Secondary and Tertiary Structure Prediction of Proteins: A Bioinformatic Approach

Minu Kesheri, Swarna Kanchan, Shibasish Chowdhury and Rajeshwar Prasad Sinha

Abstract Correct prediction of secondary and tertiary structure of proteins is one of the major challenges in bioinformatics/computational biological research. Predicting the correct secondary structure is the key to predict a good/satisfactory tertiary structure of the protein which not only helps in prediction of protein function but also in prediction of sub-cellular localization. This chapter aims to explain the different algorithms and methodologies, which are used in secondary structure prediction. Similarly, tertiary structure prediction has also emerged as one of developing areas of bioinformatics/computational biological research owing to the large gap between the available number of protein sequences and the known experimentally solved structures. Because of time and cost intensive experimental methods, experimentally determined structures are not available for vast majority of the available protein sequences present in public domain databases. The primary aim of this chapter is to offer a detailed conceptual insight to the algorithms used for protein secondary and tertiary structure prediction. This chapter systematically illustrates flowchart for selecting the most accurate prediction algorithm among different categories for the target sequence against three categories of tertiary structure prediction methods. Out of the three methods, homology modeling which is considered as most reliable method is discussed in detail followed by strengths and limitations for each of these categories. This chapter also explains different practical and conceptual problems, obstructing the high accuracy of the protein structure in each of the steps for all the three methods of tertiary structure prediction. The popular hybrid methodologies which further club together a number of features such as structural alignments, solvent accessibility and secondary structure information are also discussed. Moreover, this chapter elucidates about the Meta-servers that generate consensus result from many servers to build a protein

M. Kesheri · R.P. Sinha
Laboratory of Photobiology and Molecular Microbiology, Centre of Advanced Study in Botany, Banaras Hindu University, Varanasi 221005, India

S. Kanchan (✉) · S. Chowdhury
Department of Biological Sciences, Birla Institute of Technology and Science, Pilani, Rajasthan 333031, India
e-mail: swarnabioinfo@gmail.com

Q. Zhu and A.T. Azar (eds.), *Complex System Modelling and Control Through Intelligent Soft Computations*, Studies in Fuzziness and Soft Computing 319,
DOI 10.1007/978-3-319-12883-2_19

model of high accuracy. Lastly, scope for further research in order to bridge existing gaps and for developing better secondary and tertiary structure prediction algorithms is also highlighted.

Keywords Secondary structure prediction · Tertiary structure prediction · Ab initio folding/modeling · Threading · Homology modeling · CASP

Abbreviations

PSS	Protein secondary structure
SSE	Secondary structure elements
UniProtKB	UNIversal PROTein resource KnowledgeBase
TrEMBL	Translated European molecular biology laboratory
PDB	Protein data bank
NMR	Nuclear magnetic resonance
FM	Free modelling
TBM	Template based modelling
GOR	Garnier-Osguthorpe-Robson
NNSSP	Nearest-neighbor secondary structure prediction
ANN	Artificial neural networks
SVM	Support vector machines
SOV	Segment overlap
CASP	Critical assessment of protein structure prediction
EVA	EValuation of automatic protein structure prediction
FR	Fold recognition
BLAST	Basic local alignment search tool
PSI-BLAST	Position specific iterative basic local alignment search tool
MEGA	Molecular evolutionary genetics analysis
PHYLIP	PHYLogeny inference package
GROMACS	GROningen machine for chemical simulations
AMBER	Assisted model building and energy refinement
CHARMM	Chemistry at HARvard molecular mechanics
GDT	Global displacement test
PROCHECK	PROtein structure CHECK
PROSA	PROtein structure analysis
MAT	MonoAmine transporters
HMM	Hidden Markov model
CPU	Central processing unit
RPS-BLAST	Reversed position specific BLAST

1 Introduction

Proteins are the building blocks of all cells in the living creatures of all kingdoms. Proteins are produced by the process of translation. In this process, transcribed gene sequence or mRNA is translated into a linear chain of amino acids which are called proteins. To characterize the structural topology of proteins, primary, secondary, tertiary and quaternary structure levels have been proposed. In the hierarchy, protein secondary structure (PSS) plays an important role in modeling of the protein structures because it represents the local conformation of amino acids into regular structures. There are three basic secondary structure elements (SSEs): alpha-helices, beta-strands and coils. Alpha helices are corkscrew-shaped conformations where the amino acids are packed tightly together. Beta sheets are made up of two or more adjacent strands connected to each other by hydrogen bonds, extended so that the amino acids are stretched out as far from each other to form beta strand. There are also two main categories of the beta-sheet structures: if strands run in the same direction then, called parallel-sheet whereas, if they run in the opposite direction then, called anti-parallel beta-sheet. Several approaches have been taken in order to devise tools for predicting the secondary structure from the protein sequence alone. Moreover, secondary structure itself may be sufficient for accurate prediction of a protein's tertiary structure (Przytycka et al. 1999). Therefore, many researchers employ PSS as a feature to predict the tertiary structure (Gong and Rose 2005), function (Lisewski and Lichtarge 2006) and sub-cellular localization of proteins (Nair and Rost 2003, 2005; Su et al. 2007).

Proteins have a precise tertiary structure that directs their function. Determining the structures of various proteins would aid in our understanding of the mechanisms of protein functions in biological systems. Prediction of protein structure from amino acid sequences has been one of the most challenging tasks in computational biology/bioinformatics for many years (Baker and Sali 2001; Skolnick et al. 2000). Currently, only biophysical experimental techniques such as X-ray crystallography and nuclear magnetic resonance are able to provide precise protein tertiary structures. There are 17,473,872,940 protein sequences in the latest release of UNIversal PROTein resource KnowledgeBase (UniProtKB)/Translated European Molecular Biology Laboratory (TrEMBL) as of 22nd April 2014, whereas the Protein Data Bank (PDB) contained only 99,624 protein structures till then. This is achieved as a result of an increase in large-scale genomic sequencing projects and the inability of proteins to crystallize or crystals to diffract well. This gap has widened too much over the last decade, despite the development of dedicated high-throughput X-ray crystallography pipelines (Berman et al. 2000). Solving the protein structure by Nucleic Magnetic Resonance (NMR) is limited to small and soluble proteins only. Moreover, X-ray crystallography and NMR are costly and time consuming methods for solving the protein structure. A list of the number of different types of molecules in PDB and their experimental methods by which the structure is determined is listed in Table 1. Therefore, the computational prediction of structure of proteins is

Table 1 Current PDB holdings (as on April 22nd, 2014)

Experimental methods	Molecule types				
	Proteins	Nucleic acids	Protein/NA complexes	Other	Total
X ray	82,406	1,516	4,287	4	88,213
NMR	9,129	1,078	206	7	10,420
Electron microscopy	523	52	173	0	748
Hybrid	59	3	2	1	65
Other	155	4	6	13	178
Total	92,272	2,653	4,674	25	99,624

highly needed to fill the gap between the protein sequences available in public domain databases and their experimentally solved structures.

