence, differential population dynamics and exchange rates (k_{ex}) can be calculated as suggested by Boswell et al. Hence, ZZ- exchange could be used to study the rigid protein molecules showing a small change in the chemical shifts (conformations) after ligand binding (Boswell and Latham 2018).

A temperature dependent ZZ- exchange can be performed to deduce thermodynamic parameters like changes in free energy (ΔG), enthalpy (ΔH), entropy (ΔS), off rate (k_{off}) directly using Eyring equation (Eyring 1935). AMP- activated protein kinase (AMPK) is an important protein for eukaryotes and has interesting role in energy metabolism (Gooley et al. 2018). The β-subunit of this protein exists in two isoforms (β1 & β2) that show differential glycogen binding. To understand this differential binding kinetics, infer thermodynamic parameters and determine affinity of β1 & β2 isoforms Mobbs et al. used ZZ-exchange experiments (Mobbs et al. 2015).

9.2.5 Saturation Transfer Difference NMR (STD NMR)

STD NMR is a method of choice in recent years for screening the compound libraries, where the molecules show weak binding and are in a fast exchange regime. In this method a selected proton region of ^{1}H spectrum is saturated using the spin diffusion technique followed by subsequent transfer of magnetization to ligand from the protein through inter-molecular NOEs. Then selectively saturated spectrum of protein with signal (I_{SAT}) is subtracted from the spectrum without protein saturation having signal intensity (I_0) (Cala et al. 2014; Cala and Krimm 2015; Viegas et al. 2011). The difference spectrum ($I_{STD} = I_0 - I_{SAT}$) obtained, denotes only the signal from the bound ligand that has transferred magnetization. By subtracting *on-resonance* and *off-resonance* spectrum, the signal from unbound molecules or any other impurity is nullified due to its presence in both the spectrum. In this approach, the magnetization transfer is proportional to inter molecular distance (< 5 Å) between the protein and the ligand, hence a strongly bound ligand will show a strong signal as compared to other molecules (Cala et al. 2014; Yang et al. 2016).

This method can also be employed to discern the binding kinetics and understand differential binding of a molecule to proteins. Viegas et al. modified and performed STD experiment from earlier published work of Wang et al. (Viegas et al. 2011; Wang et al. 2004) to distinguish binding of 6-CH_3- tryptophan, and 7-CH_3-tryptophan to Human Serum Albumin (HSA). The spectrum recorded for individual ligands and the mixture was interpreted for the ligand binding studies (Fig. 9.3). Intense signal from 6-CH_3- tryptophan as compared to 7-CH_3- Trp, is in agreement to the hypothesis that 6-CH_3- tryptophan ($K_d = 37\mu M$) is strong binder in comparison to 7-CH_3- tryptophan (non-binder). Moving ahead STD-NMR has also been used for Differential Epitope Mapping and has been abbreviated as (DEEP-STD) NMR by Nepravishta et al. (Nepravishta et al. 2019). In this study, authors characterized the structure and pharmacophore having weak binding without use of chemical shift assignments.

9.2.6 Water-LOGSY

Water-LOGSY is a standard acronym for water-ligand observed via gradient spectroscopy, and is typically used for performing high throughput ligand screening (Cala et al. 2014; Yang et al. 2016). The ligand molecules are in excess to the protein hence, the concentration required for the ligand molecule is considerably higher i.e., in milli-molar range. Like STD-NMR, this method can also characterize the interaction in fast exchange regime, while the basic difference lies in the pathway followed for magnetization transfer (Raingeval et al. 2019; Yang et al. 2016). Unlike STD-NMR, magnetization transfer is through the transiently bound water molecules around the protein to the ligand. The spectrum obtained is measure of average weighted population in terms of bound and unbound states, wherein positive signals arise from interacting protons and negative signals from non-interacting protons (Cala et al. 2014; Raingeval et al. 2019).

Understandably, the bulk magnetization from water molecule can be transferred via three different routes to the protons. These routes can be, (a) direct transfer of magnetization from water to protons in binding site, (b) chemical exchange among the excited proton of water, and (c) easily exchangeable protein protons or through inter molecular dipole-dipole cross relaxation of the complex, and (d) transfer from water trapped in hydrophilic cavities of protein (Huang and Leung 2019; Raingeval

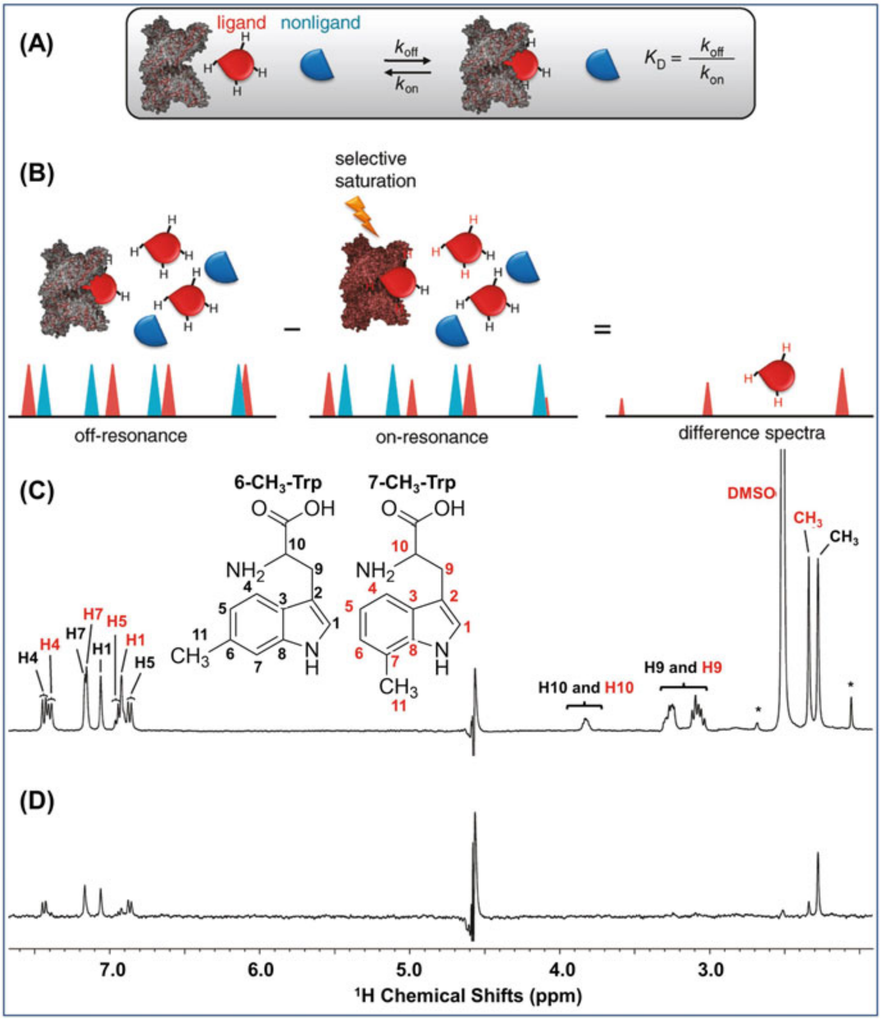


Fig. 9.3 Sequential procedural outline for STD-NMR based ligand screening experiment (A) Depiction of the exchange occurring between free and complex form. (B) Scheme of the STD- NMR experiment. (C) Reference spectrum of the mixture of molecules (6-CH3- Tryptophan and 7-CH3-Tryptophan). (D) STD-NMR spectrum. The Asterisk (*) represent impurities in spectra of mixture that are absent in STD- NMR spectrum. [Adapted with permission from Viegas et al. (Viegas et al. 2011), Copyright (2011) American Chemical Society]

et al. 2019). Out of all the four mechanisms listed, the magnetization transfer by chemical exchange between labile protons, results in “false positives’, thereby reducing the precision of technique. However, anomalies could be dealt by recording a spectrum of ligand(s) in absence of protein.

While demonstrating the particular benefit of the water –LOGSY, Geist et al. illustrated the use of fragment based drug designing approach to screen the potential ligands (Geist et al. 2017). The authors could determine the binding modes with respect to binding residues of low affinity binders by LOGSY titration method. They validated use of LOGSY- titration by studying the binding of two ligands (6-(3,4-Dimethoxyphenyl)-3-methyl [1,2,4] triazolo[4,3-b]- pyridazine and 1-Methyl-5-phenoxy-1H-pyrazole-4-carboxylic Acid Ethyl Ester) with bromodomain1 of bromodomain containing protein 4 (Brd4-BD1)(Geist et al. 2017). At times, use of Water- LOGSY experiment may lead to low sensitivity and false positives due to aggregation or denaturation of protein (Chappuis et al. 2015). Hence, hyperpolarized water could be used to increase the sensitivity of water-LOGSY experiment tremendously. Addition of hyperpolarized water can significantly reduce the span of reaction up to few seconds contrary to around 30 minutes, required without hyperpolarization (Chappuis et al. 2015). Moreover, this method can be used to verify protein sample viability and enhance proton signal of protein in low concentrations (around 20μM) in a short span of time.

9.2.7 Translational Diffusion Measurements

Inherently molecules tend to tumble in solution; this tumbling motion relies on the hydro- dynamicity of the molecule, oligomerization tendency, size, viscosity of the medium etc. The diffusion coefficient derived using Stokes Einstein relationship, suggests that any change in the size of the molecule will result in altered diffusion rates/coefficients (Cohen et al. 2005). Hence, protein-ligand interaction, if any, can be determined by the application of Stokes-Einstein equation (Gordon and Perugini 2016). Ability to distinguish subtle changes in physical properties like shape and size due to aggregation/binding, makes this technique powerful enough to resolve two different populations.

Certain proteins have enough instability to undergo self oligomerization. These oligomers have characteristic properties i.e. the molecular weight of oligomer is integral multiple of its monomer. This oligomerization tendency can be fatal for the organisms and result in impaired biomolecular responses. Ragona et al. by use of diffusion NMR (DOSY) characterized the binding of rhodamine 6G with fibroin protein. The diffusion coefficient under observation during the course of reaction suggests that rhodamine 6G inhibits the self-oligomerization of fibroin protein (Ragona et al. 2018). Hence, this approach can be further extended to other protein systems for understanding oligomer formation upon ligand binding or vice-versa, as per the desirable parameter under study. Also, Snyder et al. applied DOSY for screening a ligand library abbreviated as CoLD-CoP (Cluster of Ligand Diffusion Coefficient Pairs). This CoLD-CoP was used complementarily with DOSY to find N-acetyl-glucosamine and tryptamine as suitable binder for lysozyme and tartrate as binder of wheat germ acid phosphatase (Snyder et al. 2015). This method was also used to characterize ligand binding for neurotensin1 which is an important receptor of human central nervous system (Huber et al. 2017).

9.2.8 Merits and Demerits of NMR Spectroscopy

In general, NMR provides information about the interactions occurring in slow and fast time scale regime. However, interactions following intermediate exchange are cumbersome to characterize through NMR. Using NMR Spectroscopy, binding site, structural details, thermodynamic parameters, stability aspects, kinetic parameters, and processes like allostery/cooperativity can be studied in detail. Further the experiments involving the NOE measurements, can even ascertain the binding moieties at atomic level. Overall NMR with its versatility and multidimensional methods can prove highly beneficial for studying protein-ligand interactions. Though NMR is capable of characterizing a majority of the protein-ligand interactions, the experimental setup (instrument cost, isotope labeling and sample preparation) and the expertise (sample handling, data acquisition and analysis), makes it a tedious process. Different methods and the flexible experiments allow for the use of this technique for understanding a wide variety of interactions and associated parameters. Protein Sample concentration requirements is also a bottleneck for NMR measurements.

9.3 Surface Plasmon Resonance

Surface Plasmon Resonance (SPR) has differential optical refraction as its underlying principle. It was first developed in 1980s, and since then used for quantification of bimolecular interaction (Kooyman et al. 1990; Schuetz et al. 2017; Szabo et al. 1995). The technique gained popularity due to real time rate kinetics determination, high sensitivity and label free detection. Precise detection of analyte binding for a wide variety of molecules like proteins, small molecules, ions etc.; made it a technique of choice among pharmaceutical stalwarts. Recently, SPR is being used to derive information pertaining to thermodynamics, analytical stoichiometry, lead identification and kinetic profiling (Masson 2017; Schuetz et al. 2017). Thus, the information obtained by this technique is quite convincing to understand certain complex interactions at low concentration of analytes.

9.3.1 Theoretical Aspects of Surface Plasmon Resonance

Briefly, the instrument utilizes the basic phenomenon of change in refractive index and total internal reflection (TIR) (Yang et al. 2018). The surface of detection chip (linker layer) is mostly fabricated by binding matrix that can vary as per the end user requirement. Next to the linker layer is gold layer coated over the glass support (Fig. 9.4). SPR in phenomenon occurs when a plane polarized light is incident on gold layer (sandwiched between glass and the buffer), under TIR condition, electrons are released as evanescent energy wave of oscillating plasmon. Plasmon contributes to the electric field in gold surface due to their natural charge. Any binding interaction thereafter, undoubtedly causes angular shift (inset Fig. 9.4a) of reflected light whilst change in refractive index, subsequently affecting the

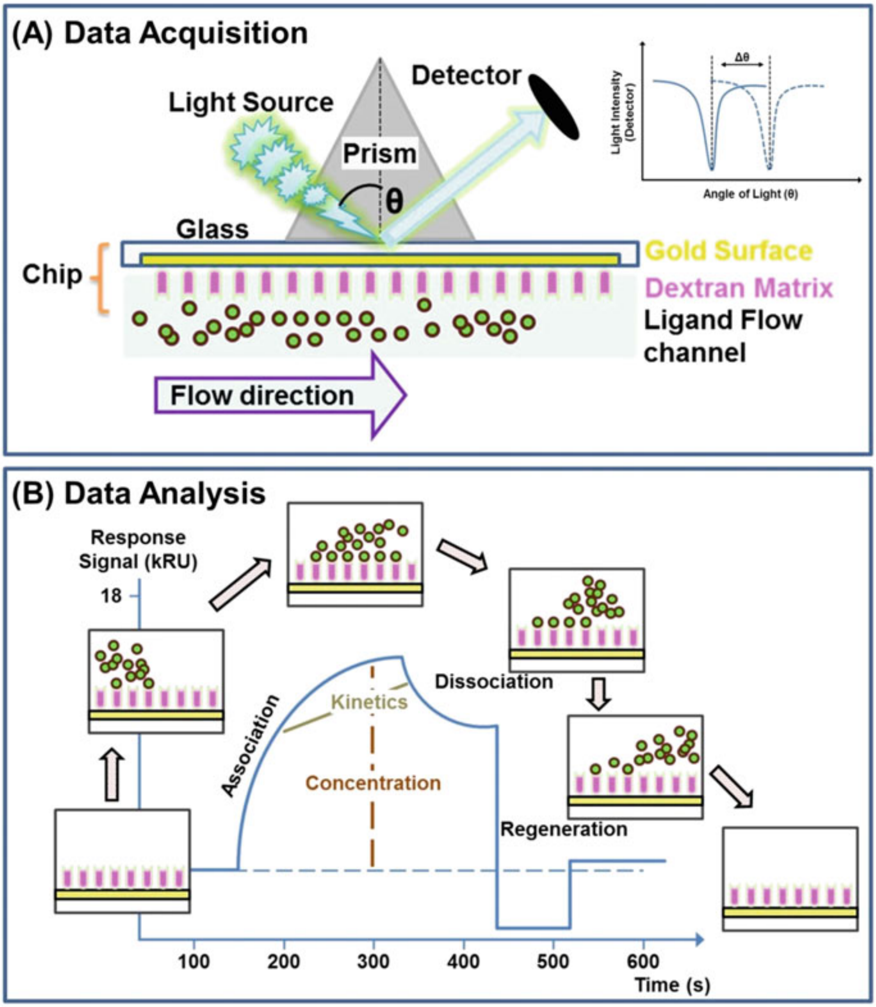


Fig. 9.4 Illustrated work flow of Surface Plasmon Resonance technique: (**a**) Schematic of internal optics arrangement; (**b**) Theoretical sensorgram depicting the major phases; (i) Buffer injection, (ii) Association, (iii) Dissociation, and (iv) Regeneration of the matrix

detectable electric field (Korshunova et al. 2019; Nguyen et al. 2015). The SPR signal is interpreted as response unit (RU) which relates to the bound mass of 1 picogram per millimeter square for 1 response unit and angular shift of 10^{-4} degrees. The functional curve obtained as function of response unit with time is denoted as sensorgram. Typically, a sensorgram has four regions (Fig. 9.4b): (i) Buffer injection phase, (ii) Association Phase, (iii) Dissociation Phase and (iv) Regeneration Phase.

A ligand binding reaction can be dissected as initial phase of buffer injection, when only buffer is introduced in system to get baseline. Next phase is the injection of the analyte; during this phase analyte binds to the ligand and results in analyte-ligand complex formation that increases response units. At the equilibrium condition, when steady state is reached, the rate of association and dissociation is equal. Subsequently, the molecules dissociate which is observed as decrease in response unit (RU). Finally, the chip can be regenerated by stripping off the probe restoring the response to baseline (Nguyen et al. 2015). The information derived from SPR sensorgram includes rate of reaction (including association, dissociation and equilibrium constant) and binding stoichiometry for (semi)-qualitative calculation of kinetics (Juhász et al. 2016; Nico and Fischer 2010). Precise information regarding underlying dynamics of biological systems, makes it indispensible for pharmaceutical industry, for selection and designing therapeutic as well as active pharmaceutical ingredients (API).

With recent developments in the surface modification technologies, matrix layers that are available presently are dextran (various concentrations), Histidine, Liposomes, Streptavidin, Gold, Carboxymethyl, 3D coating of hydrogel etc. (Table 9.2). Hydrogels are preferable now a days because of their binding specificity, improved binding capacity and flexible coupling mechanism (Li et al. 2015; Wang et al. 2010). By use of coating techniques, like amine linking and other covalent linking methods, linker is attached to gold surface. Further in a typical experiment the ligand remains immobilized on the linker, while the binding partner (analyte) flow over the surface along with the buffer. Thus, any binding interaction changes the mass and the refractive index, thereby producing a signal. Generally, for protein-ligand interaction the change in refractive index is correlated linearly to the number of sites occupied by the ligand molecules. At steady state binding condition, when all the binding sites are occupied, the response units (RU) are determined by:

$$\mathrm{RU_{max}} = \mathrm{nRU_L}(\mathrm{MW_A}/\mathrm{MW_L})$$

Where n is number of binding site(s) on ligand, $\mathrm{MW_A}$ and $\mathrm{MW_L}$ denoting molecular weight of analyte and ligand, respectively (Perspicace et al. 2009).

SPR has inherent advantage over other techniques as it enables real time determination of reaction kinetics with least dependence on molecule labeling. It can characterize molecular binding in sub micro-molar concentrations ranging between 10^{-6}μM and 10μM (Van Der Merwe 2001). Though it has simple sample preparation procedure and time required for analysis is very less, the technique cannot measure k_{on} values faster than ~10^6 m^{-1} s^{-1} due to mass transport limitation (Van

Table 9.2 Summary of the sensor chips available till date with brief description regarding their characteristics and applications

Sensor chip	Surface modification	Characteristics and applications
CM5	Most versatile dextran based chip. For immobilization of molecules via –NH2, –SH, –CHO, –OH or –COOH groups	Excellent stability, versatile chip, suitable for most of the applications
CM4	Similar to CM5 with lower charge due to low carboxylation	Suitable to address background binding. Can be used to study crude samples for kinetic application.
CM3	Similar to CM5, but with reduced length of coated dextran	Improved sensitivity while studying large molecules as the binding occurs near to chip surface
CM7	Similar to CM5, but with more density of dextran	Has higher immobilization capacity with high signal- to –noise ratio. Suitable for small molecules
C1	Flat carboxy methylated surface	Used for large/ multivalent interacting partner where dextran matrix is undesirable.
SA	Dextran matrix with pre-immobilized streptavidin	Can capture biotinylated ligands (carbohydrates, proteins, peptides, DNA fragments etc.)
NTA	Dextran matrix is pre-immobilized with NTA (Nitrilotriacetic acid)	Used for capturing Histidine tagged molecules, generally proteins. Can be reused and recycled. Can be used for steric orientation
HPA	Hydrophobic surface consisting of long chain alkane thiol molecules attached to gold surface	For studying membrane associated interactions
L1	Carboxy methylated dextran matrix modified with lipophilic substance	For capturing liposomes with retention of lipid bilayer structure
Au	Untreated gold surface	For creating novel and unique surface chemistries
PEG	Uses standard EDC/NHS chemistry for PEG immobilization	Alternative chip for ligands that non-specifically binds to dextran matrix. Can be used for multivalent large molecules
Protein A	Dextran matrix with MabSelect SuRe ligand on surface	For site directed capture of antibody. Binds specifically to heavy chain within fc region.
Protein G	Dextran matrix pre-immobilized with protein G (GammaBind G, type-2)	Orientation specific binding of heavy chain within fc region
Protein L	Dextran matrix pre-immobilized with protein L	Captures a wide range of antibody fragments, like fab, single chain variable fragments (scFv), kappa light chain subtypes (1,3 and 4)

The description and the salient features of the chip(s) are as per the catalogue from GE Healthcare, United States

Der Merwe 2001). SPR has unique feature to calculate binding affinity using Van't hoff analysis with precise temperature control and reproducibility. A substantial contemplation is required while immobilization of protein on chip because proteins often undergo conformations and rotational/translational entropy changes. But, the major concern to ponder upon is ability to distinguish specific interactions among non-specific interactions. Thus, to differentiate former from latter combined use of SPR with another detection technique like Surface Enhanced Raman Scattering (SERS) with suitable signal enhancers like Nile blue, pyridine, silver nanoparticles etc. (Deganutti et al. 2017; Dror et al. 2011; Frederick et al. 2007; Pan et al. 2013) is suggested (Radić et al. 1997).

9.3.2 Probing Protein-Ligand Interactions Using SPR

SPR technique has been used to characterize protein- protein interaction (Douzi 2017), protein- ligand interaction (Douzi 2017; Du et al. 2016), DNA-RNA hybridization (Diao et al. 2018; Wang et al. 2016) and other bi-molecular interactions (Ding et al. 2017; Kaur et al. 2016) establishing it as a versatile technique. For example, GPCR17 is an important GPCR receptor, responsible for maturation of oligodendrocyte precursor and myelinating ability. Conventionally, the state of GPCR (activated or inhibited) is determined by radiolabelled assay using [^{35}S] GTPγS. However, this assay does not provide any information about underlying molecular interaction. Thus, two engineered variants of GPR17 were immobilized on chip to measure the direct binding with two ligands Cangrelor (an antagonist) and Asinex1 (an agonist) (Capelli et al. 2019). Importantly, perseverance of the native fold for the immobilized protein was confirmed using combined data from X-ray crystallization study and Molecular dynamics study.

Another important interaction put forth from perspective of ligand binding is formation of the transient intermediates (Zhan et al. 2016). These intermediates are often point of inflection where a critical energy threshold is required to reach the final state. These transition states are short-lived phases where the ligand molecule is just temporarily bound. There exists different entropic and enthalpic components of the transition state that a ligand must overcome to undergo (un)-binding, culminating into a final state. Mechanistically, set of driving forces that dictate the net energy change during an event of binding and unbinding are desolvation (Dror et al. 2011; Pan et al. 2013), conformational entropy loss (Frederick et al. 2007) and intermolecular interactions (Radić et al. 1997; Schmidtke et al. 2011) to name a few and are unique for a particular ligand. Thus, with an aim to understand how the interactions with similar affinities can have different kinetics, Deganutti et al., studied the impact of solvation and desolvation energy on transition state properties of adenosine A_{2A} ligand binding kinetics (Deganutti et al. 2017). The authors co-related the enthalpic and entropic changes for six different ligands to their *on* and *off* rates obtained through SPR to understand the transition state energetics of the interaction. Another work by Gohlke et al. elucidates the use of SPR in addition to

NMR spectroscopy, ITC and computational modeling to find promising compounds for inhibition of KRas signaling; thus, yielding a critical starting point to decrease KRas dependent oncogenesis. KRas is member of Ras GTPase family and has been linked to over 20% of human cancers. These biophysical techniques (SPR, ITC and NMR) were the central drivers for establishment of Structure- Activity Relationship (SAR) and to study binding affinities for interaction of small molecules with KRas (Gohlke et al. 2018).

Recently, a breakthrough discovery in the form of instrument optics led to multiplexed analysis by the conversion of collimated light beam into a line shaped beam. The new angular scanning adjustment was named as Kretschmann configuration (Hatta et al. 1984; Yang et al. 2018). Due to this configuration the scan area could be reduced to 1 × 10 mm and full SPR curve could be acquired in 3.6 s, thus making the system amenable to multi-plexing and high throughput with low input. The authors could detect 16 samples simultaneously; and have tried immobilization of several proteins on chip and their antibody binding, as a proof of concept showing excellent reproducibility in a multiplex regime (Lakayan et al. 2019). Further SPR has been used to study ligand (un)- binding to the proteins (Chen et al. 2019; Drescher et al. 2018; Šakanovič et al. 2019; Sawant et al. 2016; Taghipour et al. 2018), and correlated with the targeted molecular dynamics simulations (Taghipour et al. 2018; Wolf et al. 2019), to develop a suitable computational binding model. The technique has also found its application in studies related to cellular receptors (Dorion-Thibaudeau et al. 2016; Schasfoort et al. 2018; Shanehbandi et al. 2017; Winter et al. 2018) and involving multiplexing (Day and Korolik 2018; Geuijen et al. 2017; Kumar 2017).

9.3.3 Merits and Demerits of SPR

SPR technique with the advanced instrumental work flow enables researcher to carry out cumbersome assays with enhanced precision, accuracy and reproducibility. The robotic sample insertion mechanism and the user friendly analysis platform provide a distinct edge. Increased price per reaction due to use of specialized chips (fabricated with gold and/or polymers) is a major concern for the small scale institutions. However, the technique finds its major application in industrial setup for API production, monoclonal antibody production etc. Certainly, SPR has its own merits but lately techniques like Bio-layer inferometry (BLI) and Microscale Thermophoresis (MST) are picking up popularity due to high cost of SPR setup and sensor chips. Overall SPR has tremendous potential and worth beyond its weight in gold, for understanding the kinetics and dynamic binding paradigm of protein-ligand interactions.

9.4 Isothermal Titration Calorimetry

For drug discovery and identification of lead molecules, understanding of thermodynamics and their binding kinetics is necessary. The extent/binding capacity determine the feasibility and spontaneity of the interaction providing the underlying insights. Hence, to characterize the binding capacity of small molecules as well as macromolecular assemblies to their interactive-counterparts, Isothermal Titration Calorimetry (ITC) is an excellent technique (Falconer 2016; Hansen et al. 2016). Owing to high-throughput results with enhanced precision, robustness, and unambiguity, this label independent technique is a tool for general application in commercial setup. ITC is a technique that can characterize real-time interactions and distinguish specific and non-specific binding to discern structure-function relationships with no limitation to optics and molecular weight of molecules (Du et al. 2016).

9.4.1 Theoretical Aspects of Isothermal Titration Calorimetry

The analytical chamber of ITC comprises of adiabatic lining outside enveloping two identical cells known as sample cell and reference cell made up of material with low resistivity (Fig. 9.5a). Both the cells are maintained at identical state and a ligand (L) binding event to a receptor protein (P) leads to change in the heat of the system. The change in heat energy can be exothermic or endothermic in nature. The amount of energy required to nullify the change in heat energy is recorded as change in Gibbs free energy for the complex (PL) formation that can be used to study the binding kinetics (Du et al. 2016) as follows:

$$\mathrm{P} + \mathrm{L} \rightleftharpoons \mathrm{PL}$$

$$\Delta \mathrm{G}_{\mathrm{binding}} = -\mathrm{RTln}(\mathrm{K_a})$$

Where $\Delta G_{\text{binding}}$ is the free energy change of the binding, R is the gas constant (8.314 J mol^{-1} K^{-1}), T is the absolute temperature and K_a is the association constant. The dissociation constant (K_d) is just the reciprocal of equilibrium constant. For a reaction that has equi-molar concentration of both the reactants, the equation can be the dissociation/association constant can be understood as:

$$\mathrm{K_a} = [\mathrm{PL}]/[\mathrm{P}][\mathrm{L}] \text{ or } \mathrm{Kd} = [\mathrm{P}][\mathrm{L}]/[\mathrm{PL}]$$

Additionally, for a binary complex $\Delta G_{\text{binding}}$, can also be given by the change in enthalpy (ΔH) and entropy(ΔS).

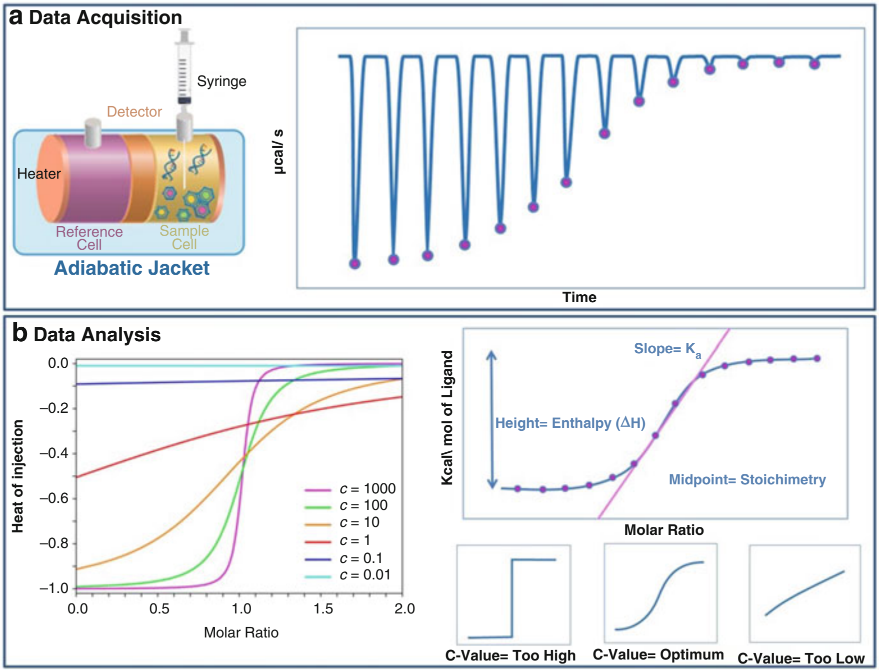


Fig. 9.5 Work flow diagram of Isothermal Calorimetry (**a**) Schematic depiction of ITC device and the raw data obtained. (**b**) Systematic data interpretation of ITC thermogram, and the relevance of data with respect to c-value

$$\Delta G_{\mathrm{binding}} = \Delta \mathrm{H} - \Delta \mathrm{S}$$

Hence, another relation could be derived that is:

$$-\mathrm{RTln}(\mathrm{K_a}) = \Delta \mathrm{H} - \Delta \mathrm{S}$$

Heat change being phenomenal to any chemical reaction could be measured by ITC in the form of ΔG, ΔS, ΔH, $\mathrm{K_a}$, and the stoichiometry (n) in a single experiment at one given temperature. The heat capacity change ΔC_p can be determined by performing measurement in varying temperatures. The heat capacity ΔC_p is defined as the amount of heat needed for a temperature change of 1 K in the system. It is denoted as the partial derivative of change in enthalpy ΔH° relative to the temperature T keeping the pressure constant.

$$\Delta Cp = \left(d\Delta \mathrm{H}^{\circ}/\mathrm{dT}\right)\mathrm{p}$$

Thus, utilizing above equation, estimation of ΔH (i.e. measure of displaced water molecules) while complex formation at various temperatures is necessary. By sequential addition of ligand into the protein solution, stoichiometry could be calculated in a step wise manner (Falconer 2016). Initially after few titrations, the binding sites get occupied while giving the measurement for ΔH. Later on, after saturation, other parameters like binding constant and stoichiometry are inferred. At the end of the titration, background heat of dilution parameter is obtained to account for the change in energy due to addition of ligand at every titration step. The experimental data obtained from ITC is a thermogram, which consists of peaks corresponding to the power consumption to maintain a zero temperature difference between the sample cell and the reference cell, after each ligand injection. Hence, to obtain value for total heat associated with reaction, the peaks can be integrated. Subsequently interpreted data is fitted into a suitable binding model as per the pre-conceived idea of binding propensity or as per the statistical analysis to obtain the best fit. Viability of the results obtained by thermogram(s) is ascertained by the c-value (Rühmann et al. 2015). c- value is described as:

$$\mathrm{c} = \mathrm{nK_aM_{tot}}$$

Where n describes the stoichiometry, K_a refers to the association constant and M_{tot} (mol L^{-1}) to the total concentration of the macromolecule. Practically, the c-value between 10 and 500 are considered good for evaluation purpose due to their sigmoidal nature. Values above 500 seem rectangular and below ten results in a flat curve (Fig. 9.5b), these values are not reliable as there is no distinct point of inflection (Rühmann et al. 2015). This limitation can be overcome by displacement titration strategy to get reliable results for high as well as low affinity binders. Another source of error is the buffer mismatch, reactions with incompatible buffer leads to false positives by means of excessive/less heat exchange. Hence, a due consideration towards buffer compatibility is required.

9.4.2 Applications of ITC

Since 1990s when calorimeter was commercially used by Wiseman et al., there were discrepancies while correlating thermodynamics data to structural data of the complexes (Wiseman et al. 1989). Thus, to enhance the limit of detection, several developments in hardware, software and data analysis techniques were performed. Several attempts were made to bring together data of different interaction in the form of repositories (Chen et al. 2002; Liu et al. 2007; Olsson et al. 2008). A landmark development in this regard is the development of kinITC (Dumas et al. 2016) method for measuring the kinetics of the reaction that progresses in multiple steps like RNA folding (Burnouf et al. 2012). Further, interaction as crucial as zinc binding to a histidine residues of a polymer was characterized with precision (Enke et al. 2017). In this recent decade, ITC has been invariably used to characterize interaction between protein and small molecules illustrating diverse application of the technique. For example, Cardiac Troponin C is a Ca^{2+} and Mg^{2+} binding protein responsible for myocardial force production and hence, the binding kinetics related to these divalent ions is important. Thus, to characterize the sequential binding of Mg^{2+} after interaction of Ca^{2+} and establish the impact of Ca^{2+} binding on heart contractility, ITC was used (Rayani et al. 2018).

Use of ITC for studying interaction of warfarin (anti- coagulant) with HSA and BSA, shows versatility of this technique by excluding the dependence of binding on ionic strength and buffer agents. Herein, the authors also show that BSA- warfarin interaction is not a preferable model for Protein-ligand interaction, due to involvement of more than one binding site (Ràfols et al. 2018). In the sequence of events, recently Panov et al. showed the use of ITC methods to decipher the binding site of PMR-116 and CX-5461 molecules to the RNA polymerase 1 (RNAP1). The levels of rRNA synthesis are dictated by the transcription by RNAP1, hence inhibitors specific to RNAP1, shall emerge as a novel anti- cancer approach (Panov and Alsahafi 2019). With an understanding towards the unexplored role of ITC towards different types of binding and limitations encountered so far, Dumas et al. developed new mathematical methods that can forego the effect of the dilutions (Dumas et al. 2016; Meyer et al. 2019). Typically used methods for the titration studies are Single injection method (SIM) and Multiple injection method (MIM), in case of MIM, the sample volume is increased considerably. Thus, to account for the heat of dilution a parallel with the concept of gauge invariance is proposed (Dumas et al. 2016).

Simultaneously, there were instances of platform upgrade that enable researchers to understand conventional isotherms. Most of the models do not account for the hydrophobic interactions that are main guiding force for reactions as well as aggregation. AFFI-NImeter platform, was developed for studying supra- molecular complexes using experimental data (Piñeiro et al. 2019). Subsequently, a new software ICITC (Buurma and Haq 2007; Buurma and Haq 2008) was developed by saeed et al. to analyze complexes with coupled equilibrium, using the phenomenon of mass balance equation and simulated annealing (Kirkpatrick 1984). Another upgraded version of ICITC with user- friendly graphical user interface has also been developed and available as I2CITC (Saeed and Buurma 2019). Until recently, ITC was only used to characterize different type of protein-protein interactions (Jaiswal et al. 2019; Santofimia-Castaño et al. 2018), Proteins with DNA (Amano et al. 2019), RNA and small molecules (Umuhire Juru et al. 2019; Vogel and Suess 2016); but by considering nano-particles as small molecules, Prozeller et al. quantified the adsorption of protein molecules on the surface of nanoparticles (Prozeller et al. 2019).

9.4.3 Merits and Demerits of ITC

Although the merits of ITC are enormous in terms of experimental excellence, versatility in terms of type of interactions characterized and minimal sample dependency, still the biggest limitation observed is in terms of high throughput. Traditionally, ITC is considered time consuming, laborious and low throughput, hence its use is confined to validation studies. With recent advent of new techniques and tools like array based nano-calorimeters and automated robotic instruments, this technique is making its place on arena of high throughput screening.

9.5 Fluorescence Spectroscopy

Fluorescence spectroscopy is a very sensitive technique and relies on small sample amount for characterization of Protein- ligand interaction. The emergence of advanced techniques like Förster Energy transfer (FRET), Fluorescence Life-time Scanning (FLS), and Fluorescence quenching have enabled researchers to understand the ensemble change in molecules relative to changes in micro-environment of fluorophore (Ghatak et al. 2015; Kyrychenko 2015; Nag et al. 2006; Vargas-Uribe et al. 2015). Binding of a ligand to the protein or vice-versa can result in change of emission spectral pattern. The changes can be either associated to spectral intensities or spectral shifts. The intensity enhancement of the spectrum is known as hyperchromic shift, whereas decline in intensity is called as hypochromic shift. Spectral shift towards shorter wavelength is known as hypsochromic shift/Blue shift, and towards higher wavelength is considered as bathochromic shift/Red shift) (Fig. 9.6). Owing to the fast response and relatively less expensive instrumentation, this technique can be used to design fluorescent reporter based sensors for chemo-sensing (Wong et al. 2017).

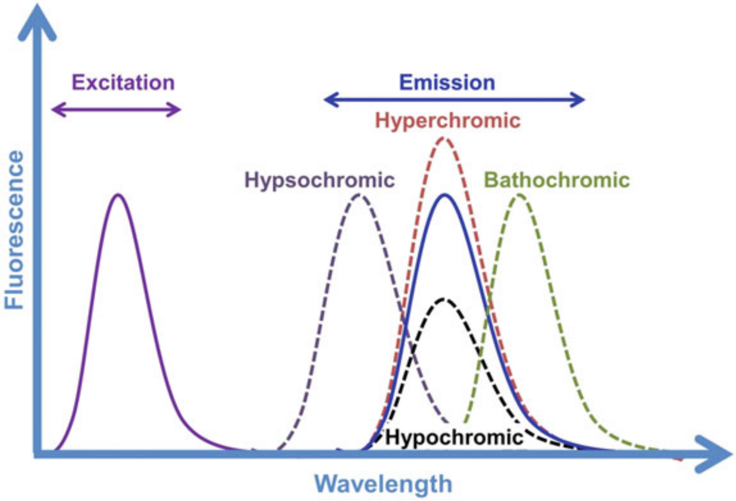


Fig. 9.6 Schematic diagram showing the various types of possible shifts in the wavelength due to protein-ligand interaction

9.5.1 Theoretical Aspects of Fluorescence Spectroscopy

Fluorescence is a type of luminescence where the incident photons excite the electrons to attain a higher vibrational energy state. This higher state is denoted as first singlet state (S1) and second singlet state (S2) as per their energy state (Fig. 9.7). There is a short time period for which this excited state persists, which is known as life-time of fluorophore, typically in range of 10^{-8}–10^{-4} s. Measurement of this particular time frame during which the electron is under transition is working principle for Fluorescence Lifetime Scanning (FLS). FLS measurements are based on time correlated single photon counting, hence, unaffected by variation in intensity. During the said transition, the electrons undergo non-radiative relaxation thereby releasing energy. This energy, if detected by a charge coupled diode, forms the basis for Fluorimetry (Burkhardt and Schwille 2006). The emitted energy is utilized to excite another fluorophore present in solution, and this forms the basis for Förster Energy transfer (FRET) and fluorophores are called FRET pairs (Fig. 9.8). For any two fluorophores to act as FRET pair, the pre-requisite is emission and excitation spectral overlap between the fluorophores. Recent advancement in this direction is use of zero mode waveguides (ZMWs). ZMW are photonic nanostructures that can create a highly confined optical observation volumes up to 20 zeptoLitres (zepto = 10^{-21}). This technique is being used for real-time biophysical characterization of individual molecules.

Fluorimetry can only be used for the sake of quick decisive binding analysis of protein with ligand on the basis of fluorescence quenching. The only pre-condition

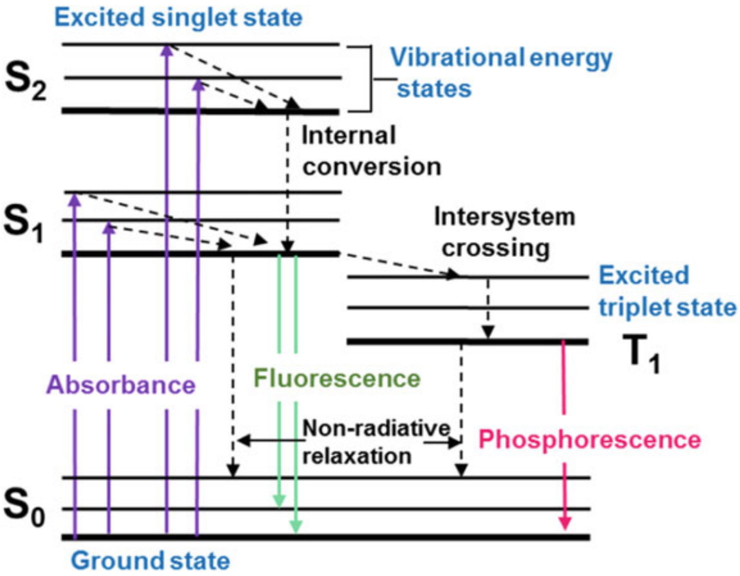


Fig. 9.7 Jablonski Diagram showing the permissible electronic transitions and the processes associated with them

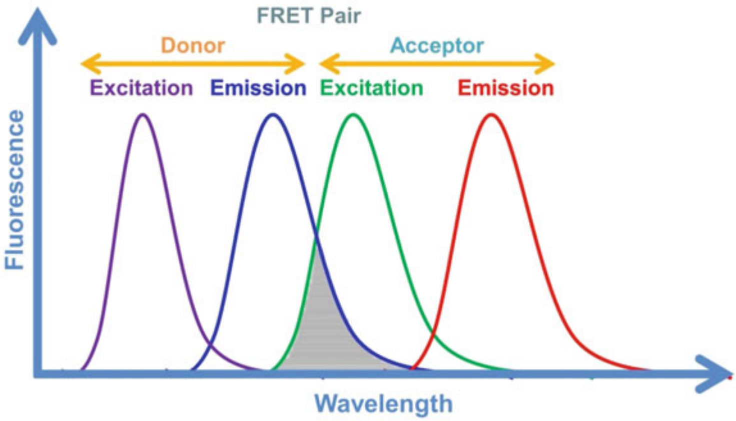


Fig. 9.8 Schematic diagram of Donor and Acceptor fluorophore showing the spectral overlap necessary for the FRET process

is that, either protein or ligand must be fluorescent and have considerable quantum yield (ratio of the photons emitted vs. photons absorbed). The underlying phenomenon(s) of fluorescence quenching are: ground state complex formation, molecular rearrangement, excited state interaction and energy state transition. Fluorescence quenching can be classified into static and dynamic/collisional quenching. Collisional quenching is outcome of diffusive event where fluorophore comes in contact with the quencher during its life-time and results in return of fluorophore to ground state without emission of photon. Contrary to this, in a static mode, an abrupt formation of stable non-fluorescent ground state results in permanent quenching. Albeit, there are also instances where both static and dynamic quenching is observed simultaneously. In such cases, normal Stern-Volmer plot assumes an upward curvature.

