Molecular Modeling of Proteins: Methods, Recent Advances, and Future Prospects

Apoorv Tiwari, Ravendra P. Chauhan, Aparna Agarwal, and P. W. Ramteke

Abstract

The three-dimensional modeling of protein structure is a reliable approach for understanding the biochemical functions along with the dynamics of protein interactions, which provides useful applications in developing drug molecules for curing diseases as well as certain other applications in biological and agricultural sciences. The conventional laboratory methods such as NMR and X-ray crystallography which are standard approaches for analysis of different proteins of interest are labor-intensive, expensive, and time-consuming. To address these challenges, the bioinformatics tools and approaches may open up new avenues for investigating the protein structures and functions. In the recent past, molecular modeling has been successfully used in various projects for 3D structure prediction of some therapeutically important proteins having applications ranging from medicine to agriculture. The approach of molecular modeling is based on the understanding of algorithms of protein structure prediction. This chapter illustrates the salient features of molecular modeling methods for a reliable and accurate structure prediction of the proteins in the field of drug designing.
2.1 Introduction

Molecular modeling is an important approach in computational biology. For addressing the problems at the interface of biological and physical sciences, molecular modeling starting from its inception with applications of physics and computer science has been used rampantly by biological scientists. The advancements in computational biology have made molecular modeling a reality (Kumar and Chordia 2017). The current chapter discusses different methods available for protein three-dimensional structure predictions. As we know that proteins determine the biological functions of a cell and also are considered the building blocks of the cells. The dynamic processes including replication, maintenance, defense, and reproduction in living beings are encoded in the form of protein structures and functions (Berg et al. 2002).

There are 20 amino acids determined by the genetic codes. Proteins are the polypeptides that are made up of amino acids (Lodish et al. 2000). There are 20 regular amino acids in the configuration of the polypeptide chain, which determines the structure and function of the protein. Proteins are technically the end products that are decoded from cellular DNA information. Proteins are the main structural and transporter elements in a cell, which function as the workhorse of the cell unit and biocatalyst (Alberts et al. 2002). Interestingly, the functions that protein monitors in a cell are determined by genetic code. The protein structure and function depend on the genetic code encoded by DNA molecule(s), which is the building block of a gene. Different sequences of amino acids assemble to give rise to specific proteins with particular functions based on their three-dimensional configuration. The folded form or confirmation of a protein is directly dependent on the protein’s linear amino acid sequence (Hooft et al. 1996, 1997).

2.1.1 Amino Acids

Amino acids are the basic units of the proteins, and each amino acid has a variable side chain (Berg et al. 2002). Multiple amino acids combine to form a long chain that is tied together with a peptide bond. The biochemical reactions govern the peptide bond formation where water (H₂O) molecule is extracted by joining the NH₂ of one and COOH of a neighboring amino acid in the polypeptide chain. A simple linear sequence of amino acids in a protein is known as the primary structure (Alberts et al. 2002). Amino acids have both polar and non-polar side chains (Lodish et al. 2000). The polarity of the side chains determines the amino acid and protein structure or conformation. The hydrogen bonding involves polar side chains, while the charged
side chains of amino acid can participate in the formation of ionic bonds. Weak interactions are formed between the hydrophobic side chains to give rise to the van der Waals interactions. Cysteine is the only amino acid involved in the formation of disulfide covalent bonds, which is formed within or between two polypeptide chains that provide stability to the proteins (He et al. 2006; Neil and Bulleid Ellgaard 2011).

The 3D structure of a protein is governed by folding and intra-molecular bonding of the amino acids. The folding of the peptide chain is determined by hydrogen bonding between two neighboring amino acids. The specific patterns of this folding are termed as alpha helices and beta sheets, which are involved in the formation of a secondary structure of a protein (Voet and Vote 1990). Finally, multiple polypeptide chains join together and create the quaternary structure of a protein (Brown et al. 2003). The most energy-efficient and stable configuration of protein lies in its final form. A given protein attains its final form through a rigorous process of undergoing a variety of formations with the folding ability, which is unique and compact. A protein fold is stabilized by thousands of non-covalent bonds. The form and stability of proteins are determined by the chemical forces between a protein and its surrounding medium. Proteins inserted in the cell membranes have some hydrophobic chemical groups on their surface (Hooft et al. 1997).

2.1.2 Basic Principles of Protein Structure

The 3D structure of a protein is an atomic model of interaction of a large number of atoms (Wooley and Lin 2005). The 3D structure of protein represents a complex level of the group of atoms. In this type of arrangement, four different levels of protein structures exist, which are known as the primary, secondary, tertiary, and quaternary structures. Usually, two additional levels of intervention between secondary and tertiary structures are known as super-secondary structures and domains. The amino acid sequence itself does not directly encode disulfide bonds and other rare types of covalent bonds formed between side chains (Bailey 2018) (Fig. 2.1).

The secondary structure of protein results due to the folding of a protein sequence in a systematic form, and this fold is stabilized by the contribution of repetitive hydrogen bonding (Fig. 2.2). For the first time, Linus Pauling and Robert Corey described the chemical nature and secondary structures of proteins. The secondary structure includes alpha helix (α-helix) on the right, parallel, and antiparallel beta (β-) plated sheets and turns (Serafini 1989).

Tertiary structure is formed by a stable and compact packing of elements of secondary structure (Breda et al. 2008). The folding results in the complete three-dimensional polypeptide structure, which depends on the sequence of amino acids and atomic details (Berg et al. 2002). However, this process leads to the formation of a hydrophobic core for soluble proteins that are represented by the polar residues, which maintain its side chains inside the protein, and as a result, hydrophilic residues get exposed to the solvent. There are different types of protein folds available in nature, and two of the most common protein fold classification are SCOP and CATH (Csaba et al. 2009). The tertiary structure view for nsp10/nsp16 complex of SARS
coronavirus has been shown in Fig. 2.3. The nsp16 causes sequence-dependent methylation and for successful methylation of viral mRNAs, nsp16 requires the interaction of nsp10 to initiate its methyltransferases activity (Chen et al. 2011). The nsp16 may serve as a potential drug target against corona disease. A potential inhibitor of nsp16 can serve as a therapeutic agent for the treatment of this disease.

The quaternary structure of the protein is formed by two or more identical or different polypeptide chains. Since two or more subunits are present, hence such proteins are termed as oligomers. The characterization illustrating how subunits of the native protein are arranged is based on the quaternary structure. The oligomeric subunits are held together by non-covalent forces, and therefore, can undergo rapid transformations affecting the biological activity of the protein. Hemoglobin, allosteric enzymes, actin, and tubulin are some examples of oligomeric proteins (Berg et al. 2011).