Historically, protein structure prediction was classified into three categories: (i) Ab initio modeling (Liwo et al. 1999; Zhang et al. 2003; Bradley et al. 2005; Klepeis et al. 2005; Klepeis and Floudas 2003) (ii) Threading or Fold recognition (Bowie et al. 1991; Jones et al. 1992; Xu and Xu 2000; Zhou and Zhou 2005; Skolnick et al. 2004) and (iii) Homology or Comparative modeling (Šali and Blundell 1993; Fiser et al. 2000). Threading and comparative modeling build protein models by aligning query sequences onto solved template structures by X-ray crystallography or NMR. When close templates are identified, high-resolution models could be built by the template-based methods. If templates are absent from the PDB, the models need to be built from scratch, i.e. ab initio modeling.

Nowadays, these prediction categories are clubbed into two major groups: free modeling (FM) involving Ab initio folding and template-based modeling (TBM), which includes comparative/homology modeling and threading. These predicted models must be checked for protein structure quality validation by various programmes available.

This chapter is broadly divided under 9 sections which are further divided into sub-headings wherever required. Section 2.1 describes about amino acid propensity based secondary structure prediction method. Section 2.2 discusses about template based secondary structure predictions and the accuracy obtained by these methods. Section 2.3 explains the secondary structure prediction methods based on machine learning approaches. Ab initio folding/modeling and its limitations are is described in Sect. 3.1. Threading and Homology modeling methods with their strengths and their weakness are explained in Sects. 3.2 and 3.3 respectively. Hybrid and Meta-Servers which aid in accuracy of protein models are described in Sects. 4 and 5. Section 6 describes about the protein structure prediction community, Critical Assessment of protein Structure Prediction (CASP). Section 7 describes about the various application of protein models generated by the three major prediction methods. Future prospects of protein secondary and tertiary structure prediction

methodologies or algorithms as well as key steps which need to be improved are discussed in Sect. 8. Finally, Sect. 9 provides a comprehensive conclusion for the entire chapter.

2 Secondary Structure Prediction

2.1 Amino Acid Propensity Based Prediction

Early prediction methods as proposed by Chou and Fasman (1974) and the Garnier-Osguthorpe-Robson (GOR) (Garnier et al. 1978) rely on the propensity of amino acids that belong to a given secondary structure. These are simple and direct methods, devoid of complex computer calculations, that utilize empirical rules for predicting the initiation and termination of helical regions in proteins. The relative frequencies of each amino acid in each secondary structure of known protein structures are used to extract the propensity of the appearance of each amino acid in each secondary structure type. Propensities are then used to predict the probability that amino acids from the protein sequence would form a helix, a beta strand, or a turn in a protein. These methods have introduced the conditional probability of immediate neighbor residues for computation. The web-servers based on Chou and Fasman (1974) and GOR showed prediction accuracy between 60–65 %. However the updated, GOR V algorithm which is available as web-server at http://gor.bb.iastate.edu/ combines information theory, bayesian statistics and evolutionary information and has reached an accuracy of prediction to 73.5 % (Sen et al. 2005).

2.2 Template Based Prediction

This method uses the information from database of proteins with known secondary structures to predict the secondary structure of a query protein by aligning the database sequence with the query sequence and finally assigning the secondary structures to the query sequence. The nearest-neighbor method belongs to this category. This category is reliable if both sequences have good identical or homologous regions as compared to a threshold value. The two most successful template-based methods are Nearest-neighbor Secondary Structure Prediction (NNSSP) (Yi and Lander 1993) and PREDATOR (Frishman and Argos 1997). The accuracy of these methods lies in the range 63–68 % (Runthala and Chowdhury 2013).

2.3 Sequence Profile Based Method

This method uses the machine learning algorithms to predict the secondary structure of the query protein. Artificial Neural Networks (ANNs), Support Vector Machines (SVMs) and Hidden Markov Models (HMMs) are the most widely used machine learning algorithms that come under this category (Jones 1999a; Karplus et al. 1998; Kim and Park 2003; Chandonia and Karplus 1995). Currently, most effective PSS prediction methods are based on machine learning algorithms, such as PSIPRED (McGuffin et al. 2000), SVMpsi (Kim and Park 2003), PHD (Rost et al. 1994), PHDpsi (Przybylski and Rost 2002), Porter (Pollastri and McLysaght 2005), JPRED3 (Cole et al. 2008), STRIDE (Heinig and Frishman 2004), SPARROW (Bettella et al. 2012) and SOPMA (Geourjon and Deléage 1995) and which employ Artificial Neural Network (ANN) or Support Vector Machines (SVM) learning models. In addition to protein secondary structure, these servers also make predictions on Solvent Accessibility and Coiled-coil regions etc. These programmes or web-servers are listed in Table 2. These methods have an accuracy ranging 72–80 %, depending on the method, the training and the test datasets.

Two types of errors are most prevalent in secondary structure prediction of proteins. One of these errors is called local errors which occur when a residue is wrongly predicted. Second type of error is called structural error, which occur when the structure is altered globally. Sometimes, errors that alter the function of a protein should be avoided whenever possible. Q3 is the most commonly used measures of local errors, whereas the Segment Overlap (SOV) Score (Zemla et al. 1999) is the most well known measure for structural errors. These measures have been adopted by various communities in these research areas e.g. CASP (Moult et al. 1995) and EVA (Eyrich et al. 2001). Good secondary structures lay the foundation for better prediction of tertiary structures of proteins. The following section provides an insight into the methods for predicting the tertiary structures of proteins.

3 Tertiary Structure Prediction

As discussed in introduction, tertiary structure prediction methods are categorized into three major methods to model a target protein sequence. Flowchart for selecting the most accurate prediction algorithm/method among these three categories for the target sequence is schematically represented in Fig. 1.

Table 2 List of sequence profile-based web servers and programmes for secondary structure prediction along with the webpage URL and the programme description

S. no.	Name of the web server/group (URL)	Description of the web server/group
1	PSIPRED (McGuffin et al. 2000) [http://bioinf.cs.ucl.ac.uk/psipred/]	A simple and accurate secondary structure prediction server, incorporating two feed-forward neural networks which perform an analysis on output obtained from PSI-BLAST
2	PORTER (Pollastri and McLysaght 2005) [http://distill.ucd.ie/porter/]	A server which relies on bidirectional recurrent neural networks with shortcut connections, accurate coding of input profiles obtained from multiple sequence alignments, second stage filtering by recurrent neural networks
3	PHD (Rost et al. 1994) [http://npsapbil.ibcp.fr/cg-bin/npsa_automat.pl?page=/NPSA/npsa_phd.html]	An automated server which uses evolutionary information from multiple sequence alignment to predict the secondary structure prediction of proteins
4	JPRED3 (Cole et al. 2008) [http://www.compbio.dundee.ac.uk/www-jpred/]	Jpred incorporates the Jnet algorithm in order to make more accurate predictions. In addition to protein secondary structure Jpred also makes predictions on solvent accessibility and coiled-coil regions
5	STRIDE (Heinig and Frishman 2004) [http://webclu.bio.wzw.tum.de/stride/]	This server implements a knowledge-based algorithm that makes combined use of hydrogen bond energy and statistically derived backbone torsional angle information
6	SPARROW (Bettella et al. 2012) [http://agknapp.chemie.fu-berlin.de/sparrow/]	This server uses a hierarchical scheme of scoring functions and a neural network to predict the secondary structure
7	SOPMA (Geourjon and Deléage 1995) [http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_sopma.html]	A web-server which improved their prediction accuracy when combined with PHD secondary structure prediction method