Both type of quenching (dynamic/static) rely on proximity of quencher and fluorophore in a micro-environment; hence numerous applications can be thought of for quenching technique. For instance, inaccessibility of a quencher to the fluorophore buried deep into the 3D fold of protein leads to no fluorescence signal. However, if somehow the tertiary structure is lost, the quenching process will occur forming the basis for the protein (un)-folding studies. Inherently, in case of proteins, presence of aromatic amino acids (tryptophan, tyrosine and phenyl alanine) serves as intrinsic fluorophores (Mocz and Ross 2013; Sarkar and Mishra 2018). Otherwise, additional external (extrinsic) fluorescent molecule could be appended to the protein under consideration to study the ensemble features of protein (Sarkar and Mishra 2018). Most commonly used intrinsic and extrinsic fluorophores have been listed in Table 9.3. Quenching data obtained from purely dynamic quenching can be analyzed using the Stern-Volmer relationship to obtain the quenching constant depicted as follows:

Table 9.3 List of fluorophores used regularly to probe protein-ligand interactions

Type of fluorophore	Mode of binding	Wavelength (nm)	Uses	References
Intrinsic fluorophores				
Tryptophan	–	Ex- 279, Em- 348	To study the structural changes and ligand binding, protein folding and elucidation of protein conformation changes	(Hevekerl et al. 2016; Plotnikova et al. 2016)
Tyrosine	–	Ex- 274, Em- 303	To study the structural changes and ligand binding, protein folding processess.	(Partlow et al. 2016; Raj et al. 2020)
Phenyl Alanine	–	Ex- 257, Em- 282	To study the structural changes and ligand binding.	(Hao et al. 2017; Zhang and Fitzpatrick 2016)
Extrinsic fluorophores				
Dansyl derivatives	Amine coupling	Ex- 336-340, Em- 520	Can be used to study binding site, and ligand binding where tryptophan is present at active site.	(Abe et al. 1999; Pecht et al. 1971)
5-Iodoacetamide fluorescein	Through thioether bond	Ex- 494 Em- 518	For assessment of ligand binding and structural modulation.	(Du and Strieter 2018; Johnson et al. 2019; Zare et al. 2016)
Fluorescein isothiocyanate (FITC)	Amine coupling	Ex- 494 Em- 518	To study protein-ligand interaction.	(Breen et al. 2016; Huang and Leung 2016)
Tetramethylrhodamine isothiocyanate (TRITC)	Amine coupling	Ex- 547 Em- 572	As a bio-analytical tool.	(Cengiz 2020)
Acrylodan	Through thioether bond	Solvent dependent	For probing ligand binding to the membrane receptors.	(Coates et al. 2020; Kivi et al. 2016; Sahni et al. 2017)

	Alexa Fluors	Amine coupling	Ex- 346-679 Em- 442-702	For probing ligand- receptor binding and structural changes.	(Croucher et al. 2006; Kozma et al. 2013)
	8-Anilinonnathaline-1-sulphonic acid (ANS)	Non- covalent binding	Ex- 380 Em- 470	To probe the conformational changes and hydrophobicity.	(Collini et al. 2003; Gasymov et al. 2008)
	6-p-Toluidino-2-naphthalene sulfonic acid (TNS)	Non- covalent binding	Ex- 318 Em- 442-702	To probe the conformational changes and hydrophobicity.	(Aziz et al. 2007; Lampe and Atkins 2006; McClure and Edelman 1966)

$$\frac{I_f^o}{I_f} = 1 + k_q \tau_0 [Q] \tag{9.1}$$

where I_f^o and I_f denotes the intensity of the flourophore in the absence and presence of quencher, respectively. The amount of quenching observed at a particular concentration [Q] is depicted by Quenching constant (k_q). For a fluorescent molecule the lifetime of the emissive state in presence of the quencher is denoted by τ_0. Further equation can be modified as follows on a double logarithmic scale:

$$\log\left(\frac{I_f^o - I_f}{I_f}\right) = n \ \log\,[Q] + \ \log Ka \tag{9.2}$$

Herein, n denotes the number of binding site that can be occupied by ligand and K_a corresponds to binding constant. Also in case of static quenching eq. 9.1 is reduced to:

$$\frac{I_f^o}{I_f} = 1 + k_s \ [Q] \tag{9.3}$$

Which is identical to the equation of dynamic quenching except for the fact the quenching constant is equivalent to association constant in static quenching.

9.5.2 Applications of Fluorescence Spectroscopy

Mostly the studies depicted by internal fluorophore like tryptophan(s) are preliminary, since due to multiple presences the exact residue undergoing interaction is unpredictable. Thus, the interaction must be confirmed using another robust technique. Recently, Sindrewicz et al. utilized tryptophan fluorescence to determine the strength of galectin-ligand interaction (Sindrewicz et al. 2019). Out of 12 human galectins, three representative members from each of the three sub-types were tested for affinity towards three carbohydrate molecules (galactose, N- acetyl- lactosamine and lactose). Seldom in absence of tryptophan, tyrosine can also be used as probe as depicted by Raj et al. while studying the binding interaction between 18β-Glycyrrhetinic acid with hup protein of *Helicobacter pylori* (Raj et al. 2020). In a recent report Winkie et al. showed the use of fluorescence spectroscopy to understand the recognition mechanism and catalytic role of methionine sulfoxide reductase enzyme in stereospecific reduction of oxidized methionine residues (Winkie et al. 2019). In absence of an internal fluorophore, an external fluorophores like Alexa Fluor, Dansyl derivatives, and ANS (8-Anilinonnathaline-1-sulphonic acid, a small naphthalene based organic molecule) can be used. Despite presence of internal fluorophore ANS can be used, incase internal fluorophore is buried or to determine surface properties. Binding of ANS results in fluorescence, indicating the presence of a hydrophobic pocket and scope for utilizing naphthalene based compounds as potent inhibitors (Eling and DiAugustine 1971; Gulati et al. 2019; Joshi et al. 1989). Similarly, ANS has been used to study the conformational switching of Calcium binding protein of *Entameoba histolytica* from "closed" to "open" state up on ion binding (Mukherjee et al. 2007). Recent studies have reported the binding kinetics of platinum group half-sandwich organometallic complexes comprising of pyrazole based (Rao et al. 2017), Fluorenone schiff base derivative complexes (Kollipara et al. 2020) and triazolo based ligands (Shadap et al. 2018) with the DNA using Fluorescence spectroscopy.

9.5.3 Merits and Demerits of Fluorescence Spectroscopy

Fluorescence spectroscopy is simple, reliable and small scale technique to understand protein-ligand binding with limited scope. The interaction could be studied in real- time using an inexpensive instrument. For a single reaction, time taken to scan is very less and the associated cost per reaction is low. The parameters that could be studied include stoichiometry, free energy change, binding constant, number of binding sites, etc. Moreover, the concerns regarding the purity of the sample are very less as the obtained spectrum is dependent on the particular flurophore under scrutiny. Often due to high concentration or overlap within excitation wavelength and emission spectrum, inner filter effect may be observed. However, this anomalous effect can be compensated by avoiding spectral overlap, by using low concentration of analyte (optical density below 0.1) or through internal correction settings of instrument using Raman correction procedure based on absorbance (Larsson et al. 2007). The parameters obtained rely on titration set which take long time for acquisition and have huge scope of experimental errors. Moreover, a careful account of Raman and Rayleigh scatter peak of water is necessary to avoid the false positives. Also, the internal filter settings of instrument may result in altered fluorescence intensities.

9.6 Concluding Remarks

In conclusion, all the four techniques discussed have quite decisive role in characterization of protein-ligand binding events. An elaborate comparison between all the four techniques has been summarized in Table 9.4. Keeping in view the applicability, versatility, affordability and the informational output, all the four techniques have mutual exclusivity; hence, it is difficult to choose amongst these four. Constant pondering upon the major trade-offs like cost effectiveness and the amount of derived information, determines the inclination towards a technique. In comparison to all the techniques NMR spectroscopy is best due to atomic level information by use of different methods; but the pre-requisites like infrastructure, affordability and

Table 9.4 Comparison of the information obtained from different techniques and the pre-requisites for an experiment

Parameters		NMR Spectroscopy	SPR	ITC	Fluorescence Spectroscopy
Studied parameters in single reaction		The parameters are derived from a set of reaction.	K_a, k_{on}, k_{off}	K_a, n, ΔH, ΔS, ΔG	Parameters like K_a and n are derived from a set of reaction
Sensitivity of technique		$10^{-9} < M < 10^{-2}$	$10^{-7} < M < 10^{-3}$	$10^{-12} < M < 10^{-2}$	$10^{-9} < M < 10^{-2}$
Medium of analysis		In solution	On the surface of chip	In solution	In solution
Sample volume	Protein	Depends on availability of the sample tubes and analyte (protein/ligand) under observation. (100–500μL)	30μL	200μL- 1400μL	Depending on the instrument, ranges from 0.3–800μL
	Ligand		30μL	40–300μL	
Sample Conc.	Protein	Depends on the nature of the protein.	Depends on the experiment and immobilization	As low as 10μg	Depends on the quantum yield of the fluorophore; generally concentration (~0.1 optical density) is recommended
	Ligand	5μM–100μM	Greater than 1 nM for molecules less than 10 kDa	Up to 100μM	
Required purity of sample		Highly pure sample is required	Well purified sample is recommended	Highest possible degree of purification is required	Well purified sample is recommended
Requirement of additional modification		No modification is required. Isotopic labeling is required for protein based methods	No	No	May be required in case the analyte is non-fluorescent
Time spent on an experiment		1 h–2 h	5–15 minutes for 15 samples, increased throughput up to 5000 samples per day with automation	8–12 samples in 8 h for manual loading; 42 samples in 24 h for automatic sampler	1–5 minutes per reaction. Results are obtained by collective use of titration gradient

The data provided is on the basis of the manuals of the devices- MicroCal (Malvern Instruments Limites, UK) and Biacore (GE Healthcare, United States) experimental expertise are major set-backs. Considering the robust high throughput as demanded by the industrial perspective, and the comprehensive details needed on an academic research perspective, all these four biophysical techniques discussed can be regarded as best for elucidating protein-ligand interactions.

Acknowledgements KMP acknowledge the receipt of Grants CRG/2018/001329 and SERB-SB/YS/LS-380/2013 from SERB-DST, and DBT-IYBA fellowship – BT/07/IYBA/2013-19. NA acknowledges the receipt of INSPIRE fellowship from Department of Science and Technology, India and VC is thankful to Department of Biotechnology (DBT-India) for JRF fellowship.

References

Abboud, M. I., Hinchliffe, P., Brem, J., Macsics, R., Pfeffer, I., Makena, A., et al. (2017). 19F-NMR reveals the role of mobile loops in product and inhibitor binding by the São Paulo Metallo-β-lactamase. *Angewandte Chemie International Edition, 56*, 3862–3866.

Abe, Y., Fukui, S., Koshiji, Y., Kobayashi, M., Shoji, T., Sugata, S., et al. (1999). Enantioselective binding sites on bovine serum albumin to dansyl amino acids. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology, 1433*, 188–197.

Agarwal, N., Nair, M. S., Mazumder, A., & Poluri, K. M. (2018). Characterization of nanomaterials using nuclear magnetic resonance spectroscopy. In *Characterization of nanomaterials* (pp. 61–102). Amsterdam: Elsevier.

Aguirre, C., Cala, O., & Krimm, I. (2015). Overview of probing protein-ligand interactions using NMR. *Current Protocols in Protein Science, 81*, 17–18. 17.18. 11-17.18. 24.

Ahmed-Belkacem, A., Colliandre, L., Ahnou, N., Nevers, Q., Gelin, M., Bessin, Y., et al. (2016). Fragment-based discovery of a new family of non-peptidic small-molecule cyclophilin inhibitors with potent antiviral activities. *Nature Communications, 7*, 1–11.

Amano, R., Furukawa, T., & Sakamoto, T. (2019). ITC measurement for high-affinity aptamers binding to their target proteins. In *Microcalorimetry of biological molecules* (pp. 119–128). New York: Springer.

Aziz, A., Santhoshkumar, P., Sharma, K. K., & Abraham, E. C. (2007). Cleavage of the c-terminal serine of human αA-crystallin produces αA1-172 with increased chaperone activity and oligomeric size. *Biochemistry, 46*, 2510–2519.

Balazs, A. Y. S., Carbajo, R. J., Davies, N. L., Dong, Y., Hird, A. W., Johannes, J. W., et al. (2019). Free ligand 1D NMR conformational signatures to enhance structure based drug design of a mcl-1 inhibitor (AZD5991) and other synthetic macrocycles. *Journal of Medicinal Chemistry, 62*, 9418–9437. https://doi.org/10.1021/acs.jmedchem.9b00716

Barelier, S., & Krimm, I. (2011). Ligand specificity, privileged substructures and protein druggability from fragment-based screening. *Current Opinion in Chemical Biology, 15*, 469–474.

Bartoschek, S., Klabunde, T., Defossa, E., Dietrich, V., Stengelin, S., Griesinger, C., et al. (2010). Drug design for G-protein-coupled receptors by a ligand-based NMR method. *Angewandte Chemie International Edition, 49*, 1426–1429.

Becker, W., Adams, L. A., Graham, B., Wagner, G. E., Zangger, K., Otting, G., et al. (2018). Trimethylsilyl tag for probing protein–ligand interactions by NMR. *Journal of Biomolecular NMR, 70*, 211–218.

Billeter, M., Wagner, G., & Wüthrich, K. (2008). Solution NMR structure determination of proteins revisited. *Journal of Biomolecular NMR, 42*, 155–158.

Bocquet N Markovic-Mueller, S., Cheng, R., Botte, M., AbdulRahman, W., Huber, S., et al. (2018) Structure based drug discovery on membrane protein targets.

Boswell, Z. K., & Latham, M. P. (2018). Methyl-based NMR spectroscopy methods for uncovering structural dynamics in large proteins and protein complexes. *Biochemistry, 58*, 144–155.

Bray, P., Emerson, J., Lee, D., Feller, S., Bain, D., & Feil, D. (1991). NMR and NQR studies of glass structure. *Journal of Non-Crystalline Solids, 129*, 240–248.

Breen, C. J., Raverdeau, M., & Voorheis, H. P. (2016). Development of a quantitative fluorescence-based ligand-binding assay. *Scientific Reports, 6*, 25769.

Burkhardt, M., & Schwille, P. (2006). Electron multiplying CCD based detection for spatially resolved fluorescence correlation spectroscopy. *Optics Express, 14*, 5013–5020.

Burnouf, D., Ennifar, E., Guedich, S., Puffer, B., Hoffmann, G., Bec, G., et al. (2012). kinITC: A new method for obtaining joint thermodynamic and kinetic data by isothermal titration calorimetry. *Journal of the American Chemical Society, 134*, 559–565.

Buurma, N. J., & Haq, I. (2007). Advances in the analysis of isothermal titration calorimetry data for ligand–DNA interactions. *Methods, 42*, 162–172.

Buurma, N. J., & Haq, I. (2008). Calorimetric and spectroscopic studies of Hoechst 33258: Self-association and binding to non-cognate DNA. *Journal of Molecular Biology, 381*, 607–621.

Cala, O., Guillière, F., & Krimm, I. (2014). NMR-based analysis of protein–ligand interactions. *Analytical and Bioanalytical Chemistry, 406*, 943–956.

Cala, O., & Krimm, I. (2015). Ligand-orientation based fragment selection in STD NMR screening. *Journal of Medicinal Chemistry, 58*, 8739–8742.

Capelli, D., Parravicini, C., Pochetti, G., Montanari, R., Temporini, C., Rabuffetti, M., et al. (2019). Surface Plasmon resonance as a tool for ligand binding investigation of engineered GPR17 receptor, a G protein coupled receptor involved in myelination. *Frontiers in Chemistry, 7*.

Cengiz, N. (2020). Glutathione-responsive multifunctionalizable hydrogels via amine-epoxy "click" chemistry. *European Polymer Journal, 123*, 109441.

Chakraborty, S., Mohan, P. K., & Hosur, R. V. (2012). Residual structure and dynamics in DMSO-d6 denatured dynein light chain protein. *Biochimie, 94*, 231–241.

Chandel, T. I., Zaman, M., Khan, M. V., Ali, M., Rabbani, G., Ishtikhar, M., et al. (2018). A mechanistic insight into protein-ligand interaction, folding, misfolding, aggregation and inhibition of protein aggregates: An overview. *International Journal of Biological Macromolecules, 106*, 1115–1129.

Chappuis, Q., Milani, J., Vuichoud, B., Bornet, A., Gossert, A. D., Bodenhausen, G., et al. (2015). Hyperpolarized water to study protein–ligand interactions. *The Journal of Physical Chemistry Letters, 6*, 1674–1678.

Chen, L., Wang, D., Lv, D., Wang, X., Liu, Y., Chen, X., et al. (2019). Identification of eupatilin and ginkgolide B as p38 ligands from medicinal herbs by surface plasmon resonance biosensor-based active ingredients recognition system. *Journal of Pharmaceutical and Biomedical Analysis, 171*, 35–42.

Chen, X., Lin, Y., Liu, M., & Gilson, M. K. (2002). The binding database: Data management and interface design. *Bioinformatics, 18*, 130–139.

Coates, C., Kerruth, S., Helassa, N., & Török, K. (2020). Kinetic mechanisms of fast glutamate sensing by fluorescent protein probes. *Biophysical Journal, 118*, 117–127.

Cohen, Y., Avram, L., & Frish, L. (2005). Diffusion NMR spectroscopy in supramolecular and combinatorial chemistry: An old parameter—New insights. *Angewandte Chemie International Edition, 44*, 520–554.

Collini, M., D'Alfonso, L., Molinari, H., Ragona, L., Catalano, M., & Baldini, G. (2003). Competitive binding of fatty acids and the fluorescent probe 1-8-anilinonaphthalene sulfonate to bovine β-lactoglobulin. *Protein Science, 12*, 1596–1603.

Croucher, D., Saunders, D. N., & Ranson, M. (2006). The Urokinase/PAI-2 complex a new high affinity ligand for the endocytosis receptor low density lipoprotein receptor-related proteiN. *Journal of Biological Chemistry, 281*, 10206–10213.

Dalvit, C., & Knapp, S. (2017). 19F NMR isotropic chemical shift for efficient screening of fluorinated fragments which are racemates and/or display multiple conformers. *Magnetic Resonance in Chemistry, 55*, 1091–1095.

Dalvit, C., & Vulpetti, A. (2018). Ligand-based fluorine NMR screening: Principles and applications in drug discovery projects. *Journal of Medicinal Chemistry, 62*, 2218–2244.

Damodaran, S., & Kinsella, J. E. (1981). The effects of neutral salts on the stability of macromolecules. A new approach using a protein-ligand binding system. *Journal of Biological Chemistry, 256*, 3394–3398.

Dansereau, S., Burz, D. S., & Shekhtman, A. (2019). Primary drug screening by in-cell NMR spectroscopy. *In-cell NMR Spectroscopy*, 249–271.

Day, C. J., & Korolik, V. (2018). Identification of specific ligands for sensory receptors by small-molecule ligand arrays and surface plasmon resonance. In *Bacterial chemosensing* (pp. 303–317). New York: Springer.

Deganutti, G., Zhukov, A., Deflorian, F., Federico, S., Spalluto, G., Cooke, R. M., et al. (2017). Impact of protein–ligand solvation and desolvation on transition state thermodynamic properties of adenosine a 2A ligand binding kinetics. *In Silico Pharmacology, 5*, 16.

Diao, W., Tang, M., Ding, S., Li, X., Cheng, W., Mo, F., et al. (2018). Highly sensitive surface plasmon resonance biosensor for the detection of HIV-related DNA based on dynamic and structural DNA nanodevices. *Biosensors and Bioelectronics, 100*, 228–234.

Ding, X., Cheng, W., Li, Y., Wu, J., Li, X., Cheng, Q., et al. (2017). An enzyme-free surface plasmon resonance biosensing strategy for detection of DNA and small molecule based on nonlinear hybridization chain reaction. *Biosensors and Bioelectronics, 87*, 345–351.

Dolphin, A. C., & Lee, A. (2020). Presynaptic calcium channels: Specialized control of synaptic neurotransmitter release. *Nature Reviews Neuroscience*, 1–17.

Dorion-Thibaudeau, J., St-Laurent, G., Raymond, C., De Crescenzo, G., & Durocher, Y. (2016). Biotinylation of the Fcγ receptor ectodomains by mammalian cell co-transfection: Application to the development of a surface plasmon resonance-based assay. *Journal of Molecular Recognition, 29*, 60–69.

Douzi, B. (2017). Protein–protein interactions: Surface plasmon resonance. In *Bacterial protein secretion systems* (pp. 257–275). New York: Springer.

Drescher, D. G., Selvakumar, D., & Drescher, M. J. (2018). Analysis of protein interactions by surface plasmon resonance. In *Advances in protein chemistry and structural biology* (Vol. 110, pp. 1–30). Amsterdam: Elsevier.

Dror, R. O., Pan, A. C., Arlow, D. H., Borhani, D. W., Maragakis, P., Shan, Y., et al. (2011). Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proceedings of the National Academy of Sciences, 108*, 13118–13123.

Du, J., & Strieter, E. R. (2018). A fluorescence polarization-based competition assay for measuring interactions between unlabeled ubiquitin chains and UCH37• RPN13. *Analytical Biochemistry, 550*, 84–89.

Du, X., Li, Y., Xia, Y.-L., Ai, S.-M., Liang, J., Sang, P., et al. (2016). Insights into protein–ligand interactions: Mechanisms, models, and methods. *International Journal of Molecular Sciences, 17*, 144.

Dumas, P., Ennifar, E., Da Veiga, C., Bec, G., Palau, W., Di Primo, C., et al. (2016). Extending ITC to kinetics with kinITC. In *Methods in enzymology* (Vol. 567, pp. 157–180). Amsterdam: Elsevier.

Eling, T., & DiAugustine, R. (1971). A role for phospholipids in the binding and metabolism of drugs by hepatic microsomes. Use of the fluorescent hydrophobic probe 1-anilinonaphthalene-8-sulphonate. *Biochemical Journal, 123*, 539–549.

Enke, M., Jehle, F., Bode, S., Vitz, J., Harrington, M. J., Hager, M. D., et al. (2017). Histidine–zinc interactions investigated by Isothermal Titration Calorimetry (ITC) and their application in self-healing polymers. *Macromolecular Chemistry and Physics, 218*, 1600458.

Eyring, H. (1935). The activated complex in chemical reactions. *The Journal of Chemical Physics, 3*, 107–115.

Falconer, R. J. (2016). Applications of isothermal titration calorimetry–the research and technical developments from 2011 to 2015. *Journal of Molecular Recognition, 29*, 504–515.

Feng, C., Kovrigin, E. L., & Post, C. B. (2019). NmrLineGuru: standalone and user-friendly GUIs for fast 1D NMR lineshape simulation and analysis of multi-state equilibrium binding models. *Scientific Reports, 9*, 1–14.

Fielding, L. (2003). NMR methods for the determination of protein-ligand dissociation constants. *Current Topics in Medicinal Chemistry, 3*, 39–53.

Fielding, L. (2007). NMR methods for the determination of protein–ligand dissociation constants. *Progress in Nuclear Magnetic Resonance Spectroscopy, 51*, 219–242.

Frederick, K. K., Marlow, M. S., Valentine, K. G., & Wand, A. J. (2007). Conformational entropy in molecular recognition by proteins. *Nature, 448*, 325–329.

Furukawa, A., Konuma, T., Yanaka, S., & Sugase, K. (2016). Quantitative analysis of protein–ligand interactions by NMR. *Progress in Nuclear Magnetic Resonance Spectroscopy, 96*, 47–57.

Gao, J., Liang, E., Ma, R., Li, F., Liu, Y., Liu, J., et al. (2017). Fluorine pseudocontact shifts used for characterizing the protein–ligand interaction mode in the limit of NMR intermediate exchange. *Angewandte Chemie International Edition, 56*, 12982–12986.

Garrett, A. B. (1962). The Bohr atomic model: Niels Bohr. *Journal of Chemical Education, 39*, 534.

Gasymov, O. K., Abduragimov, A. R., & Glasgow, B. J. (2008). Ligand binding site of tear lipocalin: Contribution of a trigonal cluster of charged residues probed by 8-anilino-1-naphthalenesulfonic acid. *Biochemistry, 47*, 1414–1424.

Gee, C. T., Arntson, K. E., Koleski, E. J., Staebell, R. L., & Pomerantz, W. C. (2018). Dual labeling of the CBP/p300 KIX domain for 19F NMR leads to identification of a new small-molecule binding site. *Chembiochem, 19*, 963–969.

Geist, L., Mayer, M., Cockcroft, X.-L., Wolkerstorfer, B., Kessler, D., Engelhardt, H., et al. (2017). Direct NMR probing of hydration shells of protein ligand interfaces and its application to drug design. *Journal of Medicinal Chemistry, 60*, 8708–8715.

Geuijen, K. P., van Wijk-Basten, D. E., Egging, D. F., Schasfoort, R. B., & Eppink, M. H. (2017). Rapid buffer and ligand screening for affinity chromatography by multiplexed surface Plasmon resonance imaging. *Biotechnology Journal, 12*, 1700154.

Ghatak, C., Rodnin, M. V., Vargas-Uribe, M., McCluskey, A. J., Flores-Canales, J. C., Kurnikova, M., et al. (2015). Role of acidic residues in helices TH8–TH9 in membrane interactions of the diphtheria toxin T domain. *Toxins, 7*, 1303–1323.

Gohlke, A., Bower, J., Brown, P. N., Cameron, K. S., Drysdale, M., Goodwin, G., et al. (2018). A central role for biophysics in cancer drug discovery-development of candidate small molecule inhibitors in mutant KRas. *Biophysical Journal, 114*, 30a–31a.

Gooley, P. R., Koay, A., & Mobbs, J. I. (2018). Applications of NMR and ITC for the study of the kinetics of carbohydrate binding by AMPK β-subunit carbohydrate-binding modules. In *AMPK* (pp. 87–98). New York: Springer.

Gordon, S. E., & Perugini, M. A. (2016). Protein-ligand interactions. In *Analytical ultracentrifugation* (pp. 329–353). New York: Springer.

Gossert, A. D., & Jahnke, W. (2016). NMR in drug discovery: A practical guide to identification and validation of ligands interacting with biological macromolecules. *Progress in Nuclear Magnetic Resonance Spectroscopy, 97*, 82–125.

Gulati, K., Gangele, K., Kumar, D., & Poluri, K. M. (2019). An inter-switch between hydrophobic and charged amino acids generated druggable small molecule binding pocket in chemokine paralog CXCL3. *Archives of Biochemistry and Biophysics, 662*, 121–128.

Gulati, K., & M Poluri, K. (2016). An overview of computational and experimental methods for designing novel proteins. *Recent Patents on Biotechnology, 10*, 235–263.

Gulati, K., & Poluri, K. M. (2019). Role of engineered proteins as therapeutic formulations. In *Pharmaceutical biocatalysis: Fundamentals, enzyme inhibitors, and enzymes in health and diseases* (p. 159). Singapore: Jenny Stanford Publishing.

Guo, J., & Zhou, H.-X. (2016). Protein allostery and conformational dynamics. *Chemical Reviews, 116*, 6503–6515.

Hansen, L. D., Transtrum, M. K., Quinn, C., & Demarse, N. (2016). Enzyme-catalyzed and binding reaction kinetics determined by titration calorimetry. *Biochimica et Biophysica Acta (BBA)-General Subjects, 1860*, 957–966.

Hao, C., Xu, G., Feng, Y., Lu, L., Sun, W., & Sun, R. (2017). Fluorescence quenching study on the interaction of ferroferric oxide nanoparticles with bovine serum albumin. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 184*, 191–197.

Hatta, A., Suzuki, Y., & Suëtaka, W. (1984). Infrared absorption enhancement of monolayer species on thin evaporated Ag films by use of a Kretschmann configuration: Evidence for two types of enhanced surface electric fields. *Applied Physics A, 35*, 135–140.

Hevekerl, H., Tornmalm, J., & Widengren, J. (2016). Fluorescence-based characterization of non-fluorescent transient states of tryptophan–prospects for protein conformation and interaction studies. *Scientific Reports, 6*, 35052.

Huang, R., & Leung, I. K. (2016). Protein-directed dynamic combinatorial chemistry: A guide to protein ligand and inhibitor discovery. *Molecules, 21*, 910.

Huang, R., & Leung, I. K. (2019). Protein–small molecule interactions by WaterLOGSY. In *Methods in enzymology* (Vol. 615, pp. 477–500). Amsterdam: Elsevier.

Huber, S., Casagrande, F., Hug, M. N., Wang, L., Heine, P., Kummer, L., et al. (2017). SPR-based fragment screening with neurotensin receptor 1 generates novel small molecule ligands. *PLoS One, 12*.

Jabar, S., Adams, L. A., Wang, Y., Aurelio, L., Graham, B., & Otting, G. (2017). Chemical tagging with tert-butyl and trimethylsilyl groups for measuring intermolecular nuclear overhauser effects in a large protein–ligand complex. *Chemistry–A European Journal, 23*, 13033–13036.

Jaiswal, N., Agarwal, N., Kaur, A., Tripathi, S., Gahlay, G. K., Arora, A., et al. (2019). Molecular interaction between human SUMO-I and histone like DNA binding protein of helicobacter pylori (Hup) investigated by NMR and other biophysical tools. *International Journal of Biological Macromolecules, 123*, 446–456.

Johnson, O. T., Kaur, T., & Garner, A. L. (2019). A conditionally fluorescent peptide reporter of secondary structure modulation. *Chembiochem, 20*, 40–45.

Jones, L. H., & Bunnage, M. E. (2017). Applications of chemogenomic library screening in drug discovery. *Nature Reviews Drug Discovery, 16*, 285.

Joseph, P. R. B., Poluri, K. M., Sepuru, K. M., & Rajarathnam, K. (2015). Characterizing protein–glycosaminoglycan interactions using solution NMR spectroscopy. In *Glycosaminoglycans* (pp. 325–333). Amsterdam: Springer.

Joshi, U. M., Rao, P., Kodavanti, S., Lockard, V. G., & Mehendale, H. M. (1989). Fluorescence studies on binding of amphiphilic drugs to isolated lamellar bodies: Relevance to phospholipidosis. *Biochimica et Biophysica Acta (BBA)-Lipids and Lipid Metabolism, 1004*, 309–320.

Juhász, A., Csapó, E., Ungor, D., Tóth, G. K., Ls, V., & Dékány, I. (2016). Kinetic and thermodynamic evaluation of kynurenic acid binding to GluR1270–300 polypeptide by surface Plasmon resonance experiments. *The Journal of Physical Chemistry B, 120*, 7844–7850.

Kaur, G., Paliwal, A., Tomar, M., & Gupta, V. (2016). Detection of Neisseria meningitidis using surface plasmon resonance based DNA biosensor. *Biosensors and Bioelectronics, 78*, 106–110.

Kay, L. E. (2011). NMR studies of protein structure and dynamics-a look backwards and forwards. *Journal of Magnetic Resonance (San Diego, Calif: 1997), 213*, 492–494.

Keeler, J. (2011). *Understanding NMR spectroscopy*. Washington, D.C.: Wiley.

Kirkpatrick, S. (1984). Optimization by simulated annealing: Quantitative studies. *Journal of Statistical Physics, 34*, 975–986.

Kivi, R., Solovjova, K., Haljasorg, T., Arukuusk, P., & Järv, J. (2016). Allosteric effect of adenosine triphosphate on peptide recognition by 3′ 5′-cyclic adenosine monophosphate dependent protein kinase catalytic subunits. *The Protein Journal, 35*, 459–466.

Kollipara, M. R., Shadap, L., Banothu, V., Agarwal, N., Poluri, K. M., & Kaminsky, W. (2020). Fluorenone Schiff base derivative complexes of ruthenium, rhodium and iridium exhibiting efficient antibacterial activity and DNA-binding affinity. *Journal of Organometallic Chemistry*, 121246.

Kooyman, R., De Bruijn, H., Eenink, R., & Greve, J. (1990). Surface plasmon resonance as a bioanalytical tool. *Journal of Molecular Structure, 218*, 345–350.

Korshunova, A., Lopanskaia, I., & Gudimchuk, N. (2019). Modern approaches to analysis of protein–ligand interactions. *Biophysics, 64*, 495–509.

Kozma, E., Jayasekara, P. S., Squarcialupi, L., Paoletta, S., Moro, S., Federico, S., et al. (2013). Fluorescent ligands for adenosine receptors. *Bioorganic & Medicinal Chemistry Letters, 23*, 26–36.

Kubinyi, H. (2006). Chemogenomics in drug discovery. In *Chemical genomics* (pp. 1–19). New York: Springer.

Kumar, P. K. (2017). Systematic screening of viral entry inhibitors using surface plasmon resonance. *Reviews in Medical Virology, 27*, e1952.

Kuzuyama, T., Takahashi, S., Takagi, M., & Seto, H. (2000). Characterization of 1-deoxy-D-xylulose 5-phosphate reductoisomerase, an enzyme involved in isopentenyl diphosphate biosynthesis, and identification of its catalytic amino acid residues. *Journal of Biological Chemistry, 275*, 19928–19932.

Kyrychenko, A. (2015). Using fluorescence for studies of biological membranes: A review. *Methods and Applications in Fluorescence, 3*, 042003.

Lakayan, D., Tuppurainen, J., Suutari, T. E., van Iperen, D. J., Somsen, G. W., & Kool, J. (2019). Design and evaluation of a multiplexed angular-scanning surface plasmon resonance system employing line-laser optics and CCD detection in combination with multi-ligand sensor chips. *Sensors and Actuators B: Chemical, 282*, 243–250.

Lampe, J. N., & Atkins, W. M. (2006). Time-resolved fluorescence studies of heterotropic ligand binding to cytochrome P450 3A4. *Biochemistry, 45*, 12204–12215.

Larsson, T., Wedborg, M., & Turner, D. (2007). Correction of inner-filter effect in fluorescence excitation-emission matrix spectrometry using Raman scatter. *Analytica Chimica Acta, 583*, 357–363.

Legchenko, A. (2013). The basics of NMR. In *Magnetic resonance imaging for groundwater* (pp. 15–44). New York: Wiley.

Li, D., DeRose, E. F., & London, R. E. (1999). The inter-ligand Overhauser effect: A powerful new NMR approach for mapping structural relationships of macromolecular ligands. *Journal of Biomolecular NMR, 15*, 71–76.

Li, S., Yang, M., Zhou, W., Johnston, T. G., Wang, R., & Zhu, J. (2015). Dextran hydrogel coated surface plasmon resonance imaging (SPRi) sensor for sensitive and label-free detection of small molecule drugs. *Applied Surface Science, 355*, 570–576.

Liu, T., Lin, Y., Wen, X., Jorissen, R. N., & Gilson, M. K. (2007). BindingDB: A web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Research, 35*, D198–D201.

Liu, X., Golden, L. C., Lopez, J. A., Shepherd, T. R., Yu, L., & Fuentes, E. J. (2019). Conformational dynamics and cooperativity drive the specificity of a protein-ligand interaction. *Biophysical Journal, 116*, 2314–2330.

Liu, X., Speckhard, D. C., Shepherd, T. R., Sun, Y. J., Hengel, S. R., Yu, L., et al. (2016). Distinct roles for conformational dynamics in protein-ligand interactions. *Structure, 24*, 2053–2066.

Masson, J.-F. (2017). Surface plasmon resonance clinical biosensors for medical diagnostics. *ACS Sensors, 2*, 16–30.

McClure, W. O., & Edelman, G. M. (1966). Fluorescent probes for conformational states of proteins. I. Mechanism of fluorescence of 2-p-toluidinylnaphthalene-6-sulfonate, a hydrophobic probe. *Biochemistry, 5*, 1908–1919.

Meyer, B., da Veiga, C., Dumas, P., & Ennifar, E. (2019). Thermodynamics of molecular machines using incremental ITC. In *Microcalorimetry of biological molecules* (pp. 129–140). New York: Springer.

Meyer, B., & Peters, T. (2003). NMR spectroscopy techniques for screening and identifying ligand binding to protein receptors. *Angewandte Chemie International Edition, 42*, 864–890.

Mobbs, J. I., Koay, A., Di Paolo, A., Bieri, M., Petrie, E. J., Gorman, M. A., et al. (2015). Determinants of oligosaccharide specificity of the carbohydrate-binding modules of AMP-activated protein kinase. *Biochemical Journal, 468*, 245–257.

Mocz, G., & Ross, J. A. (2013). Fluorescence techniques in analysis of protein–ligand interactions. In *Protein-ligand interactions* (pp. 169–210). New York: Springer.

Mukherjee, S., Mohan, P. K., & Chary, K. V. (2007). Magnesium promotes structural integrity and conformational switching action of a calcium sensor protein. *Biochemistry, 46*, 3835–3845.

Mureddu, L., & Vuister, G. W. (2019). Simple high-resolution NMR spectroscopy as a tool in molecular biology. *The FEBS Journal, 286*, 2035–2042.

Nag, N., Ramreddy, T., Kombrabail, M., Mohan, P. K., D'souza, J., Rao, B., et al. (2006). Dynamics of DNA and portein-DNA complexes viewed through time-domain fluorescence. In *Reviews in fluorescence 2006* (pp. 311–340). Springer.

Nepravishta, R., Walpole, S., Tailford, L., Juge, N., & Angulo, J. (2019). Deriving ligand orientation in weak protein–ligand complexes by DEEP-STD NMR spectroscopy in the absence of protein chemical-shift assignment. *Chembiochem, 20*, 340–344.

Nguyen, H. H., Park, J., Kang, S., & Kim, M. (2015). Surface plasmon resonance: A versatile technique for biosensor applications. *Sensors, 15*, 10481–10510.

Nico, J., & Fischer, M. J. (2010). Surface plasmon resonance: A general introduction. In *Surface Plasmon resonance* (pp. 1–14). New York: Springer.

Olsson, T. S., Williams, M. A., Pitt, W. R., & Ladbury, J. E. (2008). The thermodynamics of protein–ligand interaction and solvation: Insights for ligand design. *Journal of Molecular Biology, 384*, 1002–1017.

Orts, J., Griesinger, C., & Carlomagno, T. (2009). The INPHARMA technique for pharmacophore mapping: A theoretical guide to the method. *Journal of Magnetic Resonance, 200*, 64–73.

Pan, A. C., Borhani, D. W., Dror, R. O., & Shaw, D. E. (2013). Molecular determinants of drug–receptor binding kinetics. *Drug Discovery Today, 18*, 667–673.

Panov, K., & Alsahafi, S. (2019). Investigation of molecular mechanisms of polymerase I (pol-l) inhibitor PMR-116 using isothermal titration Calorimetry (ITC). *Access Microbiology, 1*.

Partlow, B. P., Bagheri, M., Harden, J. L., & Kaplan, D. L. (2016). Tyrosine templating in the self-assembly and crystallization of silk fibroin. *Biomacromolecules, 17*, 3570–3579.

Pecht, I., Maron, E., Arnon, R., & Sela, M. (1971). Specific excitation energy transfer from antibodies to Dansyl-labeled antigen: Studies with the "loop" peptide of hen egg-white lysozyme. *European Journal of Biochemistry, 19*, 368–371.

Pellecchia, M., Meininger, D., Dong, Q., Chang, E., Jack, R., & Sem, D. S. (2002). NMR-based structural characterization of large protein-ligand interactions. *Journal of Biomolecular NMR, 22*, 165–173.

Perspicace, S., Banner, D., Benz, J., Müller, F., Schlatter, D., & Huber, W. (2009). Fragment-based screening using surface plasmon resonance technology. *Journal of Biomolecular Screening, 14*, 337–349.

Piñeiro, Á., Muñoz, E., Sabín, J., Costas, M., Bastos, M., Velázquez-Campoy, A., et al. (2019). AFFINImeter: A software to analyze molecular recognition processes from experimental data. *Analytical Biochemistry, 577*, 117–134.

Plotnikova, O., Mel'Nikov, A., Mel'Nikov, G., & Gubina, T. (2016). Quenching of tryptophan fluorescence of bovine serum albumin under the effect of ions of heavy metals. *Optics and Spectroscopy, 120*, 65–69.

Poluri, K. M., & Gulati, K. (2016). *Protein engineering techniques: Gateways to synthetic protein universe*. Singapore: Springer.

Poluri, K. M., & Gulati, K. (2017). World of proteins: Structure-function relationships and engineering techniques. In *Protein engineering techniques* (pp. 1–25). Singapore: Springer.

Poluri, K. M., Joseph, P. R. B., Sawant, K. V., & Rajarathnam, K. (2013). Molecular basis of glycosaminoglycan heparin binding to the chemokine CXCL1 dimer. *Journal of Biological Chemistry, 288*, 25143–25153.

Post, C. B. (2003). Exchange-transferred NOE spectroscopy and bound ligand structure determination. *Current Opinion in Structural Biology, 13*, 581–588.

Prozeller, D., Morsbach, S., & Landfester, K. (2019). Isothermal titration calorimetry as a complementary method for investigating nanoparticle–protein interactions. *Nanoscale, 11*, 19265–19273.

Radić, Z., Kirchhoff, P. D., Quinn, D. M., McCammon, J. A., & Taylor, P. (1997). Electrostatic influence on the kinetics of ligand binding to acetylcholinesterase distinctions between active center ligands and fasciculin. *Journal of Biological Chemistry, 272*, 23265–23277.

Ràfols, C., Amézqueta, S., Fuguet, E., & Bosch, E. (2018). Molecular interactions between warfarin and human (HSA) or bovine (BSA) serum albumin evaluated by isothermal titration calorimetry (ITC), fluorescence spectrometry (FS) and frontal analysis capillary electrophoresis (FA/CE). *Journal of Pharmaceutical and Biomedical Analysis, 150*, 452–459.

Ragona, L., Gasymov, O., Guliyeva, A. J., Aslanov, R. B., Zanzoni, S., Botta, C., et al. (2018). Rhodamine binds to silk fibroin and inhibits its self-aggregation. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 1866*, 661–667.

Raingeval, C., Cala, O., Brion, B., Le Borgne, M., Hubbard, R. E., & Krimm, I. (2019). 1D NMR WaterLOGSY as an efficient method for fragment-based lead discovery. *Journal of Enzyme Inhibition and Medicinal Chemistry, 34*, 1218–1225.

Raj, R., Agarwal, N., Raghavan, S., Chakraborti, T., Poluri, K. M., & Kumar, D. (2020). Exquisite binding interaction of 18β-Glycyrrhetinic acid with histone like DNA binding protein of helicobacter pylori: A computational and experimental study. *International Journal of Biological Macromolecules, 161*, 231–246.