**Fig. 2.1** Primary structure of a protein

**Fig. 2.2** Secondary structure of the protein
The protein function depends on the manner how the protein surface interacts with the other molecules through bonding. Structurally similar proteins similarly interact with certain molecules and thus, represent a protein family (Alberts et al. 2002). The structural similarity of proteins within a family contributes a similar function to the family. Proteins with similar amino acid sequences belong to the same protein family, and their protein sequence remains conserved during evolution. Proteins with similar functions have a similar set of amino acid residues that interacts and binds with the substrate or signal (Cohen et al. 2009).

### 2.2 Explosion of Protein Related Data

Many databases are available with the sequence level information of the protein. Protein Data Bank (PDB) is an important database of protein structure. Currently, PDB has become the most popular protein databank with an archive containing more than 1,60,000 structures determined by different experimental methods like nuclear magnetic resonance (NMR) spectroscopy, crystallography, nuclear and electron cryo-microscopy (3D-EM), etc. PDB data and resources are very useful in the development of an experimental method, training, and testing of predictive models and drug discovery projects (Horiuchi et al. 2000). In recent years, different database resources have been developed which provide information about the classification and clustering of proteins, structural characterization, localization, phosphorylation, family and domain, active site, binding related information, protein disorder, conformational diversity, pathway, structure, function, etc. (Burley et al. 2017).

**Fig. 2.3** Tertiary structure view: nsp10/nsp16 complex of SARS coronavirus (PDB: 3R24)
2.3 Protein Structure Determination

Several methods, including X-ray crystallography, NMR spectroscopy, and electron microscopy, are currently used to determine the 3D structure of a protein, and each method has its uniqueness and limitations (Kendrew et al. 1958). The user has access to several pieces of information for each of these methods to generate the final atomic model (Wang and Wang 2017) and this could be due to the X-ray diffraction pattern in X-ray crystallography (Callaway 2015). In electron microscopy, this could be the image representing the shape of the molecule. However, only experimental information is not enough to build an accurate atomic model, additional molecular structural knowledge is often required is the amino acid sequence of a protein along with geometrical features such as bond lengths and bond angles (Rankin et al. 2014). The creation of a consistent protein model requires a set of experimental data and modeling related parameters (Carroni and Saibil 2016).

2.3.1 X-Ray Crystallography

For X-ray crystallography, the protein is purified and crystallized under suitable conditions, and then subjected to an intensive X-ray beam (Burley et al. 2019). The diffraction of an X-ray beam by the protein crystals into one or other patterns is examined to determine the distribution of electron density. Finally, a map of the electron density is generated and interpreted to determine the location of each atom. The electron density map is analyzed to locate the position of each atom in 3D space. X-ray crystallography is a powerful tool that can provide coordinate information of each atom, which can illustrate the position of each atom in a protein. It is a challenging method with limitations on certain proteins but an excellent method to study and determine the structures of rigid crystals (Wang and Wang 2017). It is difficult to study the structure of a flexible protein using X-ray crystallography. The accuracy of this method depends on the quality of the crystals used for X-ray crystallography. Resolution and R-value are the two important parameters used to represent the accuracy of crystallographic structure (Haywood 1997).

X-ray crystallography is the most practical method for determining the structure of the biomolecules. Some of the salient features that this method offers include:

1. Accuracy of models for atomic resolution, also the method enables the user to solve relatively large structures and complexes.
2. Different solvents crystallize the same protein into different conformations. Thus, the method facilitates the study of the whole mechanism mediated by a single protein using XRD. Prominent examples include viral capsid structures and the ribosome, each made up of tens of thousands of atoms (Haywood 1997).

X-ray crystallography can resolve the structure of protein and protein–ligand complexes with good accuracy. But this method has some major limitations as given below:
1. The data provide only one protein position or confirmation but not dynamic behavior.
2. Contacts between molecules of crystal and the dense packaging can affect the structures.
3. The procedure is time-consuming.
4. Challenging for the analysis of highly hydrophobic or flexible proteins.
5. Difficulty in determining hydrogen positions, which requires very high resolution, thus, unfortunately, limits the reliability of this method.

### 2.3.2 NMR Spectroscopy

NMR spectroscopy is used to determine the 3D structure of molecules. However, during this procedure, the molecule should be pure and placed in a robust magnetic field. A distinguishable set of measured resonances can be analyzed to provide a list of nearby atomic nuclei to describe the composition of atoms that are linked together (Serdyuk et al. 2007). This list of restraints is used for model creation that indicates the location of each atom. NMR spectroscopy is used for determining the structure of proteins in solution and requires aqueous crystal. It is the first method used for the study of flexible protein structures (Snyder et al. 2005). A typical NMR structure includes a set of protein structures that are consistent with the list of experimental restraints observed.

The advantages of NMR spectroscopy are as given below:

1. The sample does not need to be in crystalline condition (which limits the applications of X-ray crystallography).
2. It provides better dynamics of the molecule.

There are some limitations associated with NMR spectroscopy that are mentioned here:

1. The technical accuracy of NMR is less as compared to X-ray crystallography.
2. MNR generates a good result for small size molecules, proteins with less than 300 residues.

### 2.3.3 3D Electron Microscopy

Electron microscopy is also used to determine the 3D structures of large molecular assemblies, often referred to as 3DEM. A beam of electrons and an electron lens system are used to directly image the biomolecule (Orlova and Saibil 2011). In a limited number of cases, electron diffraction from 2D or 3D crystals can be used for determining the 3D structures with an electron microscope (Rabl et al. 2010). Finally, 3DEM techniques are in advance importance for studying the biological assemblies inside the cryo-preserved cells and tissue by using electron tomography.
In terms of molecular and atomic details, both single-particle 3DEM and electron diffraction methods now provide structures with resolution limits comparable to macromolecular crystallography (i.e., enabling visualization of amino acid side chains, surface water molecules, and non-covalently bound ligands). Cryo-electron tomography provides slightly lower resolution structural information (protein domains and structural elements). In the calendar year 2016, the PDB deposition of the 3DEM structures for the first time exceeded that of NMR spectroscopy (Jonic 2016).

To investigate the very large macromolecular assemblies where lower resolution is normal, 3DEM data are increasingly being combined with information from X-ray crystallography, NMR spectroscopy, mass spectroscopy, chemical cross-linking, fluorescence resonance energy transfer, and various computational techniques to sort out the atomic details. This practice of using multiple experimental approaches is referred to as integrative or hybrid methods (I/HM) (Jonic 2016). This approach of integration has been proved much useful for investigating multimolecular structures such as complexes of ribosomes, t-RNA, protein factors, and muscle actomyosin structures, among others. A prototype data repository known as PDB-Dev, operating in parallel with the PBD, is now available for archiving of I/HM structures and data (Dever and Green 2012; Masters and Beyreuther 1998).