3.1 Ab Initio Folding/Modeling

This method is simply based on elementary fundamentals of energy and geometry (Moult and Melamud 2000). Ab initio structure prediction seeks to predict the native conformation of a protein from the amino acid sequence alone. Ab initio prediction of protein structures makes no use of information available in databases mainly PDB (Nanias et al. 2005). The goal of this method is to predict the structure of a protein based entirely on the laws of physics and chemistry. It is assumed that the actual native state of a protein sequence has the lowest free energy. It means that the protein native state conformation is basically a model at the global minima of

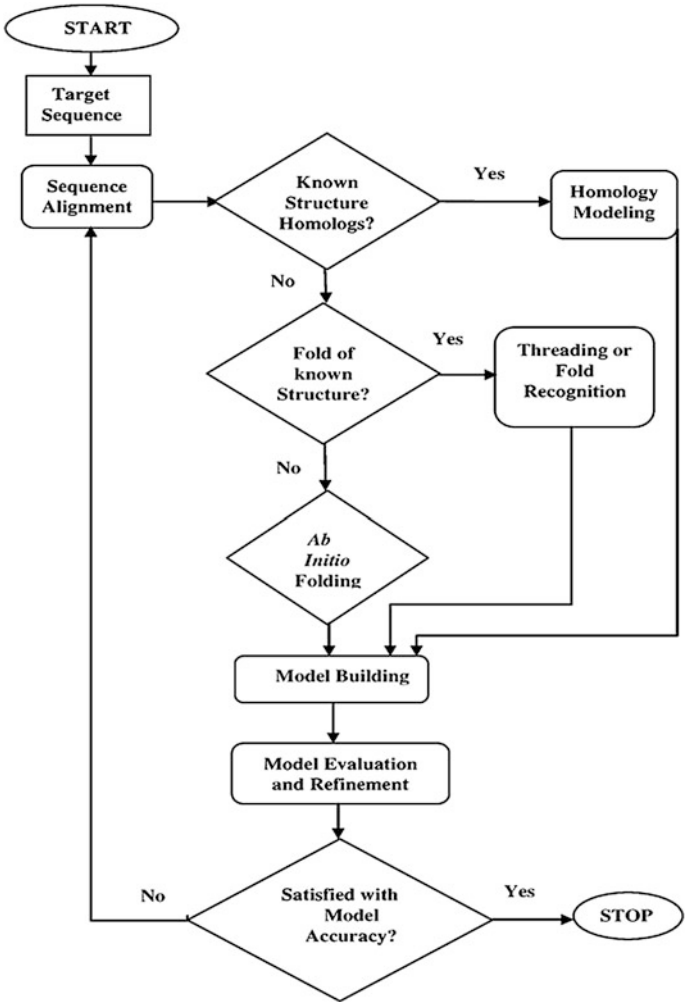

Fig. 1 Flowchart for selecting the most accurate algorithm for prediction of the target sequence against three categories of tertiary structure prediction

the energy landscape. Hence, ab initio algorithm actually searches the entire possible conformational space of a target sequence, in order to find the native state among all conformations.

For example, if we consider only three allowed conformations per residue, then a protein of 200 residues can have 3^{200} different conformations (Runthala and Chowdhury 2013). Hence, searching this huge conformational space will be extremely challenging task. This is the most difficult category of protein structure prediction among all the three different methods of structure prediction which

Table 3 List of web servers for modeling protein structure by ab initio folding method along with the webpage URL and the programme description

S. no.	Name of the web server/group (URL)	Description of the web server/group
1	ROBETTA (Kim et al. 2004; Bradley et al. 2005) [http://robetta.bakerlab.org]	This web-server provides ab initio and comparative models of protein domains. Domains having no sequence similarity with PDB sequences are modeled by Rosetta de novo protocol
2	QUARK (Xu and Zhang 2012) [http://zhanglab.ccmb.med.umich.edu/QUARK/]	De novo protein structure prediction web server aims to construct the correct protein 3D model from amino acid sequence by replica-exchange Monte Carlo simulation under the guide of an atomic-level knowledge-based force field
3	PROTINFO (Hung et al. 2005) [http://protinfo.compbio.washington.edu]	De novo protein structure prediction web server utilizes simulated annealing for 3D structure generation and different scoring functions for selection of final five conformers
4	SCRATCH (Cheng et al. 2005) [http://www.igb.uci.edu/servers/psss.html]	This server utilizes recursive neural networks, evolutionary information, fragment libraries and energy to build protein 3D model
5	BHAGEERATH (Jayaram et al. 2006) [http://www.scfbio-iitd.res.in/bhageerath]	Energy based methodology for narrowing down the search space and thus helps in building a good protein 3D model

completely predicts a new fold (Skolnick and Kolinski 2002; Floudas et al. 2006). With increasing protein size, the conformational space to be searched increases sharply, this makes the ab initio modeling of larger proteins extremely difficult (Zhang and Skolnick 2004).

Currently, the accuracy of ab initio modeling is limited to small proteins having length less than 50 amino acid residues. Ab initio structure prediction requires an efficient potential function to find the conformation of the modeled protein near to native state protein structure with lowest free energy. Ab initio structure prediction is challenging because current potential functions have limited accuracy. Few popular web servers for modeling of the protein structure by ab initio folding/modeling method are listed in Table 3.

3.2 Fold Recognition (FR) or Threading

Fold recognition or threading method aims to fit a target sequence to a known structure in a library of folds and the model built is evaluated using residue based contact potentials (Floudas 2007). Although fold recognition will not yield equivalent results as those from X-ray crystallography or NMR yet, it is a comparatively

fast and inexpensive way to build a close approximation of a structure from a sequence without involving the time and costs of experimental procedures. Fold Recognition (FR) was reserved for methods which did not rely on sequence searching and where the sequence identity between target and template was below the so-called "twilight zone" spanning between 25–30 %. The rationale behind the threading method is that total number of experimentally solved 3D structure deposited in PDB database doesn't have a new fold. The nature has limited number of basic folds which form the framework of most of the protein structures available in PDB. Generally, similar sequence implies similar structure but the reverse is not true. Similar structures are often found for proteins for which no sequence similarity to any known structure can be detected (Floudas et al. 2006). Using fold recognition or threading, we are able to identify proteins with known structures that share common folds with the target sequences. Fold recognition methods work by comparing each target sequence against a library of potential fold templates using energy potentials and/or other similarity scoring methods. For such comparison, we first need to define a library of potential folds. Once the library is defined, the target sequence will be fitted into each library entry and an energy function is used to evaluate the fit between the target sequence and the library entries to determine the best possible templates. The template with the lowest energy score is then assumed to best fit the fold of the target protein.

Fold recognition methods also includes various properties of structural environment of the amino acid residue. Structural environments are more conserved than the actual type of residue, therefore in the absence of homology, a fold could be predicted by measuring the compatibility of a sequence with template folds in terms of amino acid preferences for certain structural environments. These amino acid preferences for structural environment provide sufficient information to choose among the folds. The amino acid preferences for three main types of structural environment comprise of the solvent accessibility, the contact with polar atoms and the secondary structure. The main limitation of this method is high computational cost, since each entry in the whole library of thousands of possible folds needs to be aligned in all possible ways to select the fold(s). Another major bottleneck is the energy function used for the evaluation of alignment. It is not reasonable to expect to find the correct folds in all cases with a single form of energy function. Few popular web servers for modeling the protein structure by threading method are listed in Table 4.