Rao, A. B. P., Gulati, K., Joshi, N., Deb, D. K., Rambabu, D., Kaminsky, W., et al. (2017). Synthesis and biological studies of ruthenium, rhodium and iridium metal complexes with pyrazole-based ligands displaying unpredicted bonding modes. *Inorganica Chimica Acta, 462*, 223–235.

Rayani, K., Muñoz, E., Spuches, A., Van Petegem, F., & Tibbits, G. (2018). Binding of calcium and magnesium to cardiac troponin C assessed through Isothermal Titration Calorimetry (ITC). *Journal of Molecular and Cellular Cardiology, 124*, 86.

Rega, M. F., Wu, B., Wei, J., Zhang, Z., Cellitti, J. F., & Pellecchia, M. (2011). SAR by interligand nuclear Overhauser effects (ILOEs) based discovery of acylsulfonamide compounds active against Bcl-xL and Mcl-1. *Journal of Medicinal Chemistry, 54*, 6000–6013.

Rühmann, E., Betz, M., Fricke, M., Heine, A., Schäfer, M., & Klebe, G. (2015). Thermodynamic signatures of fragment binding: Validation of direct versus displacement ITC titrations. *Biochimica et Biophysica Acta (BBA)-General Subjects, 1850*, 647–656.

Saeed, I. Q., & Buurma, N. J. (2019). Analysis of isothermal titration calorimetry data for complex interactions using I2CITC. In *Microcalorimetry of biological molecules* (pp. 169–183). New York: Springer.

Sahni, N., Chaudhuri, R., Hickey, J. M., Manikwar, P., D'Souza, A., Metters, A., et al. (2017). Preformulation characterization, stabilization, and formulation design for the acrylodan-labeled glucose-binding protein SM4-AC. *Journal of Pharmaceutical Sciences, 106*, 1197–1210.

Šakanovič, A., Hodnik, V., & Anderluh, G. (2019). Surface Plasmon resonance for measuring interactions of proteins with lipids and lipid membranes. In *Lipid-protein interactions* (pp. 53–70). Springer.

Sánchez-Pedregal, V. M., Reese, M., Meiler, J., Blommers, M. J., Griesinger, C., & Carlomagno, T. (2005). The INPHARMA method: Protein-mediated interligand NOEs for pharmacophore mapping. *Angewandte Chemie International Edition, 44*, 4172–4175.

Santofimia-Castaño, P., Rizzuti, B., Abián, O., Velázquez-Campoy, A., Iovanna, J. L., & Neira, J. L. (2018). Amphipathic helical peptides hamper protein-protein interactions of the intrinsically disordered chromatin nuclear protein 1 (NUPR1). *Biochimica et Biophysica Acta (BBA)-General Subjects, 1862*, 1283–1295.

Sarkar, I., & Mishra, A. K. (2018). Fluorophore tagged bio-molecules and their applications: A brief review. *Applied Spectroscopy Reviews, 53*, 552–601.

Sarkar, S., Gulati, K., Kairamkonda, M., Mishra, A., & Poluri, K. M. (2018). Elucidating protein-protein interactions through computational approaches and designing small molecule inhibitors against them for various diseases. *Current Topics in Medicinal Chemistry, 18*, 1719–1736.

Sarkar, S., Gulati, K., Mishra, A., & Poluri, K. M. (2020). Protein nanocomposites: Special inferences to lysozyme based nanomaterials. *International Journal of Biological Macromolecules, 151*, 467–482.

Sawant, K. V., Poluri, K. M., Dutta, A. K., Sepuru, K. M., Troshkina, A., Garofalo, R. P., et al. (2016). Chemokine CXCL1 mediated neutrophil recruitment: Role of glycosaminoglycan interactions. *Scientific Reports, 6*, 33123.

Schasfoort, R., Abali, F., Stojanovic, I., Vidarsson, G., & Terstappen, L. W. (2018). Trends in SPR cytometry: Advances in label-free detection of cell parameters. *Biosensors, 8*, 102.

Schmidtke, P., Luque, F. J., Murray, J. B., & Barril, X. (2011). Shielded hydrogen bonds as structural determinants of binding kinetics: Application in drug design. *Journal of the American Chemical Society, 133*, 18903–18910.

Schuetz, D. A., de Witte, W. E. A., Wong, Y. C., Knasmueller, B., Richter, L., Kokh, D. B., et al. (2017). Kinetics for drug discovery: An industry-driven effort to target drug residence time. *Drug Discovery Today, 22*, 896–911.

Shadap, L., Joshi, N., Poluri, K. M., Kollipara, M. R., & Kaminsky, W. (2018). Synthesis and structural characterization of arene d6 metal complexes of sulfonohydrazone and triazolo ligands: High potency of triazolo derivatives towards DNA binding. *Polyhedron, 155*, 302–312.

Shanehbandi, D., Majidi, J., Kazemi, T., Baradaran, B., Aghebati-Maleki, L., Fathi, F., et al. (2017). Immuno-biosensor for detection of CD20-positive cells using surface plasmon resonance. *Advanced Pharmaceutical Bulletin, 7*, 189.

Sindrewicz, P., Li, X., Yates, E. A., Turnbull, J. E., Lian, L.-Y., & Yu, L.-G. (2019). Intrinsic tryptophan fluorescence spectroscopy reliably determines galectin-ligand interactions. *Scientific Reports, 9*, 1–12.

Śledź, P., & Caflisch, A. (2018). Protein structure-based drug design: From docking to molecular dynamics. *Current Opinion in Structural Biology, 48*, 93–102.

Snyder, D. A., Chantova, M., & Chaudhry, S. (2015). Analysis of ligand–protein exchange by clustering of ligand diffusion coefficient pairs (CoLD-CoP). *Journal of Magnetic Resonance, 255*, 44–50.

Szabo, A., Stolz, L., & Granzow, R. (1995). Surface plasmon resonance and its use in biomolecular interaction analysis (BIA). *Current Opinion in Structural Biology, 5*, 699–705.

Taghipour, P., Zakariazadeh, M., Sharifi, M., Dolatabadi, J. E. N., & Barzegar, A. (2018). Bovine serum albumin binding study to erlotinib using surface plasmon resonance and molecular docking methods. *Journal of Photochemistry and Photobiology B: Biology, 183*, 11–15.

Tang, W., & Yengo, C. M. (2018). Inter-filament co-operativity is crucial for regulating muscle contraction. *The Journal of Physiology, 596*, 17.

Umuhire Juru, A., Patwardhan, N. N., & Hargrove, A. E. (2019). Understanding the contributions of conformational changes, thermodynamics, and kinetics of RNA–small molecule interactions. *ACS Chemical Biology, 14*, 824–838.

Usui, M., Furihata, K., Utsumi, H., Kato, T., & Tashiro, M. (2017). Applicability of the NMR-based screening methods with 19F detection to the fluorinated compound. *J Res Anal, 3*, 34–36.

Van Der Merwe, P. A. (2001). Surface plasmon resonance. *Protein-ligand interactions: hydrodynamics and calorimetry, 1*, 137–170.

Vargas-Uribe, M., Rodnin, M. V., Öjemalm, K., Holgado, A., Kyrychenko, A., Nilsson, I., et al. (2015). Thermodynamics of membrane insertion and refolding of the diphtheria toxin T-domain. *The Journal of Membrane Biology, 248*, 383–394.

Viegas, A., Manso, J., Nobrega, F. L., & Cabrita, E. J. (2011). Saturation-transfer difference (STD) NMR: A simple and fast method for ligand screening and characterization of protein binding. *Journal of Chemical Education, 88*, 990–994.

Vogel, M., & Suess, B. (2016). Label-free determination of the dissociation constant of small molecule-aptamer interaction by isothermal titration calorimetry. In *Nucleic acid aptamers* (pp. 113–125). San Diego, CA: Springer.

Wallerstein, J., & Akke, M. (2019). Minute additions of DMSO affect protein dynamics measurements by NMR relaxation experiments through significant changes in solvent viscosity. *ChemPhysChem, 20*, 326–332.

Wang, Q., Liu, R., Yang, X., Wang, K., Zhu, J., He, L., et al. (2016). Surface plasmon resonance biosensor for enzyme-free amplified microRNA detection based on gold nanoparticles and DNA supersandwich. *Sensors and Actuators B: Chemical, 223*, 613–620.

Wang, Y., Huang, C.-J., Jonas, U., Wei, T., Dostalek, J., & Knoll, W. (2010). Biosensor based on hydrogel optical waveguide spectroscopy. *Biosensors and Bioelectronics, 25*, 1663–1668.

Wang, Y. S., Liu, D., & Wyss, D. F. (2004). Competition STD NMR for the detection of high-affinity ligands and NMR-based screening. *Magnetic Resonance in Chemistry, 42*, 485–489.

Waudby, C. A., Ramos, A., Cabrita, L. D., & Christodoulou, J. (2016). Two-dimensional NMR lineshape analysis. *Scientific Reports, 6*, 24826.

Winkie, M. L., So, J., Schlessman, J. L., & Smith, V. F. (2019). Characterization of protein-ligand interactions in MsrA from E. coli the. *FASEB Journal, 33*, 631.647–631.647.

Winter, G., Vogt, A., Glatting, G., Kletting, P., & Beer, A. (2018). Characterization of the receptor binding kinetics of PSMA-specific peptides by surface Plasmon resonance spectroscopy. *Journal of Nuclear Medicine, 59*, 1128–1128.

Wiseman, T., Williston, S., Brandts, J. F., & Lin, L.-N. (1989). Rapid measurement of binding constants and heats of binding using a new titration calorimeter. *Analytical Biochemistry, 179*, 131–137.

Wolf, S., Amaral, M., Lowinski, M., Vallée, F., Musil, D., Güldenhaupt, J., et al. (2019). Estimation of protein–ligand unbinding kinetics using non-equilibrium targeted molecular dynamics simulations. *Journal of Chemical Information and Modeling, 59*, 5135–5147.

Wong, J. K.-H., Todd, M. H., & Rutledge, P. J. (2017). Recent advances in macrocyclic fluorescent probes for ion sensing. *Molecules, 22*, 200.

Yamaoki, Y., Nagata, T., Sakamoto, T., & Katahira, M. (2020). Recent progress of in-cell NMR of nucleic acids in living human cells. *Biophysical Reviews*, 1–7.

Yang, H., Huang, Y., He, J., Li, S., Tang, B., & Li, H. (2016). Interaction of lafutidine in binding to human serum albumin in gastric ulcer therapy: STD-NMR, WaterLOGSY-NMR, NMR relaxation times, Tr-NOESY, molecule docking, and spectroscopic studies. *Archives of Biochemistry and Biophysics, 606*, 81–89.

Yang, L., Wang, J., L-z, Y., Hu, Z.-D., Wu, X., & Zheng, G. (2018). Characteristics of multiple Fano resonances in waveguide-coupled surface plasmon resonance sensors based on waveguide theory. *Scientific Reports, 8*, 1–10.

Yu, Z., Li, P., & Merz Jr., K. M. (2017). Using ligand-induced protein chemical shift perturbations to determine protein–ligand structures. *Biochemistry, 56*, 2349–2362.

Zare, H., Moosavi-Movahedi, A., Salami, M., Sheibani, N., Khajeh, K., & Habibi-Rezaei, M. (2016). Autolysis control and structural changes of purified ficin from Iranian fig latex with synthetic inhibitors. *International Journal of Biological Macromolecules, 84*, 464–471.

Zhan, S., Shi, C., Ou, H., Song, H., & Wang, X. (2016). A real-time de-noising method applied for transient and weak biomolecular interaction analysis in surface plasmon resonance biosensing. *Measurement Science and Technology, 27*, 035702.

Zhang, J., & Nussinov, R. (2019). *Protein allostery in drug discovery*. Singapore: Springer.

Zhang, S., & Fitzpatrick, P. F. (2016). Identification of the allosteric site for phenylalanine in rat phenylalanine hydroxylase. *Journal of Biological Chemistry, 291*, 7418–7425.

Chapter 10
In Silico Approach in Drug Design and Drug Discovery: An Update

Neetu Jabalia, Atul Kumar, Vinit Kumar, and Reshma Rani

Abstract Drug design and drug discovery is a passionate, prolonged, and interdisciplinary multistep process. The traditional drug discovery process is costly and time-consuming which demands hard work. The in silico approach has become indispensible for the drug discovery process. The in silico methods have been proved beneficial in the estimation of biological activities of chemical compounds against a target. Moreover, it has been used to examine the binding affinities towards the target and also used in the prediction of physicochemical properties depending upon their molecular and structural features of a broad range of chemical compounds. Besides, computational screening seems an important method to limit the cost, time, and manpower in comparison to the traditional drug discovery process. This chapter summarizes the importance and applications of various aspects of in silico approaches along with the examples where in silico methods have been applied in the drug discovery process.

Keywords Drug design · Drug discovery · In silico modeling · Ligand-protein · Molecular docking and simulation

N. Jabalia · R. Rani (✉)
Amity Institute of Biotechnology, Amity University Noida, Noida, Uttar Pradesh, India
e-mail: rrani@amity.edu

A. Kumar
Amity Institute of Engineering & Technology, Amity University Greater Noida, Noida, Uttar Pradesh, India

V. Kumar (✉)
Amity Institute of Molecular Medicine and Stem Cell Research, Amity University Noida, Noida, Uttar Pradesh, India
e-mail: vkumar25@amity.edu

S. K. Singh (ed.), *Innovations and Implementations of Computer Aided Drug Discovery Strategies in Rational Drug Design*,
https://doi.org/10.1007/978-981-15-8936-2_10

10.1 Introduction

Drug design, discovery, and development is an intense, costly, time consuming and an interdisciplinary endeavor. It is a consecutive linear process that starts with the identification of target and potent lead candidate, preparation of several analogs, optimization via in vitro, and in vivo preclinical studies to generate a drug-like molecule for drug development. In general, the drug discovery process aims to identify new drug molecules that can be specific and selective for a target to minimize the side effects. To fasten the drug discovery process and to minimize the cost of a new drug to launch in the market, in silico techniques play a significant role in the identification of a target as well as a lead molecule. Pedero Miramontes, a mathematician from the National Autonomous University of Mexico coined the term in silico' to study the biological experiments performed entirely on a computer (Gupta et al. 2011). In silico means "experiments performed on a computer or via computer simulation" which help in the identification of drug-like molecules and targets by using bioinformatics tools. In silico methods offer several implications in drug discovery process such as (1) analysis of target structure with an active site for binding, (2) evaluation of their drug-likeness, (3) investigation of binding affinity of drug molecules with target (4) ranking of the best-scored molecules and (5) their further optimization. Concisely, in silico methods offer many advantages to deliver cost-effective and more quickly new drug candidates. The most important and frequent role of in silico methods in the drug discovery process are (1) virtual screening; (2) de novo drug design; (3) in silico ADME/T prediction and (4) determination of protein-ligand binding affinity (Fig. 10.1).

1. *Virtual screening:* Pharmaceutical industry extensively employs virtual screening methods which comprise computational screening of libraries of large chemical compounds with molecular diversity against a three-dimensional binding cavity present in the target. This method is also known as structure-based virtual screening.
2. *De novo drug design*: *De novo* designing plays an important role in the identification of lead candidates in the pharmaceutical industry. In this method, the target is used without ligand to generate novel chemical compounds that can efficiently and specifically bind with the target. There are many algorithms currently

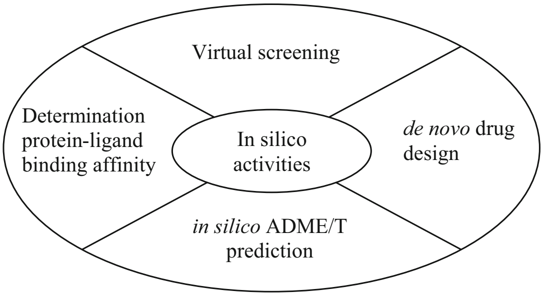

Virtual screening
Determination protein-ligand binding affinity
In silico activities
de novo drug design
in silico ADME/T prediction

Fig. 10.1 In silico activities in the drug discovery process

available that can be used to predict the initial binding interactions can be predicted to develop the desired ligands.

3. *In silico ADME/T prediction*: Computational tools can be used to predict the drug-like properties which differentiate a drug from the lead molecules at an early stage of the drug development process. A lead candidate should exhibit (1) less molecular complexity and low molecular weight with a lesser number of rings and rotatable bonds, (2) less hydrophobic nature with lower clogP and LogD, and (3) lower polarizability with lower CMR values. A lead should have well established structure-activity relationship (SAR) where compounds of similar structural diversity should possess a similar binding affinity with the same target. The lead should possess good absorption, distribution, metabolism, and excretion (ADME) properties.
4. *Determination of protein-ligand binding affinity*: Various computational tools are used to predict the protein-ligand binding affinity or protein-ligand binding energy.

There are many targets available in the database (https://www.rcsb.org/) generated via crystallography, NMR, and bioinformatics. The structural analysis of the active site or binding pocket has been proved to be immensely helpful to predict the potential drug molecules that can bind within the active sites with high specificity. Serious research on various life-threatening diseases such as different types of cancers, tuberculosis, various viral diseases including AIDS, malaria, bacterial diseases, and others warrants the need for new drug discovery with minimal side effects. Drug discovery has become one of the frontier areas of research across the globe (https://www.mordorintelligence.com/industry-reports/drug-discovery-market). It has been estimated that the drug discovery market grows exponentially from 2019 to 2027 in terms of revenue generation.

Pharmaceutical industries and academia have been focusing on multidisciplinary aspects of computational tools. Therefore, powerful molecular docking, modeling, and simulation have incorporated in different drug discovery programs for the analysis of various interactions in the drug-target complex. The combination of in silico and experimental methods has a great impact on the development of the most promising lead candidates. Researchers are actively using in silico methods as a powerful tool for the development of structure-activity relationships (SAR) (Hughes et al. 2011).

In the modern drug design approach, in silico methods are being extensively used to predict the most prominent conformation of a molecule in the binding cavity of the target and also in the estimation of the binding affinity of the molecules with the target. Today, various in silico programs and docking algorithms are available to the researchers, and therefore, a strong understanding of these in silico programs, their pros and cons are of fundamental importance to generate effective strategies and to build successful results. Robust and sophisticated computational tools are an essential part of drug design, thus more handy and refined computational tools are in demand. Moreover, the accurate and precise in silico and experimental methods offer up-to-date information on the various aspects of intermolecular interactions between drug-receptor or ligand-target (Weigelt 2010). With the help of various methods of different approaches generated pharmacodynamics (potency, affinity, efficacy, selectivity) and pharmacokinetic (ADMET: absorption, distribution, metabolism, excretion and toxicity) data which play an essential role in the drug discovery process (Lipinski et al. 2012). This chapter comprises of modern in silico methods, various dimensions of in silico drug design, its various programs, and recent advancements in the drug discovery process.

10.2 Drug Discovery Approaches

The main aim of drug discovery and development is the identification of compounds with improved therapeutic values for their application in the treatment of various diseases with minimal side effects to enhance the quality of life. In drug discovery and pharmacology, the compounds used are small organic molecules that can specifically bind to the target (biomolecules). The drug discovery process can be considered in two ways as (1) traditional drug discovery approach and (2) modern drug discovery approach.

10.2.1 Traditional Drug Discovery Approach

Earlier in the 1960s drug discovery approaches in the pharma industry were based on the screening of thousands of small molecules (synthetic as well as natural) for biological activity (Reddy and Parrill 1999). Once, a potential lead molecule was identified by then hundreds of lead analogs or related compounds were synthesized and these compounds were again evaluated for biological screening. Traditional drug discovery was expensive and associated with risk factors and has been estimated the total expense for a single novel drug was in millions. The traditional methods have various pros and cons because of the reason why a molecule/compound is active or inactive in a certain condition and how it can be improved. Further, it was unable to provide any surety that which particular active lead molecule is specific for a given human receptor protein, due to this reason clinical trials were halted. A large molecular library was used to get one lead compound and after that structural optimization of that lead molecule was performed to enhance potency and other properties. In this traditional drug discovery process, the major problem arises when to proceed from screening to synthesis (Young et al. 1997).

10.2.2 Limitation of Traditional Drug Discovery

To obtain the novel compounds with improved bioavailability and minimal side effects, extensive experiments, screening, in vitro and in vivo testing and ADMET profiling are essential. The traditional drug discovery approach required intense labor, a time-consuming multi-step process to develop novel molecules for in vivo biological screening and ADMET profiling to deliver a drug (AlQaraghuli et al. 2017). To overcome the drawbacks of the traditional drug discovery process new highly interdisciplinary approaches came into existence.

10.2.3 Modern Drug Discovery Approach

Recently, an additional component i.e. in silico approach for computer-aided drug design (CADD) has been used that offers additional benefits to the drug discovery process (Fig. 10.2). Use of in-silico chemistry and molecular modeling in drug discovery process has gained significant attention from the researcher. Besides drug discovery the in silico approach has a significant impact in molecular biology, nanotechnology, agriculture, biochemistry, environmental chemistry. Therefore, in silico approach offers several benefits to minimize the cost and time and to improve the affinity, specificity, selectivity, and ADMET properties of the drug. Of course, modern technologies are multi-step and highly interdisciplinary which are being used in the drug discovery process have the potential to minimize many challenges. It can select and represent a drug or drug-like molecules from the database (a list of thousands/millions of molecules). A minor modification in a molecular structure of a drug like molecules can significantly enhance or reduce the affinity, therapeutic effect, ADMET properties, half-life and side effects, which can be predicted by in silico methods (Young 2009). With the help of in silico methods, one can identify those molecules failed to enter into a clinical trial or not approved by the FDA for a particular disease.

With the aim of the improvement of efficacy and selectively having improved ADME properties, in silico approach can be used to improve the basic scaffold of an existing drug. A new drug can be designed based on the pharmacological properties varying with the modification of molecular structure, by structure-activity relationship and by investigation of the drug-receptor interactions applying in silico approach (Gunjan et al. 2013).

In-Silico Drug Design (Computer-Aided Drug Design)

The in silico term was coined first time in 1989 like another Latin phrase in vivo, in vitro, and in situ (Plewczynski et al. 2014). In general, in silico drug design and drug discovery means the rational drug design using computational methods,

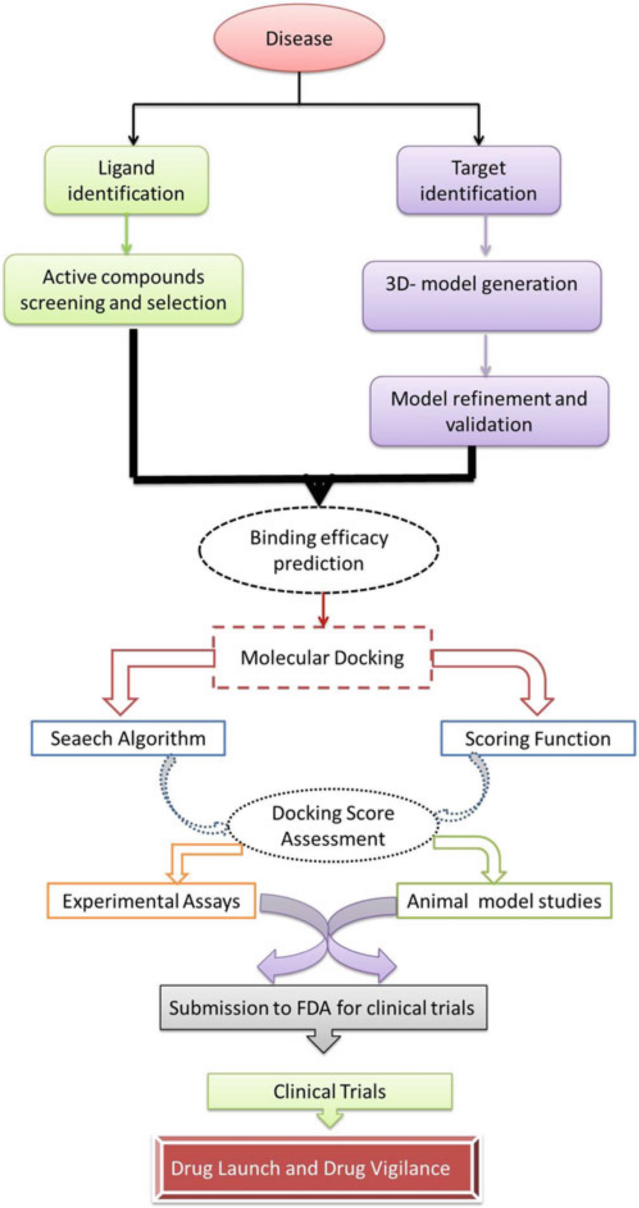


Fig. 10.2 Schematic representation of computer-aided drug design

however, drug discovery is a multidisciplinary process. It needs an understanding of several aspects of the biochemistry of genes, signaling pathways, proteins involved in diseases, or that cause the diseases. Based on the resulting information researcher can design the structure of molecules that can modulate the role of that particular

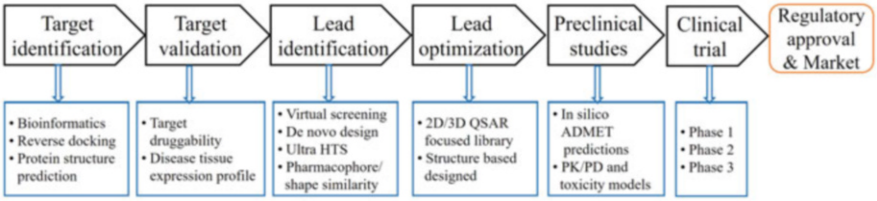

Target identification
Target validation
Lead identification
Lead optimization
Preclinical studies
Clinical trial
Regulatory approval & Market
• Bioinformatics
• Reverse docking
• Protein structure prediction
• Target druggability
• Disease tissue expression profile
• Virtual screening
• De novo design
• Ultra HTS
• Pharmacophore/ shape similarity
• 2D/3D QSAR focused library
• Structure based designed
• In silico ADMET predictions
• PK/PD and toxicity models
• Phase 1
• Phase 2
• Phase 3

Fig. 10.3 Outline of in silico drug design and discovery process

gene, signaling pathways, and proteins involved in a particular disease. Figure 10.3 depicts that both the computational and experimental methods are essential to speed up the drug discovery process (Young 2009). Therefore, in the modern era, in silico/ Computer-Aided Drug Design (CADD) plays an essential role in accelrating the process of discovery and development of new drugs.

The rational drug discovery has progressed at a significant level with help of in silico methods, simultaneously several innovations in hardware and software came into existence (Macalino et al. 2018). In comparison to traditional drug discovery techniques, in silico/CADD techniques resulted in many success stories (Table 10.1) which inspired the researchers to use the application of in silico/CADD both in academia and industry (Kokkonen et al. 2015).

The routine practice of computational tools needs the accessibility of huge amounts of information on proteins and ligand structures; protein functions and experts grasp information related to bonds (intermolecular forces) and bond energies essential for the strong interaction between ligand and target. Various computational tools incorporate several benefits at different stages of the drug discovery process such as target and its hit identification, SAR, hit to lead optimization and prediction of pharmacokinetic and pharmacodynamics properties. If in silico methods are included in the multidisciplinary drug discovery process, the chance of success of new drug increases. Therefore, CADD offers extreme advantages to the researchers to bypass random screening of billions of molecules. According to the industry expert's opinion, approximately 30% of time and cost can be saved in the drug discovery process using in silico approach. Nowadays various methodologies adopted in silico approach become an essential part of the drug discovery process. Therefore, in silico methodologies offered a advantages by reducing (1) the time and cost of new drugs and (2) the use of animals in in vivo studies. It also helps the medicinal chemist in the identification of new targets, new drug-like molecules, and their ADME properties.

To date, significant progress has been made in silico approaches to carry out new experimental procedures for the functional and structural analysis of target and ligand. Among the various methods of in silico approach, the Comparative Molecular Similarity Indices Analysis (CoMSIA) and Comparative Molecular Field Analysis (CoMFA) both are preferred ligand-based (LB) approaches for fast virtual screening (VS) and the prediction of biological activities of a class of ligands. Further, Molecular Dynamics (MD)-quantitative structure-activity relationship

Table 10.1 Lists of drug discovered by CADD

S. No.	Drug name	Status	Route	Approval year	Therapeutics action	References
1	Captopril	Approved	Oral	1981	Antihypertensive	Cushman et al. (1977), Ondetti et al. (1977)
2	Dorzolamide	Approved	Ophthalmic	1995	Carbonic anhydrase inhibitor	Baldwin et al. (1989)
3	Saquinavir	Approved, investigational	oral	1995	HIV protease inhibitor	Roberts (1990)
4	Zanamivir	Approved, investigational	Respiratory (inhalation)	1999	Inhibit neuraminidase.	Ryan et al. (1994), Hayden et al. (1996)
5	Aliskiren	Approved, investigational	Oral	2007	Human renin inhibitors	Goschke et al. (1997)
6	Boceprevir	Approved, withdrawn	Oral	Phase III clinical trials	Hepatitis C virus (HCV) inhibitor	Njoroge et al. (2008)
7	Nolatrexed	Investigational	NA	Phase III clinical trials	Recurrent or unresectable liver cancer	Webber et al. (1993)
8	TMI-005	Investigational	NA	Phase II clinical trials	Active rheumatoid arthritis	Levin et al. (2006)
9	LY-517717	Investigational	NA	Phase II clinical trials	Serine protease inhibitor	Agnelli et al. (2007)
10	Rupintrivir	Investigational	NA	NA	Rhinovirus 3C protease inhibitor	Patick (2006)
11	NVP-AUY922	Investigational	NA	Phase I clinical trials	Inhibitor for HSP90	Eccles et al. (2008), Caldwell et al. (2008)
12	Nelfinavir	Approved	Oral	1997	HIV-1 protease inhibitor	Perez et al. (2007)
13	Raltitrexed	Approved, investigational	Intravenous	1998	Thymidylate synthase inhibitor	Anderson (2003)
14	Amprenavir	Approved, Investigational	Oral	1999	HIV protease inhibitor	Wlodawer and Vondrasek (1998), Clark (2006)
15	Raltegravir	Approved	Oral	2007	HIV protease inhibitor	Anker and Corales (2008), Summa et al. (2008)

NA not available

(QSAR) MD-QSAR models include a combination of Molecular Dynamics (MD) and relative computed descriptors. MD-QSAR models provide the enhanced power to researchers to predict the biological activities. After structural and functional analysis of three dimensional (3D) structure of ligand-target complex, structure-based drug design (SBDD) such as SB pharmacophore models has been selected for the selection of novel ligands for the target. Simultaneously, MD simulations exist as one of the main and efficient computational tools to investigate the dynamics and thermodynamics of the ligand-target complex. In contrast, if the crystal structure, PDB code of target protein is not present in database or protein is unknown, homology modeling techniques can be used to generate a 3D model by employing the amino acid sequence available. Besides, quantum mechanics/molecular mechanics (QM/MM) calculations are useful to investigate the enzyme kinetics and mechanism of action of the enzyme of the target and ligand complex. Recently, QM/MM technique came into existence as a powerful tool by a combination of molecular dynamics resulting in QM/MM-MD technique to better characterization of enzymatic mechanism that generates the high number of QM calculations but also creates a computational burden. With the aim of reduction of the computational burden by maintaining the valuable maximum information, modern theoretical methods have been used for the selection of relevant configurations. Concisely, these computational tools make the drug discovery cycle shorten and cost-effective. Significant improvement in software, hardware, and biological database, and chemical database, algorithm design made the computational tools more valuable in preclinical research (Young 2009).

The following are many computational approaches useful to calculate or to predict the binding interactions and to expedite the drug design process.

Structure-Based Drug Design (SBDD)

The SBDD systematically uses the structural data including molecular targets such as receptors, proteins, and enzymes generated by experimental techniques and by employing computational homology modeling (Manda et al. 2009). These molecular targets are mostly the biomolecules (proteins and enzymes) in metabolic or cell signaling pathways of the biological system. These targets are overexpressed or associated with a particular disease. Further, chemical compounds to generate drug-like molecules are designed to interrupt these molecular targets. These chemical compounds are thought to inhibit, restore, or otherwise alter the structure and functions of molecular targets related to a particular disease. The first aim of the use of computer tools is to design ligands having appropriate electrostatic, structural, and functional features to obtained drug/ligand with a high binding affinity towards the molecular target. The 3D structures of molecular targets available in the database assist an attentive examination of their structural and functional analysis of binding cavity, sub-pockets, clefts, and their electrostatic properties including charge distribution. Recently, the SBDD method used for drug design offers to design ligands having essential structural features for the modulation of the molecular target

(Blaney 2012). Selective modification on the interface of the high-affinity ligand towards validated target ultimately resulted in the generation of a potential ligand with high affinity and improved cellular response (Urwyler 2011). SBDD is not a simple and single approach of drug design but a process of combination of various computational tools.

Target Determination

Drug Target: In drug design process identification of a valid target is one of the essential step and drug target is a biomolecule which is closely associated with cellular processes, cell signaling or metabolic pathways that are specific to a disease process. For example, the human lactate dehydrogenase A (*h*LDHA) enzyme is a viable target to discover an anticancer drug that can selectively kill the cancer cells with minimal side effects (Rani and Kumar 2017). The *h*LDHA enzyme is the last enzyme in tumor glycolysis which catalyzes the reduction of pyruvate to lactate coupled with oxidation of NADH to NAD^+. It is highly overexpressed in many cancers including head and neck, lung, stomach, intestine ovarian, breast, brain, and kidney (Maftouh et al. 2014). In SBDD, the determination of the 3D structure of a target is the initial step which is usually determined by the X-ray crystallography, Cryo-Electron microscopy, and NMR (Rani and Kumar 2016; Kumar and Rani 2019; Rani 2019; Kumar et al. 2017) Further, homology modeling and protein folding have also been used for target identification.

Homology Modeling

The homology modeling is used to build the 3D-structure of the protein sequence known in the genome. It is a cost-effective, easy to use and safe tool, which assists in determination, investigation, explanation, and identification of structural, functional, and molecular properties by using three-dimensional structures of proteins. Since different models yield different results, it is necessary to have a small number of standard models that apply to very large systems. Therefore, these 3D models of the target can be generated by using online tools that are freely available such as SWISS-MODEL Workspace, software such as Autodock, and commercial packages such as Patch dock and Schrodinger. For example, Bansal and Jabalia generated a 3D model (Fig. 10.4) of sodium-dependent serotonin transporter of *Homo sapiens* by using the SWISS-MODEL Workspace (Bansal and Jabalia 2017). This sodium-dependent serotonin transporter of *Homo sapiens* is responsible for the neurological disorder.

Protein Folding

In search of the target identification another is protein folding. Unfortunately, this process is difficult and is started with the primary sequence only and runs a calculation that tries an incredibly large number of conformers. Further, this is an assumption based effort to figure out the correct shape of the protein with the lowest energy conformer.

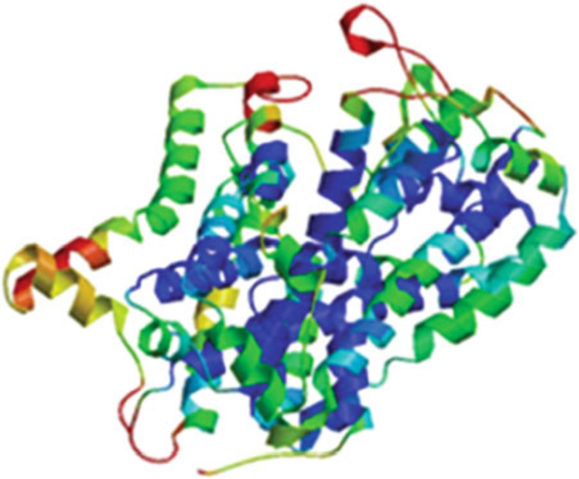


Fig. 10.4 Generation of three dimensional model of sodium-dependent serotonin transporter of *Homo sapiens* using SWISS-MODEL (Source: Bansal H, Jabalia N. Asian J Pharm Clin Res. 2017;10(8): 299–303)

Ligands Search

For the potential ligand search, the virtual screening method is an efficient method mainly based on the availability of the existing knowledge. This process requires some compounds with biological activities with known inhibitors from the previous studies. The virtual screening method is an efficient process to find out potential ligand from commercially available compounds present in a database with a low molecular weight that can strongly bind with the target molecules. For example, Zinc database and Drug banks are available and researchers can build their database for a specific use.

Molecular Docking

After the target identification and ligand search, molecular docking is the next step which is used to predict the binding affinity of the ligand and the target molecules. These target biomolecules play significant roles in various biological functions and form a stable complex with a ligand (Engel et al. 2008). Molecular docking plays a central role in rational drug design. Apart from the binding affinity, molecular docking is used to predict the ligand conformation and orientation within the binding cavity of the target. Further, virtual screening based on molecular docking has emerged a fundamental part of various modern SBDD approaches. Molecular docking has become one of the most growing applications in pharmaceutical/medicinal chemistry. Hence, virtual screening combined with molecular docking becomes a useful endeavor to assess existing various docking programs (Table 10.2) (Lengauer and Rarey 1996).

Table 10.2 Various free and commercial docking tools and software

S. No.	Software	Description	Link
1	Dock	It is fragment based package using chemical complementary method and shape. It is able to create various possible orientation of the ligand	http://dock.compbio.ucsf.edu/
2	AutoDock	Free, Suit of automated docking tools. Applicable to calculate the binding affinity of small molecule to 3D receptor/target	http://autodock.scripps.edu/
3	Gold	It is able to predict the binding affinity of flexible drug with flexible protein	https://www.ccdc.cam.ac.uk
4	V life	It is an integrated software package for CADD	http://www.vlifesciences.com
5	ICM	It is based on stochastic algorithm and is applicable to optimization of all flexible ligand in receptor filed	https://www.molsoft.com
6	Glide	It is a grid-based docking program which is used to investigate a systematic search of ligand positions, orientations and conformations with in target	https://www.schrodinger.com/glide
7	Swissdock	Swissdock is bale to predict molecular interactions between the ligand and target	http://www.swissdock.ch/docking
8	Arguslab	Molecular modeling, graphics, and drug design program	http://www.arguslab.com
9	1-ClickDocking	Predicts binding orientation and affinity of a ligand to a target	https://mcule.com/apps/1-click-docking/
10	COPICAT	It is a web service for discovery of lead compound	http://copicat.dna.bio.keio.ac.jp/
11	Sanjeevini	Active site-directed drug design	http://www.scfbio-iitd.res.in/sanjeevini/sanjeevini.jsp
12	Insight II	Graphical molecular modeling program	http://www.serc.iisc.ac.in/software/insight-ii/
13	Sybyl	Drug and molecular design environment with comprehensive tools for molecular modeling	https://omictools.com/sybyl-x-tool
14	FTDOCK	Ligand and receptor docking program	https://omictools.com/ftdock-tool

Molecular Modeling

Molecular modeling techniques are applied to calculate the number of molecules interacted with the other targets. The modeling of biomolecular complexes (target-ligand complex) by using the known target and their constituents (substrate and cofactor) with computational docking is growing rapidly as a powerful computational process in structural biology. Molecular modeling is useful, particularly where limited experimental information is available which describes the interface of the complex (Tikhonova et al. 2008). The binding ability of ligand towards a target directs an important part of the dynamics of protein that may affect the biological function negatively or positively. This implication offers an important role in rational drug design. Moreover, the capability of the target to bind with another large biomolecule such as proteins and nucleic acids generates supra-molecular complexes. These generated supra-molecular structures also play a crucial role in several biological functions and in controlling biological pathways. For example, *protein-protein interactions* (PPIs) have been widely studied and are an essential part of almost all processes in a cell, cell development, cell to cell interactions, and metabolic processes. These PPIs are specific and take place in defined binding positions in the protein. These PPIs offer various applications in the prediction of protein function, drug ability of small molecules (Goktepe and Kodaz 2018; Van Dijk et al. 2006). Docking is charecterized into rigid docking, semi-flexible docking, and flexible docking based on the flexibility of protein and ligand. In rigid docking, protein and ligand both are considered rigid while in case of semi-flexible docking protein is fixed and the ligand is flexible. Flexible docking is considered when both protein and ligand are flexible. Further, the search algorithm positions the ligands in various locations, orientations, and conformations in the binding cavity of the target and may be classified as stochastic, systematic, and deterministic. The accuracy of the docking result cannot determine by the search algorithm but the scoring functions determine. The scoring function is useful to investigate the accuracy weather the position and orientation of the ligand in the binding cavity of the target molecule is energetically favorable or not. Therefore, scoring functions is responsible for calculating the binding energy of the preferred ligand-target complex. One of the effective and widely used algorithms is the Monte Carlo search algorithm which was built around a random number generator. The tabu algorithm generates the same results with fewer iterations than the Monte Carlo search algorithm by discarding the duplicate work because it can track the positions that have already been sampled and avoids these positions to be sampled.

Scoring Functions

Various scoring functions can be used to calculate the binding affinity of the ligand towards the target to generate the score. This docking score has an impact on the ranking of the chemical compounds obtained by a virtual screening experiment. Although these scoring functions have ranged from calculating binding energy to estimate the free energy of protein-ligand complex by a simple shape and electrostatic complementarities, only a few can address the thermodynamic process during the binding process (Guedes et al. 2018). The main objective of the scoring function is to define the correct poses from incorrect poses hence, various assumptions and simplifications are adopted instead of calculating the binding affinity between receptor and ligand complex (Meng et al. 2011). Scoring functions can be categorized in force-field-based functions, empirical scoring functions, and knowledge-based scoring functions in addition to Consensus scoring functions (Kitchen et al. 2004).

(vi-a) Force-field-based functions

(vi-b) Empirical scoring functions
(vi-c) Knowledge-based scoring functions
(vi-d) Consensus scoring functions
(vi-e) Hybrid scoring functions

(vi-a) Force field-based functions: It constitutes energy terms from the classical force field including receptor-ligand complex interactions energies and internal ligand energy. Poisson–Boltzmann (PB) or the related Generalized Born (GB) is used to measuring continuum solvation (Gilson et al. 1997; Zou and Kuntz 1999). Force field based functions have a great impact in the drug discovery process. In 1995, a strong correlation was attained between nonbonded interaction energies computed with a modified MM2X force field host molecular mechanics program (OPTIMOL) and in vitro IC_{50} values of 33 inhibitors against human immunodeficiency virus (HIV)-1 protease (Holloway et al. 1995). The AMBER and CHARMM force field are widely used and have well-defined parameters. Although, AMBER has well-defined parameters such as bond length, bond angle, van der Waals interactions, and partial charge for proteins and nucleic acids but have limited parameters towards modification in proteins and nucleic acids. Comparatively, CHARMM force field has advanced features and has an all-atom force field for proteins, nucleic acids, lipids, and carbohydrates. The AMBER (Weiner et al. 1984, 1986; Cornell et al. 1995) and CHARMM (Brooks et al. 1983) nonbonded terms are used as a scoring function in several docking programs. The field-based scoring functions include GOLD (Verdonk et al. 2003) and AutoDock (Morris et al. 1998), DOCK and DockThor (de Magalhães et al. 2014) which are widely used in in silico drug discovery process. Besides, the extensions of force-field-based scoring functions consider the hydrogen bonds, salvations, and entropy contributions. Along with these many software programs such as DOCK, GOLD (Verdonk et al. 2003) and AutoDock (Morris et al. 1998) have been commonly used for prediction of binding energy calculation.

(vi-b) Empirical scoring functions: The major goal of empirical scoring functions is to measure the binding energy of noncovalent protein-ligand interactions based on the sum of localized and chemically intuitive interactions even without interpreting the 3D model of complex (Meng et al. 2011). Following empirical scoring functions such as LUDI (Böhm 1992); ChemScore (Eldridge et al. 1997); PLP (Gehlhaar et al. 1995); ID-Score (Li et al. 2013); and GlideScore (Friesner et al. 2006) have been used for calculation of binding energy.

