

Protein structure prediction

Shikha Agnihotry¹, Rajesh Kumar Pathak², Dev Bukhsh Singh³, Apoorv Tiwari⁴ and Imran Hussain⁵

¹Department of Computational Biology and Bioinformatics, Jacob Institute of Biotechnology and Bio-Engineering, Sam Higginbottom University of Agriculture, Technology and Sciences, Allahabad, India, ²School of Agricultural Biotechnology, Punjab Agricultural University, Ludhiana, India,

³Department of Biotechnology, Institute of Biosciences and Biotechnology, Chhatrapati Shahu Ji Maharaj University, Kanpur, India, ⁴Bioinformatics Sub-DIC, Molecular Biology & Genetic Engineering, College of Basic Science and Humanities, G.B. Pant University of Agriculture and Technology, Pantnagar, India, ⁵School of Life and Allied Health Sciences, Glocal University, Saharanpur, India

11.1 Introduction

Proteins are important molecules that play a significant role in different biological processes. The protein structure prediction is primarily based on sequence and structural homology. Protein structure prediction or modeling is very important as the function of a protein is mainly dependent on its 3D structure. Similarly, the 3D structure of a protein depends on its amino acid composition. A small variation in the protein sequence may bring a great extent of structural variation in the native structure. The accurate knowledge of protein 3D structure is very important but it is very hard to decipher the native structure of a protein that resides under physiological conditions of the body. X-ray crystallography and nuclear magnetic resonance spectroscopy (NMR) methods are mainly used for determining the structure of protein and protein-ligand complexes. These experimental structure determination techniques are very time taking, costly, and complex (Baker & Sali, 2001). As a result, with the development of a computational algorithm and computational facilities, theoretical knowledge of protein structure, dynamics, and folding have been used to generate a model from amino acid sequences.

X-ray crystallography, NMR spectroscopy, and electron microscopy technique have their own limitations, merits, and demerits. In X-ray crystallography, X-rays are passed through the protein crystal and then the pattern of diffraction is recorded and analyzed to gather information about the coordinates of different atoms in the molecule using computerized frameworks. NMR spectroscopy is considered good for the structure determination of small molecules, or small size protein–ligand complexes. The benefit of the NMR method is that the protein can be studied in an aqueous environment that may look like its real physiological state. The NMR strategy is restricted to small size proteins having a length of less than 150 amino acids. Electron microscopy is suitable for the investigation of large macromolecules or even cell organelles and can assist the building of a tertiary model of a single particle (Cheng, 2015).

For a long time, the algorithm behind the amino acid sequence folding into a unique protein three-dimensional (3D) stays one of the extraordinary unsolved issues. Protein sequence and structural data analysis from known and verified structures have explored the many issues related to amino acid properties, and its arrangement in the form of a helix, sheet, and coil, side chain, conformations, and other structural issues related to the packing of sequences into 3D protein model. Based on these statistics and principles, many new hypotheses about molecular evolution and structure prediction techniques have been developed (Hermoso Pulido, 2015). Several tertiary structure prediction tools have been designed that are based on homology modeling, threading/fold recognition, and ab initio approach.

11.2 Protein structure prediction

The structure and sequence information of protein is increasing continuously in the Protein Data Bank (PDB). The PDB database, one of the vast protein structure databases, is the main source of template for homology modeling (Gupta, Mohanty, & Bhatnagar, 2016). There is a big gap between the available sequence information and the existing known structures, and this is a significant issue in structural biology that needs to be addressed. Protein structures are more

conserved than the sequences. Two or more similar sequences of proteins can adopt a similar fold. It means that there is a degeneracy in protein sequences, as a result of a large number of protein sequences, there is a limited number of available structural folds. Many closely or evolutionary-related proteins have nearly similar structural folds due to a good extent of similarity among their protein sequences.

In protein modeling, the sequences whose model has to be prepared are referred to as target protein, while another protein based on which the target model is built is known as template. The structure prediction may be based on a template or without a template. In template-based modeling, a homologous template structure is used to model the target protein. Here, structural regions conserved between target and template are identified and their coordinates are assigned from template to target (Fiser, 2010). For template-based modeling, the structure of the template must be known and the template should share a good extent of similarity with the target. Homology modeling is based on the template-based approach (Fiser, 2004). Generally, we prefer to use the template-based approach but may employ other methods of modeling in case the template is not known, or does not share enough similarity with the target. Tertiary structure prediction can be used for modeling of the proteins having known sequences and unknown structures, understanding the protein folding, proteins engineering to incorporate new functions, and drug designing (Bhasin & Raghava, 2006).

Domains are discrete structural and/or functional units in a protein. Typically, domains are responsible for a particular interaction or function. Most of the structure prediction methods consider the domain level information while building a model. It has been suggested that to make models, a target sequence first needs to be categorized in multiple domains (Andersen et al., 2001). Domains are attached or linked to each other by the peptides known as linker peptides. In the multidomain proteins, two different or evolutionary unrelated or distant proteins may share the same conserved domains (Marchler-Bauer et al., 2005). For some of these important reasons, at a domain level, protein structure prediction is generally performed (Roberts, Alberts, Johnson, Walter, & Hunt, 2002). For performing domain level structure prediction, it is essential to get a primary image of the composition of the target protein domain and also of outside globular domain regions, such as linker, signal, low-complexity, transmembrane, and disordered regions (Wright & Dyson, 2015).

A domain is predicted using different approaches that are typically based on the multiple sequence alignments (MSAs), frequency of amino acids present in linker and domain, size of the domain, sequence comparison methods, and predicted secondary structure (Nagarajan & Yona, 2004). CASP6 can appropriately assign domains to more than 80% of the multidomain target residues. Phyre2 and DomCut programs for the domain homology and interdomain linker predictions, respectively (Bakhtin, van der Maaten, Johnson, Gustafson, & Girshick, 2019; Suyama & Ohara, 2003). The organization of proteins into different classes of secondary structure, such as alpha, beta, alpha/beta, coil, and others, is a useful primary approach for protein classification. These classes can be predicted using protein's amino acid composition.