3.3 Homology Modeling or Comparative Modeling

Comparative or homology protein structure modeling builds a three-dimensional model for a protein of unknown structure (the target) based on one or more related proteins of known structure. The necessary conditions for getting a useful model are

Table 4 List of web servers for modeling the protein structure by threading or fold recognition method along with the webpage URL and the description of programme

S. no.	Name of the web server/group [URL]	Description of the web server/group
1	I-TASSER (Zhang et al. 2005) [http://zhanglab.ccmb.med.umich.edu/I-TASSER/]	3D models are built based on multiple-threading alignments by LOMETS and iterative template fragment assembly
2	SPARKS^{X} (Yang et al. 2011) [http://sparks-lab.org/yueyang/server/SPARKS-X/]	This server employs significantly improved secondary structure prediction, real value torsion angle prediction and solvent accessibility prediction to model a more accurate protein structure
3	LOOPP (Teodorescu et al. 2004) [http://cbsuapps.tc.cornell.edu/loopp.aspx]	A fold recognition program based on the collection of numerous signals to build the target structure
4	PROSPECT (Xu and Xu 2000) [http://compbio.ornl.gov/structure/prospect]	PROSPECT is based on scoring function, which consists of four additive terms: (i) a mutation term, (ii) a singleton fitness term, (iii) a pairwise-contact potential term, and (iv) alignment gap penalties
5	MUSTER (Wu and Zhang 2008) [http://zhanglab.ccmb.med.umich.edu/MUSTER/]	Muster generates sequence-template alignments by combining sequence profile-profile alignment with multiple structural information
6	PHYRE2 (Kelley and Sternberg 2009) [http://www.sbg.bio.ic.ac.uk/~phyre2/html/page.cgi?id=index]	A server which uses profile–profile matching algorithms to build the protein model

(a) Detectable similarity (Greater than or equal to 30 %) between the target sequence and the template structures and
(b) Availability of a correct alignment between them.

Homology or Comparative modeling is a multistep process that can be summarized in following six steps:

3.3.1 Template Search, Selection and Alignment

Template search is generally done by comparing the target sequence with the sequence of each of the structures in the PDB database. The performance depends on the sensitivity of the comparison of target and template sequences by various programmes e.g. FASTA which is available at http://www.ebi.ac.uk/Tools/sss/fasta/ while, BLAST and PSI-BLAST (Altschul et al. 1997) are available at http://blast.ncbi.nlm.nih.gov/Blast.cgi. The simplest template selection rule is to select the structure with the highest sequence similarity with the target sequence. The quality of a template increases with its overall sequence similarity with the target and

decreases with the number and length of gaps in the alignment. Multiple sequence alignment by various freely available programmes e.g. ClustalW (Larkin et al. 2007) Mufft (Katoh et al. 2002), Kalign (Lassmann and Sonnhammer 2005), Probcons (Do et al. 2005) etc. and a development of phylogenetic tree by freely available programmes e.g. MEGA (Tamura et al. 2013) and PHYLIP etc. can help in selecting the template from the subfamily that is closest to the target sequence. HHpred (http://toolkit.tuebingen.mpg.de/hhpred) is one of the best servers which the can even detect very distant relationships between the target sequence and the solved PDB structures significantly. This is the first server that is based on the pairwise comparison of profile Hidden Markov Models (HMMs) (Söding et al. 2005).

The similarity between the 'environment' of the template and the environment in which the target needs to be modeled should also be considered. The quality of the experimentally determined structure is another important factor in template selection whereby high resolution X-ray crystal structure is more preferred for template selection than that of low resolution crystal structure. Multiple templates rather than selecting a single template, generally increases the model accuracy. A good protein structure model depends on alignment between the target and template.

3.3.2 Alignment Correction in Core Regions

An accurate alignment can be calculated automatically using standard sequence-sequence alignment methods, for example, Blast2seq (http://blast.ncbi.nlm.nih.gov/Blast.cgi) and dynamic programming based Needle global sequence alignment (https://www.ebi.ac.uk/Tools/psa/emboss_needle/). In low sequence identity cases, the alignment accuracy is the most important factor which affects the quality of the predicted model. Alignments can be improved by including structural information from the template protein structure. Gaps should be avoided in core regions mainly in secondary structure elements (which are found to be conserved in most cases), buried regions and between two residues that are far in space. It is important to inspect and edit the alignment manually by many tools e.g. Bioedit (www.mbio.ncsu.edu/bioedit/bioedit.html) etc., especially if the target-template sequence identity is low.

3.3.3 Backbone, Loop and Side-Chain Modeling

Creating the backbone is essential for modeled protein structure. For backbone, we simply copy the coordinates of those template residues that show up in the alignment with the model sequence. If two aligned residues differ, only the backbone coordinates (N, Cα, C and O) can be copied. If they are the same, we can also include the coordinates of side chain amino acid residues.

In comparative modeling, target sequences often have few inserted residues as compared to the template structures. Thus, no structural information about these inserted regions could be obtained from the template structures. These regions are called surface loops. Loops often play an important role in defining the functional specificity of a given protein structure, forming the active and binding sites for drug molecules. The accuracy of loop modeling is a major issue for comparative models for applications such as protein-ligand docking i.e. structure based drug design. There are two main classes of loop modeling methods:

(a) Database search approaches, where a small loop of 3–10 amino acid residues are searched in a database of known protein structures and if such loops fit the criteria of lowest energy, such loops are selected and added to the model structure. All major molecular modeling programs and servers support this approach e.g. Modeller (Šali and Blundell 1993), Swiss-Model (Guex and Peitsch 1997).
(b) The conformational search approaches mainly depend on an efficient energy function to choose the loop with lowest energy. If required, energy of the selected loop is minimized using Monte Carlo or molecular dynamics simulations by AMBER, and GROMACS techniques in order to arrive at the best loop conformation with lowest energy.

Side chain modeling is also one of the essential components in structure prediction of proteins. When we compare the side-chain conformations (rotamers) of residues that are conserved in structurally similar proteins, we copy coordinates of conserved amino acid residues entirely from the template to the model. But when we have different residues, side chains are added to each amino acid and their all possible rotamers are searched to find the most stable (having least energy) rotamer from rotamer library.

3.3.4 Model Refinement

One of the major limitations of computational protein structure prediction is the deviation of predicted models from their experimentally derived true, native structures. Refinement of the protein model is required, if there is problem in structural packing of side chains, loops, and secondary structural elements in the target model. For any error in backbone or side chain packing, energy minimization is done which requires an enormous precision in the energy function. At every minimization step, a few big errors (like bumps, i.e., too short atomic distances) are removed while many small errors might be introduced which lead to another distortion in the structure. In energy minimization, force fields must be fast to handle these large molecules efficiently. Refinement of the low resolution predicted models to high resolution structures are close to the native state, however, it has proven to be extremely challenging. There are various programmes e.g. GROMACS (http://

www.gromacs.org/), AMBER (www.amber.scripps.edu), and CHARMM (http://www.charmm.org/) which are freely as well commercially available for protein model refinement by correcting the overall protein structural geometry. One of the recently developed refinement methods called 3Drefine is computationally inexpensive and consumes only few minutes of CPU time to refine a protein of typical length of 300 amino acid residues (Bhattacharya and Cheng 2013).