(vi-c) Knowledge-based scoring functions: Knowledge based scoring function is a statistical interpretation of interacting atom pairs from receptor-ligand complexes with available 3D-structures where pairwise-atom data is converted into a pseudopotential, that elucidates the geometries of protein-ligand. Knowledge based scoring functions including DrugScore (Velec et al. 2005), Bleep (Mitchell et al. 1999), SMoG (DeWitte and Shaknovich 1996) and potential of mean force (PMF) (Muegge 2006) have been used depending upon varying atom size and type of energy function.

(vi-d) Consensus scoring: Consensus scoring function is the sum-up of various scores to predict the docking conformation (Charifson et al. 1999). It combines many

scoring functions to calculate the binding affinity leads to prominent hit rates. The major objective of consensus scoring is to improve the binding confirmations, poses, and enrichments in virtual screening (Feher 2006). An example of consensus scoring is CScore (Clark et al. 2002) which includes various scoring function programs such as ChemScore, PMF, GOLD, DOCK, and FlexX.

(vi-e) Hybrid scoring functions: Recently a new form of scoring functions termed as hybrid scoring functions a mixture of force-field, contact- and knowledge-based descriptors, has been developed. The example of hybrid scoring functions are GalaxyDock BP2 Score (empirical, knowledge-based, and force-field based) (Baek et al. 2017), DockTScore from the DockThor program (empirical and force-field based) (Guedes et al. 2016) and SMoG2016 (empirical and knowledge-based) (Debroise et al. 2017).

Ligand Based Drug Design

In ligand-based drug design (LBDD), the 3D structure of the target is unknown but the structure and functions of ligand that binds to the target are known. Further, that ligand can be used to develop drug design and pharmacophore models or chemical compounds. These designed molecules have structural features that are necessary to show the interaction with the target. In general, LBDD is pharmacophore-based and quantitative-structure activity relationships (QSARs) based approach. It has been considered that in LBDD, that compounds showed high similarity in structure may show the same biological function and interaction with the target (Macalino et al. 2015).

Molecular Dynamics (MD) Simulations

Molecular dynamics simulation (MD) is one of the main tools in the theoretical study of biological targets. MD is frequently used in the investigation of structural, dynamics, and thermodynamics behavior of biological targets and their complex with the ligand. MD simulations integrats the classical equations of motion generating a phase space trajectory for a system of N particles(González 2011). The MD simulation method is useful to calculate the complete information in detail with the time-dependent manner by investigating the fluctuations and conformational changes of target/biomolecules (proteins, carbohydrates and nucleic acids). Moreover, MD simulations provide valuable information about the possible interactions of target-ligand complex at the atomic level to create conformational ensembles for prediction of target flexibility as well as ligand flexibility in both the SBDD and LBDD studies. This computational tool illustrates an alternative method based on Newton's equations of motion for every atom and position of the atom with a small increment of time (McCammon et al. 1977).

Force Field in MD Simulations

A force field (FF) is a mathematical expression that describes the dependence of the energy of a system on the particle coordinate system. Molecules are a set of atoms that are linked together by simple elastic forces and the FF replaces the true potential with a simplified model valid in the region being simulated (González 2011). Classical MD simulations are used to study ligand binding interactions, enzymatic-reaction mechanisms, protein folding, and un-folding and protein-protein interactions (Lopes et al. 2015). Earlier, the primary goal of FF with the development of molecular mechanics (MM) was utilized for small organic molecules to measure the molecular structures, enthalpies of isolated molecules, and vibrational spectra. Later on, new features were upgraded which is used today. Allinger's group developed the best examples such as MM2 (Allinger 1977), MM3 (Allinger et al. 1989), and MM4 (Allinger et al. 1996). Other Good examples are the Dreiding (Mayo et al. 1990) and Universal (UFF) (Rappe et al. 1992) force fields, that contain parameters for all the atoms in the periodic table. CHARMM-FF is a very popular force field (Mackerell et al. 1998) and earlier CHARMM was utilized for protein later after development it can be used for nucleic acids, lipids, and carbohydrates. AMBER-FF is another common one (Cornell et al. 1995) which is particularly focused on important dihedral angles, besides it also supports the most common biomolecules. There are few other very popular force fields such as GROMOS (Oostenbrink et al. 2004), OPLS (Jorgensen et al. 1996), and COMPASS (Sun 1998) have been used in the drug discovery process. Besides, AMOEBA is a fully functional FF that has been used for proteins (Shi et al. 2013).

Site Identification by Ligand Competitive Saturation (SILCS)

SILCS considered a novel method in CADD that has been employed in identification of the binding sites of representative chemical entities on the entire surface of the receptor. Thus information gained can be utilized for computational fragment-based drug design to discover new ligands. The all-atom explicit-solvent MD simulations has been involved to identify 3D functional-group binding interaction between target the small organic solvent i.e. propane, methanol and etc. (Yu and MacKerell 2017).

SILCS-Pharm

SILCS-Pharm is an alternative method for docking based virtual screenings. It is a target-based pharmacophore virtual screening method which can quickly filter a database to discover potential small molecule that can significantly bind with target. Further, a pharmacophore model is characterized by chemical features which are spatially distributed and are crucial for potential specific interaction of ligand to target (Trosset and Cavé 2019).

Similarity Search

The similarity search method is the most significant, direct, and rapid method in optimization of lead to hit compounds that can have the potential to enter into further drug development process. Besides, it can be used in exploration of new potential compounds that are structurally similar with similar and improved physicochemical properties (Trosset and Cavé 2019).

Lead Optimization Using Structure-Activity Relationship (SAR)

The SAR models have been widely explored in hit to lead optimization. It can be employed to discover new compound having improved biological activity using the data of multiple hits for one target (Trosset and Cavé 2019).

Single Step Free Energy Perturbation (SSFEP)

A higher level computational method i.e. Free Energy Perturbation (FEP) has been proposed with the aim of prediction of increased accuracy in quantification of binding free energy or binding affinity with the modification in a chemical compound. Moreover, this may be in single step known as single-step FEP (SSFEP) to save computational time (Trosset and Cavé 2019).

Pharmacophore-Based Approaches

Although the first step to drug design is a 3-dimensional structure of an enzyme or a complex in-case if the structure of the biological receptor is unknown then various methods can be utilized to identify active analogs. The active or inactive analogs can be applied as a working model as per the need for biological activity also referred to as pharmacophore (Reddy and Parrill 1999). Quantitative structure-activity relationship (QSAR) modeling is the essential cheminformatics technique to predict the biological activities of compounds based on a mathematical and statistical approach. It shows a wide range of applications in the development of hit to lead compound employing virtual screening, prediction of drug-like property, and also chemical risk assessment.

Simply, a QSAR model can be represented by

Function (Chemical compounds) = Biological Activity (predicted)

There are various evolving quantitative methods *i.e.* 2D QSAR, 3D QSAR, and neural networks that utilize active compounds. The Comparative Molecular Field

Table 10.3 Types of QSAR, subtypes of QSAR and their description

Type of QSAR	Sub-types of QSAR	Description
2D QSAR	Hansch analysis	Hansch extends the concept of linear free energy relationships (LFER) to describe the effectiveness of a biologically active molecule
	Free Wilson analysis	It is a structure-activity based methodology which includes the contribution of all various structural fragments essential for overall biological activity
	Statistical methods • Discriminant analysis • Cluster analysis • Principle component analysis • Quantum mechanical methods	Statistical methods are the mathematical foundation for the development of QSAR models
3D-QSAR	• Molecular shape analysis (MSA) • Molecular topological difference (MTD) • Comparative molecular movement analysis (COMMA) • Hypothetical active site lattice (HASL) • Self-organizing molecular field analysis (SOMFA) • Comparative molecular field analysis (COMFA) • Comparative molecular similarity indices (COMSIA)	3D-QSAR used to predict the quantitative relationship of various chemical compounds, their 3D structural features with their biological activity
4D-QSAR	–	The 4D-QSAR is the extended part of 3D, involves conformational and alignment freedom to 3D-QSAR models development
5D-QSAR	–	In addition to 3D and 4D, the 5-D QSAR can signify a communal of up to six different induced-fit models
6D-QSAR	–	The 6D-QSAR models able to investigate the solvation models

Analysis (CoMFA) is one of the is majorly used 3D-QSAR technique. CoMFA represents a significant achievement due to its ability to develop a three-dimensional quantitative model that relates steric and electrostatic fields to biological activity. There are different types of QSAR mentioned (Table 10.3).

Molecular Descriptors

Molecular descriptors used in QSAR Molecular descriptors have been defined as a "numerical representation of chemical information encoded within a molecular structure via mathematical procedure" (Chandrasekaran et al. 2018) (Fig. 10.5).

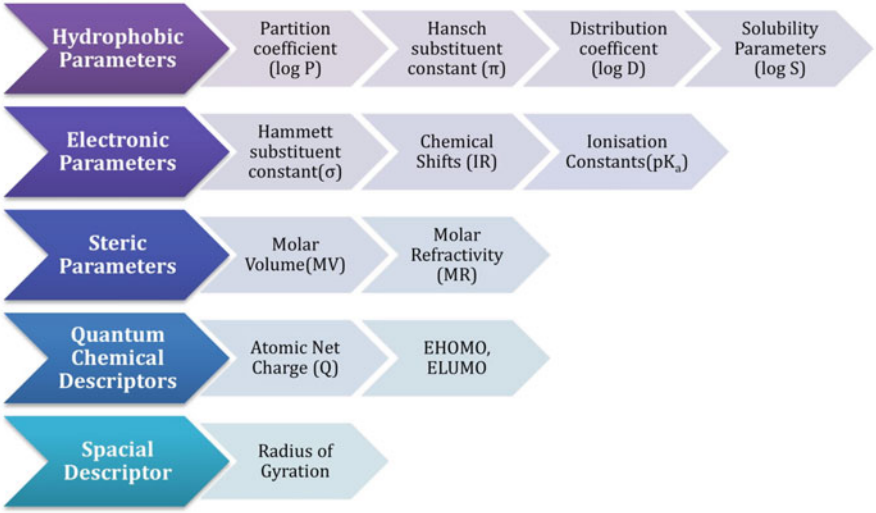

Hydrophobic Parameters
Partition coefficient (log P)
Hansch substituent constant (π)
Distribution coefficent (log D)
Solubility Parameters (log S)
Electronic Parameters
Hammett substituent constant(σ)
Chemical Shifts (IR)
Ionisation Constants(pK$_a$)
Steric Parameters
Molar Volume(MV)
Molar Refractivity (MR)
Quantum Chemical Descriptors
Atomic Net Charge (Q)
EHOMO, ELUMO
Spacial Descriptor
Radius of Gyration

Fig. 10.5 QSAR molecular descriptors

Applications of QSAR

QSAR has been applied effectively over many decades to identify predictive models for the activity of bioactive agents. A QSAR model has been developed for the identification of potential drug-like molecules against chloroquine susceptible strain (3D7) of Plasmodium *falciparum* which is responsible for malaria. This QSAR model is based on 0D, 1D, and 2D Dragon descriptors, fragments descriptors (ISIDA-2D) coupled with support vector machines (SVM) method (Zhang et al. 2013). A QSAR study has been performed on quinolizidinyl an acetylcholinesterase inhibitor by developing QSAR models based on various statistical approaches such as principal component regression (PCR), multiple linear regression (MLR), least absolute shrinkage and selection operator (LASSO) and partial least squares (PLSs) (Ghasemi et al. 2013). A hologram QSAR models have been employed to investigate the inhibitory effect of a series of 36 small molecules against acetyl/ butyrylcholinesterase (AChE/BChE) enzymes for the identification of potential molecules for Alzheimer's disease (AD) (De Souza et al. 2012). Further, hologram-QSAR, 2D-QSAR, and 3D-QSAR models have been developed for calculation of antitumor activity ofbenzo[*c*]phenanthridine (BCP) based derivatives by targeting topoisomerase I (TOP-I) (Thai et al. 2012). Moreover, binary QSAR models have been applied for the identification of new structurally similar compounds with anti-schistosomal activity. The schistosomiasis is caused by flatworms (Kuntz et al. 2007; Neves et al. 2016). Besides, a binary QSAR models using SVM and Naïve Bayesian have been employed to recognize the inhibitory effect against neuraminidase a validated protein target for treating the influenza virus (Lian et al. 2015).

10.2.4 Recent Developments of In Silico Approach and Applications

In Drug Discovery

In silico approaches are viable, significant, and essential for drug design and discovery in the identification of potential molecules possessing a wide range of biological activities such as anticancer, antibacterial, antimicrobial, and anti-inflammatory, etc. The demand for new drugs with minimal side effects is increasing exponentially thus market of the pharmaceutical industry is increasing. Pharmaceutical industries have made huge investments in this therapeutic area. Although besides these efforts, new drug discovery research always remains challenging, hence, therapeutic innovations have not yet achieved the expected clinical outcome (Magalhaes et al. 2018).

In Phenotypic and Target-Based Approach

In silico method is used in the drug discovery process by following the phenotypic and target-based approach. Recently, a computational method based on two machine learning models has been proposed to integrate the data by using the probabilistic framework by following the phenotypic and target-based approaches. The first model is used to predict the targets (proteins) for a ligand by using the CGBVS model which is skilled in the investigation on ligand-target protein interaction. Moreover, the second one is proficient in the statistical selection of significant proteins related to a phenotype. This proposed model offers to execute target deconvolution by following in silico approach which is hard to solve experimentally.

In Silico Prediction of Potential Drug-Like Compounds from Plants

In silico methods not only play a significant role in drug design but also the identification of drug-like compounds from the plant. Recently, In silico approach has been applied to identify potential drug-like compounds extracted from *Anethumsowa* L. root to discover anticancer agents. Anethumsowa L. is a herb, used as a spice in Asia and Europe to add flavor and taste. The molecular docking investigations of the phytochemicals isolated from *Anethumsowa* L. root suggest that it showed significant anticancer properties. Further, in silico study showed their significant drug-like properties that prove their therapeutic potential. Current investigation offers new revenues towards effective drug development against cancer (Saleh-e-In et al. 2019).

In-Silico Study in 3D Adenosine Receptors with Antagonists

Recently, ligand-based and structure-based -molecular modeling techniques have been widely explored in a wide range of applications. In silico tools have been applied to design potent and selective antagonists for each adenosine receptor (AR) subtype considering a mass of drug targets. The homology modeling, QSAR, pharmacophore models, and molecular docking combining with more accurate free energy calculation methods can critically utilized to investigate structure-based virtual screening of AR antagonists.

In Metal-Organic Frameworks

Not only in drug discovery but in silico tools are useful in the development of Metal-organic frameworks (MOFs). MOFs are crystalline nanoporous material with extended-network exhibiting a wide range of applications such as in separations, sensing and gas storage. Molecular modeling and molecular simulations have been applied in the simulation of adsorption of gas in MOFs. This study suggests the discovery of performant MOFs for adsorption and storage of gases such as methane, hydrogen, oxygen, xenon and carbon dioxide. These MOFs also used in chemical warfare agent capture, and xylene enrichment. This paves a new platform for the discovery of computational material that has real applications (Sturluson et al. 2019).

10.3 Conclusions

All the methods discussed above play a significant role in drug discovery and several advancements have been in these methods. Concisely, molecular docking and molecular modeling tools showed various biological applications, therefore, many significant efforts have been made for further advancement and to understand the molecular docking and modeling process (van Dijk et al. 2006). In research, various struggles focused on the elimination of compounds with serious side effects in the early stage of drug discovery to select novel drug-like candidates that can enter into the clinical trial. These in silico tools have become an integral part of the drug design and discovery process because these are significant tools to save time, money, and manpower in a preclinical trial. Moreover, these are helpful in the management of a large number of data (Lappano and Maggiolini 2011).

Acknowledgments Atul Kumar is thankful to Amity University Greater Noida. Neetu Jabalia, Reshma Rani, and Vinit Kumar are thankful to Amity University Noida for technical support and DST-SERB for financial assistance.

Conflict of Interest Statement The author(s) declare no conflicts of interest to declare.

References

Agnelli, G., Haas, S., Ginsberg, J. S., Krueger, K. A., Dmitrienko, A., & Brandt, J. T. (2007). A phase II study of the oral factor Xa inhibitor LY517717 for the prevention of venous thromboembolism after hip or knee replacement. *Journal of Thrombosis and Haemostasis, 5*, 746–753.

Allinger, N. L. (1977). Conformational analysis. 130. MM2. a hydrocarbon force field utilizing V1 and V2 torsional terms. *Journal of the American Chemical Society, 99*, 8127–8134.

Allinger, N. L., Chen, K., & Lii, J.-H. (1996). Improved force field (MM4) for saturated hydrocarbons. *Journal of Computational Chemistry, 17*, 642–668.

Allinger, N. L., Yuh, Y. H., & Lii, J.-H. (1989). Molecular mechanics. the MM3 force field for hydrocarbons. 1. *Journal of the American Chemical Society, 111*, 8551–8565.

AlQaraghuli, M. M., Alzahrani, A. R., Niwasabutra, K., Obeid, M. A., & Ferro, V. A. (2017). Where traditional drug discovery meets modern technology in the quest for new drugs. *Annals of Pharmacology and Pharmaceutics, 11*, 1–5.

Anderson, A. C. (2003). The process of structure-based drug design. *Chemistry & Biology, 10*, 787–797.

Anker, M., & Corales, R. B. (2008). Raltegravir (MK-0518), a novel integrase inhibitor for the treatment of HIV infection. *Expert Opinion on Investigational Drugs, 17*, 97–103.

Baek, M., Shin, W.-H., Chung, H. W., & Seok, C. (2017). GalaxyDock BP2 score: A hybrid scoring function for accurate protein–ligand docking. *Journal of Computer-Aided Molecular Design, 31*, 653–666.

Baldwin, J. J., Ponticello, G. S., Anderson, P. S., Christy, M. E., Murcko, M. A., Randall, W. C., et al. (1989). Thienothiopyran-2-sulfonamides: Novel topically active carbonic anhydrase inhibitors for the treatment of glaucoma. *Journal of Medicinal Chemistry, 32*, 2510–2513.

Bansal, H., & Jabalia, N. (2017). In silico characterization and molecular modeling of sodium dependent serotonin transporter protein from Homo sapiens. *Asian Journal of Pharmaceutical and Clinical Research, 10*, 299–303.

Blaney, J. (2012). A very short history of structure-based design: How did we get here and where do we need to go? *Journal of Computer-Aided Molecular Design, 26*, 13–14.

Böhm, H. J. (1992). The computer program LUDI: A new method for the de novo design of enzyme inhibitors. *Journal of Computer-Aided Molecular Design, 6*, 61–78.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., & Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry, 4*, 187–217.

Caldwell, J. J., Davies, T. G., Donald, A., McHardy, T., Rowlands, M. G., Aherne, G. W., et al. (2008). Identification of 4-(4-aminopiperidin-1-yl)-7H-pyrrolo[2,3d]pyrimidines as selective inhibitors of protein kinase B through fragment elaboration. *Journal of Medicinal Chemistry, 51*, 2147–2157.

Chandrasekaran, B., Abed, S. N., Al-Attraqchi, O., Kuche, K., & Tekade, R. K. (2018). Computer-aided prediction of pharmacokinetic (ADMET) properties. In *Dosage form design parameters* (Vol. 1, pp. 731–755). Cambridge, MA: Academic Press.

Charifson, P. S., Corkery, J. J., Murcko, M. A., & Walters, W. P. (1999). Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *Journal of Medicinal Chemistry, 42*(25), 5100–5109.

Clark, D. E. (2006). What has computer-aided molecular design ever done for drug discovery? *Expert Opinion on Drug Discovery, 1*, 103–110.

Clark, R. D., Strizhev, A., Leonard, J. M., Blake, J. F., & Matthew, J. B. (2002). Consensus scoring for ligand/protein interactions. *Journal of Molecular Graphics and Modelling, 20*(4), 281–295.

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., et al. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society, 117*(19), 5179–5197.

Cushman, D. W., Cheung, H. S., Sbo, E. F., & Ondetti, M. A. (1977). Design of potent competitive inhibitors of angiotensin converting enzyme. Carboxyalkanoyl and mercaptoalkanoyl amino acids. *Biochemistry, 16*, 5484–5491.

de Magalhães, C. S., Almeida, D. M., Barbosa, H. J. C. & Dardenne, L. E. (2014). A dynamic niching genetic algorithm strategy for docking highly flexible ligands. *Information Sciences, 289*, 206–2243.

De Souza, S. D., De Souza, A. M., De Sousa, A. C., Sodero, A. C., Cabral, L. M., Albuquerque, M. G., et al. (2012). Hologram QSAR models of 4-[(diethylamino) methyl]-phenol inhibitors of acetyl/butyrylcholinesterase enzymes as potential anti-Alzheimer agents. *Molecules, 17*(8), 9529–9539.

Debroise, T., Shakhnovich, E. I., & Chéron, N. (2017). A hybrid knowledge-based and empirical scoring function for protein–ligand interaction: SMoG2016. *Journal of Chemical Information and Modeling, 57*, 584–593.

DeWitte, R. S., & Shakhnovich, E. I. (1996). SMoG: de Novo design method based on simple, fast, and accurate free energy estimates. 1 Methodology and supporting evidence. *Journal of the American Chemical Society, 118*, 11733–11744.

Eccles, S. A., Massey, A., Raynaud, F. I., Sharp, S. Y., Box, G., Valenti, M., et al. (2008). NVP-AUY922: A novel heat shock protein 90 inhibitor active against xenograft tumor growth, angiogenesis, and metastasis. *Cancer Research, 68*, 2850–2860.

Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V., & Mee, R. P. (1997). Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *Journal of Comput Aided Molecular Design, 11*(5), 425–445.

Engel, S., Skoumbourdis, A. P., Childress, J., Neumann, S., Deschamps, J. R., Thomas, C. J., et al. (2008). A virtual screen for diverse ligands: Discovery of selective G protein-coupled receptor antagonists. *Journal of the American Chemical Society, 130*, 5115–5123.

Feher, M. (2006). Consensus scoring for protein-ligand interactions. *Drug Discovery Today, 11* (9-10), 421–428.

Friesner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., et al. (2006). Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *Journal of Medicinal Chemistry, 49*, 6177–6196.

Gehlhaar, D. K., Moerder, K. E., Zichi, D., Sherman, C. J., Ogden, R. C., & Freer, S. T. (1995). De novo design of enzyme inhibitors by Monte Carlo ligand generation. *Journal of Medicinal Chemistry, 38*(3), 466–472.

Ghasemi, G., Arshadi, S., Rashtehroodi, A. N., Nirouei, M., Shariati, S., & Rastgoo, Z. (2013). QSAR investigation on quinolizidinyl derivatives in Alzheimer's disease. *Journal of Computational Medicine, 13*, 1–8.

Gilson, M. K., Given, J. A., & Head, M. S. (1997). A new class of models for computing receptor-ligand binding affinities. *Chemistry & Biology, 4*, 87–92.

Goktepe, Y. E., & Kodaz, H. (2018). Prediction of protein-protein interactions using an effective sequence based combined method. *Neurocomputing, 303*, 68–74.

González, M. A. (2011). Force fields and molecular dynamics simulations. *École thématique de la Société Française de la Neutronique, 12*, 169–200.

Goschke, R., Cohen, N. C., Wood, J. M., & Maibaum, J. (1997). Design and synthesis of novel 2,7-dialkyl substituted 5(S)-amino-4(S)-hydroxy8-phenyl-octanecarboxamides as in vitro potent peptidomimetic inhibitors of human renin. *Bioorganic & Medicinal Chemistry Letters, 7*, 2735–2740.

Guedes, I. A., Barreto, A. M. S., Miteva, M. A., & Dardenne, L. E. (2016). Development of empirical scoring functions for predicting protein-ligand binding affinity. *Sociedade Brasileira de Bioquímica e Biologia Molecular*, 1–174.

Guedes, I. A., Pereira, F. S., & Dardenne, L. E. (2018). Empirical scoring functions for structure-based virtual screening: Applications, critical aspects, and challenges. *Frontiers in Pharmacology, 9*, 1089.

Gunjan, K., Dinesh, S., Yogesh, V., & Vishal, S. (2013). A review on drug designing, methods, its applications and prospects. *IJPRD, 5*, 15–30.

Gupta, P. K., Agrawal, P., Shivakumar, N., & Hiremath, S. B. (2011). In Silico modelling and drug design – A review. *Int Res J Pharm, 2*, 15–17.

Hayden, F. G., Treanor, J. J., Betts, R. F., Lobo, M., Esinhart, J., & Hussey, E. K. (1996). Safety and efficacy of the neuraminidase inhibitor GG167 in experimental human influenza. *Journal of the American Medical Association, 275*, 29529.

Holloway, M. K., Wai, J. M., Halgren, T. A., Fitzgerald, T. M. D., Vacca, J. P., Dorsey, B. D., et al. (1995). A priori prediction of activity for HIV-protease inhibitors employing energy minimization in the active site. *Journal of Medicinal Chemistry, 38*, 305–317.

Hughes, J. P., Rees, S., Kalindjian, S. B., & Philpott, K. L. (2011). Principles of early drug discovery. *British Journal of Pharmacology, 162*, 1239–1249.

Jorgensen, W. L., Maxwell, D. S., & Tirado-Rives, J. (1996). Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society, 118*, 11225–11236.

Kitchen, D. B., Decornez, H., Furr, J. R., & Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nature Reviews Drug Discovery, 3* (11), 935–949.

Kokkonen, P., Kokkola, T., Suuronen, T., Poso, A., Jarho, E., & Lahtela-Kakkonen, M. (2015). Virtual screening approach of sirtuin inhibitors results in two new scaffolds. *European Journal of Pharmaceutical Sciences, 76*, 27–32.

Kumar, V., Kumar, A., & Rani, R. (2017). Regulation/inhibition of human lactate dehydrogenase A: An innovative and potential approach for anti-cancer drugs development. In *Topics in anti-cancer research* (Vol. 6, pp. 114–142). Sharjah, UAE: Bentham Science Publishers.

Kumar, V., & Rani, R. (2019). Lactate dehydrogenase enzyme: An old enzyme but new viable target offers new hope in cancer therapeutics. In *lactate dehydrogenase (LDH): Biochemistry, function and clinical significance* (Vol. 1). Hauppauge, NY: Nova Publishers.

Kuntz, A. N., Davioud-Charvet, E., Sayed, A. A., Califf, L. L., Dessolin, J., Arnér, E. S. J., et al. (2007). Thioredoxin glutathione reductase from *Schistosoma mansoni*: An essential parasite enzyme and a key drug target. *PLoS Medicine, 4*(6), e206.

Lappano, R., & Maggiolini, M. (2011). G protein-coupled receptors: Novel targets for drug discovery in cancer. *Nature Reviews Drug Discovery, 10*, 47.

Lengauer, T., & Rarey, M. (1996). Computational methods for biomolecular docking. *Current Opinion in Structural Biology, 6*, 402–406.

Levin, J. I., Chen, J. M., Laakso, L. M., Du, M., Schmid, J., Xu, W., et al. (2006). Acetylenic TACE inhibitors. Part 3. Thiomorpholine sulfonamide hydroxamates. *Bioorganic & Medicinal Chemistry Letters, 16*, 1605–1609.

Li, G.-B., Yang, L.-L., Wang, W.-J., Li, L.-L., & Yang, S.-Y. (2013). ID-score: A new empirical scoring function based on a comprehensive set of descriptors related to protein–ligand interactions. *Journal of Chemical Information and Modeling, 53*, 592–600.

Lian, W., Fang, J., Li, C., Pang, X., Liu, A.-L., & Du, G.-H. (2015). Discovery of influenza A virus neuraminidase inhibitors using support vector machine and Naïve Bayesian models. *Molecular Diversity, 20*, 439–451.

Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (2012). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews, 64*, 4–17.

Lopes, P. E., Guvench, O., & MacKerell, A. D. (2015). Current status of protein force fields for molecular dynamics simulations. In *Molecular modeling of proteins* (pp. 47–71). New York: Humana Press.

Macalino, S. J., Basith, S., Clavio, N. A., Chang, H., Kang, S., & Choi, S. (2018). Evolution of in silico strategies for protein-protein interaction drug discovery. *Molecules, 23*, 1963.

Macalino, S. J., Gosu, V., Hong, S., & Choi, S. (2015). Role of computeraided drug design in modern drug discovery. *Archives of Pharmacal Research, 38*, 1686–1701.

Mackerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., et al. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The Journal of Physical Chemistry, 102*, 3586–3616.

Maftouh, M., Avan, A., Sciarrillo, R. C., Granchi, L. G., Leon, R. R., Funel, N., et al. (2014). Synergistic interaction of novel lactate dehydrogenase inhibitors with gemcitabine against pancreatic cancer cells in hypoxia. *British Journal of Cancer, 110*, 172–182.

Magalhaes, L. G., Ferreira, L. L., & Andricopulo, A. D. (2018). Recent advances and perspectives in cancer drug design. *Anais da Academia Brasileira de Ciências, 90*, 1233–1250.

Manda, I. S., Moudgil, M. N., & Mandal, S. K. (2009). Rational drug design. *European Journal of Pharmacology, 625*, 90–100.

Mayo, S. L., Olafson, B. D., & Goddard III, W. A. (1990). Dreiding: A generic force field for molecular simulations. *The Journal of Physical Chemistry, 94*, 8897–8909.

McCammon, J. A., Gelin, B. R., & Karplus, M. (1977). Dynamics of folded proteins. *Nature, 267* (5612), 585–590.

Meng, X. Y., Zhang, H. X., Mezei, M., & Cui, M. (2011). Molecular docking: A powerful approach for structure-based drug discovery. *Current Computer-Aided Drug Design, 7*(2), 146–157.

Mitchell, J. B. O., Laskowski, R. A., Alex, A., & Thornton, J. M. (1999). Bleep-potential of mean force describing protein-ligand interactions: I. Generating potential. *Journal of Computational Chemistry, 20*(11), 1165–1176.

Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., et al. (1998). Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry, 19*(14), 1639–1662.

Muegge, I. (2006). PMF Scoring Revisited. *Journal of Medicinal Chemistry, 49*(20), 5895–5902.

Neves, B. J., Dantas, R. F., Senger, M. R., Melo-Filho, C. C., Valente, W. C. G., De Almeida, A. C. M., et al. (2016). Discovery of new anti-schistosomal hits by integration of QSAR-based virtual screening and high content screening. *Journal of Medicinal Chemistry, 59*(15), 7075–7088.

Njoroge, F. G., Chen, K. X., Shih, N. Y., & Piwinski, J. P. (2008). Challenges in modern drug discovery: A case study of boceprevir, an HCV protease inhibitor for the treatment of hepatitis C virus infection. *Accounts of Chemical Research, 41*, 50–59.

Ondetti, M. A., Rubin, B., & Cushman, D. W. (1977). Design of specific inhibitors of angiotensin-converting enzyme: New class of orally active antihypertensive agents. *Science, 196*, 441–444.

Oostenbrink, C., Villa, A., Mark, A. E., & Van Gunsteren, W. F. (2004). A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *Journal of Computational Chemistry, 25*, 1656–1676.

Patick, A. K. (2006). Rhinovirus chemotherapy. *Antiviral Research, 71*, 391–396.

Perez, M. A., Fernandes, P. A., & Ramos, M. J. (2007). Drug design: New inhibitors for HIV-1 protease based on Nelfinavir as lead. *Journal of Molecular Graphics and Modelling, 26*, 634–642.

Plewczynski, D., Philips, A., Grotthuss, M. V., Rychlewski, L., & Ginalski, K. (2014). HarmonyDOCK: The structural analysis of poses in protein-ligand docking. *Journal of Computational Biology, 21*, 247–256.

Rani, R. (2019). Small molecules inhibitors of the Plasmodium Falciparum LDH enzyme and their therapeutic applications. In *Lactate Dehydrogenase (LDH): Biochemistry, function and clinical significance* (Vol. 1, pp. 141–165). Hauppauge, NY: Nova Publishers.

Rani, R., & Kumar, V. (2016). Recent Update on Human Lactate Dehydrogenase Enzyme 5 (*h*LDH5) Inhibitors: A Promising Approach for Cancer Chemotherapy. *Journal of Medicinal Chemistry, 59*, 487–496.

Rani, R., & Kumar, V. (2017). When will small molecule LDH inhibitors realize their potential in the cancer clinic? *Future Medicinal Chemistry, 9*(11), 1113–1115.

Rappe, A. K., Casewit, C. J., Colwell, K. S., Goddard III, W. A., & Skiff, W. M. (1992). UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society, 114*, 10024–10035.

Reddy, M. R., & Parrill, A. L. (1999). *Chapter 1: Overview of rational drug design* (pp. 1–11). Washington, DC: ACS.

Roberts, N. A. (1990). Rational design of peptide-based HIV proteinase inhibitors. *Science, 248*, 358–361.

Ryan, D. M., Ticehurst, J., Demsey, M. H., & Penn, C. R. (1994). Inhibition of influenza virus replication in mice by GG167 (4-guanidino-2,4dideoxy-2,3-dehydro-N-acetylneuraminic acid) is consistent with extracellular activity of viral neuraminidase (sialidase). *Antimicrobial Agents and Chemotherapy, 38*, 2270–2275.

Saleh-e-In, M. M., Roy, A., Al-Mansur, M. A., Hasan, C. M., Rahim, M. M., Sultana, N., et al. (2019). Isolation and in silico prediction of potential drug-like compounds from Anethum sowa L. root extracts targeted towards cancer therapy. *Computational Biology and Chemistry, 78*, 242–259.

Shi, Y., Xia, Z., Zhang, J., Best, R., Wu, C., Ponder, J. W., et al. (2013). The polarizable atomic multipole-based AMOEBA force field for proteins. *Journal of Chemical Theory and Computation, 9*(9), 4046–4063.

Sturluson, A., Huynh, M. T., Kaija, A. R., Laird, C., Yoon, S., Hou, F., et al. (2019). The role of molecular modelling and simulation in the discovery and deployment of metal-organic frameworks for gas storage and separation. *Molecular Simulation, 45*, 1082–1121.

Summa, V., Petrocchi, A., Bonelli, F., Crescenzi, B., Donghi, M., Ferrara, M., et al. (2008). Discovery of raltegravir, a potent, selective orally bioavailable HIV-integrase inhibitor for the treatment of HIV-AIDS infection. *Journal of Medicinal Chemistry, 51*, 5843–5855.

Sun, H. (1998). COMPASS: An ab initio force field optimized for condensed-phase application – Overview with details on alkane and benzene compounds. *Journal of Physical Chemistry, 102*, 7338–7364.

Thai, K. M., Bui, Q. H., Tran, T. D., & Huynh, T. N. (2012). QSAR modeling on benzo [c] phenanthridine analogues as topoisomerase I inhibitors and anti-cancer agents. *Molecules, 17*(5), 5690–5712.

Tikhonova, I. G., Sum, C. S., Neumann, S., Engel, S., Raaka, B. M., Costanzi, S., et al. (2008). Discovery of novel agonists and antagonists of the free fatty acid receptor 1 (FFAR1) using virtual screening. *Journal of Medicinal Chemistry, 51*, 625–633.

Trosset, J. Y., & Cavé, C. (2019). In silico target druggability assessment: From structural to systemic approaches. In *Target identification and validation in drug discovery* (pp. 63–88). New York: Humana Press.

Urwyler, S. (2011). Allosteric modulation of family C G-protein-coupled receptors: From molecular insights to therapeutic perspectives. *Pharmacological Reviews, 63*, 59–126.

Van Dijk, A. D., Kaptein, R., Boelens, R., & Bonvin, A. M. (2006). Combining NMR relaxation with chemical shift perturbation data to drive protein–protein docking. *Journal of Biomolecular NMR, 34*, 237–244.

Velec, H. F. G., Gohlke, H., & Klebe, G. (2005). DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *Journal of Medicinal Chemistry, 48*(20), 6296–62303.

Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., & Taylor, R. D. (2003). Improved protein-ligand docking using GOLD. *Proteins, 52*(4), 609–623.

Webber, S. E., Bleckman, T. M., Attard, J., Deal, J. G., Kathardekar, V., Welsh, K. M., et al. (1993). Design of thymidylate synthase inhibitors using protein crystal structures: The synthesis and biological evaluation of a novel class of 5-substituted quinazolinones. *Journal of Medicinal Chemistry, 36*, 733–746.

Weigelt, J. (2010). Structural genomics-Impact on biomedicine and drug discovery. *Experimental Cell Research, 316*, 1332–1338.

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., et al. (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society, 106*, 765–784.

Weiner, S. J., Kollman, P. A., Nguyen, D. T., & Case, D. A. (1986). An all atom force field for simulations of proteins and nucleic acids. *Journal of Computational Chemistry, 7*, 230–252.

Wlodawer, A., & Vondrasek, J. (1998). Inhibitors of HIV-1 protease: A major success of structure-assisted drug design. *Annual Review of Biophysics and Biomolecular Structure, 27*, 249–284.

Young, D. C. (2009). *Computational drug design: A guide for computational and medicinal chemists*. Hoboken, NJ: Wiley.

Young, S. S., Sheffield, C. F., & Farmen, M. J. (1997). Optimum utilization of a compound collection or chemical library for drug discovery. *Chem Inf Comput Sci, 37*, 892–899.

Yu, W., & MacKerell, A. D. (2017). In P. Sass (Ed.), *Computer-aided drug design methods, antibiotics* (Vol. 1520, pp. 85–106). New York: Springer.

Zhang, L., Fourches, D., Sedykh, A., Zhu, H., Golbraikh, A., Ekins, S., et al. (2013). Discovery of novel antimalarial compounds enabled by QSAR-based virtual screening. *Journal of Chemical Information and Modeling, 53*, 475–492.

Zou, X., & Kuntz, I. D. (1999). Inclusion of solvation in ligand binding free energy calculations using the generalized-born model. *Journal of the American Chemical Society, 121*, 8033–8043.

Links

(https://www.rcsb.org/).

(https://www.mordorintelligence.com/industry-reports/drug-discovery-market).

(https://www.globenewswire.com/news-release/2020/01/09/1968249/0/en/Global-Drug-Discovery-Market-2019-2027-Rising-Demand-for-Specialty-Medicines.html).

Chapter 11
Biological Implications of Polyethylene Glycol and PEGylation: Therapeutic Approaches Based on Biophysical Studies and Protein Structure-Based Drug Design Tools

Neha Raina, Amit Kumar Singh, and Asimul Islam

Abstract Polyethylene glycol (PEG) is one of the most extensively used biocompatible polymer. PEG-modification improves the original properties of conjugates and thus being exploited in different fields. PEGs demonstrated their ability to bind DNA, dyes and proteins, in solution and in solid phase via amine and thiol groups. Covalent linkage of PEG to drug molecules improves water-solubility, bioavailability, pharmacokinetics, immunogenic properties, and biological activities. Entrapment of drugs into the PEG vesicle offers substantial benefits in the treatment of many diseases including type 2 diabetes over conventional injection-based therapies. Therapeutic enzymes are conjugated with PEG for targeted therapy of diseases in which the native enzyme was inefficient. PEG has been most extensively investigated polymers for gene delivery due to its capability to form stable complexes by electrostatic interactions with nucleic acids. Many PEG-enzymes conjugates have already obtained FDA approval for clinical implications. PEGylated copolymers have least cytotoxicity and cell-compatibility concern, high efficiency, safety and biocompatibility and thus considered as an attractive polymer for gene and drug delivery system. For instance, many tissue engineering applications, PEG and its derivatives are likely to precise control of cell behaviour in growing tissues. For this application numerous bioresponsive and intelligent biomaterials are developed and extensively used in bone and tissue regeneration. PEG-derived hydrogels increase gene expression of bone-specific markers, secretion of bone-related matrix, and mineralization and may have a potential impact on bone-engineering therapies.

N. Raina · A. K. Singh (✉)
Department of Biotechnology, Sharda University, Greater Noida, Uttar Pradesh, India
e-mail: amitk.singh@sharda.ac.in

A. Islam (✉)
Centre for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, New Delhi, India
e-mail: aislam@jmi.ac.in

S. K. Singh (ed.), *Innovations and Implementations of Computer Aided Drug Discovery Strategies in Rational Drug Design*,
https://doi.org/10.1007/978-981-15-8936-2_11

PEG-coated poly(amidoamine) exhibits low toxicity to human corneal epithelial cells and effectively used as antimicrobial agents.

Keywords Polyethylene glycol · Drug delivery system · Molecular imaging · TISSUE engineering · PEGylation · Bioconjugation · Gene therapy · Nanocarriers · Biomaterials

11.1 Introduction

Proteins are linear chain of amino acid residues that fold into three-dimensional structures to carry out a wide variety of function inside the cell. About 5–30% functional regions of proteins are disordered lacking well defined ensemble but majority of protein domain fold into ordered 3-dimensional conformations due to physical interaction within the chain (Dawson et al. 2017; Jones and Cozzetto 2015; Mitchell et al. 2018). The structural features of proteins, in turn, determine the broad range of functions from binding specificity, forming structure inside the cell, to catalysis of biochemical reactions, signal transduction or transport. Due to recent advance in high-throughput sequencing technology, gap is quickly growing between number of known protein sequence and number of those with experimentally characterized function. There are more than 60 million protein sequence deposited in the UniProt database (UniProt 2015), but less than 0.8% of these sequences have the function manually annotated in SwissProt (Boutet et al. 2016). Automated *in silico* protein function prediction thus become crucial for making use of the recent explosion of genomic sequencing data. In this chapter, we further explore the use of structural modification of protein and its application in medicine, gene delivery, pharmaceutical industries, cosmetics, food industries and bone and tissue engineering**.** Structural modification of protein mainly included here is PEGylation of protein surface.

Polyethylene glycol (PEG) is a semicrystalline polymer of ethylene glycol, have routinely used in biological research as crowding agent, in drug delivery. PEG is a best polymer of choice in drug delivery systems because of its tunable properties and well-established safety profile. Recently polymers of different sizes of PEGs are used on the basis of green chemistry principle for extraction of organic chemicals in food and pharmaceutical industry because of its inert, hydrophilic and hydrophobic properties (Guin and Gruebele 2019; Zinov'eva et al. 2020). In addition, it possesses the prime requisites for the selection of any ingredient in formulation development of drug delivery carrier. PEGylation technique is commonly used to improve the properties of biomolecules including, proteins, peptides, enzymes, antibody fragments, oligonucleotides, small synthetic drugs, etc. (D'souza and Shegokar 2016). Upon covalent attachment of PEG to any biomolecule, it increases the half-life, solubility, stability and reduced immunogenicity of that molecule. The hydrated PEG chain protects the conjugated biomolecules from proteases and thus reduces nonspecific degradation and improved solubility and stability. All these advantages of PEGylation are exploited in the pharmaceutical industry routinely.

PEGs are amphiphilic polymers composed of repeating ethylene glycol subunits, and its number is represented by the whole integer *n*. Each ethylene glycol residue has a molecular mass of 44 Da, and n × 44 Da represents the number average molecular weight of the PEG chain. Chemically, PEGs are inert, nontoxic substance with hydroxyl groups at both termini that can be chemically activated for diverse function. In common practice, linear PEG chains, branched PEGs and polymer of more than one PEG monomers are joined either linearly or branched crosslink which are differ in their relative chemical reactivity and specificity (Suzuki et al. 1984).

PEG conjugation to biomolecules is currently a common practice to achieve persistent clinical responses and improved biological features due to thermal and mechanical stability, lower antigenicity and immunogenicity, reduced enzymatic degradation, improved solubility, optimal pharmacokinetics and pharmacodynamic properties, enhanced circulating half-life in body, decreased clearance and enhanced potency (Bailon and Berthold 1998; Zhang et al. 2014). The improved biological activity is attributed to the changes in electrostatic-binding properties, conformational changes, steric interference, hydrophobicity, local charge distribution and pI value of proteins. In addition, PEGylation significantly alters the binding affinities to the receptors attributed to the physicochemical changes, resulting in reduced activities in cell-based assays, in which incubation times are usually short (Inada et al. 1986). Degree of PEG-protein conjugate in the form of unmodified, mono-, di-, tripegylated have great advantages. Generally, with increase in the degree of pegylation, rate of absorption decreases which prolongs the availability of drug in circulation and receptor saturation.

Wide range of hydrogels are prepared from PEG which are being used in varieties of biomedical applications because of diverse nature of PEG and the versatility of mechanical and biological modification according to physiological requirements (Alexander et al. 2014; Mendez et al. 2018). The hydrogel formed by PEGs can release the drug over a long duration of time, meanwhile biocompatible and provide environment similar to soft tissue and allow diffusion of nutrients due to biodegradable properties with desired safety and efficacy. PEG prevents the molecules of the hydrogels from being dissolved in a swelling medium by holding the entire molecule together due to extensive cross linking. Some of the investigated PEG-based copolymers are currently used in biopharmaceutical industry and clinical research (Harris and Chess 2003). Many PEGylated drugs have been approved by FDA to address hepatitis, rheumatoid arthritis, neutropenia, various cancers and wound healing therapy (Katre 1993).

Because of enormous biological and clinical implications of PEG, this chapter aimed to provide a comprehensive discussion. We provide a brief discussion on protein structure prediction and different biological applications of PEG along with underlying mechanism of improved biochemical features.

11.2 Importance of Structure and Functional Dynamics in PEGylation

Genome sequencing efforts utilizing high-throughput technologies are yielding millions of protein encoding sequence that currently lack any functional characterization (Molloy et al. 2014). The function of a protein of interest can be inferred from other homologous protein with common ancestor which are functionally characterised as well. For this purpose, sequence and/or structure information can be used. Sequence comparisons methods were used for genome-wide sequence annotation that are based on sequence alignment to identify homologous proteins. BLAST, PROSITE and PFAM all are well-known sequence alignment tool for genome-wide functional annotation tool (Dhifli and Diallo 2016; Molloy et al. 2014). All these tools mentioned above are typically fast but restricted to identifying pair of proteins with significant sequence similarity (close homolog). Function of an uncharacterized query protein was determined when sequence alignment tool identify a homolog with known function and sequence similarity more than 30% with query. On the other hand, two proteins with identical function cannot be inferred based on sequence information alone. Sequence based functional annotation may miss detecting remote homolog which is either branching point or convergent evolution has resulted in high sequence diverge while preserving structure and function. First remote homolog identified was myoglobin and hemoglobin which have similar structure but different sequence. Evolutionary pressure is more on structure than sequence for preservation; methods that compare structure allow effectively casting wider lattice at detecting related proteins for functional annotation. Another word, three dimensional structure of a protein is highly conserved compared to the primary sequence. Thus, it is better to compare overall structure and shape of a protein and considered to be more eloquent way of assigning function of query protein. Structure-based function deduction promises to detect remote homolog and expand options for assigning function to novel protein sequences. So in order to implement PEGylation of proteins, one must go through these tools to predict the best results for the desired protein. These tools reduce the number of experiments and better strategy for PEGylation of proteins.

11.2.1 Protein Structure-Based Drug Design

Many therapeutic compounds currently available in the market were either discovered from the screening of natural or synthetic compound libraries or through serendipitously. These indiscriminate approaches entails testing large number of compound and developing countless high-throughput screening assay (Bonetta et al. 2016). Now days, a rational approach based on structure based route where the structure of target protein is determined and hypothetical ligand most of the case predicted by molecular modelling and movement of compounds were predicted by molecular dynamics simulation before the synthesis of screened molecules. These days protein structure is generally determined using three different methods (a) X-ray crystallography, (b) Nuclear Magnetic Resonance spectroscopy (NMR) and (c) cryoelectron microscopy (Cryo-EM). X-ray crystallography is oldest and commonly used technique, relies on the ability of protein to crystallise in a completely biologically unnatural condition in regular molecular array. NMR has advantage as it can perform in solution and no protein crystal required. Major problem associated with NMR is protein size, where it is unable to determine the structure of large proteins. Cryo-EM is just become available and soon be capable of providing structural information of proteins in solution and as good as X-ray crystallography. Present decade has witnessed rapid development of computer aided drug design with enough accuracy which allows frequent use in discovery of new therapeutics. Drug design based on protein structure prediction based on binding mode of small molecules and their relative affinity (Sledz and Caflisch 2018) is delivering better results.

11.2.2 Computer-Aided Drug Design

Computer aided drug design (CADD) methods have played fundamental role in drug discovery effort from many years. Nowadays, CADD become essential part of the discovery pipeline for pharmaceutical companies (Sledz and Caflisch 2018). Discovery of quite a lot of approved drugs including captopril, saquinavir, indinavir, ritonavir, and tirofiban, has benefited to a large extent from the application of CADD (Hillisch et al. 2015; Muegge et al. 2017). X-Ray structure generally provide proved quite static picture of the protein but when it is combined with computation techniques to obtain a clear idea of how protein function, the combination become awesome for drug desighning. CADD techniques are used principally for virtual screening, hit/lead optimization and design of novel compounds. In virtual screening a huge database of compounds is examined and subset of compound is picked out on the binding capacity for further in vitro testing. Further CADD is used for optimization of hit/lead compound driven by the rationalization of a structure–activity relationship. After the individuation of key elements for binding, the design of new compound can be attempted (Salmaso 2018).

11.2.3 Virtual Screening

Virtual screening is a computational method which is used to find potential ligand of interest by screening an *in silico* library of diverse chemical compounds. It acts as filter which allows to pass out unwanted ligands and retains which are of interest based on filter criteria like stereochemical similarity or stereochemical complementarity. Virtual screening identifies the potential ligands either based on similarity of

Table 11.1 Comparision of receptor based and ligand based virtual screening

Receptor based	Ligand based
Based on 3D structural information of receptor's binding site	Based on chemical or structural (3D) information of set of ligands
Molecular docking	Similarity searching & pharmacophore mapping
Uses principle of binding kinetics or interactions between receptor and ligand	Uses principle of chemical or structural similarity of ligands
It can find novel class of hit/lead	It cannot find novel class of hit/lead
Higher chances of false positives	Lesser chances of false positives

ligand (ligand based approach) or binding affinity with receptor (receptor based approach). A comparison between ligand based and receptor methods of screening are shown in Table 11.1.

This chapter mainly focuses on receptor based or structure based virtual screening in which a library of small molecules (ligands) is docked to the binding site of protein in lock and key manner. In this approach three things are required; (1) Protein structure of protein either determined experimentally in crystal form and solution form or computationally modelled using homology modelling. One or more known protein structures with close sequence identity are used as template in homology modelling to model protein of interest. (2) A library of small molecules is required for the docking algorithm to calculate binding affinity of ligand with binding site on protein of interest. (3) Finally a docking protocol is required which defines the docking parameters used in docking experiment (Fig. 11.1).

11.2.4 Molecular Dynamic Simulation

Molecular Dynamic Simulation provides insight into protein dynamics ahead of that available crystallographically and unravels novel cryptic binding sites, expanding the druggability of the target. MD-simulations are applied in the investigation of numerous dynamic properties and process which is directly applied in structural biochemistry, enzymology, molecular biology, biophysics, biotechnology and pharmaceutical chemistry. It allows scientist to study the thermodynamic and kinetic properties of proteins and other biomolecules.

11.3 PEG in Drug Delivery

PEGs are amphiphilic in nature thus easily dissolved in organic and aqueous solvents. Because of its non-toxic nature and their property of being easily eliminated by renal and hepatic pathways, PEGs are ideal choice for drug delivery system

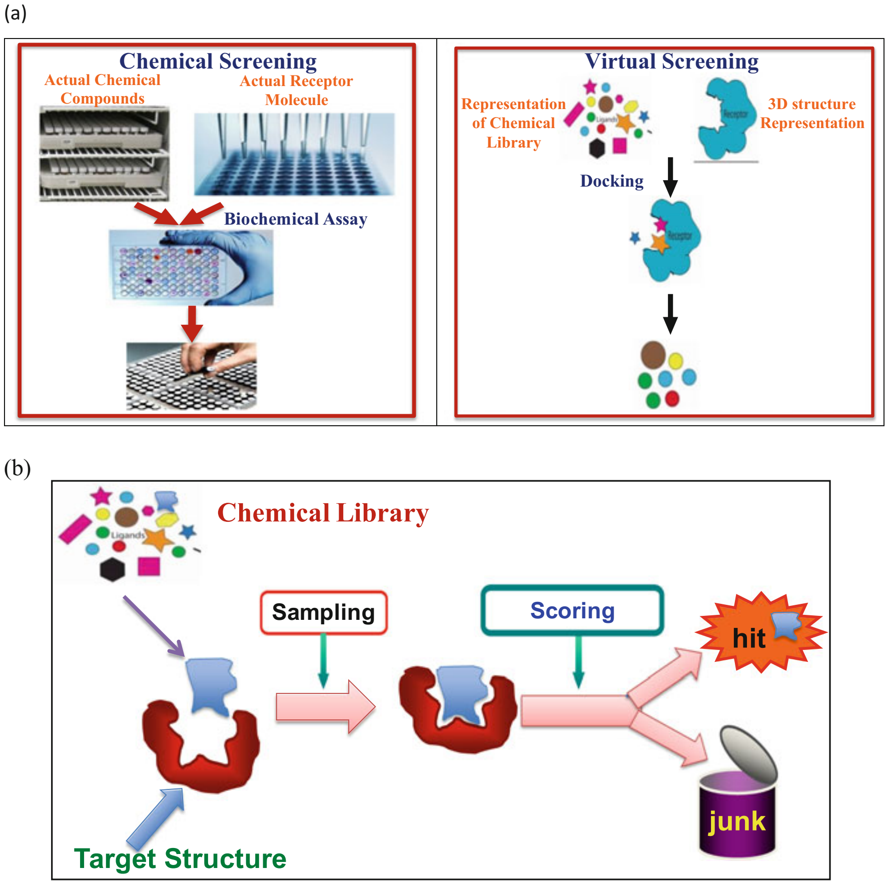


Fig. 11.1 (**a**) Diagrammatic representation and comparison between chemical screening versus virtual screening; (**b**) Diagrammatic representation of receptor based virtual screening

(Rashmi Zabihi et al. 2020). These unusual properties of PEG along with their highly hydrated polyether backbone, capable as an acceptor hydrogen bonding, a large exclusion volume make it capable to entrap versatile drugs for the targeted delivery. Various efforts are being made to develop novel strategies to conjugate PEG with biomolecules to increase its clinical applications (Yang et al. 2020b). Currently, new strategies have been undertaken to develop PEG based drug delivery system which allow the delivery of different active classes of drugs (Han et al. 2019). Clinical implications of several PEG-conjugated drugs has been routinely reported. Entrapment of IL-2, growth hormone antagonist, haemoglobin, growth factor, anticancer drugs, cytokines, enzymes, hormones, lactoferrin, antibodies and antibody fragments etc., in the PEG-based polymers are frequently used in medical industry. In addition, many proteins are conjugated to the PEGylated polymers to extended circulating plasma life and easily clearance through renal filtration because of increased aqueous solubility, decreased immunogenicity, and permits binding of the proteins to their receptors (Kolate et al. 2014a). Furthermore, due to the increased overall volume and molecular mass of PEGylated polymer, it has a greater bioavailability, and thus results in less frequent dosing (Nucci et al. 1991). Generally, PEGylation is used for the modification of therapeutic molecules by conjugation with PEG. A natural clinical implication of PEGylated proteins is in the form of antibodies (both single chain and monoclonal) modifications to achieve greater solubility and longer circulating life *in vivo*. PEGylation with longer PEG chain was more efficient than multiple PEGylation with short PEG chain to increase serum half-life (Lee et al. 1999). A new and potentially useful application of PEG mAb to an ADEPT (antibody directed enzyme prodrug therapy) system which is generally used for targeted drug delivery for the impairment of genetic diseases (AlQahtani et al. 2019a). Recent advancement in the recombinant DNA technology has improved the production of proteins at large scale. But, their clinical applications are limited because of their antigenicity. Protein PEGylation significantly reduces the antigenicity of recombinant protein and thus associated antibody formation against them in limited, contributing to their prolonged activity with increased *in vivo* life to several hundred folds. During therapeutic uses, PEG protects protein from proteolytic digestion in the body (Veronese and Mero 2008).

Many small organic molecules and anticancer agents have been conjugated to PEG which can easily be delivered to the target without any side effect (Elechalawar et al. 2020). A successful PEGylation of drug molecules can improve the pharmacokinetic and pharmacodynamic outcomes. In addition, PEGylation extend blood residence, decrease enzymatic degradation and reduce immunogenicity of drugs as well as decrease uptake by reticuloendothelial system. Entrapment of anticancer drugs into the PEG helps to passively release at target tumor site with less frequent neutropenia, alopecia and cardiotoxicity (Plosker 2008). A PEG–daunorubicin conjugates have been recently synthesized and there *in vitro* antitumor activity has been evaluated (Greenwald 2001). PEGylation improves the physicochemical properties of drug molecules, including optimal degree of modification of molecular weight, size, hydrophilicity, conformation, steric hindrance and charge which lead to altered elimination kinetics and thus, several PEGylated therapeutics have entered into the clinical trials and their successful translation to the clinical use. However, an accurate assessment of pharmacokinetic and pharmacodynamic parameters of drug like molecules in animals and patients are highly needed. A sensitive *in vivo* quantification and fate of the bound drug in the host body is required after systemic administration stability, metabolism and bioavailability of PEGylated compounds (Kolate et al. 2014b). For target-specific therapy and imaging, nanocarrier based delivery systems have recently emerged as a promising delivery system for therapeutics with great potential (Rajan et al. 2014). It was shown that incorporation of small amounts of gangliosides, glucolipids, phosphatidylinositol impart a weak negative charge on surface of liposomes which can bypass the reticuloendothelial uptake. It was found from research charge influence that the uptake for hydrophobic and neutral to positively charged particles are more prone to reticuloendothelial uptake in comparison to the negatively charged particles.

11.4 Gene Therapy

Recently, gene therapy has drawn significant attention of scientific community for the treatment of diseases associated with the non-functional and mutated genes such as haemophilia, mucopolysaccharidosis, autoimmune diseases and cancer. The outcome of gene therapy is depending on the efficiency and safety of its delivery vectors. In comparison to the gene delivery using viral particle, PEGylation and hydrogel as gene delivery vectors are less oncogenic, immunogenic, easy to prepare and specially does not have any limitation of the size of gene to be transferred. In this context, PEG has gained remarkable attention because of its desired stability in the systemic circulation (Hashemi et al. 2019). Despite a great advantage of PEG for controlling the pharmacokinetics of the carriers, but is undesirable for the cellular association of gene carriers with tumors. To address this problem, various modified gene delivery systems have been developed. Hatakeyama et al. (2007) developed an interesting gene delivery system for cancer gene therapy, using a multifunctional envelope-type nano device modified with enzymatically cleavable PEG-lipid. They prepared a cleavable PEG-lipid which is composed of a PEG/matrix metalloproteinase (MMP)-substrate to be specifically cleaved by MMP in the extra-cellular space in tumor tissues (Hatakeyama et al. 2007). As we know that MMP expression is higher in the case of tumor cells, thus such system facilitate the site directed gene delivery. Kim et al. (2006) conjugated the siRNA of vascular endothelial growth factor to PEG via a disulfide linkage. A conjugate of siRNA-PEG interacting with cationic polyethylenimine form polyelectrolyte complex micelles and consequently showed a greater stability against enzymatic degradation. Under condition similar to reductive cytosolic environment, siRNA in its intact form was released from the siRNA–PEG conjugate after cleavage of the disulfide linkage which could be potentially applied to RNAi-based anti-angiogenic treatment of cancer *in vivo*. Hashemi et al., fabricated PEG coated, calcium doped magnetic nanograin (PEG/Ca(II)/Fe3O4) as a genome expression enhancer as a promising strategy for gene therapy (Hashemi et al. 2019). The potential of large scale production and biocompatibility make PEG-based carriers increasingly attractive for gene therapy. However, many intra- and extracellular obstacles needs to get addressed.

PEG-modified adenosine deaminase has been extensively used for the enzyme replacement therapy for immunodeficiency due to adenosine deaminase deficiency. Enzyme replacement therapy by modified PEG has restored the metabolic environment and thus immune function lost due to the adenosine deaminase deficiency (Hershfield 1995). In many cases, level of functions achieved through enzyme replacement therapy has been sufficient to protect against opportunistic and life-threatening infections and diseases (Liu et al. 2009). For gene therapy against HIV in peripheral blood mononuclear cells, a cationic carbosilane–PEG dendrimers were used (Sánchez-Nieves et al. 2014). An improved performance was observed as compared to a PEG-free carbosilane dendrimer. Toxicity significantly decreased in presence of PEG moety because dendrimers and siRNA interact weakly as compared to the homodendrimer analogous. Chemical features like well-defined size and structure, flexibility, monodispersity and multivalent molecular surface, lower toxicity and lower dendriplex strength, are key properties for use of these vectors for the gene delivery to target cells (Newkome et al. 2008).

11.5 Protein Folding Studies

PEG has been extensively used in research to mimic the cellular environment to investigate the role of crowding agent in reducing misfolding and aggregation, conformational change in protein to increase stability and activity. The extent of PEG-based stabilization of proteins depends on the size of PEG, PEGylation site, structure of the PEG–protein linker, etc. The strength of a noncovalent interaction, salt-bridge and hydrogen-bond strength within a protein depends strongly on its microenvironment and increasing in nonpolar versus aqueous environments. As we know that protein folds in a crowded cellular environment but most of the studied performed on protein folding were done on a single protein (Kinjo and Takada 2002; Tsao et al. 2010; van den Berg et al. 1999). The efficiency of macromolecular crowding agent is expected to be highest hydrodynamic dimensions of a tested protein. To mimic the natural cellular milieu, many crowding agent including PEG have been extensively used as crowding agents for various proteins (Ghosh et al. 2020; Nasreen et al. 2018; Shahid et al. 2019; Shahid et al. 2017). An expected outcome of the presence of PEG as a macromolecular crowding agent is their ability to improve the conformational stability of a globular protein due to the excluded volume effects including alteration of solvent properties (Christiansen et al. 2013; Stepanenko et al. 2016; Tokuriki et al. 2004). Behaviour of protein dynamics is affected by the hydrodynamic size and concentration of inert crowder where chemical nature of crowder molecules should not play any role in the modulation of conformational properties (Fonin et al. 2019). Conformational changes of d-glucose/d-galactose-binding protein (GGBP) were studied at different PEGs (Molecular weight, 12,000, 4000, and 600) in varying concentration and a noticeable structural changes was observed (Fonin et al. 2018). All PEGs promoted compaction of GGBP and lead to the increase in ordering of its structure. These data support the previous notion that the influence of macromolecular crowders on proteins is rather complex phenomenon that extends beyond the excluded volume effects (Fonin et al. 2017).

Stepanenko et al. (2016) has demonstrated the effects of different PEGs of varying molecular masses (PEG-600, PEG-8000, and PEG-12000) on the spectral properties and unfolding-refolding processes of the super-folder green fluorescent protein. The hydrodynamic radii of PEG-600, PEG-8000, and PEG-12000 are 5.6 Å ± 0.3 Å, 24.5 Å ± 1.9 Å, and 30.9 Å ± 2.5 Å, respectively (Kuznetsova et al. 2014). According to the excluded volume theory, the greatest effect on the unfolding-refolding processes should be provided by PEG-8000, whose hydrodynamic dimensions are closest to those of sfGFP. Ferreira et al., demonstrated that PEG and UCON of similar size produces different changes in the solvent properties of water in their solutions and induced morphologically different α-synuclein aggregates (Ferreira et al. 2015; Ferreira et al. 2016). The further extended study to get deeper insights into the behaviour of proteins in a crowded environment, a similar study was conducted to determine the secondary and tertiary structure and aromatic residue solvent accessibility ten different proteins. Both PEG and UCON polymers affects secondary and tertiary structures of folded and hybrid proteins in a limited fashion with a slight unfolding. Solvent accessibility of aromatic residues was significantly increased for the majority of the proteins in the presence of UCON but not PEG. PEG also accelerated the aggregation of protein into amyloid fibrils (Breydo et al. 2015). In addition, a large number of studies on PEG has been carried out to establish the role of PEG in protein folding and dynamics (Parray et al. 2019; Shahid et al. 2017; Xiao et al. 2019).

11.6 Cosmetic Industry

PEG is one of the best investigated polymer for use in bio-related application due to its inertness, biocompatibility, biodegradability and versatility. PEG is generally used in cosmetics, food processing, pharmaceuticals, agriculture, and industrial manufacturing (AlQahtani et al. 2019b). Because of two terminal alcohol groups, the PEGs can form a series of derivatives including, mono-, di- and polyesters, ethers, amines and acetals. PEGs can favour the addition reaction to form new compounds or complexes on their ether bridges. Some common derivatives of PEG in cosmetic industry are, fatty acid esters, PEG ethers, PEG amine ethers, PEG castor oils, PEG propylene glycols, and other derivates with required properties. PEG is found in many domestic and hygiene products, because of their solubility, viscosity and low toxicity. PEGs and their fatty acid esters produce a little irritation and have extremely low acute and chronic toxicities and thus considered as a highly acceptable polymer in the cosmetic industry (Fruijtier-Pölloth 2005). Due to the abundant presence of PEG, it likely to be present at or introduced to the inflammation site. PEG fatty esters, PEGs, and PEG sorbitan fatty esters are slightly irritating to the skin of rabbits and humans and thus commonly found in the antiseptic creams. The occurrence of PEG in close vicinity to highly active immune cells may be enough to elicit the stimulation of anti-PEG antibodies (Yang and Lai 2015). PEGylated pharmaceutical products are used as an indirect molecular probe for measuring mucosal permeability. Their unique osmotic characteristics in aqueous solutions, PEG provides solubility, stability and strength to the degrading substances (Fordtran and Hofmann 2017). In addition, carcinogenicity studies with PEGs have been conducted. Interestingly, in most of studies people found no indication of a tumorigenic effect of PEGs which will further allow the effective use of PEG in cosmetics. The PEGs and their fatty acid ethers and esters produce a negligible dermal irritation along with an extremely low acute and chronic toxicities as they do not readily penetrate intact skin and thus the sensitising potential of these compounds appears to be negligible (Jang et al. 2015). Furthermore, commonly used

PEG derivatives in the cosmetic industry are free from reproductive and developmental toxicity, on genotoxicity and carcinogenic properties.

11.7 Food Industry

PEGs are widely used in medical and food industries because of their biologically inert, non-immunogenic, non-toxic and water-soluble nature. A mild to life-threatening immediate hypersensitivity for PEG are reported (Ahmed et al. 2018). PEG hypersensitivity should be considered during the diagnostic management and therapy. The Food and Drug Administration (FDA) has approved several biological application of PEGs including as a carrier, preservative of food, cosmetics and pharmaceuticals, including injectable and bowel solutions (Lim and Hong 2014). PEGs are extensively used as enhancer of solubility and oral bioavailability of compounds with poor aqueous solubility (Gullapalli and Mazzitelli 2015). The nanoparticles and essential oil loaded composite films of PEG are widely used in the food industry for packaging purposes. These films are capable of exhibiting antimicrobial effects against bacteria, and extend the shelf-life of meat. Such biocompatible films showed are manufactured commercially with increased thermal stability (Ahmed et al. 2018). The developed bionanocomposites are highly compatible to food packaging to control the spoilage and the pathogenic bacteria associated with the fresh chicken meat (Ahmed et al. 2018). Liu et at. developed EGylated chitosan modified silver nanoparticles which possess excellent antimicrobial activity against both Gram-negative and Gram-positive bacteria (Liu et al. 2017). Such antimicrobial coatings with excellent nonfouling property is available to resist bacterial attachment to the food materials for long term storage and transport. However, silver based nanoparticle coating with PEG has dramatically enhanced antibacterial property and ascribed to the synergism of PEG-based resistance of bacteria due the antibacterial property of silver. Colloidal silver nanoparticles prepared by chemical reduction using PEG as a reducing agent has been used as food simulants in apple juice.

11.8 Bone and Tissue Engineering

In recent years different types of biomaterials are developed to support the bone tissue regeneration process. A relatively new class of nanocomposite biomaterials are produced from PEG that incorporate a biopolymeric and biodegradable matrix structure render improved properties including large surface area, high mechanical strength and stability, enhanced cell adhesion, proliferation, and cell differentiation (Bharadwaz and Jayasuriya 2020). Cartilage production with the help of PEG and alginate was successfully done. Collagen, PEG encapsulated with poly-lactic or -glycolic groups and polyvinyl alcohol, are other examples of biological and synthetic polymers that have been combined with chondrocytes (Bryant and Anseth 2003). Recent development in bone tissue engineering have identified and propelled the need of PEG based biomaterials as a promising approach for effective bone regeneration because of its high tunable properties, excellent biocompatibility, controlled biodegradability and good mechanical strength (Wang et al. 2019). In bone regeneration, the biomaterial scaffold is required to be fabricated having adequate cell growth and bone tissue regeneration. Such properties may be gained by PEGs because of their high porosity, surface-to-volume ratio, and crystallinity. PEG based nanofiber mats exhibited excellent cell attachment and growth, along with adequate cytocompatibility (Scaffaro et al. 2017). PEG-based membrane incorporated with the nanocalcium phosphate was developed to reduce non-functional scar tissue defects with the help of guided bone regeneration. The addition of calcium phosphate particles increase in fiber diameter as the addition of the inorganic calcium phosphate may have reduced the conductivity of the electrospinning solution. Furthermore, osteogenic differentiation of the cells was aided by the addition of the calcium phosphate nanoparticles as they increased the concentration of calcium ions thereby facilitating better storage of the cells (Türkkan et al. 2017). For surgeries and regenerative engineering tissue adhesives are manufactured with the help of PEG to achieve a rapid crosslinking, strong wet adhesion and cohesion strengths, and minimal cytotoxicity remains a critical roadblock for clinical translation (Lu et al. 2020). Recently, Zhou et al. (2020) designed a (PEG)-based poly (urethane-urea) for bone tissue repair using cystine dimethyl ester as a cross-linker. The strength of material was further strengthened by physical blending of nanohydroxyapatite. The obtained biocompatible material showed excellent mechanical strength, biocompatibility and osteogenic capability and thus has good prospects for bone tissue repair application.

11.9 Bioimaging and Radiotherapy

PEG is an FDA-approved compound, has been extensively used to modify various anticancer agents to increase long blood circulation, and enhancing their tumor accumulation. In recent years application of PEG-based nanoparticles have received greater attention in biomedicine, particularly in diagnostic imaging. Recently, Yang et al. designed a circular aptamer-PEG structure which is capable of prolonged blood circulation, and then responds to the acidic tumor microenvironment to deeply penetrate the solid tumor and selectively recognize cancer cells for in vivo targeted bioimaging (Yang et al. 2020a). A typical AS1411 aptamer containing G-rich oligonucleotide sequences with specific recognition of nucleolin was designed to detect a non-ribosomal protein normally existing in the nucleus and also highly expressed on the surface of cancer cells. Such materials having deep tumor tissue penetration and specific targeting of cancer cells, may be employed in efficient drug delivery strategy as well as bioimaging purposes. Photothermal therapy is an efficient approach employed for cancer treatment. A novel PEG based nanomaterials were synthesized which has significant T1-weighted performance to target tumor and localize for photothermal therapy. This approach holds significant potential for the clinical application especially in the case of tumor localization and targeted therapy (Meng et al. 2020).

Pretargeting strategies have recently emerged as an attractive imaging and therapy option for cancer patients. Different types of nanostars are prepared with has high potential to accumulate in the tumor tissue via an enhanced permeability and retention and thus implemented in the development of a passive delivery of cytotoxic compounds to cancer cells. To increase the contrast between uptake in the tumor and in surrounding healthy tissues, PEG based nanostars are designed for pretargeted positron emission tomography imaging, using radioligands that are labeled with the short-living positron emitter fluorine-18 (Goos et al. 2020a). The development of a pretargeting strategy based on the passive delivery of PEG based nanostars further expand the cancer imaging and treatment options, with a minimal radiation doses to healthy cells (Goos et al. 2020b). Recently, a new type of gold-PEG based nanoparticles are designed which possess excellent X-ray absorption coefficient, utilized as a contrast agent for computed tomography scan imaging. This nanoparticle shown a reliable aqueous dispensation, low cytotoxicity, and strong X-ray absorption ability subsequently enhances contrast, having long circulation time in the blood, and a negligible in vivo toxicity. Hence, such type of nanoparticles have great potential for clinical application in CT scan imaging (Gao et al. 2020).

Encapsulation of near infrared dyes, indocyanine green in the PEG based biocompatible nano carrier has been extensively used for NIR bioimaging, photothermal and photodynamic therapy. In is interesting to note that the encapsulated dyes remained stable for over long time and slowly accumulating in the liver and spleen having a wide range with deep penetration into the living tissues, may be exploited as a promising candidate for cancer treatment and diagnosis (Yeroslavsky et al. 2020). PEG incorporated silica nanoparticles are developed for improving efficacy of targeted radiotherapy in melanoma models. PEGylated carbon dots were engineered to bind alpha melanocyte stimulating hormone cyclic peptide analogs for targeting the melanocortin-1 receptor over-expressed on melanoma tumor cells. Such quantum dots are radiochemically stable, biologically active, and exhibited high affinity cellular binding properties and internalization (Zhang et al. 2020). An ultrasmall PEGylated quantum dots, covalently encapsulating the near infra-red emitting dye Cy5 were engineered to display MC1-R targeting cyclic DOTA-α MSH peptides on their surfaces. This superior property provide selective tumor uptake and favourable biodistribution properties and improved renal clearance. The unique and tunable surface features of this targeted radiotherapeutic quantum dots are coupled with favorable pharmacokinetic properties, substantially improved treatment efficacy and demonstrated a clear survival benefit in melanoma models.

11.10 Application of PEGs in Cutting Edge Technologies

PEGs are the most successful precipitants for the protein crystallization especially in the case of high molecular weight proteins (Gorrec 2016). Various parameters related to PEG such as types, concentration and molecular mass influence the crystallizing process. PEG has been used as a component in organ preservation solutions to reduce injury from cold perfusion in animal organs (Pasut et al. 2016; Valuckaite et al. 2013). PEGs 4000 and 3350 are commonly used in the gastroenterology from a long time. A series of PEGylated polymers are developed for tissue engineering. Implementing PEG-alt-thiol conjugates, biodegradable polymeric system showed in vivo biocompatible cell matrices for tissue engineering and wound healing (Wang et al. 2008). Further applications of PEG include in plastics and resins, in printing, as ingredients of soaps and detergents, in the rubber, in the textile and leather industry, in the paper industry, in the ceramics and glass industry, petroleum, mining and metal industries, for wood preservation and as chemical intermediates. Remarkably, PEG as a kind of stable, environmentally friendly and green surfactant. Many PEG-based aqueous solution are applied in extraction, separation and preconcentration of various constituents from complex mixture because it possesses the advantages of low cost, non-volatility, biodegradation, and non-toxicity (Zhang et al. 2017). Hydrogels composed of PEG and chitosan have been extensively used in the fields of wound dressing, tissue repair, water purification, drug delivery, and bone regeneration, and coatings in dental applications. Hydrogel developed for dental application shows a superior activity in early-stage adhesion inhibition and displays remarkably long-lasting colony-suppression activity. In addition, such nanomaterial antifouling and antimicrobial functions with excellent biocompatibility (Peng et al. 2020).

11.11 Miscellaneous Applications

PEGs are widely used in medical and food industries because of their biologically inert, non-immunogenic, non-toxic and water-soluble nature. A mild to life-threatening immediate hypersensitivity for PEG are reported. PEG hypersensitivity should be considered during the diagnostic management and therapy. The Food and Drug Administration (FDA) has approved several biological application of PEGs including as a carrier, preservative of food, cosmetics and pharmaceuticals, including injectable and bowel solutions (Lim and Hong 2014). PEG has been used as a component in organ preservation solutions to reduce injury from cold perfusion in animal organs (Pasut et al. 2016; Valuckaite et al. 2013). PEGs 4000 and 3350 are commonly used in the gastroenterology from a long time. PEGylated pharmaceutical products are used as an indirect molecular probe for measuring mucosal permeability. Their unique osmotic characteristics in aqueous solutions, PEG provides solubility, stability and strength to the degrading substances (Fordtran and Hofmann

2017). A series of PEGylated polymers are developed for tissue engineering. Implementing PEG-alt-thiol conjugates, biodegradable polymeric system showed *in vivo* biocompatible cell matrices for tissue engineering and wound healing (Wang et al. 2008). Cartilage production with the help of PEG and alginate was successfully done. Collagen, PEG encapsulated with poly-lactic or -glycolic groups and polyvinyl alcohol, are other examples of biological and synthetic polymers that have been combined with chondrocytes (Bryant and Anseth 2003). PEGs are extensively used as enhancer of solubility and oral bioavailability of compounds with poor aqueous solubility (Gullapalli and Mazzitelli 2015). PEGylation is one of the best approaches for targeting of anticancer therapeutics.

PEG is one of the best investigated polymer for use in bio-related application due to its inertness, biocompatibility, biodegradability and versatility. PEG is generally used in cosmetics, food processing, pharmaceuticals, agriculture, and industrial manufacturing (AlQahtani et al. 2019b). PEG is found in many domestic and hygiene products, thus our body got repeated exposure to PEG which may causes the development of anti-PEG antibodies. Due to the abundant presence of PEG, it likely to be present at or introduced to the inflammation site. The occurrence of PEG in close vicinity to highly active immune cells may be enough to elicit the stimulation of anti-PEG antibodies (Yang and Lai 2015). Recent development in bone tissue engineering have identified and propelled the need of PEG based biomaterials as a promising approach for effective bone regeneration because of its high tunable properties, excellent biocompatibility, controlled biodegradability and good mechanical strength (Wang et al. 2019). PEGs are the most successful precipitants for the protein crystallization especially in the case of high molecular weight proteins (Gorrec 2016). Various parameters related to PEG such as types, concentration and molecular mass influence the crystallizing process.

11.12 Conclusion

PEGs are most commonly used polymers which is inert in nature, non-immunogenic, non-antigenic which enhance the solubility of hydrophobic drugs and facilitate the potential for DNA transfection, siRNA delivery and tumor targeting. PEGylated drug delivery systems are very effective nano-carriers used to deliver anticancer drugs to the tumor site. The PEGylated polymers had lower cytotoxicity and hemolytic toxicity with increased solubility, less aggregation, lower opsonization by RES and higher tumor accumulation by the enhanced permeability and retention effect. PEGylation showed a great advantage in overcoming the unfavourable characteristics of biomaterials by keeping the unique properties. PEG has also been used as gene delivery vector for the targeted delivery of gene of interest. PEG-based copolymers play a crucial role as a biomedical material for biomedical applications, because of its biocompatibility, biodegradability, thermosensitivity and easy controlled characters. PEG–based hydrogel exhibits good gelling mechanical strength and minimizes the initial burst effect of the drug.

The hydrogel developed from PEG is able to release the drug over a long duration of time, meanwhile is also biocompatible and provide environment similar to soft tissue and allow diffusion of nutrients due to biodegradable properties with desired safety and efficacy. Despite the extensive use of PEGs, future biomedical implications are the matter of concern among the scientists but still to procure the FDA approvals, a strong investigation in the clinical studies is necessary.

Acknowledgements This work was supported by the grant from the Indian Council of Medical Research (ICMR)BIC/12(16)/2014. The authors are very indebted to Sharda University, FIST Program (SR/FST/LSI-541/2012) and Jamia Millia Islamia (a Central University) for facilitating the research.

Conflict of Interest The authors declare no conflict of interest.

References

Ahmed, J., Arfat, Y. A., Bher, A., Mulla, M., Jacob, H., & Auras, R. (2018). Active chicken meat packaging based on polylactide films and bimetallic Ag-Cu nanoparticles and essential oil. *Journal of Food Science, 83*, 1299–1310.

Alexander, A., Ajazuddin Khan, J., Saraf, S., & Saraf, S. (2014). Polyethylene glycol (PEG)–poly (N-isopropylacrylamide) (PNIPAAm) based thermosensitive injectable hydrogels for biomedical applications. *European Journal of Pharmaceutics and Biopharmaceutics, 88*, 575–585.

AlQahtani, A. D., Al-Mansoori, L., Bashraheel, S. S., Rashidi, F. B., Al-Yafei, A., Elsinga, P., et al. (2019a). Production of "biobetter" glucarpidase variants to improve drug detoxification and antibody directed enzyme prodrug therapy for cancer treatment. *European Journal of Pharmaceutical Sciences, 127*, 79–91.

AlQahtani, A. D., O'Connor, D., Domling, A., & Goda, S. K. (2019b). Strategies for the production of long-acting therapeutics and efficient drug delivery for cancer treatment. *Biomedicine & Pharmacotherapy, 113*, 108750.

Bailon, P., & Berthold, W. (1998). Polyethylene glycol-conjugated pharmaceutical proteins. *Pharmaceutical Science & Technology Today, 1*, 352–356.

Bharadwaz, A., & Jayasuriya, A. C. (2020). Recent trends in the application of widely used natural and synthetic polymer nanocomposites in bone tissue regeneration. *Materials Science & Engineering. C, Materials for Biological Applications, 110*, 110698.

Bonetta, R., Ebejer, J. P., Seychell, B., Vella, M., Hunter, T., & Hunter, G. J. (2016). Role of protein structure in drug discovery. *Journal of the Malta Chamber of Scientists, 4*, 126–130.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bansal, P., Bridge, A. J., et al. (2016). UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: How to use the entry view. *Methods in Molecular Biology, 1374*, 23–54.

Breydo, L., Sales, A. E., Frege, T., Howell, M. C., Zaslavsky, B. Y., & Uversky, V. N. (2015). Effects of polymer hydrophobicity on protein structure and aggregation kinetics in crowded milieu. *Biochemistry, 54*, 2957–2966.

Bryant, S. J., & Anseth, K. S. (2003). Controlling the spatial distribution of ECM components in degradable PEG hydrogels for tissue engineering cartilage. *Journal of Biomedical Materials Research. Part A, 64*, 70–79.

Christiansen, A., Wang, Q., Cheung, M. S., & Wittung-Stafshede, P. (2013). Effects of macromolecular crowding agents on protein folding in vitro and in silico. *Biophysical Reviews, 5*, 137–145.

D'souza, A. A., & Shegokar, R. (2016). Polyethylene glycol (PEG): A versatile polymer for pharmaceutical applications. *Expert Opinion on Drug Delivery, 13*, 1257–1275.

Dawson, N. L., Lewis, T. E., Das, S., Lees, J. G., Lee, D., Ashford, P., et al. (2017). CATH: An expanded resource to predict protein function through structure and sequence. *Nucleic Acids Research, 45*, D289–D295.

Dhifli, W., & Diallo, A. B. (2016). ProtNN: Fast and accurate protein 3D-structure classification in structural and topological space. *BioDataMining, 9*, 30.

Elechalawar, C. K., Hossen, M. N., Shankarappa, P., Peer, C. J., Figg, W. D., Robertson, J. D., et al. (2020). Targeting pancreatic cancer cells and stellate cells using designer Nanotherapeutics in vitro. *International Journal of Nanomedicine, 15*, 991–1003.

Ferreira, L. A., Cole, J. T., Reichardt, C., Holland, N. B., Uversky, V. N., & Zaslavsky, B. Y. (2015). Solvent properties of water in aqueous solutions of elastin-like polypeptide. *International Journal of Molecular Sciences, 16*, 13528–13547.

Ferreira, L. A., Madeira, P. P., Breydo, L., Reichardt, C., Uversky, V. N., & Zaslavsky, B. Y. (2016). Role of solvent properties of aqueous media in macromolecular crowding effects. *Journal of Biomolecular Structure & Dynamics, 34*, 92–103.

Fonin, A. V., Darling, A. L., Kuznetsova, I. M., Turoverov, K. K., & Uversky, V. N. (2018). Intrinsically disordered proteins in crowded milieu: When chaos prevails within the cellular gumbo. *Cellular and Molecular Life Sciences, 75*, 3907–3929.

Fonin, A. V., Silonov, S. A., Sitdikova, A. K., Kuznetsova, I. M., Uversky, V. N., & Turoverov, K. K. (2017). Structure and conformational properties of d-glucose/d-galactose-binding protein in crowded milieu. *Molecules, 22*, 244.

Fonin, A. V., Stepanenko, O. V., Sitdikova, A. K., Antifeeva, I. A., Kostyleva, E. I., Polyanichko, A. M., et al. (2019). Folding of poly-amino acids and intrinsically disordered proteins in overcrowded milieu induced by pH change. *International Journal of Biological Macromolecules, 125*, 244–255.

Fordtran, J. S., & Hofmann, A. F. (2017). Seventy years of polyethylene glycols in gastroenterology: The journey of PEG 4000 and 3350 from nonabsorbable marker to colonoscopy preparation to osmotic laxative. *Gastroenterology, 152*, 675–680.

Fruijtier-Pölloth, C. (2005). Safety assessment on polyethylene glycols (PEGs) and their derivatives as used in cosmetic products. *Toxicology, 214*, 1–38.

Gao, Y., Kang, J., Lei, Z., Li, Y., Mei, X., & Wang, G. (2020). Use of the highly biocompatible Au nanocages@PEG nanoparticles as a new contrast agent for in vivo computed tomography scan imaging. *Nanoscale Research Letters, 15*, 53.

Ghosh, S., Shahid, S., Raina, N., Ahmad, F., Hassan, M. I., & Islam, A. (2020). Molecular and macromolecular crowding-induced stabilization of proteins: Effect of dextran and its building block alone and their mixtures on stability and structure of lysozyme. *International Journal of Biological Macromolecules, 150*, 1238–1248.

Goos, J. A., Cho, A., Carter, L. M., Dilling, T. R., Davydova, M., Mandleywala, K., et al. (2020a). Delivery of polymeric nanostars for molecular imaging and endoradiotherapy through the enhanced permeability and retention (EPR) effect. *Theranostics, 10*, 567.

Goos, J. A. C. M., Davydova, M., Dilling, T. R., Cho, A., Cornejo, M. A., Gupta, A., et al. (2020b). Design and preclinical evaluation of nanostars for the passive pretargeting of tumor tissue. *Nuclear Medicine and Biology, 84-85*, 63–72.

Gorrec, F. (2016). Protein crystallization screens developed at the MRC Laboratory of molecular biology. *Drug Discovery Today, 21*, 819–825.

Greenwald, R. (2001). PEG drugs: an overview. *Journal of Controlled Release, 74*, 159–171.

Guin, D., & Gruebele, M. (2019). Weak chemical interactions that drive protein evolution: crowding, sticking, and quinary structure in folding and function. *Chemical Reviews, 119*, 10691–10717.

Gullapalli, R. P., & Mazzitelli, C. L. (2015). Polyethylene glycols in oral and parenteral formulations—A critical review. *International Journal of Pharmaceutics, 496*, 219–239.

Han, S., Sun, R., Su, H., Lv, J., Xu, H., Zhang, D., et al. (2019). Delivery of docetaxel using pH-sensitive liposomes based on D-alpha-tocopheryl poly(2-ethyl-2-oxazoline) succinate: Comparison with PEGylated liposomes. *Asian Journal of Pharmaceutical Sciences, 14*, 391–404.

Harris, J. M., & Chess, R. B. (2003). Effect of pegylation on pharmaceuticals. *Nature Reviews Drug Discovery, 2*, 214–221.

Hashemi, E., Mahdavi, H., Khezri, J., Razi, F., Shamsara, M., & Farmany, A. (2019). Enhanced gene delivery in bacterial and mammalian cells using PEGylated calcium doped magnetic Nanograin. *International Journal of Nanomedicine, 14*, 9879–9891.

Hatakeyama, H., Akita, H., Kogure, K., Oishi, M., Nagasaki, Y., Kihira, Y., et al. (2007). Development of a novel systemic gene delivery system for cancer therapy with a tumor-specific cleavable PEG-lipid. *Gene Therapy, 14*, 68–77.

Hershfield, M. S. (1995). PEG-ADA replacement therapy for adenosine deaminase deficiency: An update after 8.5 years. *Clinical Immunology and Immunopathology, 76*, S228–S232.

Hillisch, A., Heinrich, N., & Wild, H. (2015). Computational chemistry in the pharmaceutical industry: From childhood to adolescence. *ChemMedChem, 10*, 1958–1962.

Inada, Y., Takahashi, K., Yoshimoto, T., Ajima, A., Matsushima, A., & Saito, Y. (1986). Application of polyethylene glycol-modified enzymes in biotechnological processes: Organic solvent-soluble enzymes. *Trends in Biotechnology, 4*, 190–194.

Jang, H.-J., Shin, C. Y., & Kim, K.-B. (2015). Safety evaluation of polyethylene glycol (PEG) compounds for cosmetic use. *Toxicological Research, 31*, 105–136.

Jones, D. T., & Cozzetto, D. (2015). DISOPRED3: Precise disordered region predictions with annotated protein-binding activity. *Bioinformatics, 31*, 857–863.

Katre, N. V. (1993). The conjugation of proteins with polyethylene glycol and other polymers: Altering properties of proteins to enhance their therapeutic potential. *Advanced Drug Delivery Reviews, 10*, 91–114.

Kim, S. H., Jeong, J. H., Lee, S. H., Kim, S. W., & Park, T. G. (2006). PEG conjugated VEGF siRNA for anti-angiogenic gene therapy. *Journal of Controlled Release, 116*, 123–129.

Kinjo, A. R., & Takada, S. (2002). Effects of macromolecular crowding on protein folding and aggregation studied by density functional theory: Statics. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics, 66*, 031911.

Kolate, A., Baradia, D., Patil, S., Vhora, I., Kore, G., & Misra, A. (2014a). PEG - a versatile conjugating ligand for drugs and drug delivery systems. *Journal of Controlled Release, 192*, 67–81.

Kolate, A., Baradia, D., Patil, S., Vhora, I., Kore, G., & Misra, A. (2014b). PEG — A versatile conjugating ligand for drugs and drug delivery systems. *Journal of Controlled Release, 192*, 67–81.

Kuznetsova, I. M., Turoverov, K. K., & Uversky, V. N. (2014). What macromolecular crowding can do to a protein. *International Journal of Molecular Sciences, 15*, 23090–23140.

Lee, L. S., Conover, C., Shi, C., Whitlow, M., & Filpula, D. (1999). Prolonged circulating lives of single-chain Fv proteins conjugated with polyethylene glycol: A comparison of conjugation chemistries and compounds. *Bioconjugate Chemistry, 10*, 973–981.

Lim, Y. J., & Hong, S. J. (2014). What is the best strategy for successful bowel preparation under special conditions? *World journal of gastroenterology: WJG, 20*, 2741.

Liu, G., Li, K., Luo, Q., Wang, H., & Zhang, Z. (2017). PEGylated chitosan protected silver nanoparticles as water-borne coating for leather with antibacterial property. *Journal of Colloid and Interface Science, 490*, 642–651.

Liu, P., Santisteban, I., Burroughs, L. M., Ochs, H. D., Torgerson, T. R., Hershfield, M. S., et al. (2009). Immunologic reconstitution during PEG-ADA therapy in an unusual mosaic ADA deficient patient. *Clinical Immunology, 130*, 162–174.

Lu, X., Shi, S., Li, H., Gerhard, E., Lu, Z., Tan, X., et al. (2020). Magnesium oxide-crosslinked low-swelling citrate-based mussel-inspired tissue adhesives. *Biomaterials, 232*, 119719.

Mendez, U., Zhou, H., & Shikanov, A. (2018). Synthetic PEG hydrogel for engineering the environment of ovarian follicles. *Methods in Molecular Biology, 1758*, 115–128.

Meng, X., Zhang, B., Yi, Y., Cheng, H., Wang, B., Liu, Y., et al. (2020). Accurate and real-time temperature monitoring during MR imaging guided PTT. *Nano Letters, 20*, 2522–2529.

Mitchell, A. L., Scheremetjew, M., Denise, H., Potter, S., Tarkowska, A., Qureshi, M., et al. (2018). EBI metagenomics in 2017: Enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic Acids Research, 46*, D726–D735.

Molloy, K., Van, M. J., Barbara, D., & Shehu, A. (2014). Exploring representations of protein structure for automated remote homology detection and mapping of protein structure space. *BMC Bioinformatics, 15*(Suppl 8), S4.

Muegge, I., Bergner, A., & Kriegl, J. M. (2017). Computer-aided drug design at Boehringer Ingelheim. *Journal of Computer-Aided Molecular Design, 31*, 275–285.

Nasreen, K., Ahamad, S., Ahmad, F., Hassan, M. I., & Islam, A. (2018). Macromolecular crowding induces molten globule state in the native myoglobin at physiological pH. *International Journal of Biological Macromolecules, 106*, 130–139.

Newkome, G. R., Moorefield, C. N., & Vögtle, F. (2008). *Dendritic molecules: Concepts, syntheses, perspectives*. Weinheim: Wiley.

Nucci, M. L., Shorr, R., & Abuchowski, A. (1991). The therapeutic value of poly (ethylene glycol)-modified proteins. *Advanced Drug Delivery Reviews, 6*, 133–151.

Parray, Z. A., Ahamad, S., Ahmad, F., Hassan, M. I., & Islam, A. (2019). First evidence of formation of pre-molten globule state in myoglobin: A macromolecular crowding approach towards protein folding in vivo. *International Journal of Biological Macromolecules, 126*, 1288–1294.

Pasut, G., Panisello, A., Folch-Puy, E., Lopez, A., Castro-Benitez, C., Calvo, M., et al. (2016). Polyethylene glycols: An effective strategy for limiting liver ischemia reperfusion injury. *World Journal of Gastroenterology, 22*, 6501–6508.

Peng, L., Chang, L., Si, M., Lin, J., Wei, Y., Wang, S., et al. (2020). Hydrogel-coated dental device with adhesion-inhibiting and colony-suppressing properties. *ACS Applied Materials & Interfaces, 12*, 9718–9725.

Plosker, G. L. (2008). Pegylated liposomal doxorubicin: A review of its use in the treatment of relapsed or refractory multiple myeloma. *Drugs, 68*, 2535–2551.

Rajan, S. S., Turovskiy, Y., Singh, Y., Chikindas, M. L., & Sinko, P. J. (2014). Poly (ethylene glycol)(PEG)-lactic acid nanocarrier-based degradable hydrogels for restoring the vaginal microenvironment. *Journal of Controlled Release, 194*, 301–309.

Rashmi Zabihi, F., Singh, A. K., Achazi, K., Schade, B., Hedtrich, S., Haag, R., et al. (2020). Non-ionic PEG-oligoglycerol dendron conjugated nano-carriers for dermal drug delivery. *International Journal of Pharmaceutics, 580*, 119212.

Salmaso, V. (2018). *Exploring protein flexibility during docking to investigate ligand-target recognition*. Padova: University of Padova.

Sánchez-Nieves, J., Fransen, P., Pulido, D., Lorente, R., Muñoz-Fernández, M. Á., Albericio, F., et al. (2014). Amphiphilic cationic carbosilane–PEG dendrimers: Synthesis and applications in gene therapy. *European Journal of Medicinal Chemistry, 76*, 43–52.

Scaffaro, R., Lopresti, F., Maio, A., Botta, L., Rigogliuso, S., & Ghersi, G. (2017). Electrospun PCL/GO-g-PEG structures: Processing-morphology-properties relationships. *Composites Part A: Applied Science and Manufacturing, 92*, 97–107.

Shahid, S., Ahmad, F., Hassan, M. I., & Islam, A. (2019). Mixture of macromolecular crowding agents has a non-additive effect on the stability of proteins. *Applied Biochemistry and Biotechnology, 188*, 927–941.

Shahid, S., Hassan, M. I., Islam, A., & Ahmad, F. (2017). Size-dependent studies of macromolecular crowding on the thermodynamic stability, structure and functional activity of proteins: In vitro and in silico approaches. *Biochimica et Biophysica Acta - General Subjects, 1861*, 178–197.

Sledz, P., & Caflisch, A. (2018). Protein structure-based drug design: From docking to molecular dynamics. *Current Opinion in Structural Biology, 48*, 93–102.

Stepanenko, O. V., Stepanenko, O. V., Kuznetsova, I. M., Uversky, V. N., & Turoverov, K. K. (2016). Peculiarities of the super-folder GFP folding in a crowded milieu. *International Journal of Molecular Sciences, 17*, 1805.

Suzuki, T., Kanbara, N., Tomono, T., Hayashi, N., & Shinohara, I. (1984). Physicochemical and biological properties of poly(ethylene glycol)-coupled immunoglobuling G. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology, 788*, 248–255.

Tokuriki, N., Kinjo, M., Negi, S., Hoshino, M., Goto, Y., Urabe, I., et al. (2004). Protein folding by the effects of macromolecular crowding. *Protein Science, 13*, 125–133.

Tsao, D., Minton, A. P., & Dokholyan, N. V. (2010). A didactic model of macromolecular crowding effects on protein folding. *PLoS One, 5*, e11936.

Türkkan, S., Pazarçeviren, A. E., Keskin, D., Machin, N. E., Duygulu, Ö., & Tezcaner, A. (2017). Nanosized CaP-silk fibroin-PCL-PEG-PCL/PCL based bilayer membranes for guided bone regeneration. *Materials Science and Engineering: C, 80*, 484–493.

UniProt, C. (2015). UniProt: A hub for protein information. *Nucleic Acids Research, 43*, D204–D212.

Valuckaite, V., Seal, J., Zaborina, O., Tretiakova, M., Testa, G., & Alverdy, J. C. (2013). High molecular weight polyethylene glycol (PEG 15-20) maintains mucosal microbial barrier function during intestinal graft preservation. *Journal of Surgical Research, 183*, 869–875.

van den Berg, B., Ellis, R. J., & Dobson, C. M. (1999). Effects of macromolecular crowding on protein folding and aggregation. *The EMBO Journal, 18*, 6927–6933.

Veronese, F. M., & Mero, A. (2008). The impact of PEGylation on biological therapies. *BioDrugs, 22*, 315–329.

Wang, J.-Z., You, M.-L., Ding, Z.-Q., & Ye, W.-B. (2019). A review of emerging bone tissue engineering via PEG conjugated biodegradable amphiphilic copolymers. *Materials Science and Engineering: C, 97*, 1021–1035.

Wang, N., Dong, A., Radosz, M., & Shen, Y. (2008). Thermoresponsive degradable poly(ethylene glycol) analogues. *Journal of Biomedical Materials Research. Part A, 84*, 148–157.

Xiao, Q., Draper, S. R. E., Smith, M. S., Brown, N., Pugmire, N. A. B., Ashton, D. S., et al. (2019). Influence of PEGylation on the strength of protein surface salt bridges. *ACS Chemical Biology, 14*, 1652–1659.

Yang, Q., & Lai, S. K. (2015). Anti-PEG immunity: Emergence, characteristics, and unaddressed questions. *Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology, 7*, 655–677.

Yang, Y., Zhu, W., Cheng, L., Cai, R., Yi, X., He, J., et al. (2020a). Tumor microenvironment (TME)-activatable circular aptamer-PEG as an effective hierarchical-targeting molecular medicine for photodynamic therapy. *Biomaterials, 246*, 119971.

Yang, Z., Guo, Q., Cai, Y., Zhu, X., Zhu, C., Li, Y., et al. (2020b). Poly(ethylene glycol)-sheddable reduction-sensitive polyurethane micelles for triggered intracellular drug delivery for osteosarcoma treatment. *Journal of Orthopaedic Translation, 21*, 57–65.

Yeroslavsky, G., Umezawa, M., Okubo, K., Nigoghossian, K., Thi Kim Dung, D., Miyata, K., et al. (2020). Stabilization of indocyanine green dye in polymeric micelles for NIR-II fluorescence imaging and cancer treatment. *Biomaterials Science, 8*(8), 2245–2254.

Zhang, F., Liu, M. R., & Wan, H. T. (2014). Discussion about several potential drawbacks of PEGylated therapeutic proteins. *Biological & Pharmaceutical Bulletin, 37*, 335–339.

Zhang, L., Cheng, Z., Zhao, Q., & Wang, M. (2017). Green and efficient PEG-based ultrasound-assisted extraction of polysaccharides from superfine ground lotus plumule to investigate their antioxidant activities. *Industrial Crops and Products, 109*, 320–326.

Zhang, X., Chen, F., Turker, M. Z., Ma, K., Zanzonico, P., Gallazzi, F., et al. (2020). Targeted melanoma radiotherapy using ultrasmall 177Lu-labeled α-melanocyte stimulating hormone-functionalized core-shell silica nanoparticles. *Biomaterials, 241*, 119858.

Zhou, Z., Wang, Y., Qian, Y., Pan, X., Zhu, J., Zhang, Z., et al. (2020). Cystine dimethyl ester cross-linked PEG-poly(urethane-urea)/nano-hydroxyapatite composited biomimetic scaffold for bone defect repair. *Journal of Biomaterials Science. Polymer Edition, 31*, 407–422.

Zinov'eva, I. V., Zakhodyaeva, Y. A., & Voshkin, A. A. (2020). Data on the extraction of benzoic, salicylic and sulfosalicylic acids from dilute solutions using PEG-based aqueous two-phase systems. *Data in Brief, 28*, 105033.

Chapter 12
Molecular Dynamics Simulation in Drug Discovery: Opportunities and Challenges

Rohit Shukla and Timir Tripathi

Abstract Drug discovery is the process used to discover new candidate medications. In the past, most drugs were discovered by identification of active-ingredients or by serendipity. Modern drug discovery is more focused and streamlined. It starts with target identification, followed by the identification of inhibitors that bind to the target and inhibit its activity. However, developing a new drug takes typically 10–12 years before it can be commercialized. Furthermore, drug discovery costs can range between several hundred million to billions of US dollars. Recent progresses in computational approaches have sped up drug discovery and development research. Computer-aided drug design (CADD) speeds up the hit-to-lead process and enables compounds to pass the barriers of preclinical testing in a short time. Molecular dynamics (MD) simulation has emerged as an important tool in the study of the conformational flexibility and dynamics of drug-target complexes. MD simulation helps to replicate the biological events in a computer simulation. It has become a routine computational tool for CADD and a revolution in the field of drug development. It provides an accurate estimate of thermodynamics and kinetics associated with drug-target interaction and binding. Development of new methods, software, and hardware has boosted the use of MD simulation among scientists working with CADD as well as in biopharmaceutical industry. Improvements in the force-field methods may further enhance the accuracy of free-energy predictions.

Keywords Protein · Force-field · Molecular dynamics simulation · Energy minimization · Drug discovery · Computer-aided drug design · Conformational flexibility · Dynamics

R. Shukla
Molecular and Structural Biophysics Laboratory, Department of Biochemistry, North-Eastern Hill University, Shillong, India

Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology (JUIT), Solan, Himachal Pradesh, India

T. Tripathi (✉)
Molecular and Structural Biophysics Laboratory, Department of Biochemistry, North-Eastern Hill University, Shillong, India

S. K. Singh (ed.), *Innovations and Implementations of Computer Aided Drug Discovery Strategies in Rational Drug Design*,
https://doi.org/10.1007/978-981-15-8936-2_12

12.1 Introduction

Nowadays, the drug discovery process has become more rapid owing to the vast use of high-throughput virtual screening (HTVS). In addition, the drug discovery process has become much more feasible for the pharmaceutical industry and is less time consuming. There is no need for labor for screening using HTVS due to process automation. Nearly 75% of the R&D cost of developing drugs arises from failures during drug discovery in pharmaceutical industry (Noble 2003). For HTVS, which is followed by molecular dynamics (MD) simulations, enormous amount of computational power, along with good software, is needed. The small molecules obtained through HTVS can be validated by MD simulations.

MD simulation was first introduced by Alder and Wainwright and developed by Karplus group in 1977 (McCammon et al. 1977). In 2013, Martin Karplus, Michael Levitt, and Arieh Warshel were awarded Nobel Prize "for the development of multiscale models for complex chemical systems", which formed the theoretical basis of MD simulations. Presently, MD simulation is vastly used in modern science, where it works as a bridge between the wet-lab experiments and theories. MD simulations are used to describe the evolution of single as well as complex systems with time by using theoretical models (Heidari et al. 2016b). Fundamentally, computer simulation is a theoretical approach that is valuable when newly developed hypotheses or predictions can subsequently be subjected to experimental validation. Most MD simulation algorithms work in parallel due to intense efforts from scientists across the globe. MD simulation data gains importance when used in conjunction with the wet-lab experiments (Feig et al. 2018b). It can describe the atomic level dynamics of biological systems (Gigosos et al. 2018a), and the data obtained after the program execution can be analyzed virtually. It provides a detailed idea of their properties at atomic-level and their kinetic behavior with respect to time; this data cannot be generated using conventional wet-lab experiments (Hollingsworth and Dror 2018). With the increase in computational power and data storage, MD simulation has been proved to be an invaluable tool in diverse research fields because of its ability to solve complex problems by relying only on fundamental first principles and using either physical or mathematical approximations (Kumar et al. 2014). It provides deeper insights into the underlying physical principles for theory validation. Sometimes, MD simulation becomes challenging due to the high degree of complexity in both experiment and theory (Gigosos et al. 2018b). MD simulations have been extensively used to study biomacromolecules, like protein, nucleic acid, etc. In recent years, such simulations have been extended to cellular scale, and simulations of an entire cell have been analyzed to understand the basic molecular principles of life (Heidari et al. 2016a).

Drug discovery process is very lengthy, and the commercialization of a drug is a costly affair. Moreover, the chances of failure during drug discovery are very high. On the basis of a 10 years research (2009–2018), Wouters et al. estimated that the research and development investment associated with developing and commercializing a drug has increased significantly in the last decade. The average cost is now

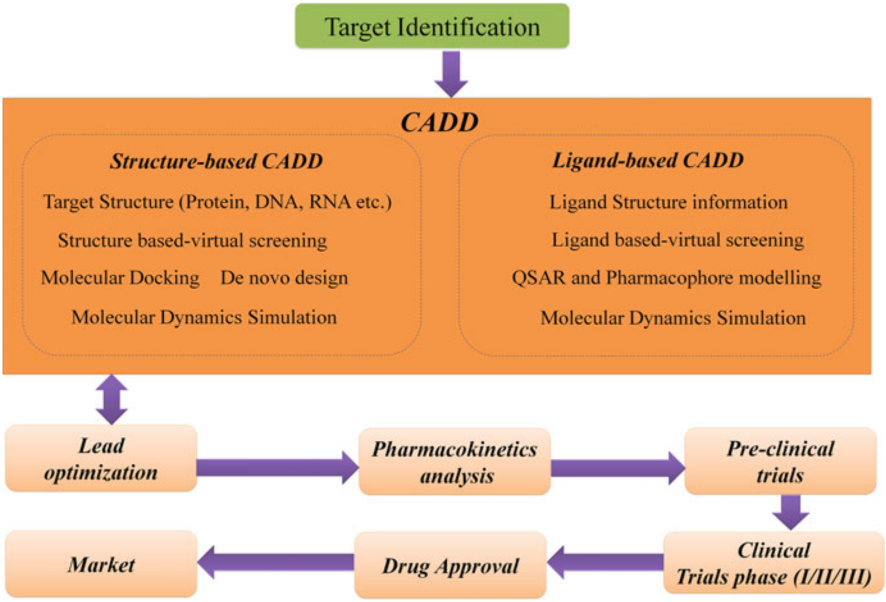

Target Identification
CADD
Structure-based CADD
Target Structure (Protein, DNA, RNA etc.)
Structure based-virtual screening
Molecular Docking
De novo design
Molecular Dynamics Simulation
Ligand-based CADD
Ligand Structure information
Ligand based-virtual screening
QSAR and Pharmacophore modelling
Molecular Dynamics Simulation
Lead optimization
Pharmacokinetics analysis
Pre-clinical trials
Clinical Trials phase (I/II/III)
Drug Approval
Market

Fig. 12.1 Overall drug discovery and development process

estimated to be ranging from $314 million to $2.8 billion (Wouters et al. 2020). Ninety percent of the drugs entering clinical trials fail to get approval of the Food and Drug Administration (FDA) and cannot reach the consumer market. Thus, using computational simulation methods in conjunction with HTVS experiments saves money as well as time of pharmaceutical industries invested in drug development. An overall protocol of drug discovery and development process is shown in Fig. 12.1.

12.2 Computer-Aided Drug Design

With the advancement in computational methodologies, computer-aided drug design (CADD) is becoming an evolutionary technology employed for drug discovery and development. It has helped in decreasing the cost of drug discovery as well as reducing the time taken for the drug discovery process. It uses silicon chip or computers to predict the lead compounds by HTVS using virtual libraries that contain millions of compounds (Loew et al. 1993). CADD can be of two types: Structure-based drug design (SBDD) and ligand-based drug design (LBDD). These approaches are briefly described below:

12.2.1 Structure-Based Drug Discovery

When the structure of a protein or a therapeutic target is known, then the most common technique used in CADD is structure-based drug discovery (SBDD). In SBDD, compounds are designed on the basis of the binding site of a therapeutic target, which is commonly a protein. SBDD uses one of the following two approaches for the identification and design of a candidate drug (that may be an antagonist, agonist, or inhibitor of a target): i) molecular docking approach, or ii) *de novo* ligand design approach. MD simulations are frequently used in SBDD to gain insights into the mechanism of ligand binding with target proteins and the conformational dynamics or fluctuations that occurred upon ligand binding (Shukla and Tripathi 2020). When membrane protein is used for drug designing then the membrane permeability must also be considered as a crucial factor (Wang et al. 2010; Hanson et al. 2015). The success rate of the SBDD is very high. By using SBDD pipeline, several compounds have passed through the clinical trials and have obtained FDA approvals to reach market much sooner than conventional drug discovery. Commercial drugs, like zanamivir, nelfinavir, and aleglitazar, etc., are few examples of drugs developed using SBDD approach (Talele et al. 2010; Schames et al. 2004).

Structure-based HTVS (SB-HTVS) against small compounds, is a widely used *in-silico* method (Shukla et al. 2018a; Shukla et al. 2018b; Shukla et al. 2018c; Shukla et al. 2018d) to search huge libraries of compounds and select eligible compounds for biological testing. In this approach, the ligands are allowed to dock with the target, and the binding energy and binding conformation of the ligands towards the target are estimated and determined. In addition to finding the best docking pose, SB-HTVS also ranks the docking results according to their predicted affinities for the entire database. SB-HTVS allows screening of millions of compounds for target complementarity within small time.

12.2.2 Ligand-Based Drug Design

Ligand-based drug design (LBDD) is an alternative method of drug designing that is employed when the target binding site or structure is not known, while the ligand structure and its IC_{50} value towards the target are known (Loew et al. 1993; Mason et al. 2001). Here, if sufficient number of active ligands with diverse activity values is known, then a 3D-pharmacophore model can be built for these sets of ligands by overlapping all of them and finding the common feature among them. This method is completely dependent on the experimental binding affinity of the previously identified ligands. Several LBDD approaches are available, among which, the more commonly used techniques are pharmacophore modeling, molecular similarity, and quantitative structure-activity relationship (QSAR) (Acharya et al. 2011).

Pharmacophore modeling is one of the major elements of drug designing employed in the absence of structural data of the target receptor. In pharmacophore modeling, common structural features of ligands, like number of hydrogen bond donors, acceptors, aromaticity, benzene rings, etc., that bind to a target, are used to perform screening of candidate molecules (Yang 2010). Pharmacophores can be used as queries for retrieving potential leads from structural databases (lead discovery), for designing compounds with desired features (lead optimization), and for assessing similarity and diversity in molecules using pharmacophore fingerprints. In molecular similarity methods, the molecular fingerprint of known ligands that bind to a target is used to identify the molecules with similar fingerprints through screening of large libraries of molecules. Molecular similarity methods allow identification of additional compounds with a higher chance of displaying similar biological activities against the same target (Bender and Glen 2004). QSAR is a bioinformatics method used to determine the relationship between structural features of ligands that bind to a target and the corresponding biological activity effect (Verma et al. 2010). QSAR methods are based on statistics that correlate the activities of target drug interactions with various molecules. The QSAR method is based on the fact that structurally similar molecules tend to show similar biological activity (Verma and Hansch 2009). QSAR is widely used in drug design and discovery to identify chemical compounds with good inhibitory effects on specific targets and with low toxicity (Amic et al. 1997, 1998; Gigosos et al. 2018b; Pronk et al. 2013).

12.3 Molecular Dynamics Simulation

12.3.1 Background

Molecular dynamics (MD) is a computer-based simulation technique used to analyze the physical motion of atoms and molecules across conformational space, and provide information on evolution of atomic positions with respect to time (Zhang et al. 2015a). MD simulations are used to describe the atomic and molecular properties of protein, drug-target interactions, solvation of compounds, and conformational changes that a protein or compound may undergo under various conditions. It is based on the method of Newtonian mechanics (NM) (also known as classical mechanics) (Zhang et al. 2015b). NM relates to the motions of large particles, while quantum mechanics (QM) relates to the motions of small particles (atoms and molecules). NM proposes the motion of particle as a continuous entity, while QM asserts that small particles exist in discrete states of motion (energies). QM incorporates nuclear and/or electronic interactions between particles; simulations are based on the Schrodinger equation, including (and within) different approximations most of which use the action principle of NM (Zhang et al. 2015b). In NM, atoms come together with nucleus and the electrons are assumed as point charge with associated mass. To understand the basis of MD simulation, the knowledge of force-field is also important. Force-fields are a set of potential energy functions that are employed to define the relation between the structure and the potential energy (Martn-Garca et al. 2015). It is a mathematical expression describing the dependence of the energy of a system on the coordinates of its particles. It also describes the building blocks required for the computation of force and energy that contains bonded and non-bonded interaction. The molecular mechanics (MM) force-field is used to compute the forces acting on each atom. MM force-fields are based on the four fundamental principles:

1. Born-Oppenheimer method: In this method, the electron speed is thought to be constant. The electrons are considered to have more motion than that of the nuclei while the nuclei are considered more stable than the electrons (Brela et al. 2016).
2. The bond length and bond angle are defined for every bond, while the steric energy of the system increases when the deviation in the bond increases. According to the force-field, the nuclear position of the atom is adjusted through a mathematical model and a new geometry with lower energy may be assigned.
3. The potential energy of the surface of the molecule is described by mathematical equations that give the best match for the experimental data. When these parameters are evaluated for a compound, then they can be generalized for similar compounds.
4. Fourth is the atom type that includes electronic distribution and hybridizations.

12.3.2 Energy Minimization

It is also known as geometry optimization. Every molecule always tends to be in the lowest energy state having the highest stability. It is a mathematical process in which the nearest local minima are searched by the activation energy barrier. This process is dependent on the initial conformation of the input structure. Several algorithms are available for the energy minimization, like Conjugant Gradient, ABNR from CHARMM-GUI, and Sander from Amber, etc. Energy minimization decreases the energy of each atom as compared to its initial structure. Before proceeding to MD simulation, it is necessary to perform energy minimization.

12.3.3 Explicit and Implicit Solvation

MD simulations are performed to study the conformational features of a protein in a cellular environment. In order to study cellular-level dynamics of a protein, it is required to create a virtual cell-like environment. Since a cell has 75% water molecules, it is necessary to add water molecules into the system to construct a cell-like environment. In the first step, a box is created. Inside the box, molecules like protein, ligand, water, ions, etc. are added. The size of the box is also important as the number of water molecules added in the box depends on the size of the protein and the box. There are two types of water models: i) the explicit water model, and ii) the implicit water model. In the case of explicit water model, we add real water molecules in the system, while in the case of implicit water model, we add water-like environment in the system using mathematical equations (Leherte and Vercauteren 2017). The implicit water model takes less computation as compared to explicit water model; most MD simulation methods use the explicit water model.

12.3.4 Periodic Boundary Conditions

Periodic boundary conditions (PBC) are used in MD simulations to avoid problems with boundary effects caused by finite size, which makes the system infinite (Linke et al. 2018). During MD simulation, inside the water box of the system, the particles might move out due to dimensionality. To avoid this error, a replica of the box is created to cover the original box from all sides, and whenever the particle moves out from the central box, another particle would enter from the adjacent replica box at the same speed. The presence of PBC means that any atom that leaves a simulation box by, say, the right-hand face, enters the simulation box by the left-hand face.

12.3.5 Computation of Long-Ranged Coulomb Interactions

The particle-mesh Ewald (PME) method is an efficient and accurate method for the evaluation of long-range electrostatic interactions in biomolecular systems that are being studied by MD simulations. The electrostatic potential comes in the category of long-range potential, and after a particular distance, they become weaker and gradually vanish. A specific cut-off value is set in each simulation for measuring the electrostatic potential. However, the cut-off does not allow smooth fall of the potential as it will become zero after certain distance and may lead to artifacts in calculation. Hence, to encounter this problem, PME method (Darden et al. 1999) is used as it can divide the long-range electrostatic interaction into two parts: first, long-range, and second, short-range. A technique called Fast Fourier Transformation is used for calculation of both the long-range and short-range electrostatic interactions; here, a cut-off value is used to switch between calculations of the two interactions.

12.3.6 Work-Flow for MD Simulation

A complete MD simulation protocol utilizes the following steps sequentially (Fig. 12.2):

1. File format conversion
2. Box construction
3. Addition of water molecules
4. Neutralization of the defined system
5. Energy minimization
6. Preparatory stages, involving heating and equilibration
7. Production run (final simulation)
8. Trajectory conversion and result analysis

The PDB protein file is converted into a specific format using software, like Gromacs and Amber. For instance, in Gromacs, the PDB file is converted into the .gro file format. This conversion generates the topology on the basis of a different force-field. The topology files are available for standard molecules, like DNA, RNA, and protein, but the topology is not defined for heteroatoms and non-standard amino

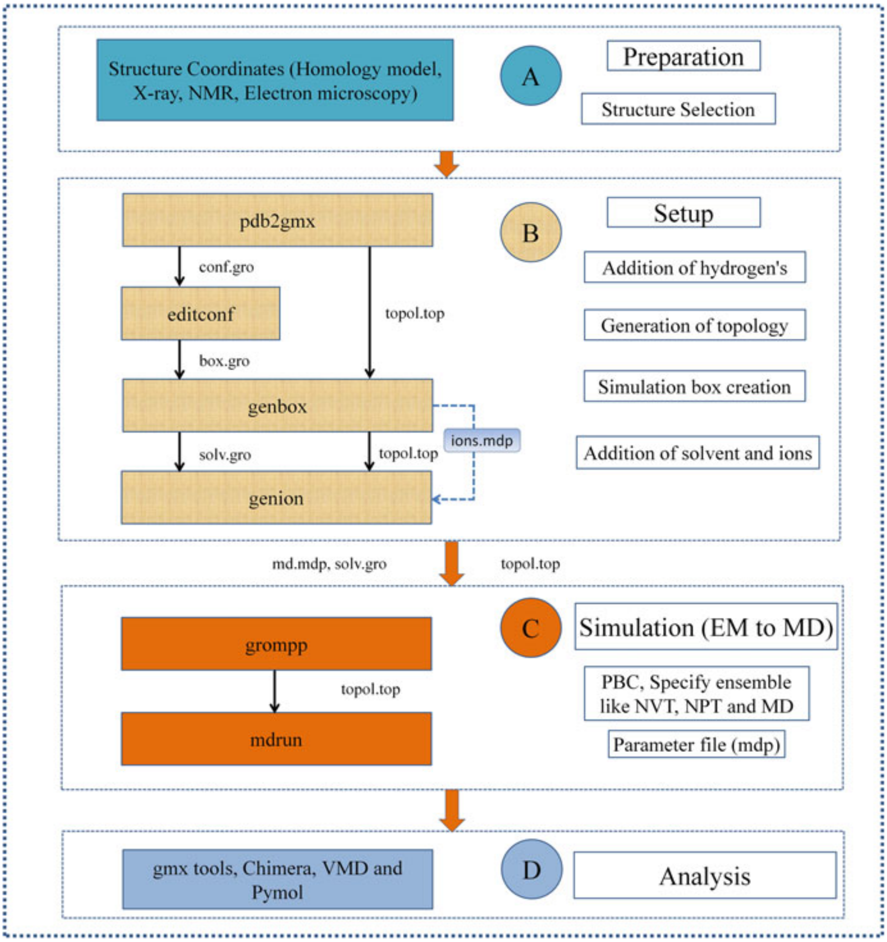


Fig. 12.2 Overall workflow of MD simulation

acids. Certain external servers are available that can generate the topology for a new ligand, such as CHARMM-GUI (Jo et al. 2008), ProDRG (van Aalten et al. 1996), automated topology builder (Stroet et al. 2018), Amber tools, etc. These servers can produce the necessary files required for ligand topology according to their atomic arrangement. After the generation of the required files, the box size is defined according to the protein to be placed inside. Then, the water molecules are filled in the box, and the system is neutralized by adding the required sodium or potassium ions. The geometry of the whole system is optimized (to remove steric clash, if any) by energy minimization. After energy minimization, the equilibration stages are defined in which the temperature linearly increases from 0 to 310 K or as per the experiment requirements. At each integration step, the velocity is re-assigned using Maxwell Boltzmann's distribution that increases the temperature. After heating to 310 K, the system is equilibrated to ensure the stability. Then, the position restraint simulation is carried out under NVT (the constant Number of particles, Volume, and Temperature) and NPT (the constant Number of particles, Pressure, and Temperature) conditions. After equilibration, final production run is initiated and the actual analyses are performed on the trajectory obtained from the production run.

12.4 Linking Wet-Lab Experiments with MD Simulations

Computational methods are becoming increasingly important and complementary to wet-lab experiments for studying the structure and function of biomolecules. MD simulations are completely based on the theoretical models; hence, a connection between experiment and simulation is required to properly evaluate the hypothesis. Wet-lab experiments provide the initial data and guide the MD simulation to obtain the mechanistic understanding that may include 3D structure of the molecule, the effect of mutation, effect of solution variables, or *in-vitro* and *in-vivo* activities (Sonkar et al. 2019; Pandey et al. 2017; Chetri et al. 2019). For instance, in case of mutational analysis, wet-lab approaches fail to explain the mechanism involved when a particular mutation causes loss in activity. MD simulation can provide information on the atomic-level changes in the structure upon mutation, for instance, the effects of mutation on active site microenvironment or the effect of mutation on the structure, dynamics, flexibility, or residue-interaction network of the protein. All this information can be obtained by comparing MD simulation trajectories of native and mutant proteins (Shukla et al. 2017a; Shukla et al. 2017b; Shukla et al. 2018e). Such data cannot be obtained by the *in-vitro* and *in-vivo* experiments. Once simulations have been carried out, comparisons with experiment are needed for validation and to follow-up on predictions for further elucidation.

12.5 Challenges in MD Simulations

The utility of MD simulations is still limited by several challenges: (1) high computational demand prohibits routine simulations greater than a μsec in length, (2) the force fields used require further refinement, (3) improved hardware and software is needed for storage, management, and dissemination of the huge trajectory data.

The experimental conditions and scales are often quite different from those of the simulations, and experiments cannot provide full atomistic resolution in nsec to psec time that simulations can provide. For instance, whilec wet-lab experiments, the protein needs to be incubated with a denaturant till unfolding equilibrium is achieved. The incubation time can range from few mins to several hours (Kalita et al. 2019). When unfolding experiments are simulated using MD, such long time-scale simulation cannot be performed as MD simulation is generally possible in nsec to μsec only. Thus, such data might not corroborate with each other.

Significant progress has been made in the development of better force-fields that can accurately describe folded proteins. However, force fields need to be further developed in order to describe intrinsically disordered proteins (IDPs). In addition, comparisons of MD simulations of IDPs and peptides with the experimental data have shown significant discrepancies that need to be addressed in the future. A force-field that can accurately describe both ordered and disordered proteins will be highly useful. It would allow MD simulations of proteins containing both ordered and disordered regions and proteins that transition between ordered and disordered states. Improvement in force-field accuracy is also required for a better representation of the molecular energy surface. In particular, the effects of charge polarization must be included in the force-field as the fields that are induced by solvent, ions, other macromolecules, and the protein itself affect electrostatic interactions. Moreover, while MD simulations can accurately predict many important molecular motions, they are not suited to systems where quantum effects are important, for instance, when transition metal atoms are involved in protein binding.

The large amount of data generated by MD simulations itself presents a challenge. At the logistical level, the storage, management, and dissemination of terabyte-scale trajectory data are still difficult even as the performance and capacity of storage resources continue to increase (Kumar et al. 2014). In most of the wet-lab research work, the primary data is publicly available; this is generally not the case for MD simulation data due to its huge volume. Public databases are available for storing biological and structural data that is obtained through experiments or from computational/bioinformatics analysis; yet there is no available resource database for storing and sharing of MD simulation data (Thibault et al. 2013). Many efforts have been made to develop a database for MD simulation data, but they failed as the trajectory files are not downloadable because of network bandwidth limitations (Feig et al. 2018a; Meyer et al. 2010; Todd 2005). It still remains unlikely in near future to efficiently transfer peta-byte scale data sets over the internet; till then, the development of an MD database does not seem possible. One way of handling the problem of huge data size is to remove less important parts of the system, like solvent data, from the file. It is also possible to store the system of interest at a coarse-grained level, even though the original simulations were carried out in atomistic detail resolution (Thibault et al. 2013). A database could be developed to store such data (Kumar et al. 2014) by increasing the hardware and software infrastructure facilities. The same challenges also apply even in a computational laboratory when there is not enough storage space to maintain all of the generated MD simulation data for direct access. For re-analysis and comparative studies of different MD simulations, data must be stored due to the high computational costs of data generation. For instance, one μsec simulation of a relatively small system running on 24 processors takes several months to complete.

12.6 Opportunities and Applications of MDS

12.6.1 MD-Derived Observables for Drug Discovery

Many observables can be obtained from MD runs (Rajendran et al. 2018). The principles of statistical mechanics allow quantitative estimation of important thermodynamic observables that can be computed from MD trajectories. These include internal energy, pressure, and heat capacity. However, the key thermodynamic quantity in drug discovery is the protein-ligand binding free energy (Ganesan et al. 2017).

12.6.2 Application in Molecular Docking and Drug Design

Molecular docking approaches are among the most commonly used approaches in the field of bioinformatics nowadays. Docking is used to understand the binding of drug/ligand/inhibitor/substrate or cofactor to the protein counterpart, which provides important details for structure-based drug designing (Kalita et al. 2017). The recognition process of drug binding to protein is a dynamic process (Koshland 1958); upon binding of drug, the protein may undergo structural rearrangements. These conformational changes play a key role in drug binding and associated energetics, providing important information on protein-drug interactions. For SB-HTVS, the structure of the target protein is downloaded from PDB and is virtually docked with drugs/ligands. Several structures of same protein could be available in the PDB in apo-form or bound to different ligands. The substrate, cofactor, inhibitor, or drugs mostly bind to the target at the same pose in same conformation. The binding of ligand may change the properties of protein *in-silico* (structure), *in vivo* or/and *in vitro* (activity). If the experimental structure of a protein is not available, it is important to have a good model structure, otherwise the docking results may not be valid and the *in silico* data may not corroborate with the *in vitro/in vivo* data. Sometimes, over-compression of a PDB structure can also be a reason for failure of ligand binding to target. This problem is mostly faced during protein-protein docking of bound and unbound structures (Gupta et al. 2020). For ligand docking to target, the ligand is set to be flexible in order to bind in a correct pose or conformation (Hawkins and Nicholls 2012; Ishikawa 2013). In the case of protein flexibility during docking, the docking methods are not well developed. Most of them use algorithms to select from a limited number of protein conformations that are either pre-computed or simulated structures (Friesner et al. 2004; Nabuurs et al. 2007; Borrelli et al. 2005). Hence, in this case, MD simulation can play an important role to evaluate dynamic stability. If the protein structure is poor, we can refine it through the MD simulation, while if the ligand does not bind properly to the protein target or falsely binds with the protein, MD simulation can be used to predict false-positive binding.

12.6.3 Application in Elucidating the Allosteric Binding Sites in Proteins

The functions of most proteins are regulated through allostery, in which an effector binds to a distal site in the protein and modulates catalytic activity (e.g., substrate-binding affinity or catalytic efficiency) at the active site. Changes in the binding affinity of substrate by allosteric effectors are thought to be mediated by conformational transitions of the proteins that may have functional consequences (e.g., an increase in substrate-binding affinity). MD simulations have been used to investigate the transition of enzyme from active to inactive state at atomic resolution. For example, several allosteric hotspot residues have been identified that can modulate the enzymatic activity of *Mycobacterium tuberculosis* isocitrate lyase (Shukla et al. 2017a; Shukla et al. 2017b; Shukla et al. 2018e). The allosteric mechanism of enzyme inactivation upon single point mutation indicates how conformational dynamics may lead to modulation of functional activity.

The concept of allostery is being studied since the beginning of research in protein biochemistry (Monod et al. 1963; Monod et al. 1965; Koshland et al. 1966). Conformation shifts involved in allostery span from small rearrangements to large quaternary shifts. It has been concluded that conformational rearrangements that are involved in allosteric transitions depend on the collective motions and dynamics of proteins (Henzler-Wildman et al. 2007). Thus, MD simulations have been extensively used as a natural tool to understand protein allostery. Although the functional consequences of allosteric binding can be directly measured, the mechanisms of action are assessed with the help of inference and simplification. Most of the data on MD simulations in this field uses simplified frameworks, like discrete-MD (Sfriso et al. 2012) or Go-Model (Sfriso et al. 2013), or popular non-simulation equivalents, like elastic network models (Kim et al. 2002; Bahar and Rader 2005; Orellana et al. 2010), which find the transition path between known experimental structures. With full atomic representations, it is common to trick the algorithm using targeted (Schlitter et al. 1993; Kruger et al. 2001; Perdih et al. 2007), or supervised MD (Deganutti et al. 2015), where the simulation is artificially driven to the desired conformation. Here, the analysis of the path could give insight into the energetic and allosteric transitions. For those cases where allosteric regulation is known to occur but one of the ends is unknown, long simulations (either alone or with enhanced conformational sampling) are required (Arkhipov et al. 2013; Klepeis et al. 2009; Anderson et al. 2014; Boczek et al. 2015). The direct use of conformational ensembles, without any conditioning, is still beyond the possibilities of current MD simulations; however, certain cases with well-defined collective motions might be feasible.

12.6.4 Application in Refining Protein Structure Predictions

Conformational sampling via MD simulations, based on atomistic force-fields, is increasingly becoming a preferred choice for structure refinement. Prediction of a 3D structure of a biomolecule is highly demanded in structural bioinformatics. For the *ab-initio* method of structure prediction, the MD simulation is a very useful technique (Dorn et al. 2014; Lindorff-Larsen et al. 2011; Piana et al. 2012, 2013; Bonneau and Baker 2001). A long time-scale simulation is mostly used to predict a well stable structure in *ab-initio* modeling. However, homology modeling (or template-based modeling) is more efficient than *ab-initio* method (Lance et al. 2010; Joo et al. 2007; Moult 2005; Roy et al. 2010; Zhang 2009; Ginalski et al. 2003; Misura and Baker 2005; Sali and Blundell 1993). In homology modeling, one or more than one structures are considered as a template on the basis of sequence identity with the target protein. Following homology modeling, a 3D structural model of the target protein is obtained. In most methods, the predicted models are subjected to energy minimization using the MM method. A strained simulation is also performed during this process (Sali and Blundell 1993). MD simulation is used to further refine the structure to obtain a more stable structure (Raval et al. 2012). A long MD simulation can be used to obtain a stable structure of the protein model that is close to its native conformation. To achieve local refinement, MD simulations on a time-scale of ~10 nsec are required, while to study major structural changes, MD simulations on at least a μs time-scale are required.

12.6.5 Application in Determination of Peptide Structures

The function of peptides is related to their unique conformational behavior; thus, accurate prediction of the peptide structure is important, particularly in peptide-based drug designing. A peptide in solution can have several distinct conformations with minor differences in free energy. Thus, the accuracy of the energy function and solvent model is important for determining conformations of peptides. MD simulation methods have been used to predict the 3D structures of many naturally occurring or designed peptides. Since the alignment of short sequences may be less reliable, use of homology modeling methods for prediction of peptide structure have low validity (Shen et al. 2014). Structures of short peptides are usually highly sensitive to their exact sequences and small variations in their sequence may lead to massive conformational alterations (McHugh et al. 2017). Compared to proteins, converged conformational sampling of small peptides can be easily achieved by MD simulations, especially using enhanced sampling methods. If the linear peptides get cyclized either by backbone or side-chain linkages, they can attain conformational stability (Damas et al. 2013; Razavi et al. 2014) and can exhibit better drug-like properties. Due to availability of their experimental structures and resemblance to protein loops, cyclic peptides are much better candidates for benchmarking structure prediction methods, compared with linear peptides. However, despite good conformational sampling, MD simulation still cannot reliably predict all peptide structures (Slough et al. 2017).

12.7 Future and Challenges of MD Simulation

MD simulations are an essential technique in drug designing that facilitates all processes from target identification till lead identification (Durrant and McCammon 2011). In the earlier sections, we have discussed how the MD simulation is used for drug design, lead optimization, understanding allostery, and target refinement. Experimental methods cannot explain atomistic level details of these issues. MD simulation can provide information on average conformational changes; this analysis is completely dependent on the accuracy of the force-field, time-scale of simulations, and analysis of MD trajectories. Nowadays, much longer simulations, in μsec time-scale, can be carried out due to the availability of extensive GPUs and TPUs and high computational power. However, msec time-scale simulation is needed to obtain reliable insight into protein dynamics, conformational transitions associated with protein function, and protein folding and aggregation processes (Daggett 2000; Day and Daggett 2003; Klepeis et al. 2009), which is yet difficult to achieve. Long time-scale simulation can provide complementary results to the experimental data. Conventional MD simulations can reveal the conformation fluctuations around an energy-stable protein conformation state. To achieve the native conformation of the protein, several enhanced sampling methods, like umbrella sampling and metadynamics simulation techniques, have been developed. Two more sampling methods have been developed that add the temperature to the system state variables and allow it to vary during the simulation. These methods include (1) simulated tempering (ST) methods, in which the temperature is treated as a dynamical variable evolving in parallel with the physical variables, and (2) replica exchange molecular dynamics (REMD), in which multiple replicas of the system are evolved at different temperatures which they exchange (Sugita and Okamoto 1999; Kastner 2011; Laio and Gervasio 2008; Bode et al. 2007; Marinari and Parisi 1992;

Nguyen et al. 2013). The ST method is more efficient than the REMD and requires low computational cost (Zhang and Ma 2008; Zhang et al. 2015c). In addition, the ST method can be used with a variety of force-fields, like all-atom to coarse-grained to QM force-field (Zhang et al. 2015c).

To obtain accurate information from MD simulations, the choice of the force-field is also a key factor as the whole simulation depends on force-fields. For the same protein, the output trajectories may be different under different force-fields (Man et al. 2017; Nguyen et al. 2011). Due to the key role of force-fields in MD simulation, continuous improvements and research for developing better force-fields are required (Freddolino et al. 2010). Furthermore, methods for trajectory analysis should also be improved for correct identification of the conformational transition pathways, especially in case of protein misfolding and aggregation studies. There are many methods for trajectory analysis, such as Markov state model (MSM) and dynamic network analysis (DNA) method (Bode et al. 2007). In the MSM method, the kinetic networks are modeled in various metastable states in conformation space to provide an ensemble of transition pathways quantitatively. As a consequence, the MSM method has been intensively applied to study protein folding and structural dynamics (Lane et al. 2013). The DNA method can reveal the communication between different domains of the protein during the conformational changes of proteins and thus, reveals the conformational changes in a protein (Kim et al. 2006). These types of trajectory analysis methods should be further improved so that they can be used in protein structure-function analysis.

As discussed in earlier sections, MD simulation plays a major role in the drug development process, including lead discovery and lead optimization when used in combination with molecular docking. MD simulations can improve the enrichment factor of HTVS by considering multiple conformations of the target and by re-ranking the hit compounds according to their binding free energy calculations. The complex, which is obtained from molecular docking, is used in MD simulation to obtain accurate binding position that can be used for lead optimization. Generally, conventional MD simulation method is used to obtain the correct binding position, while enhanced sampling simulations are used for studying the structure-function relationship of the protein, antibiotic resistant analysis, etc. (Kulczycka-Mierzejewska et al. 2018). In SB-HTVS, the protein or target binding site is in a fixed form. Thus conventional MD simulation can be directly used to understand receptor flexibility and dynamics upon ligand binding. The snapshots of the MD simulation trajectory are used along with Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA) approach to estimate the binding free energy of the protein-ligand complex, which is a major issue in lead optimization, and for ranking the lead compound (Genheden and Ryde 2015). Nowadays, metadynamics is applied as an enhanced sampling technique in SB-HTVS (Gervasio et al. 2005). Metadynamics has several advantages over conventional MD simulation, for instance, it can explore the conformational changes during ligand binding using a series of collectible variables and can determine the best binding state of ligand by using the traverse sampling among other variables. The precision of the sampling process can be adjusted by changing the bins of the variables, thereby saving computing resources.

MD simulations also offer a computational route to characterize both the structure and dynamics of protein-protein complexes; however, it is still difficult to study protein-protein association and dissociation in MD simulations. The protein-protein interaction site does not have any binding groove. MD simulations have been used to explore the molecular design of protein interaction sites. Using MDS, we can simulate the protein-protein complex with a small molecule, and then, find the binding groove for molecular docking in protein-protein complex.

12.8 Conclusion

The impact of MD simulations in the field of drug discovery has increased dramatically in recent years. Drug discovery research is an important example of an area in which computer simulations can drive wet-lab experiments. Comprehensive structure-based drug design process requires detailed understanding of the dynamic properties of the target protein; MD simulation plays a huge role in characterization of this process. At the qualitative level, MD simulations provide a variety of information to guide the ligand optimization process. At the quantitative level, MD simulation provides more accurate estimates of ligand binding affinities than other computational approaches, such as docking. Since protein-ligand interaction and the associated molecular motions are microscopic events that take place in a msec time-scale, complete understanding of the atomistic energetics and mechanics of binding is not possible using current wet-lab experimental techniques. MD simulations are helpful in providing the details that experimental approaches fail to provide. New methods and algorithms are continuously being developed and refined to improve the capabilities of MD simulation. Usefulness of MD simulations in drug discovery will exponentially grow as simulations will become faster, cheaper, more widely accessible, and more accurate.

References

Acharya, C., Coop, A., Polli, J. E., & Mackerell Jr., A. D. (2011). Recent advances in ligand-based drug design: Relevance and utility of the conformationally sampled pharmacophore approach. *Current Computer-Aided Drug Design, 7*(1), 10–22.

Amic, D., Davidovic-Amic, D., Beslo, D., Lucic, B., & Trinajstic, N. (1997). The use of the ordered Orthogonalized multivariate linear regression in a structure−activity study of Coumarin and flavonoid derivatives as inhibitors of aldose Reductase. *Journal of Chemical Information and Computer Sciences, 37*, 586.

Amic, D., Davidovic-Amic, D., Beslo, D., Lucic, B., & Trinajstic, N. (1998). QSAR of Flavylium salts as inhibitors of xanthine oxidase. *Journal of Chemical Information and Computer Sciences, 38*(5), 815–818.

Anderson, J. S., Mustafi, S. M., Hernandez, G., & LeMaster, D. M. (2014). Statistical allosteric coupling to the active site indole ring flip equilibria in the FK506-binding domain. *Biophysical Chemistry, 192*, 41–48.

Arkhipov, A., Shan, Y., Das, R., Endres, N. F., Eastwood, M. P., Wemmer, D. E., et al. (2013). Architecture and membrane interactions of the EGF receptor. *Cell, 152*(3), 557–569.

Bahar, I., & Rader, A. J. (2005). Coarse-grained normal mode analysis in structural biology. *Current Opinion in Structural Biology, 15*(5), 586–592.

Bender, A., & Glen, R. C. (2004). Molecular similarity: A key technique in molecular informatics. *Organic & Biomolecular Chemistry, 2*(22), 3204–3218.

Boczek, E. E., Reefschlager, L. G., Dehling, M., Struller, T. J., Hausler, E., Seidl, A., et al. (2015). Conformational processing of oncogenic v-Src kinase by the molecular chaperone Hsp90. *Proceedings of the National Academy of Sciences of the United States of America, 112*(25), E3189–E3198.

Bode, C., Kovacs, I. A., Szalay, M. S., Palotai, R., Korcsmaros, T., & Csermely, P. (2007). Network analysis of protein dynamics. *FEBS Letters, 581*(15), 2776–2782.

Bonneau, R., & Baker, D. (2001). Ab initio protein structure prediction: Progress and prospects. *Annual Review of Biophysics and Biomolecular Structure, 30*, 173–189.

Borrelli, K. W., Vitalis, A., Alcantara, R., & Guallar, V. (2005). PELE: Protein energy landscape exploration. A novel Monte Carlo based technique. *Journal of Chemical Theory and Computation, 1*(6), 1304–1311.

Brela, M. Z., Wãjcik, M. J., Witek, J., Boczar, M., Wrona, E., Hashim, R., et al. (2016). Born-Oppenheimer molecular dynamics study on proton dynamics of strong hydrogen bonds in aspirin crystals, with emphasis on differences between two crystal forms. *The Journal of Physical Chemistry. B, 120*(16), 3854–3862.

Chetri, P. B., Shukla, R., & Tripathi, T. (2019). Identification and characterization of glyceraldehyde 3-phosphate dehydrogenase from Fasciola gigantica. *Parasitology Research, 118*(3), 861–872.

Daggett, V. (2000). Long timescale simulations. *Current Opinion in Structural Biology, 10*(2), 160–164.

Damas, J. M., Filipe, L. C. S., Campos, S. R. R., Lousa, D., Victor, B. L., Baptista, A. M., et al. (2013). Predicting the thermodynamics and kinetics of Helix formation in a cyclic peptide model. *Journal of Chemical Theory and Computation, 9*(11), 5148–5157.

Darden, T., Perera, L., Li, L., & Pedersen, L. (1999). New tricks for modelers from the crystallography toolkit: The particle mesh Ewald algorithm and its use in nucleic acid simulations. *Structure, 7*(3), R55–R60.

Day, R., & Daggett, V. (2003). All-atom simulations of protein folding and unfolding. *Advances in Protein Chemistry, 66*, 373–403.

Deganutti, G., Cuzzolin, A., Ciancetta, A., & Moro, S. (2015). Understanding allosteric interactions in G protein-coupled receptors using supervised molecular dynamics: A prototype study analysing the human A3 adenosine receptor positive allosteric modulator LUF6000. *Bioorganic & Medicinal Chemistry, 23*(14), 4065–4071.

Dorn, M., MB, E. S., Buriol, L. S., & Lamb, L. C. (2014). Three-dimensional protein structure prediction: Methods and computational strategies. *Computational Biology and Chemistry, 53PB*, 251–276.

Durrant, J. D., & McCammon, J. A. (2011). Molecular dynamics simulations and drug discovery. *BMC Biology, 9*, 71.

Feig, M., Nawrocki, G., Yu, I., Wang, P., & Sugita, Y. (2018a). Challenges and opportunities in connecting simulations with experiments via molecular dynamics of cellular environments. *Journal of Physics Conference Series, 1036*, 012010.

Feig, M., Nawrocki, G., Yu, I., Wang, P., & Sugita, Y. (2018b). Challenges and opportunities in connecting simulations with experiments via molecular dynamics of cellular environments. *Journal of Physics Conference Series, 1036*, 012010.

Freddolino, P. L., Harrison, C. B., Liu, Y., & Schulten, K. (2010). Challenges in protein folding simulations: Timescale, representation, and analysis. *Nature Physics, 6*(10), 751–758.

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., et al. (2004). Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of Medicinal Chemistry, 47*(7), 1739–1749.

Ganesan, A., Coote, M. L., & Barakat, K. (2017). Molecular dynamics-driven drug discovery: Leaping forward with confidence. *Drug Discovery Today, 22*(2), 249–269. https://doi.org/10.1016/j.drudis.2016.11.001

Genheden, S., & Ryde, U. (2015). The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin Drug Discov, 10*(5), 449–461.

Gervasio, F. L., Laio, A., & Parrinello, M. (2005). Flexible docking in solution using metadynamics. *Journal of the American Chemical Society, 127*(8), 2600–2607.

Gigosos, M. A., Gonzãlez-Herrero, D., Lara, N., Florido, R., Calisti, A., Ferri, S., et al. (2018a). Classical molecular dynamics simulations of hydrogen plasmas and development of an analytical statistical model for computational validity assessment. *Physical Review E, 98*(3), 033307.

Gigosos, M. A., Gonzãlez-Herrero, D., Lara, N., Florido, R., Calisti, A., Ferri, S., et al. (2018b). Classical molecular dynamics simulations of hydrogen plasmas and development of an analytical statistical model for computational validity assessment. *Physical Review E, 98*(3), 033307.

Ginalski, K., Elofsson, A., Fischer, D., & Rychlewski, L. (2003). 3D-jury: A simple approach to improve protein structure predictions. *Bioinformatics, 19*(8), 1015–1018.

Gupta, S., Shukla, H., Kumar, A., Shukla, R., Kumari, R., Tripathi, T., et al. (2020). Mycobacterium tuberculosis nucleoside diphosphate kinase shows interaction with putative ATP binding cassette (ABC) transporter, Rv1273c. *Journal of Biomolecular Structure & Dynamics, 38*(4), 1083–1093.

Hanson, S. M., Newstead, S., Swartz, K. J., & Sansom, M. S. P. (2015). Capsaicin interaction with TRPV1 channels in a lipid bilayer: Molecular dynamics simulation. *Biophysical Journal, 108* (6), 1425–1434.

Hawkins, P. C., & Nicholls, A. (2012). Conformer generation with OMEGA: Learning from the data set and the analysis of failures. *Journal of Chemical Information and Modeling, 52*(11), 2919–2936.

Heidari, Z., Roe, D. R., Galindo-Murillo, R., Ghasemi, J. B., & Cheatham, T. E. (2016a). Using wavelet analysis to assist in identification of significant events in molecular dynamics simulations. *Journal of Chemical Information and Modeling, 56*(7), 1282–1291.

Heidari, Z., Roe, D. R., Galindo-Murillo, R., Ghasemi, J. B., & Cheatham, T. E. (2016b). Using wavelet analysis to assist in identification of significant events in molecular dynamics simulations. *Journal of Chemical Information and Modeling, 56*(7), 1282–1291.

Henzler-Wildman, K. A., Thai, V., Lei, M., Ott, M., Wolf-Watz, M., Fenn, T., et al. (2007). Intrinsic motions along an enzymatic reaction trajectory. *Nature, 450*(7171), 838–844.

Hollingsworth, S. A., & Dror, R. O. (2018). Molecular dynamics simulation for all. *Neuron, 99*(6), 1129–1143.

Ishikawa, Y. (2013). A script for automated 3-dimentional structure generation and conformer search from 2- dimentional chemical drawing. *Bioinformation, 9*(19), 988–992.

Jo, S., Kim, T., Iyer, V. G., & Im, W. (2008). CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry, 29*(11), 1859–1865.

Joo, K., Lee, J., Lee, S., Seo, J.-H., Lee, S. J., & Lee, J. (2007). High accuracy template based modeling by global optimization. *Proteins: Structure, Function, and Bioinformatics, 69*, 83–89. https://doi.org/10.1002/prot.21628

Kalita, J., Shukla, R., Shukla, H., Gadhave, K., Giri, R., & Tripathi, T. (2017). Comprehensive analysis of the catalytic and structural properties of a mu-class glutathione s-transferase from Fasciola gigantica. *Scientific Reports, 7*(1), 17547.

Kalita, J., Shukla, R., & Tripathi, T. (2019). Structural basis of urea-induced unfolding of Fasciola gigantica glutathione S-transferase. *Journal of Cellular Physiology, 234*(4), 4491–4503.

Kastner, J. (2011). Umbrella sampling. *Wiley Interdisciplinary Reviews: Computational Molecular Science, 1*, 932–942. https://doi.org/10.1002/wcms.66

Kim, M. K., Jernigan, R. L., & Chirikjian, G. S. (2002). Efficient generation of feasible pathways for protein conformational transitions. *Biophysical Journal, 83*(3), 1620–1630.

Kim, P. M., Lu, L. J., Xia, Y., & Gerstein, M. B. (2006). Relating three-dimensional structures to protein networks provides evolutionary insights. *Science, 314*(5807), 1938–1941.

Klepeis, J. L., Lindorff-Larsen, K., Dror, R. O., & Shaw, D. E. (2009). Long-timescale molecular dynamics simulations of protein structure and function. *Current Opinion in Structural Biology, 19*(2), 120–127.

Koshland Jr., D. E., Nemethy, G., & Filmer, D. (1966). Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry, 5*(1), 365–385.

Koshland, D. E. (1958). Application of a theory of enzyme specificity to protein synthesis. *Proceedings of the National Academy of Sciences of the United States of America, 44*(2), 98–104.

Kruger, P., Verheyden, S., Declerck, P. J., & Engelborghs, Y. (2001). Extending the capabilities of targeted molecular dynamics: Simulation of a large conformational transition in plasminogen activator inhibitor 1. *Protein Science, 10*(4), 798–808.

Kulczycka-Mierzejewska, K., Sadlej, J., & Trylska, J. (2018). Molecular dynamics simulations suggest why the A2058G mutation in 23S RNA results in bacterial resistance against clindamycin. *Journal of Molecular Modeling, 24*(8), 191.

Kumar, A., Grupcev, V., Berrada, M., Fogarty, J. C., Tu, Y. C., Zhu, X., et al. (2014). DCMS: A data analytics and management system for molecular simulation. *Journal of Big Data, 2*(1), 9. (2196-1115 (Print)).

Laio, A., & Gervasio, F. L. (2008). Metadynamics: A method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics, 71*(12), 126601.

Lance, B. K., Deane, C. M., & Wood, G. R. (2010). Exploring the potential of template-based modelling. *Bioinformatics, 26*(15), 1849–1856.

Lane, T. J., Shukla, D., Beauchamp, K. A., & Pande, V. S. (2013). To milliseconds and beyond: Challenges in the simulation of protein folding. *Current Opinion in Structural Biology, 23*(1), 58–65.

Leherte, L., & Vercauteren, D. P. (2017). Reduced point charge models of proteins: Effect of protein-water interactions in molecular dynamics simulations of ubiquitin systems. *The Journal of Physical Chemistry. B, 121*(42), 9771–9784.

Lindorff-Larsen, K., Piana, S., Dror, R. O., & Shaw, D. E. (2011). How fast-folding proteins fold. *Science, 334*(6055), 517–520.

Linke, M., Kafinger, J., & Hummer, G. (2018). Rotational diffusion depends on box size in molecular dynamics simulations. *Journal of Physical Chemistry Letters, 9*(11), 2874–2878.

Loew, G. H., Villar, H. O., & Alkorta, I. (1993). Strategies for indirect computer-aided drug design. *Pharmaceutical Research, 10*(4), 475–486.

Man, V. H., Nguyen, P. H., & Derreumaux, P. (2017). High-resolution structures of the amyloid-beta 1-42 dimers from the comparison of four atomistic force fields. *The Journal of Physical Chemistry. B, 121*(24), 5977–5987.

Marinari, E., & Parisi, G. (1992). Simulated tempering: A new Monte Carlo scheme. *Europhysics Letters, 19*(6), 451–458.

Martn-Garca, F., Papaleo, E., Gomez-Puertas, P., Boomsma, W., & Lindorff-Larsen, K. (2015). Comparing molecular dynamics force fields in the essential subspace. *PLoS One, 10*(3), e0121114.

Mason, J. S., Good, A. C., & Martin, E. J. (2001). 3-D pharmacophores in drug discovery. *Current Pharmaceutical Design, 7*(7), 567–597.

McCammon, J. A., Gelin, B. R., & Karplus, M. (1977). Dynamics of folded proteins. *Nature, 267* (5612), 585–590.

McHugh, S. M., Yu, H., Slough, D. P., & Lin, Y.-S. (2017). Mapping the sequence-structure relationships of simple cyclic hexapeptides. *s, 19*(4), 3315–3324.

Meyer, T., D'Abramo, M., Hospital, A., Rueda, M., Ferrer-Costa, C., Pérez, A., et al. (2010). MoDEL (molecular dynamics extended library): A database of atomistic molecular dynamics trajectories. *Structure, 18*(11), 1399–1409. (1878-4186 (Electronic)).

Misura, K. M., & Baker, D. (2005). Progress and challenges in high-resolution refinement of protein structure models. *Proteins, 59*(1), 15–29.

Monod, J., Changeux, J. P., & Jacob, F. (1963). Allosteric proteins and cellular control systems. *Journal of Molecular Biology, 6*, 306–329.

Monod, J., Wyman, J., & Changeux, J. P. (1965). On the nature of allosteric transitions: A plausible model. *Journal of Molecular Biology, 12*, 88–118.

Moult, J. (2005). A decade of CASP: Progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structural Biology, 15*(3), 285–289.

Nabuurs, S. B., Wagener, M., & de Vlieg, J. (2007). A flexible approach to induced fit docking. *Journal of Medicinal Chemistry, 50*(26), 6507–6518.

Nguyen, P. H., Li, M. S., & Derreumaux, P. (2011). Effects of all-atom force fields on amyloid oligomerization: Replica exchange molecular dynamics simulations of the Abeta(16-22) dimer and trimer. *Physical Chemistry Chemical Physics, 13*(20), 9778–9788.

Nguyen, P. H., Okamoto, Y., & Derreumaux, P. (2013). Communication: Simulated tempering with fast on-the-fly weight determination. *The Journal of Chemical Physics, 138*(6), 061102.

Noble, D. (2003). Will genomics revolutionise pharmaceutical R&D? *Trends in Biotechnology, 21* (8), 333–337.

Orellana, L., Rueda, M., Ferrer-Costa, C., Lopez-Blanco, J. R., Chaca, P., & Orozco, M. (2010). Approaching elastic network models to molecular dynamics flexibility. *Journal of Chemical Theory and Computation, 6*(9), 2910–2923.

Pandey, T., Shukla, R., Shukla, H., Sonkar, A., Tripathi, T., & Singh, A. K. (2017). A combined biochemical and computational studies of the rho-class glutathione s-transferase sll1545 of Synechocystis PCC 6803. *International Journal of Biological Macromolecules, 94*(Pt A), 378–385.

Perdih, A., Kotnik, M., Hodoscek, M., & Solmajer, T. (2007). Targeted molecular dynamics simulation studies of binding and conformational changes in E. coli MurD. *First published., 68*. https://doi.org/10.1002/prot.21374

Piana, S., Lindorff-Larsen, K., & Shaw, D. E. (2012). Protein folding kinetics and thermodynamics from atomistic simulation. *Proceedings of the National Academy of Sciences of the United States of America, 109*(44), 17845–17850.

Piana, S., Lindorff-Larsen, K., & Shaw, D. E. (2013). Atomic-level description of ubiquitin folding. *Proceedings of the National Academy of Sciences of the United States of America, 110*(15), 5915–5920.

Pronk, S., Pall, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., et al. (2013). GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics, 29*(7), 845–854.

Rajendran, V., Shukla, R., Shukla, H., & Tripathi, T. (2018). Structure-function studies of the asparaginyl-tRNA synthetase from Fasciola gigantica: Understanding the role of catalytic and non-catalytic domains. *The Biochemical Journal, 475*(21), 3377–3391.

Raval, A., Piana, S., Eastwood, M. P., Dror, R. O., & Shaw, D. E. (2012). Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins, 80*(8), 2071–2079.

Razavi, A. M., Wuest, W. M., & Voelz, V. A. (2014). Computational screening and selection of cyclic peptide hairpin Mimetics by molecular simulation and kinetic network models. *Journal of Chemical Information and Modeling, 54*(5), 1425–1432.

Roy, A., Kucukural, A., & Zhang, Y. (2010). I-TASSER: A unified platform for automated protein structure and function prediction. *Nature Protocols, 5*(4), 725–738.

Sali, A., & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology, 234*(3), 779–815.

Schames, J. R., Henchman, R. H., Siegel, J. S., Sotriffer, C. A., Ni, H., & McCammon, J. A. (2004). Discovery of a novel binding trench in HIV integrase. *Journal of Medicinal Chemistry, 47*(8), 1879–1881.

Schlitter, J., Engels, M., Krãger, P., Jacoby, E., & Wollmer, A. (1993). Targeted molecular dynamics simulation of conformational change-application to the T ↔ R transition in insulin. *Mol Simulat, 10*(2–6), 291–308.

Sfriso, P., Emperador, A., Orellana, L., Hospital, A., GelpA, J. L., & Orozco, M. (2012). Finding conformational transition pathways from discrete molecular dynamics simulations. *Journal of Chemical Theory and Computation, 8*(11), 4707–4718.

Sfriso, P., Hospital, A., Emperador, A., & Orozco, M. (2013). Exploration of conformational transition pathways from coarse-grained simulations. *Bioinformatics, 29*(16), 1980–1986.

Shen, Y., Maupetit, J., Derreumaux, P., & Tuffery, P. (2014). Improved PEP-FOLD approach for peptide and Miniprotein structure prediction. *Journal of Chemical Theory and Computation, 10* (10), 4745–4758.

Shukla, H., Shukla, R., Sonkar, A., Pandey, T., & Tripathi, T. (2017a). Distant Phe345 mutation compromises the stability and activity of Mycobacterium tuberculosis isocitrate lyase by modulating its structural flexibility. *Scientific Reports, 7*(1), 1058.

Shukla, H., Shukla, R., Sonkar, A., & Tripathi, T. (2017b). Alterations in conformational topology and interaction dynamics caused by L418A mutation leads to activity loss of Mycobacterium tuberculosis isocitrate lyase. *Biochemical and Biophysical Research Communications, 490*(2), 276–282.

Shukla, R., Chetri, P. B., Sonkar, A., Pakharukova, M. Y., Mordvinov, V. A., & Tripathi, T. (2018a). Identification of novel natural inhibitors of Opisthorchis felineus cytochrome P450 using structure-based screening and molecular dynamic simulation. *Journal of Biomolecular Structure & Dynamics, 36*(13), 3541–3556.

Shukla, R., Shukla, H., Kalita, P., Sonkar, A., Pandey, T., Singh, D. B., et al. (2018b). Identification of potential inhibitors of Fasciola gigantica thioredoxin1: Computational screening, molecular dynamics simulation, and binding free energy studies. *Journal of Biomolecular Structure & Dynamics, 36*(8), 2147–2162.

Shukla, R., Shukla, H., Kalita, P., & Tripathi, T. (2018c). Structural insights into natural compounds as inhibitors of Fasciola gigantica thioredoxin glutathione reductase. *Journal of Cellular Biochemistry, 119*(4), 3067–3080.

Shukla, R., Shukla, H., Sonkar, A., Pandey, T., & Tripathi, T. (2018d). Structure-based screening and molecular dynamics simulations offer novel natural compounds as potential inhibitors of Mycobacterium tuberculosis isocitrate lyase. *Journal of Biomolecular Structure & Dynamics, 36*(8), 2045–2057.

Shukla, R., Shukla, H., & Tripathi, T. (2018e). Activity loss by H46A mutation in Mycobacterium tuberculosis isocitrate lyase is due to decrease in structural plasticity and collective motions of the active site. *Tuberculosis (Edinburgh, Scotland), 108*, 143–150.

Shukla, R., & Tripathi, T. (2020). Molecular dynamics simulation of protein and protein-ligand complexes. In D. B. Singh (Ed.), *Computer-Aided Drug Design* (pp. 133–161). Singapore: Springer. https://doi.org/10.1007/978-981-15-6815-2_7. ISBN 978-981-15-6814-5.

Slough, D. P., Yu, H., McHugh, S. M., & Lin, Y. S. (2017). Toward accurately modeling N-methylated cyclic peptides. *Physical Chemistry Chemical Physics, 19*(7), 5377–5388.

Sonkar, A., Shukla, H., Shukla, R., Kalita, J., & Tripathi, T. (2019). Unfolding of Acinetobacter baumannii MurA proceeds through a metastable intermediate: A combined spectroscopic and computational investigation. *International Journal of Biological Macromolecules, 126*, 941–951.

Stroet, M., Caron, B., Visscher, K. M., Geerke, D. P., Malde, A. K., & Mark, A. E. (2018). Automated topology builder version 3.0: Prediction of solvation free enthalpies in water and hexane. *Journal of Chemical Theory and Computation, 14*(11), 5834–5845.

Sugita, Y., & Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters, 314*(1), 141–151.

Talele, T. T., Khedkar, S. A., & Rigby, A. C. (2010). Successful applications of computer aided drug discovery: Moving drugs from concept to the clinic. *Current Topics in Medicinal Chemistry, 10*(1), 127–141.

Thibault, J. C., Facelli, J. C., & Cheatham III, T. E. (2013). iBIOMES: Managing and sharing biomolecular simulation data in a distributed environment. *Journal of Chemical Information and Modeling, 53*(3), 726–736. (1549-960X (Electronic)).

Todd, M. H. (2005). Computer-aided organic synthesis. *Chemical Society Reviews, 34*(3), 247–266.

van Aalten, D. M. F., Bywater, R., Findlay, J. B. C., Hendlich, M., Hooft, R. W. W., & Vriend, G. (1996). PRODRG, a program for generating molecular topologies and unique molecular descriptors from coordinates of small molecules. *Journal of Computer-Aided Molecular Design, 10*(3), 255–262.

Verma, J., Khedkar, V. M., & Coutinho, E. C. (2010). 3D-QSAR in drug design--a review. *Current Topics in Medicinal Chemistry, 10*(1), 95–115.

Verma, R. P., & Hansch, C. (2009). Camptothecins: A SAR/QSAR study. *Chemical Reviews, 109* (1), 213–235.

Wang, Y., Shaikh, S. A., & Tajkhorshid, E. (2010). Exploring Transmembrane diffusion pathways with molecular dynamics. *Physiology, 25*(3), 142–154.

Wouters, O. J., McKee, M., & Luyten, J. (2020). Estimated Research and Development investment needed to bring a new medicine to market, 2009-2018. *JAMA, 323*(9), 844–853. https://doi.org/10.1001/jama.2020.1166

Yang, S. Y. (2010). Pharmacophore modeling and applications in drug discovery: Challenges and recent advances. *Drug Discovery Today, 15*(11–12), 444–450.

Zhang, C., & Ma, J. (2008). Comparison of sampling efficiency between simulated tempering and replica exchange. *The Journal of Chemical Physics, 129*(13), 134112.

Zhang, J., Xu, F., Hong, Y., Xiong, Q., & Pan, J. (2015a). A comprehensive review on the molecular dynamics simulation of the novel thermal properties of graphene. *RSC Advances, 5* (109), 89415–89426.

Zhang, J., Xu, F., Hong, Y., Xiong, Q., & Pan, J. (2015b). A comprehensive review on the molecular dynamics simulation of the novel thermal properties of graphene. *RSC Advances, 5* (109), 89415–89426.

Zhang, T., Nguyen, P. H., Nasica-Labouze, J., Mu, Y., & Derreumaux, P. (2015c). Folding atomistic proteins in explicit solvent using simulated tempering. *The Journal of Physical Chemistry. B, 119*(23), 6941–6951.

Zhang, Y. (2009). Protein structure prediction: When is it useful? *Current Opinion in Structural Biology, 19*(2), 145–155.

Chapter 13
Molecular Dynamic Simulation of Intrinsically Disordered Proteins and Relevant Forcefields

Prateek Kumar, Nitin Sharma, Amit Kumar, and Rajanish Giri

Abstract Intrinsically disordered proteins (IDPs) exist in every form of life, from bacteria to humans and viruses. They do not hold any well-defined or properly folded structure in the physiological state but have the ability to gain different structural conformations upon interactions with their physiological partners. The limitations of experimental techniques to study high structural dynamics of IDPs have led us to depend on computational simulations. The current scenario of interdisciplinary studies to understand biology with physics has been advantageous for exploring atomic-level dynamics of IDPs. To date, several physics-based forcefields have been developed that calculates the microscopic parameters of a biomacromolecule in an aqueous environment. In this chapter, we have discussed the conformational behavior of IDPs and induced structural properties through understanding the relevant forcefields for molecular dynamics simulations.

13.1 Introduction

By the end of twentieth-century and start of twenty-first century, the understanding of disordered or unstructured proteins started developing. At present, a large number of researchers from every corner of world have devoted their research to describe the proper structure and functioning of disordered regions. A large proportion of gene sequences appear to code not only for folded, globular proteins but also for long stretches of amino acids that are likely to be either unfolded in solution or adopt non-globular structures of unknown conformation (Wright and Dyson 1999). Approximately 44% of genes in humans that code for proteins contain disordered regions (Van Der Lee et al. 2014; Oates et al. 2013). Generally, these proteins or regions are termed as intrinsically disordered proteins or regions (IDPs/IDPRs).

P. Kumar · N. Sharma · A. Kumar · R. Giri (✉)
School of Basic Sciences, Indian Institute of Technology Mandi, Kamand, Himachal Pradesh, India
e-mail: rajanishgiri@iitmandi.ac.in

S. K. Singh (ed.), *Innovations and Implementations of Computer Aided Drug Discovery Strategies in Rational Drug Design*,
https://doi.org/10.1007/978-981-15-8936-2_13

These intrinsically disordered protein regions (IDPRs) can be highly conserved within various closely related families or domains of proteins in both composition and sequence (Van Der Lee et al. 2014; Chen et al. 2006). The disordered regions are partially or fully unstructured and are characterized based on various parameters. In other words, they do not possess a proper three-dimensional structure as they fail to acquire structural propensity measured through spectroscopy techniques such as X-ray, NMR, etc. (Dunker et al. 2001). The intrinsic lack of structure can confer functional advantages to a protein like IDPs provide larger interaction surface area, more conformational flexibility, and exposure to interaction prone structural motifs allows IDPs to interact with several other proteins (Babu et al. 2011).

Furthermore, distinct post-translational modifications alleviate regulation of their function and stability in a cell. Some IDPs can attain a fixed tertiary structure on interaction with other molecules known as folders. In contrast, other are called non-folders which do not possess any defined tertiary structure under any physiological conditions. They have ability to undergo partial folding on interaction with specific binding partner proteins (coupled folding and binding), whereas many others constitute flexible linkers that have a role in the assembly of macromolecular arrays (Nishimura et al. 2005). Their conformations may vary from random coils, partially extended globules to collapsed globules with different contents of secondary structure. These distinctly variable structural behaviors of IDPs led to propose multi-state protein structure theories such as trinity (collapsed, ordered, and extended disorder) and quartet (coil, pre-molten globule, molten globule, and folded structure) (Zhang et al. 2013; Dunker and Obradovic 2001).

To elucidate the structure of IDPs and detailed mechanistic insight into their function, firstly, IDPs differential conformations need to be determined. The molecular dynamics (MD) simulation is an excellent computational route for determination of proteins disordered states at atomic level. However, the peculiarities of MD simulation results depend on the accuracy of the physical model (i.e. forcefield) used (Robustelli et al. 2018). There are a number of force fields have been used for the description of folded proteins, but limited for disordered structure prediction (Nerenberg et al. 2012; Piana et al. 2015; Best et al. 2014; Mittal and Best 2010; Lindorff-Larsen et al. 2012, 2013; Beauchamp et al. 2012; Lange et al. 2010). Therefore, in this chapter, we are focused on the IDPs and how the computational method MD simulation exploring the structure disorder via different force fields.

13.2 IDPs and IDPRs: Structure-Function Relationship

The universal lock and key hypothesis for structure function paradigm changed the protein science for a longer time. The proteins 3-D structures were mapped mostly with X-ray crystallography. Despite that, most of the proteins lack complete structures and so-called missing electron density regions (Le Gall et al. 2007). These proteins and regions are unfolded, unstructured and inherent properties of proteins, hence named “intrinsically disordered proteins or intrinsically disordered protein

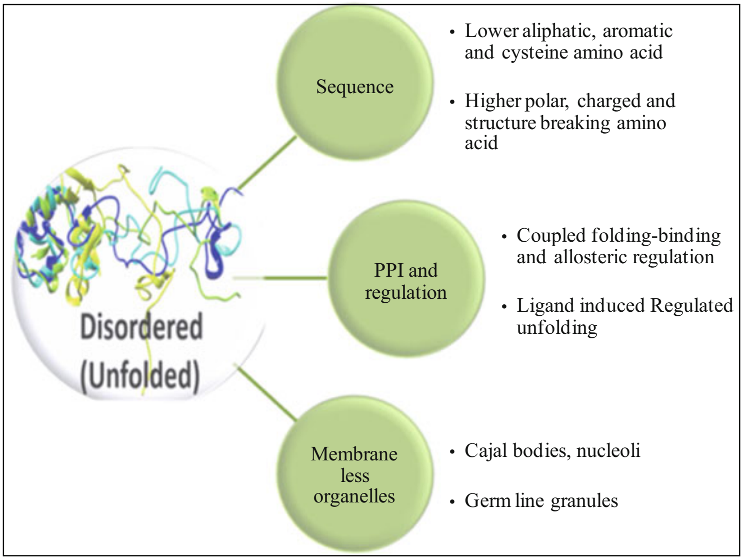

Sequence
• Lower aliphatic, aromatic and cysteine amino acid
• Higher polar, charged and structure breaking amino acid
Disordered (Unfolded)
PPI and regulation
• Coupled folding-binding and allosteric regulation
• Ligand induced Regulated unfolding
Membrane less organelles
• Cajal bodies, nucleoli
• Germ line granules

Fig. 13.1 The versatility of intrinsically disordered proteins

regions (IDPs/IDPRs)" (Dunker et al. 2013). The comparison studies of ordered and disordered proteins backbone revealed that the disorder proteins are rich in amino acid Ala, Arg, Gly, Gln, Ser, Glu, Lys, and Pro (Williams et al. 2000; Romero et al. 2001).

As it is stated that the ordered protein follows the structure-function paradigm, i.e., sequence-structure-function, whereas disordered protein follows the disorder-function paradigm (Uversky 2013). The disordered proteins are abundant in all three kingdom of life and viruses, which speculate their importance (Gadhave et al. 2020; Garg et al. 2019; Giri et al. 2016, 2020; Singh et al. 2018a; Kumar et al. 2017, 2019, 2020a; Schad et al. 2011). The very interesting properties of these IDPs are their versatility of performing functions, which can be explained by the "fly casting mechanism" (Huang and Liu 2009; Shoemaker et al. 2000) (Fig. 13.1). Moreover, IDPs can perform function either in native disorder state or can bind to a partner to acquire folding state (Tompa 2005; Uversky and Dunker 2010, 2013; Tompa and Fuxreiter 2008; Dunker et al. 2002). This functional diversity of IDPs lies in its sequence heterogeneity, which allows it to bind with different partners and thus, different conformation and functions (Oldfield et al. 2008). Further, the IDPs/IDPRs possess larger surface area and structural flexibility. Due to the structural flexibility of IDPs they tend to expose peptide regions having molecular recognition features (MoRFs), which may fold while interacting with binding partners (Kumar et al. 2017, 2020a; Mohan et al. 2006; Oldfield et al. 2005; Uversky et al. 2005; Singh et al. 2018b; Mishra et al. 2018). A classic example of this IDPs-binding partner gaining multiple conformations can be illustrated with p53 C-terminal domain (p53-CTD). The p53 CTD (residue 374–388) bound to different partners and acquire different conformations viz., cyclin A (coil), sirtuin (sheet), CBP (coil), and S100bb (helix) (Uversky 2009; Fadda and Nixon 2017; Kannan et al. 2016).

Besides the potential of IDPs/IDPRs of their multiple functions depending upon the binding partners, surrounding environment need to be considered. There are ample of IDPs which showed a change in conformation in the presence of varying pH, temperature, ions, detergent, organic solvent, crowding agents, and lipids (Kumar et al. 2020a; Uversky 2009; Lopes et al. 2013; Kjaergaard et al. 2010). The surrounding environment imparts electrostatic interaction, hydrophobic interaction, and osmophobic effect, which help IDPs to gain structural conformation (Uversky 2009).

13.3 IDPs in the Human Genome: Organizing Functions or Problems?

Despite being physiologically disordered, IDPs play crucial roles in biological activities. The abundance of IDPs in complex cellular organization displays its importance in regulatory processes (Uversky et al. 2008). These processes include molecular recognition, molecular assembly, entropic activities, and post-translational modifications. Various studies have reported the presence of IDPs in human regulatory proteins such as transcription factors and co-regulators. Eukaryotic proteins seem to use disorder for transient binding purposes (signaling and regulation), while prokaryotic proteins seem to use disorder for longer-lasting interactions, such as complex formation. A recent report suggested that functional misfolding can be induced by fugacious changes in protein environment, and structure can be reversed by restoring the environment or modifications. These induced nature and fugacious characters are important features of these IDPRs or conditionally disordered protein regions (Uversky 2015). Some interesting studies about the occurrence of IDPs in viruses have shown that it plays crucial roles in hijack of host cellular functional machinery (Gadhave et al. 2020; Garg et al. 2019; Giri et al. 2016, 2020; Xue et al. 2010; Kumar et al. 2020b).

The versatile nature of IDPs associated with folding, signaling, and many more, however, they are also implicated in many diseases. It is seen that selective mutations in IDPRs (i.e., amyloid β-peptide, α-synuclein, and huntingtin) may lead to structural complexity and enhanced aggregation propensity of these systems, which are associated with numerous neurodegenerative diseases (Babu et al. 2011; Uversky et al. 2008; Wu and Fuxreiter 2016). The IDPRs contain certain motif which are important for interaction and a slight change in these motif lead to altered cell signaling and thus cancer like diseases (Babu 2016; Hegyi et al. 2009; Colak et al. 2013).

According to the studies by disorder predictors, eukaryotic mammals are shown to contain nearly 75% of signaling proteins that contain long disordered regions with more than 30 residues, and about 25% of the predicted proteins are fully disordered in nature (Dunker et al. 2008a). Furthermore, eukaryotic proteins utilize the disorder for transient binding purposes (signaling and regulation), while prokaryotic proteins seem to use disorder for complex formation (Dunker et al. 2008b). Another example of varied length intrinsically disordered proteins are Transmembrane proteins that contain extracellular or cytosolic disorder regions (Uversky 2013; Xue et al. 2009). A total of 40% of human integral plasma proteins is predicted to contain long stretched disordered regions (Minezaki et al. 2007; Yang et al. 2008; De Biasio et al. 2008). Disordered regions usually bind to multiple targets with low affinity, which is an ideal condition for signal transduction (Dunker et al. 2002). Some recent findings mention the functioning of ordered proteins on a decrease in the percentage of their ordered structure and need partial or complete functional misfolding (Uversky 2015).

13.4 Characterization of IDP and IDPRs

In the past two decades, rapid progress in the exploration of IDPs have radically changed the understanding and importance of the field. The high occurrence of IDPs in a cellular organization has increased the demand for new perspectives in structural and functional studies. The conformational flexibility of IDPs did not allow it to accurately study with old structural techniques. Therefore, it appeals to introduce new methods to study functional aspects in IDPs (Habchi et al. 2014). The structural studies of IDPs have shed light on the critical aspect that disorder lies in the amino-acid sequence of a protein. The thorough studies of IDP sequences and structural information suggest that disordered regions show low hydrophobicity and higher net charge, and characterized by low hydrodynamic radius, high structural heterogeneity, and poor secondary structure organization (Uversky 2019). But it may have a tendency to gain structural regions in presence of natural ligands. On the basis of these structural and sequence-based data, various algorithms have been designed to predict the structural disorder propensity of protein regions using disorder predictors. These bioinformatics disorder prediction tools are commonly used to characterize the protein disorderedness. The higher proportions of hydrophilic stretch of sequences are analyzed by a web server such as DISPROT (Megan Sickmeier et al. 2007), IUPred (Zsuzsanna Dosztányi et al. 2005), PONDR (Obradovic et al. 2003), PrDOS (Ishida and Kinoshita 2007), D^2P^2 (Oates et al. 2013) and ESpritz (Walsh et al. 2012), which indicates the higher probability of disorder of those regions. Some of the test sets for structural predictions have been further confirmed by various experimental tools such as NMR, X-ray studies (Konrat 2014; Brutscher et al. 2015). This represents the high reliability of disorder predictors and improves the knowledge of the functional relevance of IDPs and IDPRs in various organisms.

13.5 Molecular Dynamics Simulations: Relevance with Structure Biology

The three-dimensional (3D) structure of biological macromolecules (e.g., proteins) or chemically synthesized polymers are essential for structural biology and applications in drug discovery. These days, the structure elucidation through X-ray, NMR, Cryo-EM techniques have been an advantage to understand their Structure-Function-Paradigm. However, there are many proteins which can not form rigid three-dimensional structures. So their thermodynamic properties, microscopic energies, and specific interaction with other molecules at the atomic-level cannot be understood well through experimental methods (Chong et al. 2017). Therefore, the characterization of proteins at the atomic level is more feasible through atomistic computational simulations rather than experiments. Molecular dynamics (MD) simulations are capable of determining conformational dynamics, structure compositions, and organization of proteins in an aqueous environment (Hollingsworth and Dror 2018). Additionally, the interaction of proteins with lipid molecules, inhibitors, with partner proteins, etc. can also be determined through MD simulations. Due to advancements in computer hardware, it is now possible to explore such macromolecules in a deeper level for longer timescale up to seconds to meet the experimental observations (Perilla et al. 2015). Various simulation packages such as Desmond (Bowers et al. 2006), Gromacs (Berendsen et al. 1995), Amber (Ponder and Case 2003), NAMD (Phillips et al. 2005), etc. are available with different optimized forcefields. Generally, a Forcefield (FF) can be explained as the interatomic potential energy of a system, which is calculated along with several parameters such as bonds, angle, torsion, dihedral, etc., defined on the atomic coordinates (Jorgensen and Tirado-Rives 2005; González 2011). The atoms which are held together by simple harmonic or elastic forces represent a molecule within the specified region for simulation (González 2011). Also, van der waal interactions and electrostatic interactions are the integral constituents of a forcefield. The overall equation that defines a forcefield is,

$$\begin{aligned} U(potential\ energy) = & \sum_{bonds} \frac{1}{2} k_b (r - r_0)^2 + \sum_{angles} \frac{1}{2} k_a (\theta - \theta_0)^2 + \sum_{torsions} \frac{V_n}{2} \\ & \times [1 + \cos(n\varnothing - \delta)] + \sum_{improper} V_{imp} + \sum_{LJ} 4\epsilon_{ij} \left(\frac{\sigma_{ij}^{12}}{r_{ij}^{12}} - \frac{\sigma_{ij}^{6}}{r_{ij}^{6}} \right) \\ & + \sum_{electrostatic} \frac{q_i q_j}{r_{ij}} \end{aligned}$$

Where, oscillations about equilibrium bond length, bond angle, torsional rotation of 4 atoms about a central bond, and nonbonded energy terms (electrostatics and Lenard-Jones (LJ)) are summed up for calculating potential energy.

13.6 MD Force Fields and Their Role in Conformation Dynamics

Several forcefields have been developed to date and are being used with different purposes for investigation in almost every field of science. For biological macromolecules, OPLS (AA) (Jorgensen and Tirado-Rives 1988), GROMOS (Berendsen et al. 1995), CHARMM (Vanommeslaeghe et al. 2009), AMBER (Ponder and Case 2003), Drude (Li et al. 2017) forcefields are optimized to deduce the conformational change, structural composition, protein aggregation, binding efficacy with respect to time in a given environment. All these forcefields have a different level of tendency to estimate structural composition. However, for disordered proteins, it is essential to be picky for selection of accurate forcefields. A wonderful comparison has been made by Ham and colleagues between GROMOS, CHARMM, AMBER, and OPLS forcefields for correct selection of forcefield to perform MD of an IDP. Among all of them, OPLS-AA has proper balanced tendency to evaluate the helical and beta property of protein (Chong et al. 2017). Also, for IDPs, OPLS, and a recently introduced CHARMM36 (Huang et al. 2016) are used to simulate disordered regions properly, which allow them to gain a proper helical or beta structure, if induces. Two IDP models amyloid-beta and p53, have been extensively used as model systems for testing of different forcefields and correlating them with experiments. Pacheco and Strodel have investigated the accuracy of five forcefields on amyloid-beta (Aβ; one of the responsible protein for Alzheimer's Disease) where CHARMM22, OPLS, Amber's 99sb, 99sb-ildn, had high accuracy with the NMR results than 99sbildn-NMR forcefield (Carballo-Pacheco and Strodel 2017). Recently, D.E. Shaw and colleagues have modified Amber-ff99sb forcefield for disordered proteins which correlate well with experimental observations. The improved forcefield ff99sb-disp accurately calculates the transition states between ordered and disordered states (Robustelli et al. 2018). Additionally, a short disordered peptide of 24 amino acids, Histatin 5, was investigated through verities of Amber and Gromos forcefields and compared with experiments (Henriques et al. 2015).

Along with a forcefield, the selection of water model for simulation is essential for precise evaluation of interaction and behavior of a protein in an aqueous environment. Water models are defined based on the interaction sites, which is centered on the nuclei of water molecule. Most commonly, water models are SPC (simple point charge; with HOH angle 109.47°), TIP3P (104.5°), and TIP4P (104.5°) for protein simulations. TIP3P and TIP4P water models are basically based on transferrable interaction potentials (TIPS) with three- and four-point charge (Ouyang and Bettens 2015). Three-point charge has three interaction sites as water molecule has three atoms, while four-point charge has an additional dummy atom to improve the electrostatic distribution (Jorgensen and Tirado-Rives 2005). Moreover, five-point and six-point charge water models are also available, which has dummy atoms representing the lone pairs and one extra site for interaction (Fig. 13.2). These water models are placed around the protein structure, which is centered in a

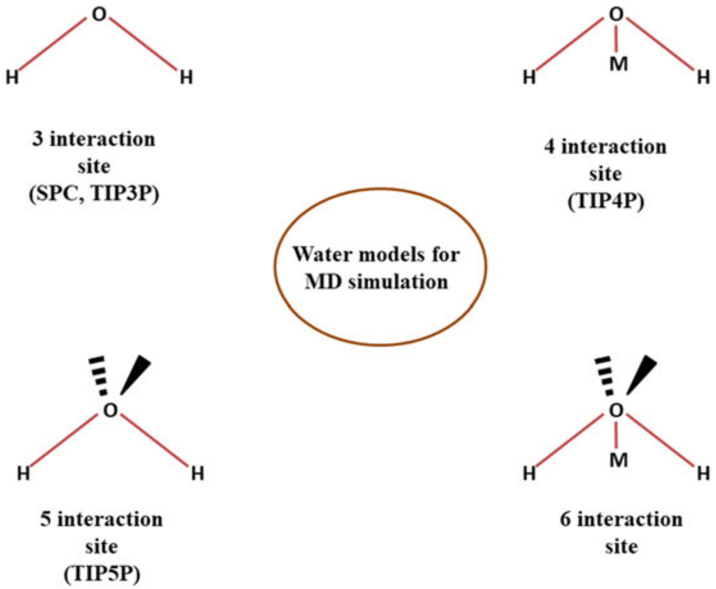

O
H
H
3 interaction site (SPC, TIP3P)
O
M
H
H
4 interaction site (TIP4P)
Water models for MD simulation
O
H
H
5 interaction site (TIP5P)
O
M
H
H
6 interaction site

Fig. 13.2 Representation of water models based on different interaction sites. Here, O and H are oxygen and Hydrogen atoms while M represents the dummy atom in water models with 4, 5, and 6 interactions sites

simulation box (e.g. cubic, dodecahedron, etc.) with a defined size in the periodic cell.

13.7 MD Simulation Terminology: Structural Conformation Assessment

13.7.1 Energy Minimization, Equilibration, and Timescale

Building a simulation system to final production run, there are a number of steps for successfully trajectory generation. Before simulating a protein, the minimized structural conformation is vital as the system might produce erroneous results due to excess heat caused by unwanted and huge forces without minimization (Mackay et al. 1989). For simulating IDPs, a very popular method, steepest descent method is used for certain steps till it converged under required energy for the system.

Afterward, the simulation setup is subjected to equilibration as the minimized system has unoptimized solvents (e.g., water) around the proteins. Generally, the system is equilibrated under two conditions in constant temperature and pressure because for simulating a protein at a temperature, the system needs to be in proper positions without any unrestrained atoms. Two ensembles NPT and NVT are

Table 13.1 Tabulation of the timescale of different simulation techniques based on their applications

Sr. no.	Simulation methods	Applications	Timescale (seconds)
1.	Quantum mechanics (QM/MD)	Atomic motion (e.g., bond stretching, vibration)	10^{-15} to 10^{-12}
2.	All-atom MD	Structural transitions, side-chain rotation, loop movement, ligand binding	10^{-12} to 10^{-6}
3.	Coarse-grained MD	Biological assembly, protein folding, protein interaction	10^{-9} to 10^{3}

commonly used for equilibration in which the number of atoms (N), pressure (P), temperature (T), and volume (V) are kept constant and the system is processed for small simulation run upto few picoseconds or nanoseconds or till the system is equilibrated.

Finally, the production run for all-atom classical MD is performed at a constant temperature and pressure, which are maintained by specific thermostats and barostats, respectively. Thermostats like Nose Hoover, Berendsen, and Langevin while barostats such as Berendsen (Berendsen et al. 1984), Martyna-Tobias-Klein (MTK) (Lippert et al. 2013), Parrinello Rahman (Parrinello and Rahman 1980, 1981) are used as per the simulation setup requirements. In the case of IDPs, Nose Hoover (Posch et al. 1986) and Berendsen thermostats are preferred, and barostats MTK or Parrinello Rahman are considered for controlling temperature and pressure at an average value.

The simulation time is another important aspect to be looked carefully. As shown in above Table 13.1, the purpose of performing any simulation should be accomplished successfully within an adequate range of time. This timescale also depends on the number of atoms in a simulation setup and computer hardware. More number of atoms require more time and vice-versa, therefore, all atoms MD demands time from nanoseconds to few microseconds. However, apart from all-atoms MD simulation, Coarse-grain (CG) MD is one of the examples where a group of atoms are considered as a single bead, which reduces the degree of freedom among the atoms and allow the system for longer simulation in comparatively less time (Kmiecik et al. 2016).

13.7.2 RMSD, RMSF and Radius of Gyration

For the conformational analysis from the simulation trajectory, Root mean square deviation (RMSD) is the most critical parameter to be investigated. RMSD is measured between two sets of atoms (backbone, c-alpha, heavy, side-chain atoms) at a given time with reference to the initial one or the desired reference set of atoms (Kuzmanic and Zagrovic 2010). Similarly, the fluctuation in residues with respect to time can also be calculated as average of all simulated frames in the trajectory, it is known as Root mean square fluctuation (Martínez 2015). Both these values deduce the conformational change of protein in the simulation environment over the time course. Another parameter, Radius of gyration (RoG) states the compactness of a protein structure. A well-folded or tightly packed structure will have lowest RoG value and vice-versa.

Many IDPs, which tend to get converted into an ordered form upon interaction with some physiological partners, are also vulnerable to different forcefields. In last two decades, the computer hardware power has been increased gradually and consequently, MD simulations have been very effective to explore the folding and unfolding of a protein in the presence of different conditions such as mixture of multiple solvents, membranes, ions, etc. at a constant or varying temperature and pH ranges. Also, various categories of simulation have made it easier to investigate the structural dynamics for longer timescale. A method where conformational swapping occurs between a number of MD replicas to obtain a conformation with minimum energy. These are known as Replica Exchange (RE) MD in which multiple replicas run simultaneously which are formed based on temperatures (lower to higher), required after selection through literature or experimental evidences (Sugita and Okamoto 1999). In next section, we have discussed the REMD with a well-suited example of p53, a tumor suppressor gene.

13.8 IDPs and Replica Exchange MD: In Perspective of p53-CTD

In our recent study, we performed a Replica exchange molecular dynamics simulation on p53-CTD using OPLS 2005 forcefield, embedded in Desmond simulation package (Bowers et al. 2006). As aforementioned, OPLS forcefield has a proper balance for alpha and beta propensity estimation in simulation. As illustrated in this study, the temperature induces changes in structural conformation of IDP (p53-CTD), which showed it's highly dynamics/flexible nature (Kumar et al. 2020a). The hydrophobic and electrostatic interactions play an important role in structural conformation of p53-CTD. The circular dichroism studies showed that the higher temperature leads to the compaction of a peptide, which is associated with the helical structural conformations (Kumar et al. 2020a; Kjaergaard et al. 2010). Previously, NMR studies showed that the temperature-induced structural conformation is associated with the random stretch of amino acid (non-helical) in a peptide (Kjaergaard et al. 2010). Our result also corroborated with theses finding where MD simulation showed that the random stretch of amino acid or non-helical regions is responsible for change in structural transformations.

The p53-CTD adopts random coil conformations and have a tendency to gain structural conformation. According to REMD analysis, the highest structural compaction occurred at 80 °C where two major helical regions were formed (Fig. 13.3). The total potential energy of p53-CTD in aqueous system was calculated to be more

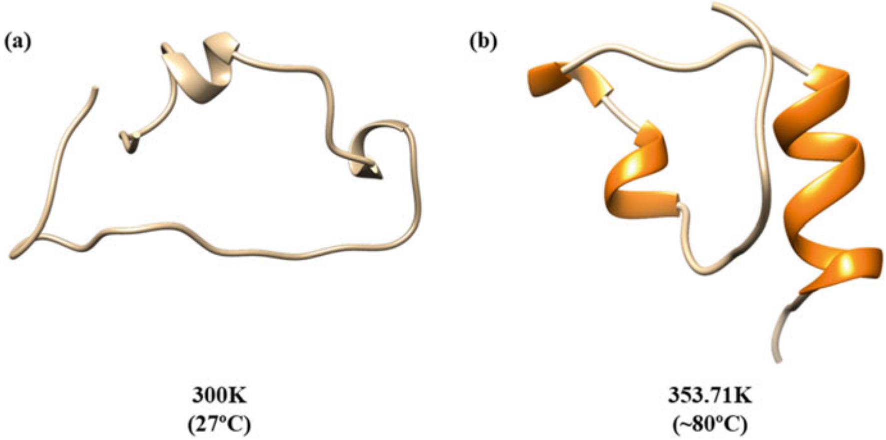


Fig. 13.3 Induction of helicity at two regions in p53-CTD at high temperature as obtained from Replica Exchange MD, using OPLS 2005 forcefield

negative (−33914.25 kcal/mol) at high temperature (353.71 K) while at 300 K, the potential energy of the system was −30382.4 kcal/mol. The p53 CTD could not fold spontaneously due to high intrinsic free energy. In the presence of binding partners, the p53-CTD possesses the minimum free energy and minimum binding energy from partner. Thus, the competitive effect of energy minimization results in different characteristic states, where each mechanism determines the sampling frequency of each characteristic state (Han et al. 2017).

13.9 Future Prospects

In general, IDPs are very challenging to study at atomistic detail with current computational simulation forcefields to achieve its accuracy with experimental models. A great advancement has been made to counter these challenges, but still, a lot of improvement needs to be done. However, Bioinformatics has made it possible to a large extent, and specifically, molecular dynamics simulations are being used extensively to explore IDPs at atomic level. The exact evaluation of structural properties of IDPs/IDPRs have been very distinguished among different forcefield. In this context, it can be seen as a high opportunity for development of new FF or improvement of current FF, which unravel the atomistic details on IDPs' conformational dynamics with respect to experimental measures. The outcome will certainly lead to a better understanding of biophysics of IDPs and pave a potential role in drug discovery.

Acknowledgements All the authors would like to thank IIT Mandi for providing facilities. The work was partially supported from DBT, Government of India (BT/11/IYBA/2018/06) to

RG. Funding. PK and NS were supported by MHRD for funding. AK was supported by DBT, Government of India (BT/11/IYBA/2018/06).

Conflict of Interest All authors declare that there is no financial competing interest.

References

Babu, M. M. (2016). The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochemical Society Transactions, 44*, 1185–1200. https://doi.org/10.1042/BST20160172

Babu, M. M., van der Lee, R., de Groot, N. S., & Gsponer, J. (2011). Intrinsically disordered proteins: Regulation and disease. *Current Opinion in Structural Biology, 21*, 432–440. https://doi.org/10.1016/j.sbi.2011.03.011

Beauchamp, K. A., Lin, Y. S., Das, R., & Pande, V. S. (2012). Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements. *Journal of Chemical Theory and Computation, 8*, 1409–1414. https://doi.org/10.1021/ct2007814

Berendsen, H. J. C., Postma, J. P. M., Van Gunsteren, W. F., Dinola, A., & Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics, 81*, 3684–3690. https://doi.org/10.1063/1.448118

Berendsen, H. J. C., van der Spoel, D., & van Drunen, R. (1995). GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications, 91*, 43–56. https://doi.org/10.1016/0010-4655(95)00042-E

Best, R. B., Zheng, W., & Mittal, J. (2014). Balanced protein-water interactions improve properties of disordered proteins and non-specific protein association. *Journal of Chemical Theory and Computation, 10*, 5113–5124. https://doi.org/10.1021/ct500569b

Bowers, K. J., Bowers, K. J., Chow, E., Xu, H., Dror, R. O., Eastwood, M. P., et al. (2006). Scalable algorithms for molecular dynamics simulations on commodity clusters. In *Proceedings of the 2006 ACM/IEEE conference on supercomputing* (pp. 43–43). https://doi.org/10.1109/SC.2006.54

Brutscher, B., Felli, I. C., Gil-Caballero, S., Hošek, T., Kümmerle, R., Piai, A., et al. (2015). NMR methods for the study of instrinsically disordered proteins structure, dynamics, and interactions: General overview and practical guidelines. *Advances in Experimental Medicine and Biology, 870*, 49–122. https://doi.org/10.1007/978-3-319-20164-1_3

Carballo-Pacheco, M., & Strodel, B. (2017). Comparison of force fields for Alzheimer's A β42: A case study for intrinsically disordered proteins. *Protein Science, 26*, 174–185. https://doi.org/10.1002/pro.3064

Chen, J. W., Romero, P., Uversky, V. N., & Dunker, A. K. (2006). Conservation of intrinsic disorder in protein domains and families: II. Functions of conserved disorder. *Journal of Proteome Research, 5*, 888–898. https://doi.org/10.1021/pr060049p

Chong, S.-H., Chatterjee, P., & Ham, S. (2017). Computer simulations of intrinsically disordered proteins. *Annual Review of Physical Chemistry, 68*, 117–134. https://doi.org/10.1146/annurev-physchem-052516-050843

Colak, R., Kim, T., Michaut, M., Sun, M., Irimia, M., Bellay, J., et al. (2013). Distinct types of disorder in the human proteome: Functional implications for alternative splicing. *PLoS Computational Biology, 9*, e1003030. https://doi.org/10.1371/journal.pcbi.1003030

De Biasio, A., Guarnaccia, C., Popovic, M., Uversky, V. N., Pintar, A., & Pongor, S. (2008). Prevalence of intrinsic disorder in the intracellular region of human single-pass type I proteins: The case of the notch ligand delta-4. *Journal of Proteome Research, 7*, 2496–2506. https://doi.org/10.1021/pr800063u

Dunker, A. K., & Obradovic, Z. (2001). The protein trinity – Linking function and disorder. *Nature Biotechnology, 19*, 805–806. https://doi.org/10.1038/nbt0901-805

Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., et al. (2001). Intrinsically disordered protein. *Journal of Molecular Graphics & Modelling, 19*, 26–59. https://doi.org/10.1016/S1093-3263(00)00138-8.

Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., & Obradović, Z. (2002). Intrinsic disorder and protein fluctuations. *Biochemistry, 41*, 6573–6582. https://doi.org/10.1021/BI012159+

Dunker, A. K., Silman, I., Uversky, V. N., & Sussman, J. L. (2008a). Function and structure of inherently disordered proteins. *Current Opinion in Structural Biology, 18*, 756–764. https://doi.org/10.1016/j.sbi.2008.10.002

Dunker, A. K., Oldfield, C. J., Meng, J., Romero, P., Yang, J. Y., Chen, J. W., et al. (2008b). The unfoldomics decade: An update on intrinsically disordered proteins. *BMC Genomics, 9*, 1–26. https://doi.org/10.1186/1471-2164-9-S2-S1

Dunker, A. K., Babu, M. M., Barbar, E., Blackledge, M., Bondos, S. E., Dosztányi, Z., et al. (2013). What's in a name? Why these proteins are intrinsically disordered. *ntrinsically Disordered Proteins, 1*, e24157. https://doi.org/10.4161/idp.24157

Fadda, E., & Nixon, M. G. (2017). The transient manifold structure of the p53 extreme C-terminal domain: Insight into disorder, recognition, and binding promiscuity by molecular dynamics simulations. *Physical Chemistry Chemical Physics, 19*, 21287–21296. https://doi.org/10.1039/c7cp02485a

Gadhave, K., Gehi, B. R., Kumar, P., Xue, B., Uversky, V. N., & Giri, R. (2020). The dark side of Alzheimer's disease: Unstructured biology of proteins from the amyloid cascade signaling pathway. *Cellular and Molecular Life Sciences*, 1–46. https://doi.org/10.1007/s00018-019-03414-9

Garg, N., Kumar, P., Gadhave, K., & Giri, R. (2019). The dark proteome of cancer: Intrinsic disorderedness and functionality of HIF-1α along with its interacting proteins. *Progress in Molecular Biology and Translational Science, 166*, 371–403. https://doi.org/10.1016/BS.PMBTS.2019.05.006

Giri, R., Kumar, D., Sharma, N., & Uversky, V. N. (2016). Intrinsically disordered side of the zika virus proteome. *Frontiers in Cellular and Infection Microbiology, 6*, 144. https://doi.org/10.3389/fcimb.2016.00144

Giri, R., Bhardwaj, T., Shegane, M., Gehi, B. R., Kumar, P., & Gadhave, K. (2020). Dark proteome of newly emerged SARS-CoV-2 in comparison with human and bat coronaviruses. *BioRxiv*. https://doi.org/10.1101/2020.03.13.990598

González, M. A. (2011). Force fields and molecular dynamics simulations. *Collection SFN, 12*, 169–200. https://doi.org/10.1051/sfn/201112009

Habchi, J., Tompa, P., Longhi, S., & Uversky, V. N. (2014). Introducing protein intrinsic disorder. *Chemical Reviews, 114*(13), 6561–6588. https://doi.org/10.1021/cr400514h

Han, M., Xu, J., & Ren, Y. (2017). Compromise in competition between free energy and binding effect of intrinsically disordered protein p53 C-terminal domain. *Molecular Simulation, 43*, 110–120. https://doi.org/10.1080/08927022.2016.1237023

Hegyi, H., Buday, L., & Tompa, P. (2009). Intrinsic structural disorder confers cellular viability on oncogenic fusion proteins. *PLoS Computational Biology, 5*, e1000552. https://doi.org/10.1371/journal.pcbi.1000552

Henriques, J., Cragnell, C., & Skepö, M. (2015). Molecular dynamics simulations of intrinsically disordered proteins: Force field evaluation and comparison with experiment. *Journal of Chemical Theory and Computation, 11*, 3420–3431. https://doi.org/10.1021/ct501178z

Hollingsworth, S. A., & Dror, R. O. (2018). Molecular dynamics simulation for all. *Neuron, 99*, 1129–1143. https://doi.org/10.1016/j.neuron.2018.08.011

Huang, Y., & Liu, Z. (2009). Kinetic advantage of intrinsically disordered proteins in coupled folding-binding process: A critical assessment of the "fly-casting" mechanism. *Journal of Molecular Biology, 393*, 1143–1159. https://doi.org/10.1016/j.jmb.2009.09.010

Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., De Groot, B. L., et al. (2016). CHARMM36m: An improved force field for folded and intrinsically disordered proteins. *Nature Methods, 14*, 71–73. https://doi.org/10.1038/nmeth.4067

Ishida, T., & Kinoshita, K. (2007). PrDOS: Prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Research, 35*, W460–W464. https://academic.oup.com/nar/article/35/suppl_2/W460/2923303. Accessed 30 May 2020

Jorgensen, W. L., & Tirado-Rives, J. (1988). The OPLS potential functions for proteins. Energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society, 110*, 1657–1666. https://doi.org/10.1021/ja00214a001

Jorgensen, W. L., & Tirado-Rives, J. (2005). Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proceedings of the National Academy of Sciences of the United States of America, 102*, 6665–6670. https://doi.org/10.1073/pnas.0408037102

Kannan, S., Lane, D. P., & Verma, C. S. (2016). Long range recognition and selection in IDPs: The interactions of the C-terminus of p53. *Scientific Reports, 6*, 1–13. https://doi.org/10.1038/srep23750

Kjaergaard, M., Nørholm, A.-B. B., Hendus-Altenburger, R., Pedersen, S. F., Poulsen, F. M., & Kragelund, B. B. (2010). Temperature-dependent structural changes in intrinsically disordered proteins: Formation of α-helices or loss of polyproline II? *Protein Science, 19*, 1555–1564. https://doi.org/10.1002/pro.435

Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A. E., & Kolinski, A. (2016). Coarse-grained protein models and their applications. *Chemical Reviews, 116*, 7898–7936. https://doi.org/10.1021/acs.chemrev.6b00163

Konrat, R. (2014). NMR contributions to structural dynamics studies of intrinsically disordered proteins. *Journal of Magnetic Resonance, 241*, 74–85. https://doi.org/10.1016/j.jmr.2013.11.011

Kumar, D., Sharma, N., & Giri, R. (2017). Therapeutic interventions of cancers using intrinsically disordered proteins as drug targets: C-Myc as model system. *Cancer Informatics, 16*, 117693511769940. https://doi.org/10.1177/1176935117699408

Kumar, P., Saumya, K. U., & Giri, R. (2019). Identification of peptidomimetic compounds as potential inhibitors against MurA enzyme of *Mycobacterium tuberculosis*. *Journal of Biomolecular Structure & Dynamics*, 1–21. https://doi.org/10.1080/07391102.2019.1696231

Kumar, A., Kumar, P., Kumari, S., Uversky, V. N., & Giri, R. (2020a). Folding and structural polymorphism of p53 C-terminal domain: One peptide with many conformations. *Archives of Biochemistry and Biophysics, 684*, 108342. https://doi.org/10.1016/j.abb.2020.108342

Kumar, D., Singh, A., Kumar, P., Uversky, V. N., Rao, C. D., & Giri, R. (2020b). Understanding the penetrance of intrinsic protein disorder in rotavirus proteome. *International Journal of Biological Macromolecules, 144*, 892–908. https://doi.org/10.1016/j.ijbiomac.2019.09.166

Kuzmanic, A., & Zagrovic, B. (2010). Determination of ensemble-average pairwise root mean-square deviation from experimental B-factors. *Biophysical Journal, 98*, 861–871. https://doi.org/10.1016/j.bpj.2009.11.011

Lange, O. F., Van Der Spoel, D., & De Groot, B. L. (2010). Scrutinizing molecular mechanics force fields on the submicrosecond timescale with NMR data. *Biophysical Journal, 99*, 647–655. https://doi.org/10.1016/j.bpj.2010.04.062

Le Gall, T., Romero, P. R., Cortese, M. S., Uversky, V. N., & Dunker, A. K. (2007). Intrinsic disorder in the protein data bank. *Journal of Biomolecular Structure & Dynamics, 24*, 325–341. https://doi.org/10.1080/07391102.2007.10507123

Li, H., Chowdhary, J., Huang, L., He, X., MacKerell, A. D., & Roux, B. (2017). Drude polarizable force field for molecular dynamics simulations of saturated and unsaturated zwitterionic lipids. *Journal of Chemical Theory and Computation, 13*, 4535–4552. https://doi.org/10.1021/acs.jctc.7b00262

Lindorff-Larsen, K., Maragakis, P., Piana, S., Eastwood, M. P., Dror, R. O., & Shaw, D. E. (2012). Systematic validation of protein force fields against experimental data. *PLoS One, 7*, e32131. https://doi.org/10.1371/journal.pone.0032131

Lindorff-Larsen, K., Piana, S., Dror, R. O., & Shaw, D. E. (2013). How fast-folding proteins fold. *Science, 334*, 517–520. https://doi.org/10.1126/science.1208351

Lippert, R. A., Predescu, C., Ierardi, D. J., Mackenzie, K. M., Eastwood, M. P., Dror, R. O., et al. (2013). Accurate and efficient integration for molecular dynamics simulations at constant temperature and pressure. *The Journal of Chemical Physics, 139*, 164106. https://doi.org/10.1063/1.4825247

Lopes, J. L. S., Orcia, D., Araujo, A. P. U., Demarco, R., & Wallace, B. A. (2013). Folding factors and partners for the intrinsically disordered protein micro-exon gene 14 (MEG-14). *Biophysical Journal, 104*, 2512–2520. https://doi.org/10.1016/j.bpj.2013.03.063

Mackay, D. H. J., Cross, A. J., & Hagler, A. T. (1989). The role of energy minimization in simulation strategies of biomolecular systems. In *Prediction of protein structure and the principles of protein conformation* (pp. 317–358). Boston: Springer. https://doi.org/10.1007/978-1-4613-1571-1_7

Martínez, L. (2015). Automatic identification of mobile and rigid substructures in molecular dynamics simulations and fractional structural fluctuation analysis. *PLoS One, 10*, e0119264. https://doi.org/10.1371/journal.pone.0119264

Megan Sickmeier, A. K. D., Hamilton, J. A., LeGall, T., Vacic, V., Cortese, M. S., Tantos, A., et al. (2007). DisProt: The database of disordered proteins. *Nucleic Acids Research, 35*, D786–D793. https://pubmed.ncbi.nlm.nih.gov/17145717/?from_single_result=17145717&expanded_search_query=17145717. Accessed 30 May 2020

Minezaki, Y., Homma, K., & Nishikawa, K. (2007). Intrinsically disordered regions of human plasma membrane proteins preferentially occur in the cytoplasmic segment. *Journal of Molecular Biology, 368*, 902–913. https://doi.org/10.1016/j.jmb.2007.02.033

Mishra, P. M., Uversky, V. N., & Giri, R. (2018). Molecular recognition features in zika virus proteome. *Journal of Molecular Biology, 430*, 2372–2388. https://doi.org/10.1016/j.jmb.2017.10.018

Mittal, J., & Best, R. B. (2010). Tackling force-field bias in protein folding simulations: Folding of Villin HP35 and Pin WW domains in explicit water. *Biophysical Journal, 99*, L26–L28. https://doi.org/10.1016/j.bpj.2010.05.005

Mohan, A., Oldfield, C. J., Radivojac, P., Vacic, V., Cortese, M. S., Dunker, A. K., et al. (2006). Analysis of molecular recognition features (MoRFs). *Journal of Molecular Biology, 362*, 1043–1059. https://doi.org/10.1016/j.jmb.2006.07.087

Nerenberg, P. S., Jo, B., So, C., Tripathy, A., & Head-Gordon, T. (2012). Optimizing solute-water van der waals interactions to reproduce solvation free energies. *The Journal of Physical Chemistry. B, 116*, 4524–4534. https://doi.org/10.1021/jp2118373

Nishimura, C., Lietzow, M. A., Dyson, H. J., & Wright, P. E. (2005). Sequence determinants of a protein folding pathway. *Journal of Molecular Biology, 351*, 383–392. https://doi.org/10.1016/j.jmb.2005.06.017

Oates, M. E., Romero, P., Ishida, T., Ghalwash, M., Mizianty, M. J., Xue, B., et al. (2013). D^2P^2: Database of disordered protein predictions. *Nucleic Acids Research, 41*, D508–D516. https://pubmed.ncbi.nlm.nih.gov/23203878/. Accessed 30 May 2020.

Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C. J., & Dunker, A. K. (2003). Predicting intrinsic disorder from amino acid sequence. *Proteins: Structure, Function, and Bioinformatics, 53*, 566–572. https://doi.org/10.1002/prot.10532

Oldfield, C. J., Cheng, Y., Cortese, M. S., Romero, P., Uversky, V. N., & Dunker, A. K. (2005). Coupled folding and binding with α-helix-forming molecular recognition elements. *Biochemistry, 44*, 12454–12470. https://doi.org/10.1021/bi050736e

Oldfield, C. J., Meng, J., Yang, J. Y., Qu, M. Q., Uversky, V. N., & Dunker, A. K. (2008). Flexible nets: Disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics, 9*, S1. https://doi.org/10.1186/1471-2164-9-S1-S1

Ouyang, J. F., & Bettens, R. P. A. (2015). Modelling water: A lifetime enigma. *CHIMIA International Journal for Chemistry, 69*, 104–111. https://doi.org/10.2533/chimia.2015.104

Parrinello, M., & Rahman, A. (1980). Crystal structure and pair potentials: A molecular-dynamics study. *Physical Review Letters, 45*, 1196–1199. https://doi.org/10.1103/PhysRevLett.45.1196

Parrinello, M., & Rahman, A. (1981). Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics, 52*, 7182–7190. https://doi.org/10.1063/1.328693

Perilla, J. R., Goh, B. C., Cassidy, C. K., Liu, B., Bernardi, R. C., Rudack, T., et al. (2015). Molecular dynamics simulations of large macromolecular complexes. *Current Opinion in Structural Biology, 31*, 64–74. https://doi.org/10.1016/j.sbi.2015.03.007

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., et al. (2005). Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry, 26*, 1781–1802. https://doi.org/10.1002/jcc.20289

Piana, S., Donchev, A. G., Robustelli, P., & Shaw, D. E. (2015). Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *The Journal of Physical Chemistry. B, 119*, 5113–5123. https://doi.org/10.1021/jp508971m

Ponder, J. W., & Case, D. A. (2003). *Force fields for protein simulations*. Hoboken: Elsevier.

Posch, H. A., Hoover, W. G., & Vesely, F. J. (1986). Canonical dynamics of the Nosé oscillator: Stability, order, and chaos. *Physical Review A, 33*, 4253–4265. https://doi.org/10.1103/PhysRevA.33.4253

Robustelli, P., Piana, S., & Shaw, D. E. (2018). Developing a molecular dynamics force field for both folded and disordered protein states. *Proceedings of the National Academy of Sciences of the United States of America, 115*, E4758–E4766. https://doi.org/10.1073/pnas.1800690115

Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J., & Dunker, A. K. (2001). Sequence complexity of disordered protein. *Proteins, 42*, 38–48. https://doi.org/10.1002/1097-0134(20010101)42:1<38::AID-PROT50>3.0.CO;2-3

Schad, E., Tompa, P., & Hegyi, H. (2011). The relationship between proteome size, structural disorder and organism complexity. *Genome Biology, 12*, R120. https://doi.org/10.1186/gb-2011-12-12-r120

Shoemaker, B. A., Portman, J. J., & Wolynes, P. G. (2000). Speeding molecular recognition by using the folding funnel: The fly-casting mechanism. *Proceedings of the National Academy of Sciences of the United States of America, 97*, 8868–8873. https://doi.org/10.1073/pnas.160259697

Singh, A., Kumar, A., Yadav, R., Uversky, V. N., & Giri, R. (2018a). Deciphering the dark proteome of Chikungunya virus. *Scientific Reports, 8*, 5822. https://doi.org/10.1038/s41598-018-23969-0

Singh, A., Kumar, A., Uversky, V. N., & Giri, R. (2018b). Understanding the interactability of chikungunya virus proteins *via* molecular recognition feature analysis. *RSC Advances, 8*, 27293–27303. https://doi.org/10.1039/C8RA04760J

Sugita, Y., & Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters, 314*, 141–151. https://doi.org/10.1016/S0009-2614(99)01123-9

Tompa, P. (2005). The interplay between structure and function in intrinsically unstructured proteins. *FEBS Letters, 579*, 3346–3354. https://doi.org/10.1016/j.febslet.2005.03.072

Tompa, P., & Fuxreiter, M. (2008). Fuzzy complexes: Polymorphism and structural disorder in protein-protein interactions. *Trends in Biochemical Sciences, 33*, 2–8. https://doi.org/10.1016/j.tibs.2007.10.003

Uversky, V. N. (2009). *Intrinsically disordered proteins and their environment: Effects of strong denaturants, temperature, pH, counter ions, membranes, binding partners, osmolytes, and macromolecular crowding*. Berlin: Springer. https://doi.org/10.1007/s10930-009-9201-4

Uversky, V. N. (2013). A decade and a half of protein intrinsic disorder: Biology still waits for physics. *Protein Science, 22*, 693–724. https://doi.org/10.1002/pro.2261

Uversky, V. N. (2015). Intrinsically disordered proteins and their (disordered) proteomes in neurodegenerative disorders. *Frontiers in Aging Neuroscience, 7*, 18. https://doi.org/10.3389/fnagi.2015.00018

Uversky, V. N. (2019). Intrinsically disordered proteins and their “mysterious” (meta)physics. *Frontiers of Physics, 7*, 10. https://doi.org/10.3389/fphy.2019.00010

Uversky, V. N., & Dunker, A. K. (2010). Understanding protein non-folding. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 1804*, 1231–1264. https://doi.org/10.1016/j.bbapap.2010.01.017

Uversky, V. N., & Dunker, A. K. (2013). The case for intrinsically disordered proteins playing contributory roles in molecular recognition without a stable 3D structure. *F1000 Biology Reports, 5*, 1. https://doi.org/10.3410/B5-1

Uversky, V. N., Oldfield, C. J., & Dunker, A. K. (2005). Showing your ID: Intrinsic disorder as an ID for recognition, regulation and cell signaling. *Journal of Molecular Recognition, 18*, 343–384. https://doi.org/10.1002/jmr.747

Uversky, V. N., Oldfield, C. J., & Dunker, A. K. (2008). Intrinsically disordered proteins in human diseases: Introducing the D 2 concept. *Annual Review of Biophysics, 37*, 215–246. https://doi.org/10.1146/annurev.biophys.37.032807.125924

Van Der Lee, R., Buljan, M., Lang, B., Weatheritt, R. J., Daughdrill, G. W., Dunker, A. K., et al. (2014). Classification of intrinsically disordered regions and proteins. *Chemical Reviews, 114*, 6589–6631. https://doi.org/10.1021/cr400525m

Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., et al. (2009). CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of Computational Chemistry, 31*, NA–NA. https://doi.org/10.1002/jcc.21367

Walsh, I., Martin, A. J., Di Domenico, T., & Tosatto, S. C. (2012). ESpritz: Accurate and fast prediction of protein disorder. *Bioinformatics, 28*, 503–509. https://pubmed.ncbi.nlm.nih.gov/22190692/. Accessed 30 May 2020

Williams, R. M., Obradovic, Z., Mathura, V., Braun, W., Garner, E. C., Young, J., et al. (2000). The protein non-folding problem: Amino acid determinants of intrinsic order and disorder. In *Biocomputing 2001* (pp. 89–100). Singapore: World Scientific. https://doi.org/10.1142/9789814447362_0010.

Wright, P. E., & Dyson, H. J. (1999). Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *Journal of Molecular Biology, 293*, 321–331. https://doi.org/10.1006/jmbi.1999.3110

Wu, H., & Fuxreiter, M. (2016). The structure and dynamics of higher-order assemblies: Amyloids, signalosomes, and granules. *Cell, 165*, 1055–1066. https://doi.org/10.1016/j.cell.2016.05.004

Xue, B., Li, L., Meroueh, S. O., Uversky, V. N., & Dunker, A. K. (2009). Analysis of structured and intrinsically disordered regions of transmembrane proteins. *Molecular BioSystems, 5*, 1688–1702. https://doi.org/10.1039/b905913j

Xue, B., W. Williams, R., J. Oldfield, C., Kian-Meng Goh, G., Keith Dunker, A., & N. Uversky, V. (2010). Viral disorder or disordered viruses: Do viral proteins possess unique features? *Protein and Peptide Letters, 17*, 932–951. https://doi.org/10.2174/092986610791498984

Yang, J. Y., Yang, M. Q., Dunker, A. K., Deng, Y., & Huang, X. (2008). Investigation of transmembrane proteins using a computational approach. *BMC Genomics, 9*, S7. https://doi.org/10.1186/1471-2164-9-S1-S7

Zhang, T., Faraggi, E., Li, Z., & Zhou, Y. (2013). Intrinsically semi-disordered state and its role in induced folding and protein aggregation. *Cell Biochemistry and Biophysics, 67*, 1193–1205. https://doi.org/10.1007/s12013-013-9638-0

Zsuzsanna Dosztányi, I. S., Csizmok, V., & Tompa, P. (2005). IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics, 21*, 3433–3434. https://pubmed.ncbi.nlm.nih.gov/15955779/?from_single_result=15955779&expanded_search_query=15955779. Accessed 30 May 2020