11.3 Method of protein structure prediction

11.3.1 Homology modeling (comparative modeling)

Homology modeling is an important and most preferred method of protein structure prediction. This method is used when a template structure with a good extent of sequence similarity with the target protein sequence is available. This method assumes that protein structures are more conserved than protein sequences. Hence, an accurate model of a protein could be designed using core modeling, loop modeling, and side chain modeling followed by energy minimization (Gromiha, Nagarajan, & Selvaraj, 2019; Singh & Pathak, 2020). The homology modeling approach is similar to the process of fragment assembly where structurally conserved regions (SCRs), or core, variable loop regions, and conformationally bioactive side chains are combined to produce a complete protein model (Fig. 11.1). The first and most important step in homology modeling is to search for a template using the target protein sequence as a query (Nikolaev et al., 2018). A template is searched using different sequence similarity or alignment-based search tools, such as BLAST and PSI-BLAST, and then looking for their known structures in the PDB database. A high-quality X-ray or NMR-verified structure is used as a template. The accuracy of the modeled protein can be much better if a template is closely related to the target or both belong to the same protein family (Pearson, 2013). Target and template sequence alignment is generated to find the similarity and compare the compositional variations between these two sequences. In the absence of a suitable template, the homology modeling approach based tools cannot be used for modeling the structure.

Homology modeling approach includes: (1) identification of suitable template for a given target sequence using BLAST search; (2) template-target alignment; (3) alignment corrections to ensure that the functional or conserved

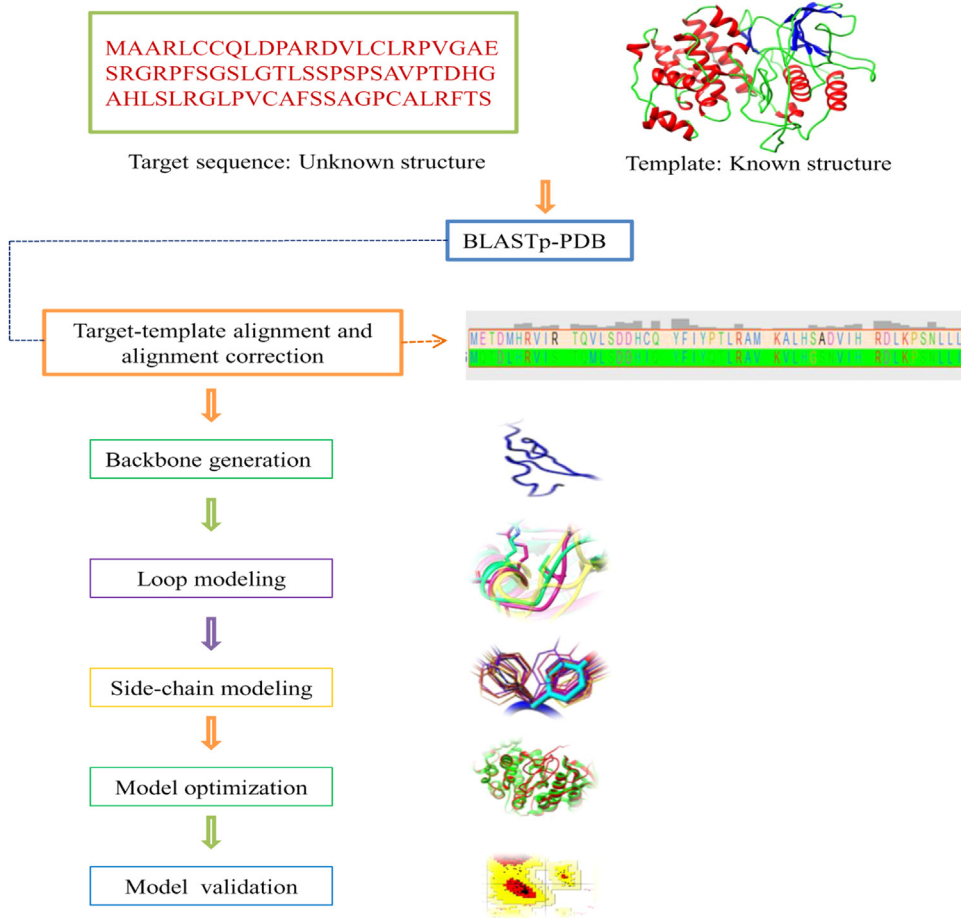


FIGURE 11.1 Protein structure prediction using a homology modeling approach.

residues are adjusted and aligned properly; (4) generation of backbones; (5) modeling the loop regions; (6) modeling the side chains using the rotamer libraries; (7) energy minimization of modeled structure; and (8) stereochemical assessment of model Ramachandran plot (Gromiha et al., 2019). The BLAST server compares a target query sequence with sequences in databases, correlates with the PDB database for availability of its structure, and identifies the sequences having the best and the high degree of similarity. BLAST is used to compare the target sequence with each sequence present in a sequence database. The program compares nucleotide or protein sequences to the sequence databases and calculates their statistical significance in terms of identity, positives, gap (indel), and e -values (Madden, 2013). A sequence with a lower e -value is preferred for similarity. PSI-BLAST is a profile-based sequence comparison method (Li et al., 2012). It uses a profile from the alignment of multiple sequences searched using protein BLAST. Profile-based method searches the evolutionarily related proteins that cannot be identified by pairwise arrangement examination techniques (Bhagwat & Aravind, 2007). MSA tools are used to align template sequences with a target to define core and variable loops based on similarity, deletions, and insertions (Zhang, Kochhar, & Grigorov, 2003). The basic steps involved in homology modeling are SCR modeling, loop modeling, and side chain modeling.

11.3.1.1 SCR or core modeling

SCR or core is a structurally conserved region that can be modeled using the information of atomic coordinates of residues belonging to SCR of a template (Muhammed & Aki-Yalcin, 2019). There may be many more SCR regions, and all these SCRs are modeled copying the coordinates of SCR from the template and then assigning them to a target structure (Dixit, Torkamani, Schork, & Verkhivker, 2009; Huang, Pei, & Grishin, 2013). In this way, a partial model of protein becomes ready that does not include the loop portion. The higher the similarity between template and target sequence, the higher will be the accuracy of the modeled structure as much portion of the modeled structure will be derived from X-ray crystallography or NMR determined high-resolution templates.