3.3.5 Model Evaluation or Validation

The predicted model must be checked for

(a) Errors or distortion in side chain packing of the modeled structure.
(b) Distortions or shifts in correctly aligned region of target with the template structures.
(c) Distortions or shifts of a region that does not align with any of the template structures.
(d) Distortions or shifts of a region that is aligned incorrectly with the template structures.

Structural model accuracy is mainly based on global distance test (GDT), which is an average percentage of model Cα atoms within a specified distance threshold to actual native conformation (Jauch et al. 2007).

$$\mathrm{GDT} = \frac{1}{4}\left(\max {}^{C}_{1A}0 + \max {}^{C}_{2A}0 + \max {}^{C}_{4A}0 + \max {}^{C}_{8A}0\right) \quad (1)$$

Equation 1 GDT score where C^{0}_{nA} is the number of atom pairs closer than distance of n = 1, 2, 4 and $8A^{0}$.

TM score is another method for validating the model accuracy to score the topological similarity of target and template structures, where the score near to 1.00 is the best predicted near-native model against the actual experimental structure for a target (Xu and Zhang 2010). MaxSub is another new and independently developed method which aims at identifying the largest subset of C(alpha) atoms of a model that superimpose 'well' over the experimental structure, and produces a single normalized score that represents the quality of the model (Siew et al. 2000).

Various programmes and web-servers are available for checking the quality of the model. One of these is Procheck (Laskowski et al. 1993) that generates the Ramachandran Plot, which illustrates the stereo chemical quality of the protein model. Few popular web servers for protein structure quality validation and their description are listed in Table 5.

Few popular web servers for modeling the protein structure by homology or comparative modeling method along with the webpage URL and description are listed in Table 6.

Table 5 List of web servers for protein structure quality validation along with the webpage URL and the programme description

S. no.	Name of the web server/group [URL]	Description of the web server/group
1	QMEAN (Benkert et al. 2009) [http://swissmodel.expasy.org/qmean/cgi/index.cgi]	Quality estimate is based on geometrical analysis of single model, and the clustering-based scoring function
2	PROSA-WEB (Wiederstein and Sippl 2007) [https://prosa.services.came.sbg.ac.at/prosa.php]	Quality is checked by generation of Z-scores and energy plots that highlight potential problems spotted in protein structures
3	PROCHECK (Laskowski et al. 1993) [http://services.mbi.ucla.edu/SAVES/]	Stereo chemical quality of a protein structure is checked by analyzing residue-by-residue geometry and overall structural geometry
4	VERIFY-3D (Bowie et al. 1991; Luthy et al. 1992) [http://services.mbi.ucla.edu/SAVES/]	Determines the compatibility of an atomic model (3D) with its own amino acid sequence (1D) by assigning a structural class based on its location and environment
5	ERRAT (Colovos and Yeates 1993) [http://services.mbi.ucla.edu/SAVES/]	This server analyzes the statistics of non-bonded interactions between different atom types and plots the value of the error function

3.3.6 Homology Models Repositories

However, there are many repositories available, which contain protein homology models generated using various automated methods that provide models which serve as starting points for biologists/experimentalists. SWISS-MODEL repository (http://swissmodel.expasy.org/repository/) is one of the databases of annotated three-dimensional comparative protein structure models generated by the fully automated homology-modelling pipeline SWISS-MODEL. Protein Model Portal (http://proteinmodelportal.org) is another repository aimed at storing manually built 3D models of proteins (Arnold et al. 2009). The most recent database is Modbase (http://modbase.compbio.ucsf.edu) which contains the datasets of comparative protein structure models, calculated by modeling pipeline ModPipe (Pieper et al. 2011).

Several additional features when clubbed to the methods for tertiary structure prediction generate hybrid methods which are used to produce more accurate protein tertiary structures. Following section discusses about these hybrid methods for the protein tertiary structure prediction.

Table 6 List of web servers for modeling the protein structure by homology modeling or comparative modeling method along with the webpage URL and the programmes description

S. no.	Name of the web server/group [URL]	Description of the web server/group
1	GENO3D (Combet et al. 2002) [http://geno3d-pbil.ibcp.fr/]	A web server which builds the model based on distance geometry, simulated annealing and energy minimization algorithms to build the protein 3D model
2	M4T (Fernandez-Fuentes et al. 2007) [http://manaslu.aecom.yu.edu/M4T/]	A fully automated comparative protein structure modeling server with two major modules, Multiple Templates (MT) and Multiple Mapping Method (MMM)
3	CPHMODELS 3.2 (Nielsen et al. 2010) [http://www.cbs.dtu.dk/services/CPHmodels/]	Protein modeling is based on profile-profile alignment guided by secondary structure and exposure predictions
4	3DJIGSAW (Bates et al. 2001) [http://www.bmm.icnet.uk/servers/3djigsaw/]	An automated server to build three-dimensional models for proteins based on homologues of known structure
5	PUDGE (Norel et al. 2010) https://bhapp.c2b2.columbia.edu/pudge/cgi-bin/pipe_int.cgi	A server that includes secondary structure predictions, domains predictions and disorder prediction to predict the high quality homology model
6	SWISS-MODEL (Guex and Peitsch 1997) [http://swissmodel.expasy.org/SWISSMODEL.html]	A fully automated protein structure homology-modeling server
7	ESYPRED3D (Lambert et al. 2002) [http://www.fundp.ac.be/sciences/biologie/urbm/bioinfo/esypred/]	This server results in good protein model by using several multiple alignment programs by combining, weighing and screening

4 Hybrid Methods for Protein Tertiary Structure Prediction

Nowadays, a number of fully automated hybrid methods are designed in order to perform rapid, completely automated fold recognition on a proteome wide scale. These hybrid methods club together a number of features such as structural alignments, solvent accessibility and secondary structure information in order to produce a protein model with high accuracy. Some such methods are discussed below.

GenTHREADER (Jones 1999b) is a fully automated hybrid method for fold recognition which uses a traditional sequence alignment algorithm to generate alignments. These generated alignments are thereafter evaluated by a method derived from threading techniques. The algorithm for GenTHREADER is divided into three stages: alignment of sequences, calculation of pair potential as well as

solvation terms and finally, evaluation of the alignment using a neural network (Jones 1999b). GenTHREADER is advantageous as apart from being very fast, it requires no human intervention in the prediction process.

FUGUE (Shi et al. 2001) is another example of hybrid server for recognizing distant homologues by sequence-structure comparison. FUGUE utilizes environment-specific substitution tables and structure-dependent gap penalties. Here scores for amino acid matching and insertions/deletions are evaluated based on the local structural environment of each amino acid residue in a known structure. Local structural environment defined in terms of secondary structure, solvent accessibility, and hydrogen bonding status, are used by FUGUE to produce a high quality 3D protein model. FUGUE also encompasses scanning database of structural profiles, calculation of the sequence-structure compatibility scores and prediction of alignment of multiple sequences against multiple structures in order to enrich the conservation/variation information (Shi et al. 2001).

123D+ (http://pole-modelisation.univ-bpclermont.fr/prive/fiches_HTML/123D+.html) is another hybrid server which combines sequence profiles, secondary structure prediction and contact capacity potential to thread a protein sequence through asset of structures.