Structural resolution is the main limitation of EM. The EM resolution is approximately 3.5 Å, which is not enough to determine the location of side chains (Wohlgemuth et al. 2008). The advantages of electron microscopy are given below:

1. EM can solve very large biological complexes that are not accessible to crystallography with X-rays.
2. It can be used as a reference for interpreting X-ray diffraction patterns.
3. EM structures and X-ray data can be combined for determining the structure of large molecules.

### 2.4 Protein Structure Prediction

Three computational methods widely used for protein structure prediction are (1) homology modeling, (2) fold recognition, (3) ab initio method. Several tools for protein structure predictions are available, which utilizes different approaches and methods for modeling and refinement of the protein structure (Table 2.1).

#### 2.4.1 Homology or Comparative Modeling

Comparative modeling is a template based modeling consists of five main steps: (a) identification of similar sequences with known structure, (b) alignment of the target sequence with template structures, (c) modeling of structurally conserved regions using templates, (d) modeling of lateral chains and loops, (e) the quality of the model being refined and evaluated by conformational sampling (Fig. 2.4). The
| S. No. | Name       | Description                                                                 | Weblink                                                                 |
|-------|------------|-----------------------------------------------------------------------------|------------------------------------------------------------------------|
| 1     | RaptorX    | Remote homology detection, protein modeling, prediction of the binding site | http://raptorx.uchicago.edu/                                           |
| 2     | ESyPred3D  | Template search, better performance of alignment, used for 3D modeling based on homology | https://www.unamur.be/sciences/biologie/urbm/bioinfo/esypred          |
| 3     | FoldX      | Energy calculations and protein modeling integrate protein fragment libraries of different length | http://foldxsuite.crg.eu                                               |
| 4     | Geno3D     | Comparative protein modeling uses geometrical restraints (dihedral angles and distances) | https://geno3d-prabi.ibcp.fr                                           |
| 5     | HHpred     | Homology detection and template search, alignment, 3D modeling, based on hidden Markov model (HMM) | https://toolkit.tuebingen.mpg.de                                      |
| 6     | LOMETS     | Local meta-threading server for tertiary structure prediction               | https://zhanglab.ccmb.med.umich.edu                                   |
| 7     | Modeller   | Comparative protein modeling, satisfaction of spatial restraints, optimization of various models of a protein | https://salilab.org                                                   |
| 8     | MOE        | Molecular operating environment (MOE), template identification, use of multiple templates, loop modeling, side-chain modeling by rotamers | https://www.chemcomp.com                                              |
| 9     | Prime      | Homology modeling, assessment, and refinement of the generated model, based on the energy function | https://www.schrodinger.com/prime                                     |
| 10    | ROBETTA    | Homology modeling and ab initio fragment assembly can model multi-chain complexes | http://robetta.bakerlab.org/                                           |
| 11    | Swiss model| Homology modeling based server, template search, local similarity/fragment assembly, model evaluation by comparison to X-ray verified structures | https://swissmodel.expasy.org/                                        |
| 12    | I-TASSER   | Combination of ab initio folding and threading methods, template-based fragment assembly | https://zhanglab.ccmb.med.umich.edu                                   |
| 13    | NovaFold   | Combination of threading and ab initio folding based on iterative assembly simulations, also predicts ligand binding site | https://www.dnastar.com/manuals/protein3d/15.3/en/topic/run-nova-applications-through-the-dnastar-website |
accuracy of comparative modeling predictions depends on the degree of sequence similarity between template and target. If the target and template sequence share a high degree of similarity, then the accuracy of the predicted model is very high. For a sequence identity of 30–50%, more than 80% of C-α atom is expected to be within 3.5 Å of their true positions while significant errors are likely to occur for less than 30% of sequence identity (Rodriguez et al. 1998; Krieger et al. 2003). Comparative modeling is based on the principle that similar evolutionary sequences have similar three-dimensional folded structures.

The goal of protein modeling is to predict the structure of a target protein from its sequence information using the related known structure as a template. This would enable the rapid use of in silico protein models in all fields, such as structure-based drug prediction, protein function investigation, network analysis, antigenic behavior, and protein structure with increased soundness or novel capacities. In the case where experimental strategies are limited, one of the best ways to obtain the auxiliary data is protein modeling. Several proteins are very large and insoluble and hence cannot be studied by NMR and X-ray diffraction. Homology modeling is one of the easiest approaches to the 3D structure prediction elaborated in this chapter (Peach et al. 1994; Blundell et al. 1987).

If we have to know the structure of any protein that contains 200 amino acids, we use the BLAST tool of NCBI to compare the sequence of this protein in the PDB database and fortunately, we found a structure B with a total of 400 amino acids which aligns 50% identical residues with structure A. In this case, we can regard structure B as a template and structure A as a target, so that we can model the protein using structural information of template protein B. The homology modeling can only be used to model 3D structures of a target protein if it shares more than 30% sequence similarity with the template (Sanchez and Sali 1997; Peitsch 1997). Homology modeling is a multi-step process that includes template search, database searches, sequence alignment, structural refinement, loop search, side-chain modeling, coordinate assignment, energy minimization, and structure validation to create a quality structure (Johnson et al. 1994).

The modeler cannot make the finest protein structure, and therefore the main task of the modeling process involves genuine thinking as to how to play between

| S. No. | Name     | Description                                                                 | Weblink                                      |
|-------|----------|------------------------------------------------------------------------------|----------------------------------------------|
| 14    | MUSTER  | Threading algorithm, sequence-template alignments by profile-profile alignment | https://zhanglab.ccmb.med.umich.edu/MUSTER/  |
| 15    | EVfold   | Calculates evolutionarily coupled residues, evolutionary couplings calculated from correlated mutations in a protein family, ranks the models on geometric criteria and clustering | http://evfold.org/                           |
Fig. 2.4  A sequence of steps involved in homology modeling
different considerably similar choices. To construct the homology models, considerable efforts and research have been done to train the computational models to make better decisions (Peitsch et al. 2000). A sequence of steps involved in homology modeling is discussed here.

2.4.1.1 Template Recognition and Initial Alignment
The percentage identity between target protein sequence and the template is calculated using searching tools such as BLAST (Altschul et al. 1990) or FASTA (Pearson 1990). Two main matrices are used to identify these hits by comparison of the query sequence to all the known structures.

1. A matrix of residue exchange: The probability of alignment of two of the 20 amino acids is determined by the elements of this matrix.
2. A matrix of alignment: The two aligned sequences are represented by the axes of this matrix.

One needs to feed the query sequence to BLAST servers available on the web, followed by the selection of the PDB database for search. Finally, a list of templates and their alignment score with the query or target protein is received.

2.4.1.2 Alignment Correction
Several templates can be used for modeling. However, it is time to acquire better alignment using more sophisticated methods. It is difficult to model regions with a low percentage of sequence identity. Another sequence of homologous proteins can be used to find a suitable solution. Multiple sequence alignment programs such as CLUSTAL W can align several related sequences (Thompson et al. 1994) and a huge amount of data can be retrieved from the resulting alignment. If only exchanges between hydrophobic residues are observed at a certain position, then there are high chances of the residue being buried. Position-specific scoring matrices referred to as profiles are derived by multiple sequence alignments (Taylor 1986). Insertions (additional residues in the model) or deletions (missing residues in the model) can be achieved by multiple sequence alignments merely at the places where we find quite different sequences.

2.4.1.3 Modeling Structurally Conserved Region (SCR) and Backbone Generation
The important step in homology modeling is to determine the regions that are structurally conserved among the structures related to templates. Structurally conserved region (SCR) or core is determined by computing the C-alpha distance matrix for each structure and then small portions of the distance matrix are compared to find the peptide segments with lower root mean squared difference (RMSD) for related structures. In this way, all SCRs are determined and related SCRs share very high sequence and structural similarity. The generation of the model starts with the alignment of target and template protein. Template-target alignment indicates the residue blocks in the target which corresponds to SCRs of template. Coordinates of
the amino acid residues of the template for structurally conserved regions between the template and target protein are taken from the template structure and assigned to the target model. There may be some varying residues within SCR region of aligned template and target, and if only their side chains differ, then the backbone (N, Ca, C, and O) coordinates of these residues are copied template and assigned to target model. Experimental protein structures are better than modeled. Choosing the template with the fewest errors is a simple way to build a good model. But what if there are two templates and each with a region that is poorly determined, but these regions are not the same, and then both templates can be used for model building using multiple templates approach. This approach is also used if there are good matches in different regions between alignments between the target sequence and templates. Multiple template modeling is done by servers like Swiss model (Peitsch et al. 2000).

2.4.1.4 Loop Modeling

During template-target alignment, gaps occur between the aligned model and the template sequence. These gaps represent the insertion/deletion between template and target. The structural fold of these gap residues or loop needs to be determined and incorporated in between the two conserved core region. It requires modification of the backbone. In the regular secondary structural elements, orientation or conformational changes are not found. Thus, it is safe to remove all the insertions or deletions within the alignment form helixes and strands and put them in turns and loops. We frequently realize different loop configurations within the template and target even while not insertions or deletions. There are the following reasons behind this problem (Krieger et al. 2003):

1. Surface loops lead to a major modification within the conformation of the template, and therefore the target.
2. Beneath the loop, the exchange of little to large side chain pushes it away.
3. The mutation of proline or glycine to the other residue in the loop.

In all cases, the residue must be placed in the loop considering the Ramachandran plot.

Two main approaches used for modeling the loop region are given here.

Knowledge-Based

Here, we search for the structure of the loops region with endpoints reminiscent from the known structures, and then the coordinates of loop structure are placed in between two cores. Most of the molecular modeling programs such as Modeller (Sali and Blundell 1993), Insight (Dayringer et al. 1986), Swiss model (Peitsch et al. 2000) or 3D-Jigsaw (Bates and Sternberg 1999) support knowledge-based approach for loop modeling.
Energy-Based
The energy function is employed to assess the loop quality and uses Monte Carlo or molecular dynamics simulation methods (Fiser et al. 2000) to generate the most accurate loop form. The energy function can be modified to generate a better loop structure that can best fit in the core (Tappura 2001). For small loops (up to 5–8 residues), the various strategies are available to predict a loop configuration that well overlaps the important structure.

2.4.1.5 Side-Chain Modeling
During the coordinate assignment in core modeling, coordinates of all amino acids are copied from template to target except those amino acids where side chain differs. At the position of the varying side chain, only the backbone coordinate of amino acid is assigned to target, and the related side chain is modeled using rotamer libraries. Rotamer libraries contain the biologically active conformation of side chains for different amino acids. As we know that all conformation of an amino acid is not biologically active, so it becomes important to determine and place the correct conformation of the side chain (Sanchez and Sali 1997).

It uses rotamers libraries derived from high-resolution X-ray structures. These rotamers are validated with a range of energy functions for their fitness (Scouras and Daggett 2010). The selection of an explicit rotamer mechanically affects the rotamers of all near residues. With a 100 residues and a median of five rotamers per residue, 5100 different mixtures would be scored already. There has been a great deal of analysis into developing strategies to create this vast search space traceable (Desmet et al. 1992). For a given backbone configuration, just one powerfully inhabited rotamer may be modeled immediately, so providing an anchor for additional versatile side chains within the surroundings. There are mainly two reasons for low prediction accuracy.

1. Flexible side chains on the surface can form several conformations.
2. Rotamers in hydrophobic packaging in the core can be easily scored, but ionic interactions on the surface, hydrogen bonds with water, and related entropic effects are challenging (Sliwoski et al. 2014).

It is vital to notice that in nearly all publications, the prediction accuracy cannot be achieved in real-world applications. The algorithms, therefore, accept the proper backbone that is not offered within the modeling of homology. The template’s backbone is commonly significantly different from the target (Fiser 2010). The rotamers should be predicted based on the wrong backbone, and the predictive accuracy, in this case, tends to be lower.

2.4.1.6 Model Optimization
The right backbone is needed for the prediction of high-precision side-chain rotamers, which relies on the rotamers and their packaging. The main approach to a tangle of this kind is the re-iterative prediction of rotamers, then the ensuing backbone shifts, and the new backbone rotamers until the process converges. This
method reduces the series of rotator predictions and steps of energy reduction (Hansen and Kay 2011).

The methods described above are used not only in the loop modeling but also for model optimization and should be applied for the entire protein structure (Hintze et al. 2016). At each minimization step, a few major errors, such as bumping due to atomic clashes, are eliminated, whereas several tiny mistakes are created. When small errors begin to accumulate, the model becomes less accurate. Better optimization of a model can be achieved by more accurate energy functions for force field calculation. Precision can be achieved by using the following approaches.

**Quantum Force Fields**
The force field calculation method should be fast and efficient to cover large protein molecules. The recent advancement in computational biology enabled methods of quantum chemistry to attain a more accurate interpretation of the charge distribution for the whole protein molecule (Liu et al. 2001).

**Self-Parameterizing Force Fields**
Force field accuracy depends on its variables (e.g., atomic loads and van der Waals radii). These variables are generally derived from small molecules quantum chemical calculations and fitting of these values to experimental data (Krieger et al. 2002). This method results in a rather expensive computer procedure. Take starting parameters for the force field, modify the parameter, minimize the energy of models, and save the new force field if the quality of the model improved otherwise return to the previous parameter of force field. This approach can increase the accuracy of the force field in the correct direction during energy minimization. A protein model can be optimized using molecular dynamics simulation, and it samples trajectory of the motions of the protein at a duration of 10 fs and generates the true folding dynamics of the protein (Adcock and McCammon 2006). It is therefore considered that during the simulation, the model will approach to real structure (Hospital et al. 2015).

### 2.4.1.7 Model Validation

Each model generated by homology contains some errors. The number of errors (for one particular method) depends primarily on two points.

1. A highly accurate protein model can be generated if the target shares very high similarity with the template. If the accuracy of the model is greater than 90%, then it can be compared accurately equivalent to an X-ray determined structure (Chothia and Lesk 1986; Sippl 1993). If the sequence identity between template and target lies below 25%, then alignment becomes meaningless for homology modeling, and the resulted model may have a high error.
2. There might be some errors in the template structure, which may result in the modeled protein. Structural errors can be estimated by the following methods:
   (a) Force field-based methods evaluate the bond angles, bond lengths, and bumps within the atoms. Lower energy models do not guarantee for the
accuracy of protein structure because sometimes the misfolded inaccurate models also achieve the low energy folds (Novotny et al. 1988).

(b) Normality indices can be used to compare the feature in the model that resembles the real structures. Many characteristics of protein structures are suitable for the analysis of normality. Many of these are based directly or indirectly on interatomic distance and contact analysis. It is important to observe the normality of torsion angles, bond angles and bond lengths and quality parameters of determined structures, but are less appropriate for model assessment (Czaplewski et al. 2000; Morris et al. 1992). Polar and non-polar residue distributions inside/outside can be used to predict misfolding in the protein model (Baumann et al. 1989). Most of the methods used to verify models can be applied to X-ray and NMR verified structures.

### 2.4.2 Fold Recognition or Threading Method

This method of prediction is used when there exists a low degree of similarity between template and target sequence as we cannot proceed for homology modeling due to low similarity (>30%). There is still no complete understanding of the relationship between the sequences, structure, and function. The only reliable fold prediction tools are currently the analogy based prediction algorithms. The threading approach is able to identify the most distant homologs and unrelated proteins with similar structures in some cases (Jaroszewski et al. 1998). The main challenge in the field of fold recognition is to develop tools to comply with structure, function, and analysis (Pruisner 1996) (Fig. 2.5).

The threading approach is used when the similarity between the target and template lies below 30% (Hendlich et al. 1990; Sippl 1993). In such cases, homology modeling may not generate a reliable model, so it is necessary to consider detailed structural parameters in the alignment (Jones et al. 1992). Threading methods consider structural information that is missed in the alignment process by sequence

![Fig. 2.5](image_url)  Prediction of potential structure through threading approach
comparison. Structural details can be included in various ways (Bowie et al. 1991). The 3D profile is another method used for threading, which is based on the structural environmental class of each amino acid residue and generates a matrix for the probability of each amino acid to stay in each environmental class (Shi et al. 2001; Bowie et al. 1991). Each amino acid has a probability to reside in a particular environmental class.

Another approach calculates the contact residue potential of the pair and maximizes the hydrophobic core score. This method identifies the core of the protein structure that is essential for maintaining the structural integrity. This can be done directly, including contact potentials in pairs (Bowie et al. 1991; Jones et al. 1992). Threading is based on environmental class and uses a dynamic programming algorithm but has some limitations related to the preservation of environmental class. Contact residue potential approach considers the formation of hydrophobic core on contacts between hydrophobic residues. Threading utilizes two approaches: (1) profile of structural environmental classes (2) the contact potentials directly in pairs.

In the 3D profile method, a template structure is represented as a descriptor string that describes the structural environment. There are three basic environment classes; (1) area of the lateral chains buried by other protein atoms, (2) fraction of the lateral chains covered by polar atoms, and (3) secondary local structure. Here, a 3D protein structure is represented in the form of an ID string, which represents each residue’s environmental class in the folded protein structure. The environment of a side chain is first classified as buried (B), partially buried (P) or exposed (E) depending on the area exposed to solvent. The buried and partially buried residue environments are further subdivided into P and Pi and B, Bi, B3, respectively (Peterson et al. 2014). The E, P, Pi, B, Bi, and B3 are the basic six environmental classes. In this way, there are a total of 18 environmental classes for all three secondary structures: helix, sheet, and coil. The 3D-ID scoring table where the pairing residue score i is given as follows with the environment j. P (i, j) represents the probability of amino acid residue i in environment j, and Pi is the overall probability of amino acid residue i in any environmental class (Ihm 2004). Here, the 3D-1D scoring table is used to generate the profile of a template structure. This is also known as sequence-structure because in this approach target and template both are represented in the form of string. The target protein is represented as a string of amino acids and the template structure represented as a string of the environmental classes (Jones 1999). The fitness score between the target and template environment class is calculated using a dynamic programming algorithm.

I-TASSER, a Yang Zhang Lab structure prediction threading method primarily identifies the structural template or fragment from the PDB subset using multiple threading approaches. Second, the initial conformations generated from the templates replica-exchange Monte Carlo simulations to produce a large number of reduced models. Third, all the models are grouped by SPICKER43 and the centroid cluster is formed by averaging the coordinates of each cluster of all decoys. Fourth, the simulation of the fragment assembly is carried out again starting from the selected cluster centroids of the cluster. Fifth, FG-MD reconstructs and refines the
all-atom structures. Finally, five full-length atomic models are produced along with models of approximate accuracy. As a comprehensive process, even for new fold targets, I-TASSER performs pretty well (Yang et al. 2016).

2.4.3 Ab Initio Methods

This approach generates a protein model from sequence information due to the unavailability of structural counterparts or structural folds. The ab initio method enables us to understand the physicochemical principle related to the nature of proteins. The accuracy of ab initio modeling is low as compared to other methods of structure prediction (Simons et al. 1999). If the target sequence does not share structural similarities with structures in the database, then protein structure can be generated by determining the configuration space of atoms in amino acids. This method utilizes the knowledge of various principles of physics, chemistry, and mathematics. The use of reduced protein representations makes the computation easy. Some of the models represent a residue using only two locations, such as backbone one and side chain (Cohen et al. 2009). Others use several sites, including heavy backbone atoms and a side link. The main driving force for protein folding is known to be hydrophobic interaction, and there is some empirical energy function for the calculation of interactions in protein. For the ab initio prediction, three factors must be established: (1) reduce the representation of proteins, (2) a potential energy function for interaction, (3) method for searching the conformation space.

Simulated annealing is used for searching the configuration space of the fragment structures. A move is taken to replace the torsion angles of a randomly selected neighbor in a randomly selected position with the current configuration. Movements that bring two atoms closer within 2.5 Å are discarded, and other movements are evaluated. Baker and colleagues predicted the tertiary structure of a protein using the sequence information of amino acid, and no template details were considered.

Most structure prediction methods currently depend on the information provided by the structures predicted by experimental methods, which is not much supportive in exploring the basic law of protein folding. Template-free methods are guided by the practical application as well also consider the fundamental principles of protein folding. Template-free methods are based on information from known structures, their development may better reflect the prediction of the theoretical and technical level of protein structure than template-based methods. ROSETTA one of the efficient ab initio modeling approaches (Han and Baker 1995; Shortle et al. 1998) is a template-free method created by the David Baker Lab, assembling a complete structure based on fragments of 3–9 residues from PDB. Similar to the template-based methods, the selection of fragments is based on the similarity between known and predicted secondary structure. Monte Carlo method simulates the assembly process with the annealing search technique. QUARK is an incredible Yang Zhang Lab fragment assembly tool. The fragments used for QUARK vary from 1 to 20 residues and the simulation of the assembly is performed under the guidance
from a knowledge-based atom level force field by Monte Carlo replica-exchange simulation.

Many other approaches, including Scrape, PROFESY, FRAGFOLD, etc. are also based on fragment assembly. The main distinction between these approaches and the template-based approaches is that they are not based on any global structural blueprint and they also do not utilize homology or structural similarities between the target and the proteins from which the fragments derive. It is more capable of modeling the target of new folds for template-free methods. However, due to the high computational requirement and low force field accuracy, it is still a major challenge for template-free methods for modeling proteins with a length of >150 residues. Prediction of contact map based on a co-evolution approach has recently shown progress to break down such a length limit of ab initio structure folding.

2.5 Evaluation and Validation of Modeled Structure

The final predicted model has to be evaluated to make sure that the structural features of the model are consistent with the physicochemical rules. This involves checking anomalies in \( \phi-\psi \) angles, bond lengths, close contacts, and so on. Another way of checking the quality of a protein model is to implicitly take these stereochemical properties into account. This is a method that detects errors by compiling statistical profiles of spatial features and interaction energy from experimentally determined structures. By comparing the statistical parameters with the constructed model, the method reveals which regions of a sequence appear to be folded normally and which regions do not. If structural irregularities are found, the region is considered to have errors and has to be further refined.

SAVEs server is a set of programs, which offers to check the model accuracy by just uploading the predicted structure. Procheck program is one of them that is able to check general physicochemical parameters such as \( \phi-\psi \) angles, chirality, bond lengths, bond angles, and so on. The parameters of the model are used to compare with those compiled from well-defined, high-resolution structures. If the program detects unusual features, it highlights the regions that should be checked or refined further. WHATIF is another comprehensive protein analysis server that validates a protein model for chemical correctness. It has many functions, including checking of planarity, collisions with symmetry axes (close contacts), proline puckering, anomalous bond angles, and bond lengths. It also allows the generation of Ramachandran plots as an assessment of the quality of the model.

Atomic non-local environment assessment (ANOLEA) is a web server that uses the statistical evaluation approach. It performs energy calculations for atomic interactions in a protein chain and compares these interaction energy values with those compiled from a database of protein X-ray structures. If the energy terms of certain regions deviate significantly from those of the standard crystal structures, then it suggests that the corresponding region has not been modeled correctly. The threshold for unfavorable residues is normally set at 5.0. Residues with scores above 5.0 are considered regions with errors. Verify3D is another server using the
statistical approach. It uses a pre-computed database containing 18 environmental profile based on secondary structures and solvent exposure, compiled from high-resolution protein structures. To assess the quality of a protein model, the secondary structure and solvent exposure propensity of each residue is calculated. If the parameters of a residue fall within one of the profiles, it receives a high score, otherwise a low score. The result is a two-dimensional graph illustrating the folding quality of each residue of the protein structure. The threshold value is normally set at zero. Residues with scores below zero are considered to have a non-favorable environment.

The assessment results can be different using different verification programs. Although the full-length protein chain of this model is declared favorable by ANOLEA, residues in the C-terminus of the protein are considered to be of low quality by Verify3D. Because no single method is clearly superior to any other, a good strategy is to use multiple verification methods and identify the consensus between them. It is also important to keep in mind that the evaluation tests performed by these programs only check the stereochemical correctness, regardless of the accuracy of the model, which may or may not have any biological meaning. Some tools predict the accuracy of the predicted model based on the Ramachandran plot of amino acid residues. It is a two-dimensional scatter plot showing torsion angles of each amino acid residue in a protein structure. The plot delineates preferred or allowed regions of the angles as well as disallowed regions based on known protein structures. This plot helps in the evaluation of the quality of a new protein model.

### 2.6 Recent Advances in Prediction Approaches

Two types of qualitatively different approaches for structural modeling are available: comparative modeling and de novo methods. Comparative modeling uses structural templates, while de novo methods model the protein without the detail of structural templates. The computational assessment of structural prediction (CASP) categorizes targets into two groups: (1) template-based modeling (TBM) and (2) free modeling (FM) (Moult et al. 2018).

Several parameters have been considered to improve the accuracy of prediction by TBM. PSI-BLAST and the profile-to-profile alignment methods have improved the accuracy of template identification and alignment. A composite structure assembly simulation utilizes the information from multiple templates, which refined the individual templates to be more similar to the native structures (Yang et al. 2015). In recent years, the availability of vast experimental sequence and structural databases has made it easier to get close homology templates for a target sequence. FM approaches are also referred to as ab initio or de novo structure prediction. Fragment assembly is also a type of FM approach (Dukka 2017). Some recent advances in the FM approach have been made, which considers the evolutionary constraints, contact information, and correlated mutation in scoring functions to improve the accuracy of the predicted protein model.
2.7 Applications

Accessibility of protein 3D structures and other structural analysis tools is facilitating the integration of an immense amount of information, which could be useful to further explore the possibilities to strengthen understanding of protein structure and function in the future. The 3D structures of a protein provide a better insight into the binding site and other functionally important regions, which could be utilized for drug designing. The availability of a target structure is the prerequisite condition to proceed for structure-based drug designing, and it also guides the changes in lead molecules. A 3D complex structure of a protein with a ligand provides better information about the residues involved in the interaction. Interaction of a receptor with a small molecule explains the mechanism of the pharmacological activity of a drug, binding affinity, and lead modification.

Computational modeling also explains how the mutation of an amino acid causes loss of function by destroying the native structure of protein required for its normal function. It can also explain the mechanism of drug resistance by depicting structural changes in the mutant target protein, which causes loss of proper binding or interaction of a drug with the target protein. Numerous forces such as hydrophobic interaction, Van der Waals, hydrogen bonding, and electrostatic are involved between the protein–ligand complexes to provide stability. Modeling of intermolecular connections in the protein–ligand complex is a very complex process due to a large number of degrees of freedom and inadequate information related to the impact of water on the binding.

2.8 Conclusion

Molecular modeling has turned into a significant and fundamental approach available to restorative scientific experts in the field of drug designing. Molecular modeling reveals the three-dimensional structures of proteins to unravel its related physicochemical properties. The protein modeling makes efficient use of computer science algorithms, theoretical science principles, and experimental information to uncover the structural and biological properties of a macromolecule. The selection of different tools for protein structure prediction depends on the nature of the problem to be addressed. This chapter has described the basic approaches and the recent advancements in methods of protein structure prediction. Required emphasis has also been given to the type of errors that may emerge and accumulate during protein modeling work. The development of a highly accurate and precise modeling tool should be the prime necessity as the applications of X-ray crystallography or NMR spectroscopy which being time-intensive do not seem to be the practical approaches for determination of the structure of every given protein. The accuracy of protein structure prediction is a crucial step because it is the basis of structure-based drug designing.

Competing Interest The authors declare that there are no competing interests.
References

Adcock SA, McCammon JA (2006) Molecular dynamics: survey of methods for simulating the activity of proteins. Chem Rev 106(5):1589–1615

Albers B, Johnson A, Lewis J, Walter P, Raff M, Roberts K (2002) Protein function in: molecular biology of the cell, 4th edn. Garland Science, New York

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410

Bailey R (2018) Learn about the 4 types of protein structure. Thought Co. thoughtco.com/protein-structure-373563

Bates PA, Sternberg MJ (1999) Model building by comparison at CASP3: using expert knowledge and computer automation. Proteins S3:47–44

Baumann G, Frommel C, Sander C (1989) Polarity as a criterion in protein design. Protein Eng 2 (5):329–334

Berg JM, Tymoczko JL, Stryer L (2002) Protein structure and function. Biochemistry 262:159

Blundell TL, Sibanda BL, Sternberg MJ, Thornton JM (1987) Knowledge-based prediction of protein structures and the design of novel molecules. Nature 326(6111):347–352

Bowie JU, Luthy R, Eisenberg D (1991) Method to identify protein sequences that fold into a known three-dimensional structure. Science 253(5016):164–170

Breda A, Valadares NF, Norberto de Souza O et al (2008) Protein structure, modelling and applications. In: Gruber A, Durham AM, Huynh C et al (eds) Bioinformatics in tropical disease research: a practical and case-study approach [internet]. National Center for Biotechnology Information, Bethesda

Brown TL, LeMay HE, Bursten BE, Burdge JR (2003) Chemistry the central science, 14th edn. Pearson, Upper Saddle River

Burley SK, Kurisu G, Markley JL, et al (2017) PDB-Dev: a prototype system for depositing integrative/hybrid structural models. Structure 25(9):1317–1318

Burley SK, Berman HM, Bhikadiya C et al (2019) RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Res 47(D1):464–474

Callaway E (2015) The revolution will not be crystallized: a new method sweeps through structural biology. Nature 525(7568):172–174

Carroni M, Saibil HR (2016) Cryo electron microscopy to determine the structure of macromolecular complexes. Methods 95:78–75

Chen Y, Su C, Ke M et al (2011) Biochemical and structural insights into the mechanisms of SARS coronavirus RNA ribose 2'-O-methylation by nsp16/nsp10 protein complex. PLoS Pathog 7 (10):e1002294

Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5(4):823–836

Cohen M, Potapov V, Schreiber G (2009) Four distances between pairs of amino acids provide a precise description of their interaction. PLoS Comput Biol 5(8):e1000470

Csaba G, Birzele F, Zimmer R (2009) Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis. BMC Struct Biol 9:23

Czaplewski C, Rodziewicz-Motowidlo S, Liwo A, Ripoll DR, Wawak RJ, Scheraga HA (2000) Molecular simulation study of cooperativity in hydrophobic association. Protein Sci 6:1235–1245

Dayringer HE, Tramontano A, Sprang SR, Fletterick RJ (1986) Interactive program for visualization and modelling of proteins, nucleic acids and small molecules. J Mol Graph 6:82–87

Desmet J, De-Maeyer M, Hazes B, Lasters I (1992) The dead-end elimination theorem and its use in protein side-chain positioning. Nature 356(6369):539–542

Dever TE, Green R (2012) The elongation, termination, and recycling phases of translation in eukaryotes. Cold Spring Harb Perspect Biol 4(7):a013706
Dukka BK (2017) Recent advances in sequence-based protein structure prediction. Brief Bioinform 18(6):1021–1032
Fiser A (2010) Template-based protein structure modeling. Methods Mol Bio (Clifton, N.J.) 673:73–94
Fiser A, Do RK, Sali A (2000) Modeling of loops in protein structures. Protein Sci 9(9):1753–1773
Han KF, Baker D (1995) Recurring local sequence motifs in proteins. J Mol Biol 251(1):176–187
Hansen DF, Kay LE (2011) Determining valine side-chain rotamer conformations in proteins from methyl 13C chemical shifts: application to the 360 kDa half-proteasome. J Am Chem Soc 133(21):8272–8281
Haywood AM (1997) Transmissible spongiform encephalopathies. N Engl J Med 337(25):1821–1828
He HT, Gursoy RN, Kupczyk-Subotkowska L, Tian J, Williams T, Siahaan TJ (2006) Synthesis and chemical stability of a disulfide bond in a model cyclic pentapeptide: cyclo(1,4)-Cys-Gly-Phe-Cys-Gly-OH. J Pharm Sci 95(10):2222–2234
Hendlich M, Lackner P, Weitckus S, Floeckner H, Froschauer R, Gottsbacher K, Casar G, Sippl JM (1990) Identification of native protein folds amongst a large number of incorrect models: the calculation of low energy conformations from potentials of mean force. J Mol Biol 216(1):167–180
Hintze BJ, Lewis SM, Richardson JS, Richardson DC (2016) Molprobity’s ultimate rotamer-library distributions for model validation. Proteins 84(9):1177–1189
Hooft RW, Vriend G, Sander C, Abola EE (1996) Errors in protein structures. Nature 381(6580):272
Hooft RW, Sander C, Vriend G (1997) Objectively judging the quality of a protein structure from a Ramachandran plot. Comput Appl Biosci 13(4):425–430
Horiuchi M, Priola SA, Chabry J, Caughey B (2000) Interactions between heterologous forms of prion protein: binding, inhibition of conversion, and species barriers. Proc Natl Acad Sci U S A 97(11):5836–5841
Hospital A, Goni JR, Orozco M, Gelpí JL (2015) Molecular dynamics simulations: advances and applications. Adv Appl Bioinforma Chem 8:37–47
Ihm Y (2004) A threading approach to protein structure prediction: studies on TNF-like molecules, Rev proteins, and protein kinases. Retrospective Theses and Dissertations 948
Jaroszewski L, Rychlewski L, Zhang B, Godzik A (1998) Fold prediction by a hierarchy of sequence, threading, and modeling methods. Protein Sci 7(6):1431–1440
Johnson RL, Lauffer E, Riddle RD, Tabin C (1994) Ectopic expression of sonic hedgehog alters dorsal-ventral patterning of somites. Cell 79(7):1165–1173
Jones DT (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 287(4):797–715
Jones DT, Taylor WR, Thornton JM (1992) A new approach to protein fold recognition. Nature 358(6381):86–89
Jonic S (2016) Cryo-electron microscopy analysis of structurally heterogeneous macromolecular complexes. Comput Struct Biotechnol J 14:385–390
Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC (1958) A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. Nature 181(4610):662–666
Kriegel E, Koraimann G, Vriend G (2002) Increasing the precision of comparative models with YASARA NOVA—a self-parameterizing force field. Proteins 47(3):393–302
Kriegel E, Nabuurs SB, Vriend G (2003) Homology modelling, methods of biochemical analysis. Struct Bioinf 44:509–524
Kumar A, Chordia N (2017) Role of bioinformatics in biotechnology. Res Rev Biosci 12(1):116
Liu H, Elstner M, Kaxiras E, Frauenheim T, Hermans J, Yang W (2001) Quantum mechanics simulation of protein dynamics on long timescale. Proteins 44(4):484–489
Lodish H, Berk A, Zipursky SL et al (2000) Molecular cell biology, 4th edn. W. H. Freeman, New York
Masters CL, Beyreuther K (1998) Alzheimer’s disease. Br Med J 316(7129):446–448
Morris AL, MacArthur MW, Hutchinson EG, Thornton JM (1992) Stereochemical quality of protein structure coordinates. Proteins 12(4):345–364
Moult J, Fidelis K, Krystafovyvch A, Schwede T, Tramontano A (2018) Critical assessment of methods of protein structure prediction (CASP)-round XII. Proteins 86(S1):7–15
Neil J, Bulleid Ellgaard L (2011) Multiple ways to make disulfides. Trends Biochem Sci 36 (9):485–492
Novotny J, Rashin AA, Brucoleri RE (1988) Criteria that discriminate between native proteins and incorrectly folded models. Proteins 4(1):419–430
Orlova EV, Saibil HR (2011) Structural analysis of macromolecular assemblies by electron microscopy. Chem Rev 111(12):7710–7748
Peach RJ, Bajorath J, Brady W, Leytze G, Greene J, Naemura J, Linsley PS (1994) Complementarity determining region 1 (CDR1)- and CDR3-analogous regions in CTLA-4 and CD28 determine the binding to B7-1. J Exp Med 180(6):2049–2058
Pearson WR (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol 183:63–98
Peitsch M (1997) Large scale protein modeling and model repository. In: Proceedings, International conference on intelligent systems for molecular biology; ISMB. International conference on intelligent systems for molecular biology, vol 5, p 234
Peitsch M, Schwede T, Guex N (2000) Automated protein modelling—the proteome in 3D. Pharmacogenomics 1(3):257–266
Peterson LX, Kang X, Kihara D (2014) Assessment of protein side-chain conformation prediction methods in different residue environments. Proteins 82(9):1971–1984
Pruisner SB (1996) Molecular biology and pathogenesis of prion diseases. Trends Biochem Sci 21 (12):482–487
Raj J, Leibundgut M, Ataide SF, Haag A, Ban N (2010) Crystal structure of the eukaryotic 40S ribosomal subunit in complex with initiation factor 1. Science 331(6018):730–736
Rankin NJ, Preiss D, Welsh P, Burgess KEV, Nelson SM, Lawlor DA, Sattar N (2014) The emergence of proton nuclear magnetic resonance metabolomics in the cardiovascular arena as viewed from a clinical perspective. Atherosclerosis 237(1):287–200
Rodriguez R, Chinea G et al (1998) Homology modeling, model, and software evaluation: three related resources. Bioinformatics 14(6):523–528
Sali A, Blindell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234(3):779–715
Sanchez R, Sali A (1997) Advances in comparative protein-structure modeling. Curr Opin Struct Biol 7(2):206–214
Scouras AD, Daggett V (2010) The Dynameomics rotamer library: amino acid side chain conformations and dynamics from comprehensive molecular dynamics simulations in water. Protein Sci 20(2):341–352
Serafini A (1989) Linus Pauling: a man and his science, 1st edn. Paragon House, New York
Serdyuk IN, Zaccai NR, Zaccai J (2007) Methods in molecular biophysics: structure, dynamics, function. Cambridge University Press, New York
Shi J, Blindell TL, Mizuguchi K (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 310 (1):243–257
Shortle D, Simons KT, Baker D (1998) Clustering of low-energy conformations near the native structures of small proteins. Proc Natl Acad Sci U S A 95(19):11158–11162
Simons KT, Bonneau R, Ruczinski I, Baker D (1999) Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins 37(S3):171–176
Sippl MJ (1993) Recognition of errors in three-dimensional structures of proteins. Proteins 17 (4):355–362
Sliwowski G, Kothiwale S, Meiler J, Lowe EW (2014) Computational methods in drug discovery. Pharmaco Rev 66(1):334–395
Snyder DA, Chen Y, Denissova NG et al (2005) Comparisons of NMR spectral quality and success in crystallization demonstrate that NMR and X-ray crystallography are complementary methods for small protein structure determination. J Am Chem Soc 127(47):16505–16511

Tappura K (2001) Influence of rotational energy barriers to the conformational search of protein loops in molecular dynamics and ranking the conformations. Proteins 44(3):167–179

Taylor WR (1986) Identification of protein sequence homology by consensus template alignment. J Mol Biol 188:233–258

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22(22):4673–4680

Voet D, Voet JJ (1990) Biochemistry. Wiley, New York

Wang HW, Wang JW (2017) How cryo-electron microscopy and X-ray crystallography complement each other. Protein Sci 26(1):32–39

Wohlgemuth I, Brenner S, Beringer M, Rodnina MV (2008) Modulation of the rate of peptidyl transfer on the ribosome by the nature of substrates. J Biol Chem 283(47):32229–32235

Wooley JC, Lin HS (2005) National Research Council (US) committee on frontiers at the interface of computing and biology; catalyzing inquiry at the interface of computing and biology. National Academies Press, Washington

Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y (2015) The I-TASSER suite: protein structure and function prediction. Nat Methods 12:7–8

Yang J, Zhang W, He B, Walker SE, Zhang H, Govindarajoo B, Virtanen J, Xue Z, Shen HB, Zhang Y (2016) Template-based protein structure prediction in CASP11 and retrospect of I-TASSER in the last decade. Proteins 84(S1):233–246