11.3.1.2 Loop modeling

Most of the time, the gaps can be seen among the alignment of target and template sequences. These gaps represent the highly variable regions due to insertions and deletions. These highly variable regions in the template and target alignment are known as loops. Loop regions belong to the frequent binding. These loop regions exist between the two SCRs. After SCR modeling, atomic coordinates of amino acids corresponding to loop regions are determined by loop searching in structure databases, and loop structure is placed between the two SCR (Jamroz & Kolinski, 2010). During loop modeling, loop residues along with some neighboring residues are taken as input sequences to search for the fold of this loop region in the known structure databases (Fiser, Do, & Sali, 2000). The atomic coordinates for loop regions are copied from the known structures and then assigned to the target. In case, the fold of a loop is not available in the structure database, and then knowledge-based computer algorithms are used to generate an optimized loop structure that can best fit between the two SCRs. During modeling, the exchange of the side chains can prompt a change in the direction and spatial arrangement due to the substitution of one amino acid by a different amino acid. The amino acids proline and glycine are an exemption from the Ramachandran plot. Proline has a limitation in the plot because of its five-membered rings while glycine has a hydrogen atom as its side chain which is hard to anticipate from the plot. This makes it hard to identify mutations that have ended up loop build up glycine or proline.

11.3.1.3 Side chain modeling

During SCR modeling, coordinates of SCR are assigned to the target except for differing amino acids. For locations where amino acids differ in template-target alignment, only backbone coordinates are taken from the template and side chains are modeled later on. Side chains are modeled using partially knowledge-based libraries of rotamers derived from high-resolution X-ray structures (Bower, Cohen, & Dunbrack, 1997). Rotamer libraries contain bioactive side chains. It is important to incorporate an active and right conformation of side chains into the build model to avoid steric clashes and structural bumping. For an accurate model, the number of side chain outliers must be zero or as small as possible.

11.3.1.4 Model accuracy and validation

Modeled protein may contain some errors. These mistakes fundamentally rely on the selection of the template, identity between template and target, and error in the template structure. The template must share a good extent of similarity with the target sequence. If sequence identity between template and target is less than 25%, then modeled protein may not be reliable and accurate. Therefore homology modeling is recommended only the condition if sequence identity between template and target is more than 25%. Modeling errors can be assessed by calculating the energy of a model on a force field. This strategy verifies whether the bond lengths and bond angle points are within a typical normal range as observed in the case of experimental structures.

Based on the energy minimization concept, the most accurate and stable protein structure will have minimum free energy at thermodynamic equilibrium. Energy-based methods attempt to find the protein with free energy at the global minimum state. The energy minimization software packages are CHARMS (Grinspun, Krysl, & Schroder, 2002), AMBER (Case et al., 2008), GROMOS (Scott et al., 1999) and ECEPP (Zimmerman, Pottle, Némethy, & Scheraga, 1977). Some issues with methods of energy minimization are: (1) requirement of high computations for the representation of protein configurations and (2) interaction potential calculations are not well satisfactory for a protein at the atomic level (Bhasin & Raghava, 2006). Accurate prediction of protein structure is the main challenge of structural genomics. There is a big gap between the number of protein sequences available and the number of known structures. Several algorithms have been developed for modeling the proteins but algorithms have their own limitation in their use and application.

11.3.2 Threading or fold recognition

Many studies suggest that the proteins having a good extent of sequence similarity or often closely connected to homologous have similar 3D structures as well as functions. The threading or fold recognition method predicts the 3D structure with high accuracy for a protein sequence even with low sequence similarity. Threading word was first coined by (Jones, Taylor, & Thornton, 1992a,b) and described that the query sequence may be “threaded” onto the backbone of the template structure. Generally, the backbone of any protein structure is considered rigid, and only the side chains of an amino acid are considered flexible. Interaction potential along with the other scores is used to predict the structure. Consequently, threading programs do not use interaction potential that required comprehensive explanations of the configuration or coordinates of all the atoms (Rykunov & Fiser, 2010). The scoring criteria are also extended only to pair interaction between the residue centers or backbone/side chain centers. Empirical energy scores can be calculated in the

known structures based on the observed interactions of amino acids rather than using the physical energy functions of the atom interactions. Protein threading having four fundamental parameters: (1) structure database, (2) energy function, (3) sequence-structure alignment algorithm, and (4) predicted reliability assessment (Fiser, 2010).

$$\text{Total energy } (E) = E_p + E_s + E_g$$

where E_p is “how preferable to put two particular residues nearby,” E_s is “how well a residue fits a structural environment,” and E_g is the alignment gap penalty.

Each residue position is classified into an environment class based on these criteria. In this way, a 3D protein structure is transformed into a one-dimensional string that reflects each residue’s environmental class in the folded protein structure. A 3D structure profile table is created, which contains score values that represent the probability of an amino acid at a location in all environmental classes. With the increasing amount of information about the template used by the method to measure the comparison score, the precision of fold recognition methods increases (Yang, Faraggi, Zhao, & Zhou, 2011). The profile-based approach that determines the structural environmental class for each residue in the template structure and develops a substitution matrix from the probability of amino acids being in each environmental class is one of the most common and effective methods among 3D structure prediction methods. Another strategy for threading is to calculate the potential of residue–residue contact pairs and optimize the hydrophobic core score. Together, the process of threading can be largely classified into two distinct aspects: (1) providing structural details as a profile of the structural environment groups and (2) directly including the pair wise contact potential (Tiwari, Chauhan, Agarwal, & Ramteke, 2020).

In nature, the unique structural folds are reasonably limited. Sequences are specifically fitted into the backbone coordinates of recognized configurations of proteins. In 3D space, matching of sequences to backbone coordinates specifically integrating particular pair interactions. The approach begins with a known 3D protein structure and specifies within the structure three main characteristics of the surrounding of each residue (Fig. 11.2) (Berg, Tymoczko, & Stryer, 2002). These characteristics are: (1) the total area of the side-chain residue that is buried by other solvent-inaccessible protein atoms; (2) the proportion of the area protected by polar atoms (O, N) or water in the side chain; and (3) the composition of the secondary structure. Based on these characteristics, a total of six basic environmental classes exist for amino acids. An amino acid can participate in all three secondary structures, such as helix, coil, and sheet. Therefore there are a total of 18 possible environmental classes for each amino acid. The major steps in the threading approach can be pointed out as given here.

- Structure template database construction: Choose protein structures as structural templates from the protein structure databases. In general, protein structures are selected from databases, such as PDB, SCOP, CATH, or FSSP (Fox, Brenner, & Chandonia, 2015).

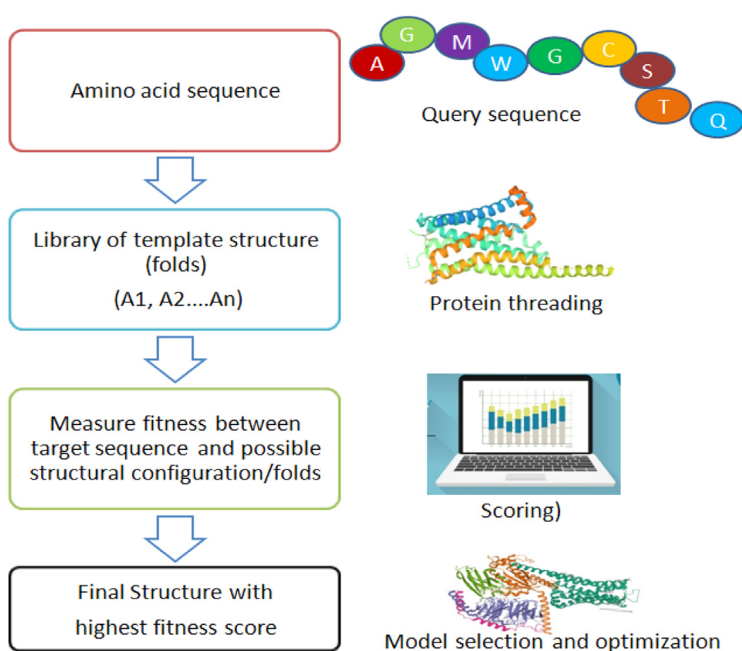


FIGURE 11.2 Twofold recognition approach for protein 3D structure prediction.

- Scoring function design: Create a scoring function to calculate the fitness between target sequences and templates based on the information of the known relationships between structures and sequences. A better scoring function should include possible mutation, pair-wise potential, compatibility of the secondary structure, environment fitness potential, and gap penalties.
- Threading alignment: Align the target sequence by optimizing the developed scoring function with each of the structure models, or fold. This step is one of the key tasks of all threading-based structural prediction programs that take into account the pair wise contact potential; otherwise, a dynamic programming algorithm will do so (Peng & Xu, 2011a,b).
- Threading prediction: The threading prediction selects the threading alignment which is statistically most probable. Then, place the backbone atoms of the target sequence at their aligned backbone locations of the chosen structural template to create a structure model for the target.
- Model assessment and validation: Manual inspection and modification are also needed to build a successful structure model using threading alignments.

11.3.3 Ab initio approach

It is a powerful computational approach to predict the native structure of a protein using amino acid sequence information when no structural homolog is detected in the structural database, that is, PDB. As we know that if the query or target protein has structural homologs or similar known structures, the modeling task is much easy and generates good quality models *via* homology modeling or threading approach. However, if no structural homolog is available for the target sequence, the modeling job is challenging; therefore the model can build from the scratch. This process is known as ab initio modeling. It is necessary for a complete solution to the problem linked to the prediction of protein structure. Besides, it may also enable us to know how proteins fold in nature based on physicochemical theory, and different conformation for a given amino acid sequence is searched (Fig. 11.3). Conformational search is done under the guidance of an energy function and generates a number of possible conformation for a protein. Physics-based or knowledge-based

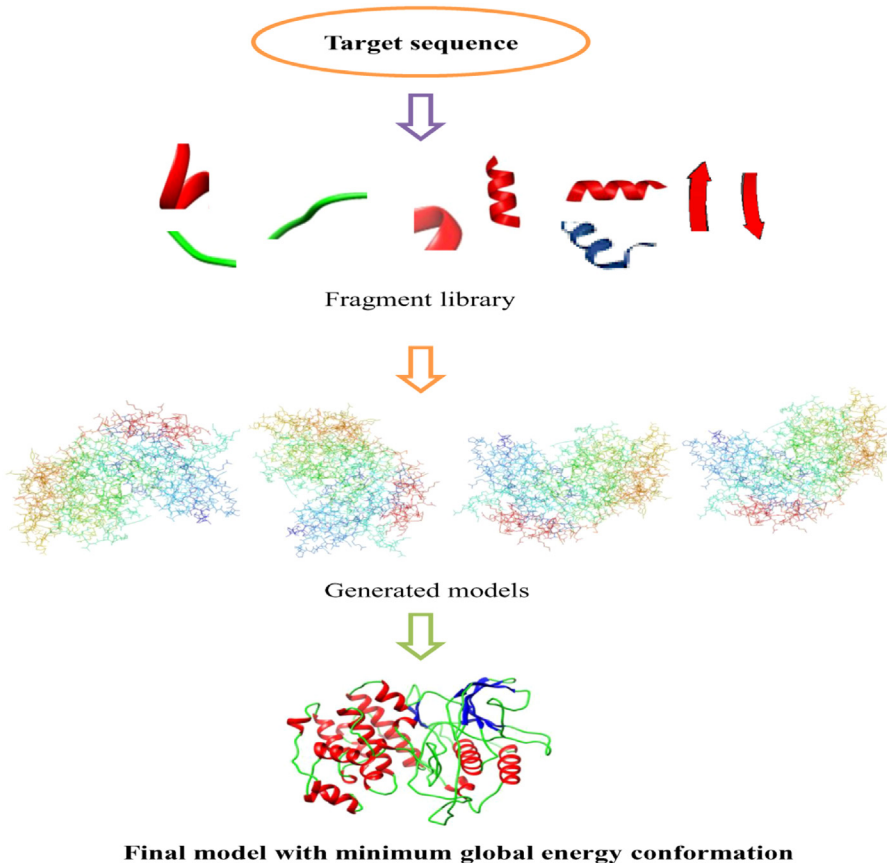


FIGURE 11.3 Basic concept behind the ab initio approach of modeling.

(hydrogen bonding, contact potential, secondary structure propensities derived from PDB structures) energy functions are used for conformational search. An efficient method is required to find the low energy states from a set of decoy structures, or conformations. The conformations that reduce the energy function are thought to be the ones that the protein will follow in its native state.

Finally, the generated structures having the lowest global energy conformation will be considered as stable structures and recommended for further research. The ab initio modeling's accuracy is currently poor, and its success is also limited (Lee, Wu, & Zhang, 2009; Singh & Pathak, 2020; Singhal & Singhal, 2018).