RaptorX (http://raptorx.uchicago.edu/) is a protein structure prediction hybrid server that excels in predicting 3D structures for protein sequences without close homologs in the PDB (Källberg et al. 2012). It predicts secondary and tertiary structures, contacts, solvent accessibility, disordered regions and binding sites for a given input sequence. Raptor X, first of all uses profile-entropy scoring method to assess the quality of information content in sequence profiles (Peng and Xu 2010). Thereafter it uses conditional random fields to integrate a variety of biological signals in a nonlinear threading score. Finally, multiple-template threading procedure (Peng and Xu 2009), which enables the use of multiple templates to model a single target sequence is used to produce a high quality protein 3D model.

MULTICOM toolbox (http://sysbio.rnet.missouri.edu/multicom_toolbox/) is another programme consisting of a set of protein structure and structural feature prediction tools. Secondary structure prediction, solvent accessibility prediction, disorder region prediction, domain boundary prediction, contact map prediction, disulfide bond prediction, beta-sheet topology prediction, fold recognition, multiple template combination and alignment, template-based tertiary structure modeling, protein model quality assessment, and mutation stability prediction are some of the functions facilitated by MULTICOM toolbox (Cheng et al. 2012).

Hybrid methods use various aspects for predicting an accurate protein tertiary structure. However Meta-servers discussed in the following section deals with generation of a consensus prediction of protein tertiary structure assembled from different servers.

5 Meta-Servers for Protein Tertiary Structure Prediction

Several meta-servers not only integrate protein structure predictions performed by various methods but also assemble and interpret the results to come up with a consensus prediction. This section deals with a comprehensive discussion of such meta-servers.

Pcons.net meta-server (Wallner et al. 2007) retrieves results from several publically available servers which are then analyzed and assessed for structural correctness using Pcons as well as ProQ, thus presenting the users a ranked list of possible models (Lundström et al. 2001). In combination of several publically available servers, Pcons.net meta-server also uses Reversed Position Specific BLAST (RPS-BLAST) to parse the sequence into structural domains by analyzing the significance and span of the best RPS-BLAST alignment.

3D-Jury (Ginalski et al. 2003) are the meta-servers which focus on the selection of high quality obtained from different servers. 3D-Jury, takes groups of models generated by a set of servers as input which are then compared with each other and a similarity score is assigned to each pair by MaxSub tool (Siew et al. 2000) followed by providing ranking to the models.

3D-SHOTGUN (Fischer 2003) meta server does not just select the best model but also refines initial models for building the protein structure model with high accuracy. 3D-SHOTGUN meta-predictor consists of three steps: (i) assembly of hybrid models, (ii) confidence assignment, and (iii) selection. 3D-SHOTGUN first assembles hybrid models from the initial models and then assigns scores to each of the assembled models by using the original models scores and the structural similarities between them. Thereby resulting a highly sensitive and ensuring a significantly higher specificity of the models than that of individual servers (Fischer 2003).

GeneSilico (Kurowski and Bujnicki 2003) is another meta-server which combines the useful features of other meta-servers available, but with much greater flexibility of the input in terms of user-defined multiple sequence alignments. However, there are several drawbacks reported in the current meta-servers including 3D-Jury (Ginalski et al. 2003) and GeneSilico (Kurowski and Bujnicki 2003). They take the initial threading inputs from remote computer which are occasionally shut down or are not available. Secondly, the instability of the algorithms of the remote servers is another drawback of these meta-servers (Wu and Zhang 2007).

LOMETS (Wu and Zhang 2007), overcomes the above drawbacks. It is one of the good performing meta-servers in which all nine individual threading servers are installed locally, which facilitates controlling and tuning of Meta-server algorithms in a consistent manner making the users able to obtain quick final consensus. It facilitates quick generation of initial threading alignments owing to the nine state of art threading programs that are installed and run in a local computer cluster, thus ensure faster results as compared to the traditional remote-server-based meta-servers. Based on TM-score, the consensus models generated from the top

Table 7 List of meta-servers for protein tertiary structure prediction along with the webpage URL and the description of the programs

S. no.	Name of the web server/group [URL]	Description of the web server/group
1	LOMETS (Wu and Zhang 2007) [http://zhanglab.ccmb.med.umich.edu/LOMETS/]	Meta server that includes locally installed threading programs FUGUE, HHpred, SPARKS. LOMETS generates the final models using a consensus approach
2	3D-Jury (Ginalski et al. 2003) [http://BioInfo.PL/Meta/]	The meta server provides access and results assessment from various remote predictors including, 3DPSSM, ESyPred3D, FUGUE, HHpred, mGenTHREADER etc.
3	GeneSilico (Kurowski and Bujnicki 2003) [https://genesilico.pl/meta2/]	The meta server provides access to various remote and local predictors including 3DPSSM, FUGUE, HHpred, mGenTHREADER, Pcons, Phyre, etc.
4	Pcons.net (Lundström et al. 2001), (Wallner et al. 2007) [http://pcons.net/]	The Pcons protocol analyzes the set of protein models and looks for recurring three-dimensional structural patterns and assigns a score
5.	3D-SHOTGUN (Fischer 2003) [http://bioinfo.pl/meta]	This meta-predictor consists of three steps: (i) assembly of hybrid models, (ii) confidence assignment, and (iii) selection

predictions by LOMETS were at least 7 % more accurate than the best individual servers. In addition to the 3D structure prediction by threading, LOMETS also provides highly accurate contact and distance predictions for the query sequences. The performance of LOMETS can be analyzed by the fact that average CPU time for a medium size protein (~200 residues) is less than 20 min when the programs are run in parallel on nine nodes of the cluster.

A List of Meta-servers for protein tertiary structure prediction along with the webpage URL and the description of the programs in Table 7.

The need for critical evaluation of various methods and developments in the field of protein structure prediction is successfully fulfilled by CASP meetings. The following section gives an overview of several agenda of CASP.

6 CASP

Protein structure prediction algorithms are constantly being developed and redefined to reach the experimental accuracy. Therefore, protein structure prediction strategies and methodologies are tested every 2 years in the Critically Assessment of techniques for protein Structure Prediction (CASP) meeting, which started since 1994. Since then, ten successful CASP meetings are over by 2012 and CASP11 is due in 2014. The participation by various research groups in the CASP are

increasing by each successive CASP meetings. The main goal of CASP is to obtain an in-depth and objective assessment of the current abilities and inabilities in the area of protein structure prediction. It critically evaluates the various protein structure programmes and servers besides assigning ranks for the same. CASP also tests the prediction accuracy of those protein sequences, whose solved experimental structures are kept undisclosed until the end of summit. Predictors/participants in CASP, fall into two categories. The first category comprises of teams of human participants who devote considerable time, usually a period of several weeks in order to model each target, to complete their work. The second category involves automatic servers with a target time period of 48 h for the completion of the assigned task (Moult 2005). Participant registration, target management, prediction collection and numerical analysis are all handled by the Protein Structure Prediction Center (http://predictioncenter.org/). The later also provides access to details of all experiments and results apart from providing a discussion forum for the CASP community. CASP also monitors progress in identification of disordered regions in proteins, and the ability to predict three-dimensional (3D) contacts which can be used as restraints during tertiary structure prediction of proteins (Moult et al. 2014). Ab initio modeling methods have also improved substantially and now we have topologically accurate models for small residues (<100 residues) having single domain non-template proteins due to regular CASP experiments (Kryshtafovych et al. 2014). Homology models vary greatly in accuracy depending on a number of factors, and for that reason CASP has encouraged the development of methods that can estimate overall accuracy of a model and accuracy at the individual amino acid level. The accuracy of homology models monitored by CASP, has improved dramatically, through a combination of improved methods. In CASP10, a new "contact-assisted" category has been introduced apart from the already existing previous categories. The idea in the CASP contact-assisted category is to investigate the extent to which experimental information is needed in order to deliver a given level of model accuracy besides encouraging the development of new methods for the same (Moult et al. 2014).

In CASP10 experiment, 114 protein sequences were released as modeling targets. Among these, 53 were designated "all groups" (human and server) targets. Finally 96 experimental structures were available for evaluation and assessment after cancellation of 18 targets (Moult et al. 2014). In CAS10, 217 groups registered, from several relevant communities. Finally, 41,740 predicted models submitted by 150 predictor groups were assessed as template-based modeling predictions where Zhang-Server, QUARK, PMS, Leecon and Zhang groups provided the most accurate models for the assessment units targets (Huang et al. 2014). Thus, CASP meeting is the best way to keep updated with the advancement in protein structure prediction strategies and methodologies.

Any development in the field of science is considered important if it has applications which are of significance to biological systems. The following section deals with various applications of the above discussed methods for protein structure prediction.

7 Applications of Protein Structure Prediction

Homology/Comparative modeling plays an essential role in structure based drug design. For example representative structures produced by in silico screening forms the basis of generation of three-dimensional structures of the remaining proteins encoded in the various genomes that can be predicted by homology modeling (Takeda-Shitaka et al. 2004). Comparative modeled proteins may be used for predicting the binding modes and affinities of different drug compounds as they interact with protein binding sites in structure-based drug design. Computational approach to this problem is usually termed as molecular docking. The goal of ligand-protein docking is to predict the predominant binding mode(s) of a ligand with a protein of known three-dimensional structure. Docking can be used to perform virtual screening on large libraries of compounds, rank the results, and propose structural hypotheses of how the ligands inhibit the target (Morris and Lim-Wilby 2008). However, it is widely accepted that docking with comparative models is more challenging and less successful than docking with crystallographic structures. Comparative models are not only useful in protein-ligand, but also useful in protein-protein docking (Vakser 1997).

Comparative models can also be used for testing and improving sequence structure alignment (Wolf et al. 1998). Based on the alignment of known structures, alignments can be well defined even for a new target sequence. Apart from the presence of functional motifs or the signature sequences, calculated electrostatic potential around the protein structure may help in predicting the protein function (Drew et al. 2011).

Protein models by comparative method can be also used to decipher important residues for biological activity as well as function of the protein. These models can be helpful in designing mutants to test hypotheses about protein functions (Boissel et al. 1993). On the basis of its primary sequence and the location of its disulfide bonds, erythropoietic hormone erythropoietin was modeled by homology modeling which predicts a four alpha-helical bundle motif, in common with other cytokines. Deletions of 5–8 residues from erythropoietin hormone erythropoietin protein within predicted alpha-helices resulted in the failure of export of the mutant protein from the cell (Boissel et al. 1993).

Comparative models can also be used to explore the substrate specificity in several enzymes. After the crystallization of the bacterial leucine transporter protein LeuT, development of 3-D computational models were used for structure-function studies on the plasmalemmal monoamine transporters (MATs). LeuT-based MAT models were used to guide elucidation of substrate and inhibitor binding pockets. Moreover, molecular dynamics simulations using these models provided insight into the conformations involved in the substrate translocation cycle (Manepalli et al. 2012).

Comparative models have been used in conjunction with virtual screening to successfully identify novel inhibitors over the past few years. Novel inhibitors of dihydrofolate reductase in *Typnosoma. cruzi* (the parasite that causes Chagas

disease) was discovered by docking into a comparative model to dihydrofolate reductase in *L. major*, a related parasite (Zuccotto et al. 2001). Since the crystal/NMR structure of various drug targets are not available so far, comparative models of drug targets could also be used for computational screening of new inhibitors for *Mycobacterium tuberculosis* drug target proteins (Gahoi et al. 2013).

Comparative modeled structure of cell receptors responsible for binding of foreign particles and thus causing diseases may also be used to study these interactions and may facilitate in investigating the mechanism. Comparative models can be also used to predict the antigenic epitopes. Mouse mast cell protease (mMCP) 1, mMCP-2, mMCP-4, and mMCP-5 models were used to predict immunogenic epitopes and surface regions that are likely to interact with proteoglycans (Sali et al. 1993).

Native PAGE results illustrated the presence of variations in number of isoforms of superoxide dismutase antioxidative enzymes in different cyanobacterial samples (Kesheri et al. 2011). Comparative modeling may be used to generate antioxidative enzymes models that may further help in studying the binding of metal cofactors with the isoforms. Comparative modeling may also be used to study the drug resistance in many vectors.

Garg et al. (2009) constructed the comparative model of dihydropteroate synthase protein which illustrated that novel point mutations at two positions may lead to sulphadoxine drug resistance in *Plasmodium falciparum*. Compararative models facilitates molecular replacement in X-ray structure/NMR models which allows refinement of a determined structure through the knowledge of already known structures.The computational prediction of protein structure also serves as an alternative to produce raw informations that may be validated by wet lab experiments. Following section produces an overview of further developments that may be made in the field of protein structure prediction.

8 Future Prospects

Homology modeling and protein threading are becoming more powerful and important for structure prediction along with the PDB growth and the improvement of prediction protocols. The error of a template-based model comes from template selection and sequence-template alignment. So, the identification of the best template is still a challenging task in protein structure prediction. However, HMM based template search algorithms like HHpred has solved this issue to some extent. Now, another big dilemma is of generation and choosing the correct alignment between target sequence and template sequence. Still, there is no set benchmark available for selection of the best alignment between the target and template sequence.

Model building is also one of the challenging task in structure prediction, in which a number of times it has been seen that side chains are not added properly in their proper conformations which mostly need structure refinement. Model

Refinement algorithms mostly don't fold a target structure to its possible native state. Model refinement is still obstructed with incorrect energy function, integrated with an additional complication of erroneous conformational search programs.

Model selection among hundreds of models generated by Modeller is still a challenging task. However, these issues have been solved to some extent by evaluating these models by various scores e.g. GDT-TS and TM Score etc. Improvement in the current algorithms is needed for the selection of the best model since till date there is no set benchmark for selection of the best model, even by top ranked servers as per CASP.

9 Conclusion

Correct prediction of secondary structure is the key to predict a good or satisfactory tertiary structure of the protein. Secondary structure not only helps in predicting the tertiary structure but also helps in predicting the function as well as sub-cellular localization of proteins. Staring from the amino acid propensity based secondary structure prediction methods, machine learning approaches has revolutionized the prediction accuracy of secondary structure from 60 to 80 %.

Tertiary structure prediction by bioinformatics or computational biology tools is always a challenging task for scientists. Ab initio folding and threading are computationally expensive methods for tertiary structure prediction which, also results in protein structural models having low accuracy. Tertiary structure prediction by ab initio folding/modelling still has a limitation due to searching a large number of conformations generated as well as absence of suitable potential functions as the number of amino acid increases. Another method is fold recognition where, the prediction accuracy is better than ab initio folding/modeling. Homology modeling, the third prediction method, has emerged as the sole method which can build the model close to X-ray crystal/NMR structure. Therefore, among the three methods, comparative or homology modeling is considered as the best method for protein structure predication with high accuracy in such cases where the sequence identity between the target and template sequence is more than 30 %. These comparative models may be used for structure based drug designing as well as virtual screening to identify novel inhibitors. Selecting the best model in homology modelling is one of the major challenging tasks to look into. In homology modeling, the major chances of error may be in loop modeling if long loop is present in the target protein molecule. Side chain modeling is another challenging area where prediction accuracy should be increased. Now a day, hybrid methods became popular because they club together a number of features such as structural alignments, solvent accessibility and secondary structure information in order to produce a protein model with high accuracy. Along with hybrid methods, several meta-servers are also available which integrate protein structure predictions performed by various methods that assemble and interpret the results to come up with a consensus model prediction. Nevertheless, we have not reached the pinnacle of that modelling

accuracy till date. However, it is interesting to discuss that, all our predictions may take a long time, while a cell takes only a few micro-seconds to fold a primary sequence into fully functional global native minima structure. Hence, further research to improve the algorithms is still needed to make the prediction close to native state or in other words close to fold adopted by the nature.

Acknowledgments Minu Kesheri is thankful to University Grant Commission, Govt. of India, New Delhi, for providing financial assistance in the form of research fellowship. Swarna Kanchan is thankful to University Grant Commission, Govt. of India, New Delhi for providing the financial support in the form of the Basic Science Research Fellowship under University Grant Commission (New Delhi) Special Assistance Programme to Department of Biological Sciences, Birla Institute of Technology and Science, Pilani, India.

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research, 25*(17), 3389–3402.

Arnold, K., Kiefer, F., Kopp, J., Battey, J. N., Podvinec, M., Westbrook, J. D., et al. (2009). The protein model portal. *Journal of Structural and Functional Genomics, 10*(1), 1–8.

Baker, D., & Sali, A. (2001). Protein structure prediction and structural genomics. *Science, 294* (5540), 93–96.

Bates, P. A., Kelley, L. A., MacCallum, R. M., & Sternberg, M. J. E. (2001). Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins: Structure, Function, and Bioinformatics, 45*(5), 39–46.

Benkert, P., Künzli, M., & Schwede, T. (2009). QMEAN server for protein model quality estimation. *Nucleic Acids Research, 37*(Web Server issue), W510–W514.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Research, 28*(1), 235–242.

Bettella, F., Rasinski, D., & Knapp, E. W. (2012). Protein secondary structure prediction with SPARROW. *Journal of Chemical Information and Modeling, 52*(2), 45–56.

Bhattacharya, D., & Cheng, J. (2013). 3Drefine: Consistent protein structure refinement by optimizing hydrogen bonding network and atomic-level energy minimization. *Proteins: Structure, Function, and Bioinformatics, 81*(1), 119–131.

Boissel, J. P., Lee, W. R., Presnell, S. R., Cohen, F. E., & Bunn, H. F. (1993). Erythropoietin structure-function relationships. Mutant proteins that test a model of tertiary structure. *Journal of Biological Chemistry, 268*(21), 15983–15993.

Bowie, J., Luthy, R., & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science, 253*(5016), 164–170.

Bradley, P., Misura, K. M. S., & Baker, D. (2005). Toward high-resolution de novo structure prediction for small proteins. *Science, 309*(5742), 1868–1871.

Chandonia, J.- M., & Karplus, M. (1995). Neural networks for secondary structure and structural class predictions. *Protein Science, 4*(2), 275–285.

Cheng, J., Li, J., Wang, Z., Eickholt, J., & Deng, X. (2012). The MULTICOM toolbox for protein structure prediction. *BMC Bioinformatics, 13*, 65.

Cheng, J., Randall, A. Z., Sweredoski, M. J., & Baldi, P. (2005). SCRATCH: A protein structure and structural feature prediction server. *Nucleic Acids Research, 33*(Web Server issue), W72–W76.

Chou, P. Y., & Fasman, G. D. (1974). Prediction of protein conformation. *Biochemistry, 13*(2), 222–245.

Cole, C., Barber, J. D., & Barton, G. J. (2008). The Jpred3 secondary structure prediction server. *Nucleic Acids Research, 36*(Web Server issue), W197–W201.

Colovos, C., & Yeates, T. O. (1993). Verification of protein structures: Patterns of non-bonded atomic interactions. *Protein Science, 2*(9), 1511–1519.

Combet, C., Jambon, M., Deléage, G., & Geourjon, C. (2002). Geno3D: Automatic comparative molecular modelling of protein. *Bioinformatics, 18*(1), 213–214.

Do, C. B., Mahabhashyam, M. S. P., Brudno, M., & Batzoglou, S. (2005). ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Research, 15*(2), 330–340.

Drew, K., Winters, P., Butterfoss, G. L., Berstis, V., Uplinger, K., Armstrong, J., et al. (2011). The Proteome folding project: Proteome-scale prediction of structure and function. *Genome Research, 21*(11), 1981–1994.

Eyrich, V. A., Martí-Renom, M. A., Przybylski, D., Madhusudhan, M. S., Fiser, A., Pazos, F., Valencia, A., Sali, A., & Rost, B. (2001). EVA: Continuous automatic evaluation of protein structure prediction servers. *Bioinformatics*, *17*(12), 1242–1243.

Fernandez-Fuentes, N., Madrid-Aliste, C. J., Rai, B. K., Fajardo, J. E., & Fiser, A. (2007). M4T: A comparative protein structure modeling server. *Nucleic Acids Research, 35*(Web Server issue), W363–W368.

Fischer, D. (2003). 3D-SHOTGUN: A novel, cooperative, fold-recognition meta-predictor. *Proteins: Structure, Function, and Bioinformatics, 51*(3), 434–441.

Fiser, A., Do, R. K. G., & Šali, A. (2000). Modeling of loops in protein structures. *Protein Science, 9*(9), 1753–1773.

Floudas, C. A. (2007). Computational methods in protein structure prediction. *Biotechnology and Bioengineering, 97*(2), 207–213.

Floudas, C. A., Fung, H. K., McAllister, S. R., Mönnigmann, M., & Rajgaria, R. (2006). Advances in protein structure prediction and de novo protein design: A review. *Chemical Engineering Science, 61*(3), 966–988.

Frishman, D., & Argos, P. (1997). Seventy-five percent accuracy in protein secondary structure prediction. *Proteins: Structure, Function, and Bioinformatics, 27*(3), 329–335.

Gahoi, S., Mandal, R. S., Ivanisenko, N., Shrivastava, P., Jain, S., Singh, A. K., et al. (2013). Computational screening for new inhibitors of M. tuberculosis mycolyltransferases antigen 85 group of proteins as potential drug targets. *Journal of Biomolecular Structure and Dynamics*, *31*(1), 30–43.

Garg, S., Saxena, V., Kanchan, S., Sharma, P., Mahajan, S., Kochar, D., et al. (2009). