Protein Analysis: From Sequence to Structure

4

Jaykumar Jani and Anju Pappachan

Abstract

Proteins are primary molecules that control most of the cellular processes. The sequence of a protein is linked to its structure which in turn is linked to its function. Understanding and integrating protein sequence, structure, and function information is necessary to address many challenging areas of Biology including protein engineering, structural biology, and drug discovery. Bioinformatics deals with protein sequences, structures, predictions, and analysis. Accessibility of these data and availability of high-throughput analysis tools will supplement experimental work in order to understand proteins better. Prediction of three-dimensional structures of proteins and studying the structural features are very necessary to understand various diseases and aid in disease diagnosis and drug discovery. In this chapter we discuss about various databases and *in silico* tools and methods related to protein sequence and structure analysis.

Keywords

Sequence · Protein structure prediction · Protein analysis · *In silico* analysis · Protein database · Homology modelling

4.1 Introduction

Proteins are the key players that control almost all activities which sustain living organisms. Even though the genome of an organism consists of information for survival, proteins are the versatile macromolecules that regulate virtually all life

J. Jani · A. Pappachan (✉)
School of Life Sciences, Central University of Gujarat, Gandhinagar, Gujarat, India
e-mail: jaykumar.jani@cug.ac.in; anju.p@cug.ac.in

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-33-6191-1_4

processes within a cell. By providing structural and catalytic support proteins regulate various dynamic process of cells. Cytoskeletal proteins are examples of structural proteins that maintain cellular integrity and overall shape. Other proteins maintain cellular homeostasis by catalyzing various processes like DNA replication, transcription, translation, metabolism, cell communication and provide defence and immunity (Cohn 1939; Nelson et al. 2008). Defective proteins results in many disease conditions like Alzheimer's disease and sickle cell anaemia to name a few (Chou 2004). Study of proteins is of interest not only to biologists but also chemists because proteins are intriguing chemical entities and analysing their structures and how they carry out various functions are of prime importance. Detailed study of protein structure and function also helps to understand their molecular mechanism and role in various diseases. Throughout the kingdom of life from bacteria to higher eukaryotes, proteins are polypeptides made up of the same ubiquitous 20 amino acids. So, understanding the chemistry of amino acid is central to understand the molecular biochemistry of proteins. How amino acids are linked to one another through various kinds of covalent and non-covalent interactions give rise to proteins of varying structures, which can be grouped under distinct protein families that perform diverse functions (Nelson et al. 2008).

Study of proteins have traditionally been carried out using *in vitro* and *in vivo* techniques. But in modern protein chemistry, *in silico* studies are equally important. The wealth of sequence and structural data that has come as an outcome of the genome projects made it necessary for protein chemists to turn to the computers as laboratories to perform virtually various experiments in order to understand proteins better. Today, there are many bioinformatics tools and databases which help to correlate protein sequences with their structure and function. Identification of protein structure through conventional biophysical techniques like X-ray crystallography, NMR and Cryo-electron microscopy can be lengthy and complex which can be made easy with the development of structural bioinformatics which deals with prediction and analysis of the three-dimensional structure of bio-macromolecules (Marco 2009). An *in silico analysis* of protein sequence and structure can both complement and supplement experimental work.

The key challenge in bioinformatics is how to retrieve and analyse meaningful data and use it to enhance our understanding of biological molecules. Different protein databases store different pieces of information and address different aspects of protein analysis. In this chapter we will discuss some of the commonly used protein databases and tools available for protein sequence and structure analysis. We provide a flow chart on how to characterize proteins computationally starting from their sequence and proceeding to their structural analysis (Fig. 4.1). We also discuss some of the recent examples of how such *in silico* analysis is helping in the structure-based drug discovery and medical biology.

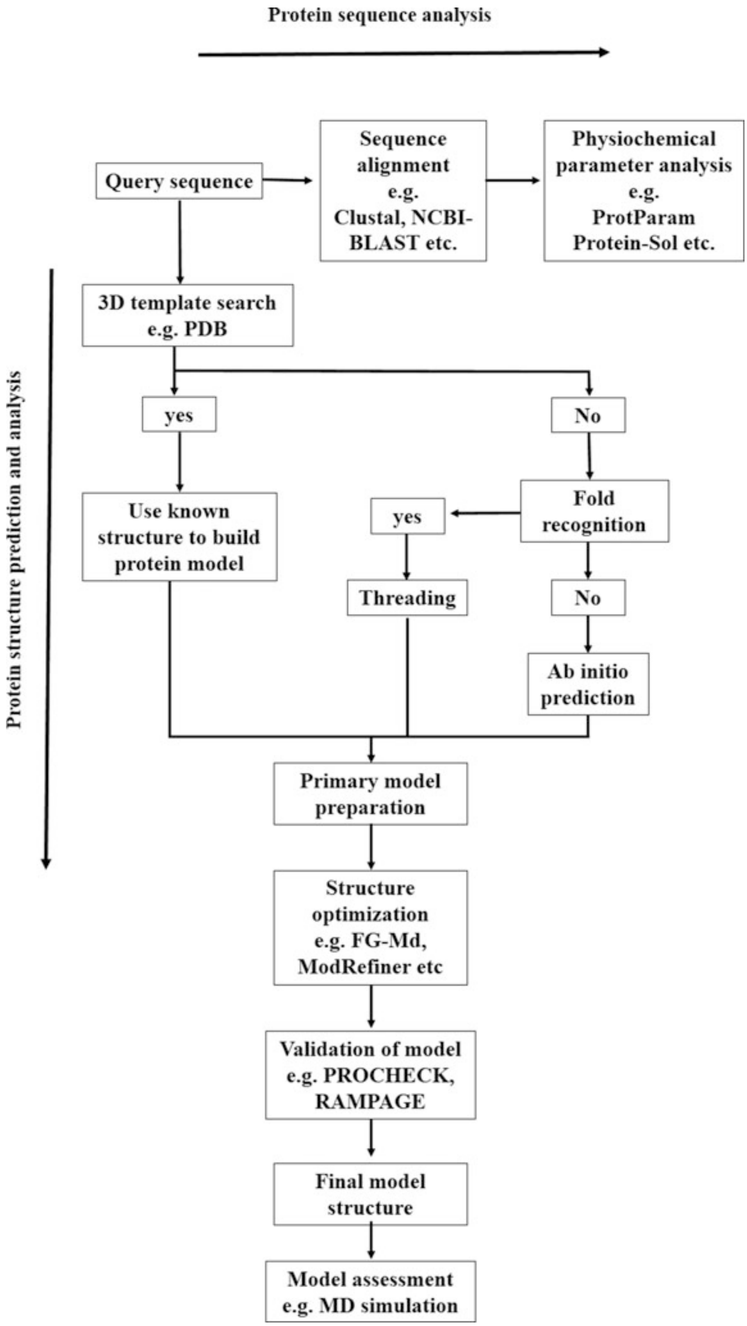
Protein sequence analysis
Query sequence
Sequence alignment e.g. Clustal, NCBI-BLAST etc.
Physiochemical parameter analysis e.g. ProtParam Protein-Sol etc.
Protein structure prediction and analysis
3D template search e.g. PDB
yes
No
Use known structure to build protein model
yes
Fold recognition
Threading
No
Ab initio prediction
Primary model preparation
Structure optimization e.g. FG-Md, ModRefiner etc
Validation of model e.g. PROCHECK, RAMPAGE
Final model structure
Model assessment e.g. MD simulation

Fig. 4.1 Flow chart on protein sequence analysis to structure prediction and analysis

4.2 Protein Structure Overview

Protein can be classified into different groups based on its structure, chemical nature, and biological role. Complex structural detail of protein can be studied at primary, secondary, tertiary, and quaternary levels of structural organization (Nelson et al. 2008).

4.2.1 Primary Structure

A linear linking of amino acids with each other as a chain via a peptide bond is represented as the primary structure of a protein. The polypeptide chain has an N-terminus and a C-terminus based on the presence of free amino or carboxyl group, respectively. The peptide bond is planar and rigid (non-rotatable) in nature as it partially shares two pairs of an electron. Whereas N-Cα and Cα-C have some freedom to rotate (−180 to +180), which helps proteins to acquire a three-dimensional structure (Nelson et al. 2008).

4.2.2 Secondary Structure

Local arrangement of some part of a polypeptide in particular conformation is referred to as protein secondary structure. The most common secondary structures are α-helices and β-strands, others are loops, turns, and coils. The secondary structure is mainly stabilized by Hydrogen bond (H-bond). The geometry has specific phi (φ) and psi (ψ) dihedral angle which can be studied by Ramachandran plot (Nelson et al. 2008).

4.2.2.1 Alpha(α) Helix

Alpha helix is most abundant in proteins compared to other helices (Kendrew et al. 1958; Pauling et al. 1951). Each helical turn is composed of 3.6 residues and has negative Φ and ψ angles (Φ = −64 +/−7 and ψ −41+/−7). The α-helix repeats itself at every 0.54 nm, the radius of helices is 0.23 nm, and residue transition distance is 0.15 nm. The hydrogen bond between the nitrogen of amide of the 1st amino acid and carboxyl oxygen atom of 5th (i + 4) amino acid is a characteristic feature of α-helix (Pauling et al. 1951).

4.2.2.2 The β-strand

It shows extended conformation compared to α-helices. The distance between adjacent amino acids is around 3.5 Å. The pattern of H-bond in the β-strand can be parallel, antiparallel, or mixed type based on the direction of the strand from the amino to carboxyl terminal (Richardson 1977).

4.2.2.3 3_{10} Helices

Compared to regular α-helices this structure is found less frequently. The H-bond pattern is i + 3 instead of i + 4 in α-helices (Taylor 1941). The backbone dihedral angles (phi and psi) for 3_{10} helices are −49 and −26, respectively (Ramakrishnan and Ramachandran 1965).

4.2.2.4 β-turns

β-turns are irregular in shape and length, they connect two β-strands and help polypeptide to change the direction. β-turns are also known as reverse turns. β-turns are usually found on the surface of protein which enables them to interact with other proteins and molecules (Venkatachalam 1968).

4.2.3 Tertiary Structure

Tertiary structure is a three-dimensional arrangement of local secondary structure in a specific conformation. This structure is supported by various interactions including hydrogen bond, hydrophobic interaction, disulphide bridges, salt bridges, and Van der Waals interaction. Other than the various interactions post-translational modifications significantly contribute to protein folding. Based on current knowledge tertiary structure of the protein can be classified into three classes which are α-protein, β-protein, α+β-protein (Nelson et al. 2008).

4.2.4 Quaternary Structure

The quaternary structure represents complex interaction among polypeptide chain, this complex is made up of multiple polypeptide subunits but operates as a single functional unit. The subunits can be same or different. The overall structure is stabilized by hydrogen bonds, salt bridges, and various other intramolecular interactions (Nelson et al. 2008).

4.2.5 Domains, Motifs, and Folds

Consideration of various domains, motifs, and folds becomes very important while predicting the structure and function of the protein as they have evolutionarily conserved sequence and are mainly found in the active site of protein which is responsible for catalysis (Nelson et al. 2008).

4.2.5.1 Domain

It is a conserved part of the polypeptide chain and can individually form its three-dimensional structure irrespective of other domains in the protein. Even it can execute its function irrespective to rest of the protein. A single protein may have more than one domain. Chimeric proteins with desired activity can be generated through protein engineering utilizing domain swapping (Nelson et al. 2008).

4.2.5.2 Motifs

Motifs are conserved sequence of amino acids found among proteins having similar catalytic activity. One motif may have more than one secondary structure element, e.g. Helix turn helix (Nelson et al. 2008).

4.2.5.3 Fold

Folds are similar to motifs and represent general protein architecture. Proteins with the same folds show the same combinations of secondary structure (Nelson et al. 2008).

4.3 Classification of Proteins Based on Protein Folding Patterns

In order to obtain structure-based information retrieval, databases have been developed deriving information from the Protein Data Bank. Similarities in protein folding pattern has been used to organize and group proteins. The prominent structural classification databases used heavily by biologists to understand protein structure are SCOP and CATH (Ghoorah et al. 2015). Such classifications are useful because they reflect both structural and evolutionary relatedness.

4.3.1 CATH (Class, Architecture, Topology, Homology)

This database groups protein based on topology, homology, class, and architecture. The topology level classification clusters proteins based on the overall shape and secondary structure. Homology based classification groups protein by their sequence identity along with protein domain similarity shared with the ancestor. Class of protein is mainly determined by their secondary structure and fold pattern and includes; all α, all β, α-β, etc. The architecture of proteins represents an overall structure and shape of a protein generated by different secondary structure organization. Architecture level classification system groups protein based on its secondary structure arrangement in three-dimensional space (Ghoorah et al. 2015; Orengo et al. 1997).

4.3.2 SCOP (Structural Classification of Proteins)

SCOP is an open-access database created in 1994. This database maintained by MCR Laboratory of Molecular Biology UK was created with a purpose to provide evolutionary information and structural similarity between all proteins with known structure. The database organizes protein structures in a hierarchy starting from domains at the lowest level. Set of domains are classified into families of homologues. Families that share common structure and function are grouped into superfamilies. Superfamilies that share a common folding topology are grouped as folds. Each fold group may belong to one of the general classes—α, β, α + β, and small proteins which often have minimal secondary structures. This database classifies protein based on Family, Superfamily, fold, IUPR (Intrinsically Unstructured Protein Region), Classes, and protein type (Ghoorah et al. 2015; Murzin et al. 1995; Andreeva et al. 2020).

4.4 Commonly Used Databases to Retrieve Protein Sequence and Structure Information

There are various sequence, structure, and composite databases which provide different information regarding proteins. Sequence databases provide protein sequence information and structure databases like PDB provide three-dimensional structural information about protein, and the composite database integrates information from various primary databases. The different composite database uses different algorithms and criteria to yield diverse information on proteins (Chen et al. 2017). Table 4.1 gives a list of commonly used protein databases.

4.4.1 Commonly Used Protein Sequence Databases

Primary databases mainly consist of experimentally derived information, for example, protein sequence, structure, etc. Commonly used primary database for proteins is PIR (Chen et al. 2017).

4.4.1.1 Protein Information Resource (PIR)

PIR was established in 1984 with the purpose to support genomic, proteomic, and system biology research. The database was developed at the National Biomedical Research Foundation (NBRF). Initially, information was obtained and compiled from Atlas of protein sequence and structure published by Margaret Dayhoff.

Table 4.1 Web-links for protein databases

Tool name	Weblink
SCOP	http://scop.mrc-lmb.cam.ac.uk/
CATH	http://www.cathdb.info/
PIR	https://proteininformationresource.org/
Swiss-Prot/Uniprot	https://www.uniprot.org/
PROSITE	https://prosite.expasy.org/
PRINT	http://130.88.97.239/PRINTS/index.php
BRENDA	https://www.brenda-enzymes.org/
Pfam	https://pfam.xfam.org/
PDB	rcsb.org

Later in 2002 PIR, along with its international partner created a single worldwide database UniProt by combining PIR-PSD, Swiss-Prot, and TrEMBL (Wu et al. 2003).

4.4.2 Structure Database

4.4.2.1 PDB

Protein Data Bank is a repository of macromolecular structures experimentally deciphered by X-ray crystallography, NMR spectroscopy, and Cryo-EM all around the world. Initially, the database was created as a joint project by Cambridge Crystallographic Data Center, UK and Brookhaven National Laboratory, the USA in 1971. In 2003 the database becomes an international organization. Now there are four members which are PDBj, PDBe, Research Collaboration for Structural Bioinformatics (RCSB), and Biological Magnetic Resonance Data Bank (BMRB) who deal with data deposition, data processing, and distribution. The information submitted to the database is reviewed manually and computationally for its authenticity. Each submitted structure is given unique four letter accession ID called PDB ID. The database can be dug by protein name, PDB ID, author name, deposition date, etc. PDB also contains information regarding protein secondary structure, experimental procedure, experimental data, and ligand information. The protein structure coordinate file can be downloaded as a .pdb file and can be visualized using structure visualization software such as Pymol, VMD, Rasmol, etc. The main purpose of the database is to provide structural information of biologically important macromolecules. Further some secondary and curated databases utilize information from PDB to predict protein structure (Berman 2008).

4.4.3 Composite Databases

Composite databases utilize information from different primary and secondary databases and use a complex combination of computational algorithms in order to provide vital information like biological role, a conserved region of the protein, active site residue, signature sequence, etc. (Chen et al. 2017). Some of them are listed below:

4.4.3.1 Swiss-Prot

Swiss-Prot is designed by EMBL (European Molecular Biology Laboratory) and Department of Medical Biochemistry at University of Geneva collectively. In 2002, Swiss-Prot became UniProt Knowledgebase (UniProtKB) with supplement information from TrEMBL and PIR protein database. Today, UniProtKB provides detailed information about protein function, structure, post-translational modification, etc., with minimum redundancy (Bairoch and Apweiler 2000).

4.4.3.2 PROSITE

PROSITE is a secondary database that contains information about conserved motifs of proteins which relates to its biological function. Multiple sequence alignment (MSA) is performed by a database to provide information related to the query sequence. When a search is made for a new protein sequence in the database it gives two types of information. First, it gives information about sequence patterns and enlists other proteins with the same pattern. Second, it gives detail about the protein family and its denoted biological role (Hulo et al. 2006).

4.4.3.3 PRINT

This database classifies protein into different families based on protein fingerprints. Fingerprints are multiple small conserved motifs identified by sequence alignment. Motifs are not necessarily present in the contiguous sequence, but they might come together in 3D space upon protein folding, which defines active site or interacting site of the protein. Thus the study of fingerprint represents protein fold and function better than single motif (Attwood et al. 2000).

4.4.3.4 BRENDA (BRaunschweig ENzyme DAtabase)

BRENDA database is specifically for enzymes and its biological pathway. It gives information about the functional and molecular properties of enzymes that have been classified by IUBMB (International Union of Biochemistry and Molecular Biology). The information available in the database is obtained by manual extraction from literature, text mining, data mining, and computational prediction. Every enzyme classified in BRENDA contains information about its biochemical reaction and kinetic property such as substrate and product of the corresponding enzyme (Schomburg et al. 2002).

4.4.3.5 Pfam

Pfam is a protein family database. Entry in Pfam is classified as family, domain, repeats, and motifs. Search can be made using protein sequence, domain, keyword, or taxonomy. As a result, it provides Pfam annotations for domain architecture, sequence alignment, interaction with other proteins, and protein structure in PDB (Finn et al. 2014).

4.5 Protein Sequence Analysis

The sequence of the protein determines the structure and the function of proteins. A thorough analysis of the protein sequence will throw light on its biological role, active site, stability, post-translational modification sites, regulatory elements, etc. Today there are several databases and tools available which predict protein features based on its sequence composition.

4.5.1 Protein Sequence Alignment

Knowledge of residue to residue correspondence between sequences will help to understand patterns of conservation and variability among sequences and infer evolutionary relationships. Two or more protein sequences share similarity if they have evolved from a common ancestor. Sequence similarity beyond a certain threshold indicates that the proteins share a common structure and biological function. Alignment of multiple protein sequences helps to understand protein features which might appear non-significant in pairwise alignment. Patterns of amino acid conservation can give information on domains, active site, and distant relationships may be detected. In short sequence alignment tools permit the researcher to predict the function of gene and protein fastly and accurately *in silico* by comparing query sequence with previously characterized protein, which could not be easily possible manually in the laboratory (Chenna et al. 2003). For a meaningful analysis, multiple sequence alignment should have both closely and distantly-related sequences. Various sequence alignment tools based on different algorithms are available. Clustal maintained by EMBL-EBI is one of the widely used multiple sequence alignment tools (Do and Katoh 2008).

4.5.1.1 Clustal

Clustal includes a series of programs commonly used in bioinformatics for sequence alignment purposes. Originally the program was developed in 1988 and managed by EMBL-EBI. There are many versions of Clustal based on the development/upgradation of an algorithm, Clustal Omega is the current standard version. All versions of Clustal perform multiple sequence alignment from a series of pairwise alignments, and assess it on the basis of scores based on a scoring matrix. These values will be used by the algorithm for distance measurement which reflects the evolutionary distance between sequences and the tool can build a phylogenetic tree using the neighbour-joining approach (Chenna et al. 2003).

4.5.1.2 Sequence Alignment in Database Searching

When complete genomes were determined, in order to identify the unknown function of many proteins coded by the genome, databases can be searched to identify their homologues by sequence alignment. The most commonly used such tool by biologists all over the world is the NCBI BLAST (Basic Local Alignment Search Tool).

BLAST

BLAST is a fast, accurate, and most commonly used method worldwide to find sequence similarity between a query sequence and sequences available in the databases. The sequence is queried against a specified database, and produces a report of those proteins in the database that are related to the query sequence. BLAST provides different options for standard and specialized data mining. Standard BLAST includes BLASTP (protein query against a protein database), BLASTN (DNA nucleotide query against DNA database), TBLASTN (protein query against translated nucleotide sequence database), BLASTX (translated nucleotide sequence against protein database), PHI-BLAST (Pattern Hit Initiated-BLAST that finds homologous protein sequences which also contains a regular pattern), and PSI-BLAST (Position-Specific Iterated-BLAST). While specialized search includes SmartBLAST (to find protein having high similarity to query sequence), PrimerBLAST (to design primer specific to the template), GlobalAlign (to compare two sequences entirely), CD-Search (to find conserved domain architecture), IgBLAST (to search immunoglobulins and T-cell receptors sequence), MOLE-BLAST (to establish the taxonomy for uncultured or environmental sequences), etc. (Altschul et al. 1990; Madden et al. 2019).

MSAs contain patterns that characterize families of proteins. There are several methods for applying MSAs of known proteins to identify related sequences in database searches, important ones being Profiles, PSI-BLAST, and Hidden Markov Model (HMM). All the three methods are useful to identify distantly-related sequences in a database search. Profiles contain conserved patterns found in a MSA of a set of homologous sequences. These patterns can be used to identify other homologous proteins by matching the query sequences from the database against the sequences in the alignment table, with higher weight to the conserved positions than variable regions. PSI-BLAST a modification of BLAST starts with a normal BLAST, then derives pattern information from MSA of initial hits and reprobes the database using the pattern. This process is iterated, by refining the pattern in successive cycles. HMM is more powerful than the other two to find distant relatives and predicting protein folding patterns. These are computational structures for describing fine patterns that define homologous protein families (Mount 2009).

4.5.2 Physicochemical Parameters from Sequence Analysis

Understanding various physiochemical parameters of protein such as molecular weight, extinction coefficient, half-life, hydropathicity index, solubility, and isoelectric point (PI) is very essential to understand protein behaviour and function. Parameters such as solubility of protein affect protein folding, interaction with macromolecules and ligands. To design novel therapeutics and to optimize recombinant protein production this prediction becomes useful. Bioinformatics tools like ProtParam and Protein-Sol are generally used for computational prediction of physicochemical properties of proteins based on sequence information (Gasteiger et al. 2005; Hebditch et al. 2017).

4.5.2.1 ProtParam

ExPASy is a bioinformatics tool which provides access to the various database in the field of life sciences like proteomics, genomics, transcriptomics, population genetics, etc. This portal is operated by Swiss Institute of Bioinformatics (SIB). ProtParam is one of many tools available on the ExPASy server which calculates various parameters of protein which are given below (Gasteiger et al. 2005).

Molecular Weight

The molecular weight of a protein is calculated by adding the average isotopic mass of each amino acid in the sequence.

Theoretical PI

The isoelectric point (PI) of protein depends on the pKa value of amino acid. The pKa value depends on the side-chain composition of amino acid. However, the pH of the solution where protein is present significantly affects the PI and solubility of the protein.

Grand Average of Hydropathicity (GRAVY)

GRAVY index is used to represent the hydrophobicity of a given protein. It gives sum of hydropathy value of each amino acid in the sequence divided by total length of the protein. The positive and negative GRAVY value represents hydrophobic and hydrophilic nature of protein, respectively. This calculation is done by hydropathy values given by Kyte and Doolittle.

Half-life

This is the predicted time required for half of the protein to degrade after its synthesis in the cellular system.

Instability Index

This parameter represents the stability of a given amino acid sequence in the test tube. If the value is lower than 40 it is considered stable and if the value is greater than 40 it is considered as unstable.

Extinction Coefficient

The extinction coefficient represents the absorbance of light by a given medium at a particular wavelength. Experimentally this value can be calculated by using the reference of known amino acid sequence. Computationally it is predicted by analysing number of aromatic amino acids in a given amino acid sequence.

4.5.2.2 Protein–Sol

Protein–sol is an online open-access tool (http://protein-sol.manchester.ac.uk). This tool predicts solubility of a given amino acid sequence, the algorithm of the tool calculates 35 features of sequence which include twenty amino acid composition scores, seven other composites, protein length, folding propensity, disorder propensity, beta-strand propensities, Kyte-Doolittle hydropathy, PI, sequence entropy, and absolute charge. If the predicted solubility score is >0.45 then the protein is predicted to be soluble, if the value is <0.45 then solubility is less (Hebditch et al. 2017).

4.6 Protein Structure Prediction

4.6.1 Secondary Structure Prediction

Local secondary structure can be predicted by utilizing information of its amino acid sequence. It is the first crucial step to tertiary structure prediction. Available methods focus to identify conserved local secondary structures such as helices, strands, and turns. These structures form at the early stage of protein folding. Thus, understanding of protein secondary structure is essential to study the protein folding process also. There are many prediction methods available which use different algorithm for secondary structure prediction. The Chou–Fasman method was considered as a breakthrough method having almost 50–60% accuracy in prediction. However, recent methods have an improvised algorithm with an increased accuracy of up to 60–65% (Kabsch and Sander 1983). Apart from the use of amino acid sequence for secondary structure prediction, consideration of microenvironment of protein and solvent accessibility of protein improvises prediction.

4.6.1.1 Chou–Fasman Method

This is one of the earliest methods developed by Peter Y Chou and Gerad D Fasman in order to predict the secondary structure of a protein. This method is based on analysis derived from data generated by X-ray crystallography. It analyses the relative frequency of each amino acid to occur at a particular position in protein secondary structures. By studying verified data it was found that each amino acid has a certain propensity to prefer one secondary structure over other or a specific position in the secondary structure, e.g. proline and glycine are found at the end of the helix. Consideration of frequency of specific amino acid rather than available chemical and physical theories for structure prediction makes this method less accurate. Nevertheless, Chen *et al.,* in 2006 improvised this method which made it predict secondary structure more accurately (Chou and Fasman 1974).

4.6.1.2 (Garnier–Osguthorpe–Robson) GOR Method

It is an information theory-based method. In addition to Chou method, it considers the conditional probability of each amino acid to form a secondary structure to predict the location of secondary structure in a given sequence. The original method is more accurate in predicting α-helices than β-strands (Garnier et al. 1978). This method has approximately 65% accuracy (Mount and Mount 2001).

4.6.1.3 Neural Network-Based Method

JPRED, SPINE, PHD, and PSIPRED are neural network-based prediction methods. This method predicts helices and sheets with higher accuracy. The commonly used neural network-based methods use a two-layer neural network prediction approach. The first layer network utilizes sequence to structure approach where it predicts the secondary structure of a protein by considering central residue utilizing a position-specific scoring matrix (PSSM) or MSA. In the second layer, it uses structure to structure approach, and filters outs output from the first layer to generate a final structure with higher accuracy. The accuracy of the predicted structure by this method is up to 70% (Lin et al. 2005).

4.6.2 Protein Tertiary Structure Prediction

The most successful approach for predicting protein tertiary structure is the template-based homology modelling. It is based on the knowledge that homologous protein sequences fold into similar three-dimensional structures. The general criteria are that two sequences must be at least 25% identical to assume structural similarity between them. To predict the three-dimensional structure of a protein, homology modelling starts with doing a database search to identify its homologues whose structures are solved. Now this structure is used as a template to predict the unknown protein structure. Then their amino acid sequences are aligned and structurally conserved regions are assigned based on closely related amino acid sequences. The atomic coordinates of these regions are then used to construct a partial model of the unknown protein. Side chains that are different between the two proteins within these regions are replaced with the correct ones taken from suitable structure libraries. In this partial model, now the gaps are filled by loop searching and modelling of the loops. At the end of this process, a complete model with certain errors in bond length, bond angle, etc., may be obtained which has to be corrected by molecular mechanics and energy minimization (Marks et al. 2012).

The main problem in three-dimensional structure prediction is the calculation of free energy and obtaining structure with the globally lowest energy. Nevertheless, due to recent advancements in technology, several automated bioinformatics tools are now available to do this. Mainly two types of approaches are used for protein model structure preparation (1) template-based and (2) template independent. Both methods have their advantages and disadvantages (Marks et al. 2012; Kc 2017). However, template-based methods are more accurate than other methods (Kc 2017; Zhang and Skolnick 2004). Few commonly used tools are explained below. Individuals can access the CAMEO website (https://www.cameo3d.org/), which is an automated server to provide continuous assessment of protein structure prediction services in order to decide on a tool for protein structure prediction (Haas et al. 2018).

4.6.2.1 Template-based Method for Predicting Tertiary Structure of Proteins

SWISS-MODEL

SWISS-MODEL is a widely used modelling tool as it is fast, accurate, and user friendly. This server consists of three integrated compounds (1) SWISS-MODEL pipeline—contains software for database related to protein modelling (2) SWISS-MODEL Workspace—provides virtual workspace and handles complex tasks during model preparation (3) SWISS-MODEL Repository—provides updated information regarding 3-D protein model of model organisms. The structure prediction process by SWISS-MODEL consists of the following steps: template searching, target-template alignment, structure building, and last, evaluation of the model. For the template searching and alignment, it uses BLAST and HHblits. If the query sequence is identical to previously known structure, then it copies coordinate information from that and builds homologous structure. However, if the structure is non-identical or has a patch of the unaligned region, it builds structure from information available in the fragment library. The final model is evaluated by QMEAN, which is knowledge-based scoring, and given as output. An optimized model can be downloaded as a PDB file (Schwede et al. 2003; Waterhouse et al. 2018).

Modeller

Modeller was developed by Andrej Sali Laboratory at the University of California, San Francisco. This tool is used for tertiary and quaternary structure prediction. It derives important information about protein structure from experimental data generated by NMR spectroscopy, site-directed mutagenesis, fluorescence spectroscopy, image reconstructions from electron microscopic studies, etc. This information is utilized to understand various parameters such as bond length, bond angle, and dihedral angle in the protein model structure building. To build modelled structure MODELLER uses following sequential steps; (i) searching for the available evaluated structure related to the query sequence, (ii) alignment of query and template sequence, (iii) model preparation, and (iv) evaluation of the final model. The DOPE method is used for model evaluation. Other than model building it also performs fold assignment, phylogenetic tree preparation, and *de novo* modelling of protein loop (Webb and Sali 2016).

I-TASSER

Developed by Yang Zhang Lab, upgraded version of I-TASSER models structure using threading method. In order to generate a protein model from the query sequence, it performs multiple steps. First, it searches for a super secondary structure related to query in PDB, using multiple threading approaches also called LOMETS [50]. Then, the different fragments of the modelled structure are combined using the Monte Carlo method. Multiple models of protein having lower energy levels are generated using Replica Exchange Monte Carlo Simulation (REMC). Coordinates of all the models are clustered by SPICKER method and average values of coordinates from all models are taken further for model preparation. Lastly, FG-MD algorithm is used to reconstruct all the atoms of the model having low free energy states. As a final output, five full-length models with atomic resolution and estimated accuracy are shown up. If in case given template does not have any previously available homologous structure for modelling, then the structure is prepared from scratch using *ab initio*-based approach by QUARK tool. QUARK is an integral part of I-TASSER structure prediction pipeline but these steps are only used when domains in the template are <300 residues (Roy et al. 2010; Xu and Zhang 2012).

4.6.2.2 Template Free Method for Predicting Tertiary Structure of Proteins

Even though the template-based method is more accurate, the template free method is very crucial for proteins that do not have a satisfactory template or have novel fold (s). The limitation of this method is low accuracy of the force field and it requires a greater computational facility for a query having >150 residues. Differences between template based and template free method is that the template free approach utilizes the basic principles of protein folding and does not need a homologous structure. Therefore this method is capable to model novel proteins even with new folds (Kc 2017). Rosetta is one of the methods which performs template free structure prediction. It generates a full-length model based on 3–9 residues fragment available from the known structure. The fragments are selected based on sequence similarity. Monte Carlo method is used for the assembly of a different fragment to give rise to the final full-length structure (Rohl et al. 2004). QUARK is another fragment-based structure prediction tool that is developed by Yang Zhang Lab, it uses 1–20 residue fragments (Xu and Zhang 2012). These fragments are assembled by REMC and atomic-level knowledge-based force fields are used to generate the final model. Other template free structure prediction methods are FRAGFOLD (Jones 2001), SCRATCH (Cheng et al. 2005), etc.

4.6.3 CASP

The Critical Assessment of protein Structure Prediction (CASP) primarily helps in advancing the methods for protein 3-D structure prediction from the amino acid sequence. It provides an opportunity to research groups to test their structure prediction method and compare it with other available methods. CASPs performs worldwide experiments at an interval of every two years which critically evaluates the current state and progress in protein structure prediction and what is the future scope for development. Till now thirteen CASPs experiments have been performed, the assessment and result of each experiment was published in Proteins: Structure, Function, and Bioinformatics journal. These analyses help the individual researcher to choose appropriate structure prediction method for their research work (Kinch et al. 2019).

4.7 Evaluation, Refinement, and Analysis of Predicted Protein Structure

4.7.1 Evaluation of Predicted Structure

Evaluation of modelled protein structure is a common step performed to ensure that the predicted structure is closest to the original structure. This is done by studying stereo-chemical properties such as bond angle, torsion angle, bond length, and planarity of bonds. The G-factor is a measurement used to study how usual or unusual are the stereo-chemical parameters of given protein model. Lower the G-factor: lower the probability of a particular conformation (Wlodawer 2017).

4.7.2 Structure Refinement

Due to the limitation of a force field and all-atom reconstruction the quality of predicted structure may not be very good. So, the refinement of a predicted structure is a necessary step in protein structure prediction. The aim of refinement is to improve the model structure quality with minor improvement of coordinates in the backbone and side-chain atoms. Refinement will help to get a structure with high stereo-chemical quality which is nearer to the native structure. Potential energy minimization (PEM) techniques and molecular dynamics help to get a structure with lower energy. FG-MD is one of the methods which performs atomic-level molecular dynamics simulation to obtain a lower energy structure without much change in overall structure (Zhang et al. 2011). Mod-Refiner which is also used for structure refinement uses Monte Carlo simulation for energy minimization. This method usually refines backbone structure first, from the primary C_α traces. After refining the backbone at minimum energy, it performs another round of simulation to reconstruct side-chain atoms and gives a final refined model with lower minimum free energy. The refined model can be validated using Ramachandran analysis (Xu and Zhang 2011; Feig 2017) to see if there are stearic clashes between the atoms in the structure. PROCHECK (Laskowski et al. 1993), RAMPAGE, and Moleman2 (Kleywegt and Jones 1996) are extensively used online tools for structure validation.

4.7.3 Structure Analysis

4.7.3.1 Molecular Dynamics Simulation

MD simulations are efficient tools to effectively understand protein structure to function relationships. How proteins function require knowledge of structure as well as dynamics. Molecular dynamics simulations provide powerful tools for exploring the conformational energy landscape accessible to these molecules, Though the method was developed in the 1950s, with the advancement in computational facilities and MD algorithms, this technique has achieved time scales close to that of biological processes and has helped us to move from the analysis of single structures, to the analysis of conformational ensembles. Biologists mainly use this method to study the conformation dynamics of protein, refinement of protein structure, and to understand the interaction of the protein with other molecules. Structure prediction studies performed through MD simulation can be tested using a community-wide experiment in CASP. GROning Machine for Chemical Simulation (GROMACS) is one of the common open-access software packages used to perform simulation of proteins, lipids, and nucleic acids. Once a complex structure of a protein with the ligand is prepared by docking or obtained from PDB repository it can be used as an input file in GROMACS. By applying script code for the different force fields, the movement of the molecule over time can be created in MD run(s). The output of the simulation can be analysed and visualized in the supplemented tool provided in the MD package (Abraham et al. 2015; Hollingsworth and Dror 2018).

4.8 Protein Interaction Studies Using *In Silico* Methods

Proteins rarely act alone. For various metabolic and regulatory processes, they may be associated with ligands or nucleic acids or other proteins. Understanding the molecular and structural basis of these interactions is very necessary for the functional elucidation of the proteins. There are several *in silico* methods to predict and characterize protein–ligand/nucleic acid/other protein interacting sites. In order to predict interaction of protein with other molecules large number of available structural data are being utilized to develop and improvise available prediction algorithms. The empirical, force field, knowledge based, and machine learning are four scoring functions currently in use (Böhm 1994). These scoring functions use different approaches to calculate binding energy of protein with another molecule. SWISS Dock (Grosdidier et al. 2011) is a commonly used tool to study protein–ligand interaction. Manual docking and simulation studies are also helpful to understand these interactions.

4.8.1 Protein–Protein Interaction (PPI)

Interacting proteins are necessary for proper functioning of various cellular processes. There are several examples like proteinase-inhibitor complexes, antigen–antibody interactions, various signalling complexes, RNA polymerase assembly, etc. Experimental study of protein–protein interaction is costly and time-consuming, and such studies can be made easy computationally. Various computational tools are available for PPI prediction, primarily all tools utilize protein sequence information for analysis (Jones and Thornton 1997). Previously studied protein and structural information are useful to identify a surface patch of protein that may be found at the interface site. PPI interaction can be studied online using fully automated tools and offline by using manual docking software. These tools give information about binding geometry and binding energy (Kangueane and Nilofer 2018). Some of the available tools are PrISE, InterPreTS, iLoops, Struct2Net which are structure-based prediction tools. PPI spider, Path2PPI, POINeT, RedNemo are PPI network prediction tools. TRI_tool, HIVsemi, ChiPPI, InterPORC are model organism-based PPI prediction tools. STRING, SPRINT, HSPPIP, BindML+, and iFrag are other PPI prediction tools (Kangueane and Nilofer 2018; Rao et al. 2014).

4.8.2 Protein DNA Interaction

Protein DNA interactions are very important for the fundamental processes like DNA replication, transcription, and translation. Its importance in epigenetic regulation is also now well recognized.

Transcription factor and histone proteins are examples of protein with multiple substrate specificity which makes them difficult to learn. However, there are numerous bioinformatics tools which predicts DNA–protein integration. Mainly two approaches are used for this prediction: sequence-based and structure-based. The structure-based approach requires protein structure to predict interaction and a sequence-based approach utilizes previously available sequence information to predict interaction (Sarai and Kono 2005). Examples of such tools are DBS-PSSM, DBS-Pred, DISIS, DISPLAR, DP-Bind, BindN, FoldX, and DNAbinder (Sarai and Kono 2005; Si et al. 2015).

4.8.3 Protein–Carbohydrate Interaction

Protein–carbohydrate interactions play a crucial role in a biological system in processes of cell signalling, inflammation, host–pathogen interaction, cell adhesion, etc. Among all carbohydrate interacting proteins, antibodies and lectins are well characterized (Chandra et al. 2006; Sacchettini et al. 2001). We have very limited information about protein–carbohydrate interaction because carbohydrates are the very diverse molecules which can adopt a wide range of conformations. The information generated through protein crystallographic methods is a limitation as it gives a snapshot of only one particular conformation in which it was crystallized (Taherzadeh et al. 2016). BALLDock, SLICK, Vina-Carb, and PROCARB are commonly used tools for protein–carbohydrate interaction prediction (Taherzadeh et al. 2016; Malik et al. 2010; Kerzmann et al. 2006).

4.9 Applications of Protein Sequence and Structure Analysis in Drug Discovery

Earlier, novel drug discovery was either by chance or a trial and error process which is usually performed by a high throughput screening method. However, advancement in protein structure prediction and docking algorithms reduced the cost and time needed for this process. Bioinformatics helps in different aspects of drug discovery and development starting from target selection to prediction of a lead compound to its improvement. Protein sequence and structure analysis is important to select a potential drug target against a disease. Knowledge about multi-protein complexes makes it possible to target specific protein–protein interaction.

Even if the tertiary structure of a potential drug target may not be available, with a predicted protein structure, we can create a hypothesis about its function, interaction with other macromolecules, and its regulatory aspect in the biological system. For

Table 4.2 Drugs currently under clinical trial for COVID-19 treatment

Target	Antiviral treatment
RNA polymerase	Remdesivir, favipiravir, ribavirin, umifenovir, galidesivir, oseltamivir, sofosbuvir, methylcobalamin
3CL protease	Lopinavir/Ritonavir, Ivermectin
PL protease	Disulfiram
Protein S	Griffithsin
Miscellaneous	Resveratrol, Loperamide, Losartan, Chloroquine, Hydroxychloroquine

example, understanding of protein structure allows us to design site-directed mutations that alter its function or multimeric status (Takeda-Shitaka et al. 2004). A detailed study of the structures will help to design and test potential ligands and for selecting structural features for combinatorial synthesis of libraries.

The importance of protein structure prediction in drug discovery is evident during the current COVID-19 pandemic situation caused by novel coronavirus later denoted as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). To identify potential drug targets and drug candidates, researchers are using computational approaches to predict protein structure and carrying out docking and simulation studies to screen a range of drug candidates against this virus *in silico* (Elmezayen et al. 2020; Joshi et al. 2020; Narayanan and Nair 2020). The shortlisted molecules are being studied both *in vitro* and *in vivo* for future use. Considering this pandemic situation Zhang Lab has provided predicted 3D model structure and its functional annotation of the COVID-19 proteins coded by the genome of SARS-CoV-2 which can be directly used for docking and simulations studies (https://zhanglab.ccmb.med.umich.edu/COVID-19/). Many structural and non-structural proteins of the virus as well as host-based drug targets are studied *in silico* for druggability. Already studied antiviral drugs for various protein targets are listed in Table 4.2 (Gil et al. 2020). Most of the mentioned drugs are under clinical trial for COVID-19 disease (Huang et al. 2020)

Antibodies are very crucial protein molecules for both basic research and pharmaceutical applications. Atomic-level structural information is required to understand the molecular specificity of antibody which further illustrates its biological importance. Several computational tools are available which deals with different antibody feature predictions. For example, Fv modelling of antibody used to study paratope, epitope, and protein docking. These tools precisely give information about residues that are involved in antigen–antibody interaction. This information further utilizes to increase or decrease antigen–antibody interaction by mutation studies *in vitro*. SAbPred is an online server that contains multiple tools used to predict antibody structure and other features (Dunbar et al. 2016).

Another important group of proteins are membrane proteins which are challenging to crystallize. Approximately 25% of the total proteins in a cell are membrane proteins and yet there are only few structures available. Since crystallization of this protein is very difficult, protein structure modelling remains the next option for structural study. There are plenty of reports where the researchers have used modelled membrane protein structures to screen for various drugs (Becker et al. 2004; Hauser et al. 2018). Even there are dedicated tools for modelling of GPCR family proteins such as GPCR-SSFE 2.0, GPCRM, and GOMoDo (Worth et al. 2017; Miszta et al. 2018; Sandal et al. 2013).

4.10 Conclusion

There are several databases of protein sequence and structures which are not only repositories of validated and annotated data, but also provide several tools to analyse these data. Once a new protein is discovered, the biological function can be understood by sequence comparisons with homologous proteins because proteins with related functions have related amino acid sequences. Such comparisons also throw light on the evolution of these proteins. Families of proteins with related functions have evolved from a common ancestor. Such proteins will show similar three-dimensional structure too which means that the three-dimensional structure of an unknown protein can be predicted by homology modelling if a homologous structure is already known. Due to the tremendous advances in our knowledge of protein folding as well as machine learning tools and algorithms, protein structure prediction methods have improved significantly in the past decade. This has facilitated the prediction of model protein structure with greater accuracy and closer to the native structure. These protein structures can be further analysed to understand their structure–function relationships. One of the major applications of such studies is in drug discovery and development. However few challenges need to be addressed for future development such as modelling of multi-domain proteins, prediction of structure involving loop-mediated interactions, simulation of macromolecular complexes, better algorithms to understand protein folding, etc. With the advancement in computational facilities and development of powerful algorithms, such *in silico* analysis of protein sequences and structures can make tremendous impact on major challenges in biology.

References

Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, Lindahl E (2015) GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 1–2:19–25

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

Andreeva A, Kulesha E, Gough J, Murzin AG (2020) The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Res 48:D376–D382

Attwood TK, Croning MD, Flower DR, Lewis AP, Mabey JE, Scordis P, Selley JN, Wright W (2000) PRINTS-S: the database formerly known as PRINTS. Nucleic Acids Res 28:225–227

Bairoch A, Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 28:45–48

Becker OM, Marantz Y, Shacham S, Inbal B, Heifetz A, Kalid O, Bar-Haim S, Warshaviak D, Fichman M, Noiman S (2004) G protein-coupled receptors: In silico drug discovery in 3D. Proc Natl Acad Sci U S A 101:11304

Berman HM (2008) The protein data bank: a historical perspective. Acta Crystallogr A 64:88–95

Böhm HJ (1994) On the use of LUDI to search the Fine Chemicals Directory for ligands of proteins of known three-dimensional structure. J Comput Aided Mol Des 8:623–632

Chandra NR, Kumar N, Jeyakani J, Singh DD, Gowda SB, Prathima MN (2006) Lectindb: a plant lectin database. Glycobiology 16:938–946

Chen C, Huang H, Wu CH (2017) Protein bioinformatics databases and resources. Methods Mol Biol (Clifton, N.J.) 1558:3–39

Cheng J, Randall AZ, Sweredoski MJ, Baldi P (2005) SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res 33:W72–W76

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD (2003) Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 31:3497–3500

Chou KC (2004) Structural bioinformatics and its impact to biomedical science. Curr Med Chem 11:2105–2134

Chou PY, Fasman GD (1974) Prediction of protein conformation. Biochemistry 13:222–245

Cohn EJ (1939) Proteins as chemical substances and as biological components. Bull N Y Acad Med 15:639

Do CB, Katoh K (2008) Protein multiple sequence alignment. Methods Mol Biol 484:379–413

Dunbar J, Krawczyk K, Leem J, Marks C, Nowak J, Regep C, Georges G, Kelm S, Popovic B, Deane CM (2016) SAbPred: a structure-based antibody prediction server. Nucleic Acids Res 44: W474–W478

Elmezayen AD, Al-Obaidi A, Şahin AT, Yelekçi K (2020) Drug repurposing for coronavirus (COVID-19): in silico screening of known drugs against coronavirus 3CL hydrolase and protease enzymes. J Biomol Struct Dyn:1–13

Feig M (2017) Computational protein structure refinement: almost there, yet still so far to go. WIREs Comput Mol Sci 7:e1307

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M (2014) Pfam: the protein families database. Nucleic Acids Res 42:D222–D230

Garnier J, Osguthorpe DJ, Robson B (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 120:97–120

Gasteiger E, Hoogland C, Gattiker A, Wilkins MR, Appel RD, Bairoch A (2005) Protein identification and analysis tools on the ExPASy server. The proteomics protocols handbook. Springer

Ghoorah AW, Devignes M-D, Alborzi SZ, Smaïl-Tabbone M, Ritchie DW (2015) A structure-based classification and analysis of protein domain family binding sites and their interactions. Biology 4:327–343

Gil C, Ginex T, Maestro I, Nozal V, Barrado-Gil L, Cuesta-Geijo M, Urquiza J, Ramírez D, Alonso C, Campillo NE, Martinez A (2020) COVID-19: drug targets and potential treatments. J Med Chem

Grosdidier A, Zoete V, Michielin O (2011) SwissDock, a protein-small molecule docking web service based on EADock DSS. Nucleic Acids Res 39:W270–W277

Haas J, Barbato A, Behringer D, Studer G, Roth S, Bertoni M, Mostaguir K, Gumienny R, Schwede T (2018) Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. Proteins 86(Suppl 1):387–398

Hauser AS, Chavali S, Masuho I, Jahn LJ, Martemyanov KA, Gloriam DE, Babu MM (2018) Pharmacogenomics of GPCR drug targets. Cell 172:41–54.e19

Hebditch M, Carballo-Amador MA, Charonis S, Curtis R, Warwicker J (2017) Protein-Sol: a web tool for predicting protein solubility from sequence. Bioinformatics (Oxford, England) 33:3098–3100

Hollingsworth SA, Dror RO (2018) Molecular dynamics simulation for all. Neuron 99:1129–1143

Huang X, Pearce R, Zhang Y (2020) De novo design of protein peptides to block association of the SARS-CoV-2 spike protein with human ACE2. Aging 12:11263

Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJA (2006) The PROSITE database. Nucleic Acids Res 34:D227–D230

Jones DT (2001) Predicting novel protein folds by using FRAGFOLD. Proteins 45:127–132

Jones S, Thornton JM (1997) Analysis of protein-protein interaction sites using surface patches. J Mol Biol 272:121–132

Joshi T, Joshi T, Sharma P, Mathpal S, Pundir H, Bhatt V, Chandra S (2020) In silico screening of natural compounds against COVID-19 by targeting Mpro and ACE2 using molecular docking. Eur Rev Med Pharmacol Sci 24:4529–4536

Kabsch W, Sander C (1983) How good are predictions of protein secondary structure? FEBS Lett 155:179–182

Kangueane P, Nilofer C (2018) Protein-protein and domain-domain interactions. Springer

Kc DB (2017) Recent advances in sequence-based protein structure prediction. Brief Bioinform 18:1021–1032

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 181:662–666

Kerzmann A, Neumann D, Kohlbacher O (2006) SLICK– scoring and energy functions for protein–carbohydrate interactions. J Chem Inf Model 46:1635–1642

Kinch LN, Kryshtafovych A, Monastyrskyy B, Grishin NV (2019) CASP13 target classification into tertiary structure prediction categories. Proteins Struct Funct Bioinform 87:1021–1036

Kleywegt GJ, Jones TA (1996) Phi/psi-chology: Ramachandran revisited. Structure 4:1395–1400

Laskowski RA, Macarthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 26:283–291

Lin K, Simossis VA, Taylor WR, Heringa J (2005) A simple and fast secondary structure prediction method using hidden neural networks. Bioinformatics 21:152–159

Madden TL, Busby B, Ye J (2019) Reply to the paper: misunderstood parameters of NCBI BLAST impacts the correctness of bioinformatics workflows. Bioinformatics 35:2699–2700

Malik A, Firoz A, Jha V, Ahmad S (2010) PROCARB: a database of known and modelled carbohydrate-binding protein structures with sequence-based prediction tools. Adv Bioinform 2010

Marco W (2009) Structural bioinformatics: from the sequence to structure and function. Curr Bioinform 4:54–87

Marks DS, Hopf TA, Sander C (2012) Protein structure prediction from sequence variation. Nat Biotechnol 30:1072–1080

Miszta P, Pasznik P, Jakowiecki J, Sztyler A, Latek D, Filipek S (2018) GPCRM: a homology modeling web service with triple membrane-fitted quality assessment of GPCR models. Nucleic Acids Res 46:W387–W395

Mount DW (2009) Using hidden Markov models to align multiple sequences. Cold Spring Harb Protoc, 2009, pdb.top41

Mount DW, Mount DW (2001) Bioinformatics: sequence and genome analysis. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536–540

Narayanan N, Nair DT (2020) Vitamin B12 may inhibit RNA-dependent-RNA polymerase activity of nsp12 from the SARS-CoV-2 virus. IUBMB Life

Nelson DL, Lehninger AL, Cox MM (2008) Lehninger principles of biochemistry. Macmillan

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) CATH – a hierarchic classification of protein domain structures. Structure 5:1093–1109

Pauling L, Corey RB, Branson HR (1951) The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci 37:205

Ramakrishnan C, Ramachandran G (1965) Stereochemical criteria for polypeptide and protein chain conformations: II. Allowed conformations for a pair of peptide units. Biophys J 5:909–933

Rao VS, Srinivas K, Sujini GN, Kumar GNS (2014) Protein-protein interaction detection: methods and analysis. Int J Proteomics 2014:147648

Richardson JS (1977) β-Sheet topology and the relatedness of proteins. Nature 268:495–500

Rohl CA, Strauss CE, Misura KM, Baker D (2004) Protein structure prediction using Rosetta. Methods Enzymol 383:66–93

Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5:725–738

Sacchettini JC, Baum LG, Brewer CF (2001) Multivalent protein− carbohydrate interactions. a new paradigm for supermolecular assembly and signal transduction. Biochemistry 40:3009–3015

Sandal M, Duy TP, Cona M, Zung H, Carloni P, Musiani F, Giorgetti A (2013) GOMoDo: a GPCRs online modeling and docking webserver. PLoS ONE 8:e74092

Sarai A, Kono H (2005) Protein-DNA recognition patterns and predictions. Annu Rev Biophys Biomol Struct 34:379–398

Schomburg I, Chang A, Schomburg D (2002) BRENDA, enzyme data and metabolic information. Nucleic Acids Res 30:47–49

Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res 31:3381–3385

Si J, Zhao R, Wu R (2015) An overview of the prediction of protein DNA-binding sites. Int J Mol Sci 16:5194–5215

Taherzadeh G, Zhou Y, Liew AW-C, Yang Y (2016) Sequence-based prediction of protein–carbohydrate binding sites using support vector machines. J Chem Inf Model 56:2115–2122

Takeda-Shitaka M, Takaya D, Chiba C, Tanaka H, Umeyama H (2004) Protein structure prediction in structure based drug design. Curr Med Chem 11:551–558

Taylor HS (1941) Large molecules through atomic spectacles. Proc Am Philos Soc:1–12

Venkatachalam CM (1968) Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. Biopolymers 6:1425–1436

Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, De Beer TAP, Rempfer C, Bordoli L, Lepore R, Schwede T (2018) SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res 46:W296–W303

Webb B, Sali A (2016) Comparative protein structure modeling using MODELLER. Curr Protoc Bioinform 54:5.6.1–5.6.37

Wlodawer A (2017) Stereochemistry and validation of macromolecular structures. Methods Mol Biol 1607:595–610

Worth CL, Kreuchwig F, Tiemann JKS, Kreuchwig A, Ritschel M, Kleinau G, Hildebrand PW, Krause G (2017) GPCR-SSFE 2.0-a fragment-based molecular modeling web tool for Class A G-protein coupled receptors. Nucleic Acids Res 45:W408–w415

Wu CH, Yeh L-SL, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu Z, Kourtesis P, Ledley RS, Suzek BE, Vinayaka CR, Zhang J, Barker WC (2003) The protein information resource. Nucleic Acids Res 31:345–347

Xu D, Zhang Y (2011) Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. Biophys J 101:2525–2534

Xu D, Zhang Y (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 80:1715–1735

Zhang Y, Skolnick J (2004) Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci U S A 101:7594–7599

Zhang J, Liang Y, Zhang Y (2011) Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. Structure 19:1784–1795