Several modeling tools, servers, and software have been developed for structure prediction based on the homology modeling, threading, and ab initio approaches (Table 11.1).

TABLE 11.1 Some protein structure prediction software.

PHD	Structure prediction and sequence analysis method	Rost, Sander, and Schneider (1994)
APSSP2	Secondary structure of protein prediction	Raghava (2002)
PsiPred	PSI-blast-based secondary structure PREDiction	McGuffin, Bryson, and Jones (2000)
JPRED	Method for secondary structure prediction	Cuff, Clamp, Siddiqui, Finlay, and Barton (1998)
SWISS-MODEL	Automated protein structure homology-modeling server	Schwede, Kopp, Guex, and Peitsch (2003)
GENO3D	3D proteins structures modeling server	Combet, Jambon, Deleage, and Geourjon (2002)
CPHMODELS	Homology modeling server	Nielsen, Lundegaard, Lund, and Petersen (2010)
HMMSTR	Hidden Markov model for local sequence–structure correlation	Byströff, Thorsson, and Baker (2000)
AMBER	Molecular dynamics simulations of proteins	Case et al. (2008)
CHARMS	Programs for molecular dynamics simulation	Grinspun et al. (2002)
I-TASSER	Based on hierarchical approach for structure prediction and function annotation of protein	Yang et al. (2015)
ROBETTA	Protein structure prediction server	Kim, Chivian, and Baker (2004)
HHpred	For remote protein structure prediction and homology detection using HMM comparison	Söding, Biegert, and Lupas (2005)
METATASSER	Protein tertiary structure prediction	Zhang, Arakaki, and Skolnick (2005)
MULTICOM	Program for protein structure prediction	Wang, Eickholt, and Cheng (2010)
Pcons	Metaserver for protein structure prediction based on conservation	Lundström, Rychlewski, Bujnicki, and Elofsson (2001)
SAM-T08	Hidden Markov model-based protein structure prediction	Karplus (2009)
THREADER	Protein fold prediction	Jones et al. (1992a,b)
RaptorX	Protein structural prediction without knowledge of close homolog in PDB (threading approach)	Peng and Xu (2011a,b)
SwissModel	Server for protein homology modeling	Schwede et al. (2003)
MODELLER	Server for protein homology modeling	Eswar et al. (2007)
PSIpred	Protein secondary structure prediction program	McGuffin et al. (2000)
JPRED	Protein secondary structure prediction server	Cuff et al. (1998)
MUSTER	3D structure prediction using threading. it uses template structures from the PDB database	Wu and Zhang (2008)

PDB, Protein Data Bank.

11.4 Evaluation of predicted protein structure

Ramachandran plot can be used for evaluating the accuracy of predicted protein structure. Several tools and servers, such as PDBsum server, MolProbity, STAN Server, and others, are used to generate Ramachandran plot. The SAVES v6.0 (<https://saves.mbi.ucla.edu/>) is a comprehensive toolkit and has five tools (ERRAT, VERIFY3D, PROVE, PROCHECK, and WHAT CHECK), which predict different types of stereochemical parameters of the protein structure. The Ramachandran plot predicts the structural stereochemical property. The PROCHECK analyzes the overall model geometry with the residue by residue geometry and provides the stereochemical quality of a predicted model. PROCHECK tool requires modeled protein file as an input and generates the Ramachandran plot (Fig. 11.4). It indicates that for the given protein 91.4% (470), 6.8% (35), 1.0% (5), and 0.8% (4) residues belong to the most favored regions, additionally allowed regions, generously allowed regions, and disallowed regions, respectively. It is expected that no more than 2% of residues should belong to the allowed region and no residue should reside in the disallowed or outlier region. For this protein, nine residues, that is, Val4, Leu23, Arg267, Ser323, Arg367, Ser412, Lys440, Asn519, and Lys594, belong to generously allowed regions and disallowed regions. These residues can be remodeled or the energy minimization can be performed to produce a model with better stereochemical properties.

The ERRAT has a database of highly refined protein structures and it plots a value graph of the position of the nine residue sliding window versus error function (Colovos & Yeates, 1993). This plot is based on nonbonded interaction statistics between different atom types that are calculated by its refined structure database. The result of the ERRAT server shows a graph between residues and error values (Fig. 11.5). The overall quality score of this input structure is 72.66% and this is not considered good. If the input structure has good resolution, then it should have a quality score of greater than 95%. More than 90% score is observed for the protein structure with a resolution of 2–3 Å. In the ERRAT graph, regions with red and yellow color represent the problematic part while the white color represents the normal part in the structure. Residues with error values more than 95% and 99% can easily be identified from the plot analysis.

The Verify3D also uses the knowledge of the structural database of proteins and it predicts the compatibility of a 3D structure with the 1D amino acid sequence based on its structure assignments, such as loops, sheets, alpha-helix, and others (Eisenberg, Lüthy, & Bowie, 1997). It compares the quality of modeled protein with high-resolution protein structures from the database. The PROVE calculates the statistical Z-score deviation of the given input structure using the highly refined protein structure of less than or equal to 2.0 Å resolution as a comparison set. Another tool, WHAT CHECK is derived from the WHAT IF structure validation tool. It checks several stereochemical parameters of each residue in the predicted model and gives a graphical output. The PROSA web server is a tool that predicts the Z-score (overall quality of the protein) and residue wise energy in the plotted form (<https://prosa.services.came.sbg.ac.at/prosa.php>). It has a PDB structure database and compares the input structure with the PDB structure. Users can know the

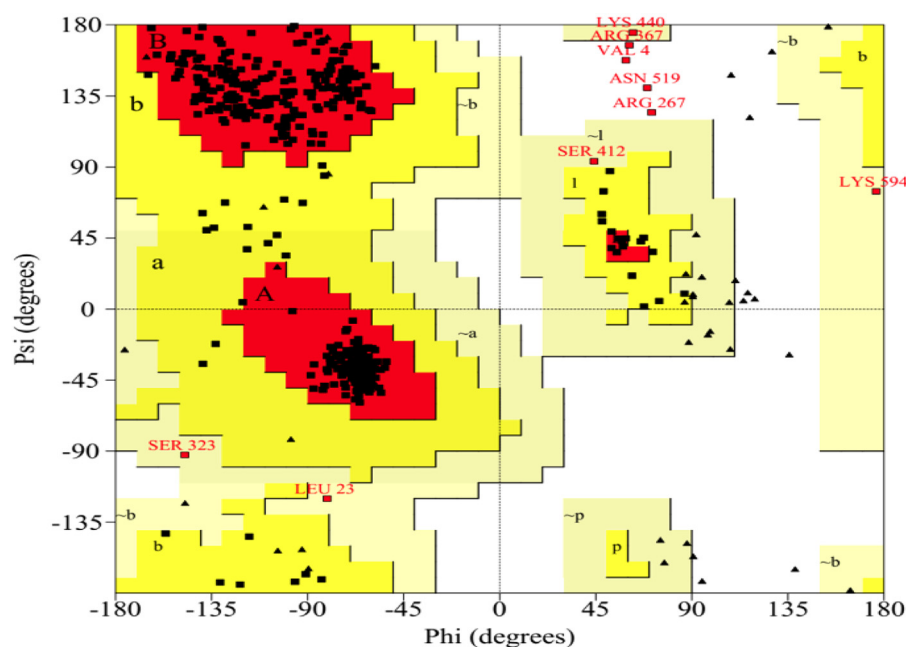


FIGURE 11.4 Validation of modeled protein using Ramachandran plot of PROCHECK analysis.

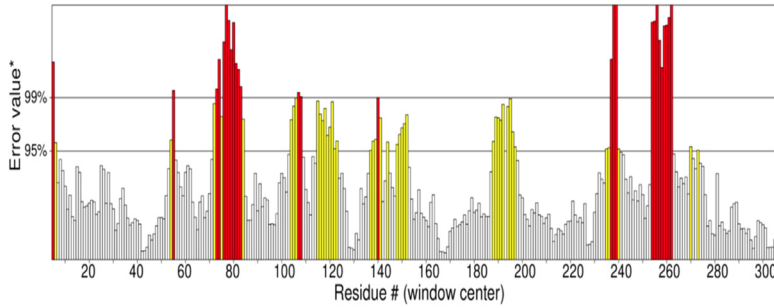


FIGURE 11.5 Validation of modeled protein using ERRAT score.

reliability of the predicted model as it indicates that modeled protein is how much close to experimentally determined structures in terms of accuracy.

11.5 Applications of structure prediction

The predicted protein model has a lot of applications in the diverse field of analysis, such as drug designing, mutational study and structural comparisons, protein unfolding study, binding site and cavity analysis, molecular docking for virtual screening, molecular dynamics simulation for understanding the dynamics of the structure, function analysis, evolutionary analysis, and others. All the applications of modeled protein are described in detail in the below section.

11.5.1 Mutation studies

3D structure of a mutant protein is required to understand the impact of the mutation on structures. The predicted protein model can be visualized and analyzed by using several structure visualization tools and comparison tools, such as Chimera and PyMol. The extent of structural changes that occurred as a result of mutation can be estimated by structural comparison of the normal and mutant protein in terms of root mean square deviation (RMSD). The difference in the orientation of side-chain residue between the native and mutant residue can be visualized. To get further insight into the impact of mutation of protein dynamics, the molecular dynamics simulation can be performed to measure the atomic level and residue wise changes/fluctuations, free energy, and structures compactness (Shukla, Shukla, Sonkar, & Tripathi, 2017). Hence, the initial requirement for such types of analysis can only be fulfilled with protein structures.

11.5.2 Unfolding studies

The protein folding and unfolding is still a mystery. But due to the availability of high-end computational powers and techniques, the unfolding dynamics of the protein can be revealed. Protein unfolding studies also require a 3D structure that can easily be prepared using prediction approaches. For this purpose, the modeled structure is solvated with the denaturant, such as Urea, GdnHCl, or others, to see their impact on denaturation, and that solvated structure can be employed for the MDS to sample the unfolding dynamics in different trajectories (Sonkar, Shukla, Shukla, Kalita, & Tripathi, 2019).

11.5.3 Binding site prediction

The binding site prediction is a critical and important step in the case of drug designing. For the binding site prediction, a 3D structure of a protein is required. Many tools measure the area and volume of cavities in a protein to decipher some insight about the binding site and other sites, such as allosteric sites and cryptic sites. In the case of in silico studies, accurate information of the binding site is necessary to understand the nature and affinity of interactions for a substrate or ligand. Tools, such as CASTp, ATLAS, and many others, are used for a cavity, or binding site predictions, and some docking tools have the facility to predict and define the binding site before the docking run.

11.5.4 Protein docking and virtual screening

Molecular docking is the key technique to characterize the protein–ligand interactions. The docking cannot be performed without a proper 3D structure (Shukla et al., 2018). The signaling cascades are formed by the interaction of

a series of proteins. Hence, characterization of protein–protein interaction plays a significant role to understand the signaling cascade mechanism. During drug discovery, a vast set of ligands can be screened for their binding affinity and interaction with a protein drug target that also requires a known structure for the target.

11.5.5 Understanding the dynamics of protein or protein–ligand complex

The molecular dynamics simulation is a computational technique to characterize the atomic-level changes in the protein or protein–ligand complex under physiological conditions (Shukla & Singh, 2020). For such types of analysis, the structure of protein or protein–ligand complex must be available. Stability and perturbation of a protein molecular system can be interpreted using the RMSD, root mean square fluctuation, hydrogen binding dynamics, the radius of gyration, and free energy parameters. After predicting the model using the modeling methods, such as homology modeling, threading, and ab initio approach, molecular dynamics simulation of the modeled protein structure can also be performed to evaluate its thermodynamic stability.

11.5.6 Structure evolution analysis

The 3D structure shares a common fold across the same species and that are more conserved as compared to the sequence. Hence, to understand the structure evolutionary studies, a proper 3D structure is required and protein modeling can play an important role here also.

11.6 Conclusion

Advancements in computational biology approaches have made significant improvements in the field of structural biology. An accurate and real model of a protein is required to interpret the function, and its interactions with natural substrates as well as with other ligands. There are many proteins whose structure cannot be determined by X-ray crystallography and NMR techniques due to practical limitations. For such cases, modeling protein structures from sequence information are the only way. Nowadays, several structure prediction tools are available, and these tools have their advantages and limitations. There is a need to generate more and more experimental data related to proteins and protein–ligand complexes to overcome the unavailability of templates for modeling and also for improving the accuracy of prediction.

Conflict of interest

The authors declare that they have no conflict of interest.

References

- Andersen, J. N., Mortensen, O. H., Peters, G. H., Drake, P. G., Iversen, L. F., Olsen, O. H., . . . Møller, N. P. (2001). Structural and evolutionary relationships among protein tyrosine phosphatase domains. *Molecular and Cellular Biology*, 21(21), 7117–7136.
- Baker, D., & Sali, A. (2001). Protein structure prediction and structural genomics. *Science (New York, N.Y.)*, 294(5540), 93–96.
- Bakhtin, A., van der Maaten, L., Johnson, J., Gustafson, L., & Girshick, R. (2019). Phyre: A new benchmark for physical reasoning. arXiv preprint arXiv:1908.05656
- Berg, J. M., Tymoczko, J. L., & Stryer, L. (2002). *Section 3.6: The amino acid sequence of a protein determines its three-dimensional structure. Biochemistry* (5th ed.). New York: W.H. Freeman.
- Bhagwat, M., & Aravind, L. (2007). *PSI-BLAST tutorial. Comparative genomics* (pp. 177–186). Humana Press.
- Bhasin, M., & Raghava, G. P. (2006). Computational methods in genome research, In *Applied Mycology and Biotechnology*, 6, 179–207.
- Bower, M. J., Cohen, F. E., & Dunbrack, R. L., Jr (1997). Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: A new homology modeling tool. *Journal of Molecular Biology*, 267(5), 1268–1282.
- Bystroff, C., Thorsson, V., & Baker, D. (2000). HMMSTR: A hidden Markov model for local sequence-structure correlations in proteins. *Journal of Molecular Biology*, 301(1), 173–190.
- Case, D. A., Darden, T. A., Cheatham, T. E., Simmerling, C. L., Wang, J., Duke, R. E., . . . Merz, K. M. (2008). *Amber 10*. University of California.
- Cheng, Y. (2015). Single-particle cryo-EM at crystallographic resolution. *Cell*, 161(3), 450–457.
- Colovos, C., & Yeates, T. O. (1993). Verification of protein structures: Patterns of nonbonded atomic interactions. *Protein Science*, 2(9), 1511–1519.
- Combet, C., Jambon, M., Deleage, G., & Geourjon, C. (2002). Geno3D: Automatic comparative molecular modelling of protein. *Bioinformatics (Oxford, England)*, 18(1), 213–214.

- Cuff, J. A., Clamp, M. E., Siddiqui, A. S., Finlay, M., & Barton, G. J. (1998). JPred: A consensus secondary structure prediction server. *Bioinformatics (Oxford, England)*, *14*(10), 892–893.
- Dixit, A., Torkamani, A., Schork, N. J., & Verkhivker, G. (2009). Computational modeling of structurally conserved cancer mutations in the RET and MET kinases: The impact on protein structure, dynamics, and stability. *Biophysical Journal*, *96*(3), 858–874.
- Eisenberg, D., Lüthy, R., & Bowie, J. U. (1997). VERIFY3D: Assessment of protein models with three-dimensional profiles. *Methods in Enzymology*, *277*, 396–404.
- Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M. Y., . . . Sali, A. (2007). Comparative protein structure modeling using MODELLER. *Current Protocols in PROTEIN SCIENCE*, *50*(1), 2–9.
- Fiser, A. (2004). Protein structure modeling in the proteomics era. *Expert Review of Proteomics*, *1*(1), 97–110.
- Fiser, A. (2010). Template-based protein structure modeling. *Methods in Molecular Biology*, *673*, 73–94.
- Fiser, A., Do, R. K., & Sali, A. (2000). Modeling of loops in protein structures. *Protein Science: A Publication of the Protein Society*, *9*(9), 1753–1773.
- Fox, N. K., Brenner, S. E., & Chandonia, J. M. (2015). The value of protein structure classification information—Surveying the scientific literature. *Proteins*, *83*(11), 2025–2038.
- Grinspun, E., Krysl, P., & Schroder, P. (2002). Charms: A simple framework for adaptive simulation. *ACM Transactions on Graphics*, *21*(3), 281–290.
- Gromiha, M. M., Nagarajan, R., & Selvaraj, S. (2019). *Protein structural bioinformatics: An overview. Encyclopedia of bioinformatics and computational biology* (pp. 445–459). Elsevier.
- Gupta, A., Mohanty, P., & Bhatnagar, S. (2016). Protein structure prediction using homology modeling: Methods and tools, In *Methods and algorithms for molecular docking-based drug design and discovery*, 339–359.
- Hermoso Pulido, T. (2015). *Design and practical usage of web biological databases for the annotation and classification of proteins*. Universitat Autònoma de Barcelona.
- Huang, I. K., Pei, J., & Grishin, N. V. (2013). Defining and predicting structurally conserved regions in protein superfamilies. *Bioinformatics (Oxford, England)*, *29*(2), 175–181.
- Jamroz, M., & Kolinski, A. (2010). Modeling of loops in proteins: A multi-method approach. *BMC Structural Biology*, *10*, 5.
- Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992a). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences: CABIOS*, *8*(3), 275–282.
- Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992b). A new approach to protein fold recognition. *Nature*, *358*(6381), 86–89.
- Karplus, K. (2009). SAM-T08, HMM-based protein structure prediction. *Nucleic Acids Research*, *37*(suppl_2), W492–W497.
- Kim, D. E., Chivian, D., & Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research*, *32*(suppl_2), W526–W531.
- Lee, J., Wu, S., & Zhang, Y. (2009). Ab initio protein structure prediction. In D. J. Rigden (Ed.), *From protein structure to function with bioinformatics*. Dordrecht: Springer, https://doi.org/10.1007/978-1-4020-9058-5_1.
- Li, W., McWilliam, H., Goujon, M., Cowley, A., Lopez, R., & Pearson, W. R. (2012). PSI-Search: Iterative HOE-reduced profile SSEARCH searching. *Bioinformatics (Oxford, England)*, *28*(12), 1650–1651.
- Lundström, J., Rychlewski, L., Bujnicki, J., & Elofsson, A. (2001). Pcons: A neural-network-based consensus predictor that improves fold recognition. *Protein Science*, *10*(11), 2354–2362.
- Madden, T. (2013). *The BLAST sequence analysis tool. The NCBI handbook* (2nd ed.). United States: National Center for Biotechnology Information.
- Marchler-Bauer, A., Anderson, J. B., Cherukuri, P. F., DeWeese-Scott, C., Geer, L. Y., Gwadz, M., . . . Lanczycki, C. J. (2005). CDD: A Conserved Domain Database for protein classification. *Nucleic Acids Research*, *33*(Suppl. 1), D192–D196.
- McGuffin, L. J., Bryson, K., & Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics (Oxford, England)*, *16*(4), 404–405.
- Muhammed, M. T., & Aki-Yalcin, E. (2019). Homology modeling in drug discovery: Overview, current applications, and future perspectives. *Chemical Biology & Drug Design*, *93*(1), 12–20. Available from <https://doi.org/10.1111/cbdd.13388>.
- Nagarajan, N., & Yona, G. (2004). Automatic prediction of protein domains from sequence information using a hybrid learning system. *Bioinformatics (Oxford, England)*, *20*(9), 1335–1360.
- Nielsen, M., Lundegaard, C., Lund, O., & Petersen, T. N. (2010). CPHmodels-3.0—Remote homology modeling using structure-guided sequence profiles. *Nucleic Acids Research*, *38*(suppl_2), W576–W581.
- Nikolaev, D. M., Shtyrov, A. A., Panov, M. S., Jamal, A., Chakchir, O. B., Kochemirovsky, V. A., . . . Ryazantsev, M. N. (2018). A comparative study of modern homology modeling algorithms for rhodopsin structure prediction. *ACS Omega*, *3*(7), 7555–7566.
- Pearson, W. R. (2013). An introduction to sequence similarity (“homology”) searching. *Current Protocols in Bioinformatics*, *42*(1), 3.1.
- Peng, J., & Xu, J. (2011a). A multiple-template approach to protein threading. *Proteins*, *79*(6), 1930–1939.
- Peng, J., & Xu, J. (2011b). RaptorX: Exploiting structure information for protein alignment by statistical inference. *Proteins: Structure, Function, and Bioinformatics*, *79*(S10), 161–171.
- Raghava, G. P. S. (2002). APSSP2: A combination method for protein secondary structure prediction based on neural network and example based learning. *CASP*, *5*, A-132.
- Roberts, K., Alberts, B., Johnson, A., Walter, P., & Hunt, T. (2002). *Molecular biology of the cell*. New York: Garland Science.

- Rost, B., Sander, C., & Schneider, R. (1994). PHD-an automatic mail server for protein secondary structure prediction. *Bioinformatics (Oxford, England)*, *10*(1), 53–60.
- Rykunov, D., & Fiser, A. (2010). New statistical potential for quality assessment of protein models and a survey of energy functions. *BMC Bioinformatics*, *11*, 128.
- Schwede, T., Kopp, J., Guex, N., & Peitsch, M. C. (2003). SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Research*, *31*(13), 3381–3385.
- Scott, W. R., Hünenberger, P. H., Tironi, I. G., Mark, A. E., Billeter, S. R., Fennel, J., ... van Gunsteren, W. F. (1999). The GROMOS biomolecular simulation program package. *The Journal of Physical Chemistry. A*, *103*(19), 3596–3607.
- Shukla, H., Shukla, R., Sonkar, A., & Tripathi, T. (2017). 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., Shukla, H., Kalita, P., Sonkar, A., Pandey, T., Singh, D. B., ... Tripathi, T. (2018). 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., & Singh, T. R. (2020). Identification of small molecules against cyclin dependent kinase-5 using chemoinformatics approach for Alzheimer's disease and other tauopathies. *Journal of Biomolecular Structure & Dynamics*, 1–13.
- Singh, D. B., & Pathak, R. K. (2020). *Computational approaches in drug designing and their applications. Experimental protocols in biotechnology* (pp. 95–117). New York, NY: Humana.
- Singhal, S., & Singhal, S. (2018). *Bioinformatics*. Pragati Prakashan. ISBN: 978-93-87151-70-3.
- Söding, J., Biegert, A., & Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, *33*(Suppl. 2), W244–W248.
- 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.
- Suyama, M., & Ohara, O. (2003). DomCut: Prediction of inter-domain linker regions in amino acid sequences. *Bioinformatics (Oxford, England)*, *19* (5), 673–674.
- Tiwari, A., Chauhan, R. P., Agarwal, A., & Ramteke, P. W. (2020). Molecular modeling of proteins: Methods, recent advances, and future prospects. In D. B. Singh (Ed.), *Computer-aided drug design*. Singapore: Springer.
- Wang, Z., Eickholt, J., & Cheng, J. (2010). MULTICOM: A multi-level combination approach to protein structure prediction and its assessments in CASP8. *Bioinformatics (Oxford, England)*, *26*(7), 882–888.
- Wright, P. E., & Dyson, H. J. (2015). Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews. Molecular Cell Biology*, *16*(1), 18–29.
- Wu, S., & Zhang, Y. (2008). MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins*, *72*(2), 547–556.
- Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite: Protein structure and function prediction. *Nature Methods*, *12*(1), 7–8.
- Yang, Y., Faraggi, E., Zhao, H., & Zhou, Y. (2011). Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics (Oxford, England)*, *27*(15), 2076–2082.
- Zhang, Y., Arakaki, A. K., & Skolnick, J. (2005). TASSER: An automated method for the prediction of protein tertiary structures in CASP6. *Proteins: Structure, Function, and Bioinformatics*, *61*(S7), 91–98.
- Zhang, Z., Kochhar, S., & Grigorov, M. (2003). Exploring the sequence-structure protein landscape in the glycosyltransferase family. *Protein Science*, *12*(10), 2291–2302.
- Zimmerman, S. S., Pottle, M. S., Némethy, G., & Scheraga, H. A. (1977). Conformational analysis of the 20 naturally occurring amino acid residues using ECEPP. *Macromolecules*, *10*(1), 1–9.

Further reading

- Singh, D. B. (Ed.). *Computer-aided drug design*. Singapore. https://doi.org/10.1007/978-981-15-6815-2_7.
- Xiang, Z. (2006). Advances in homology protein structure modeling. *Current Protein and Peptide Science*, *7*(3), 217–227.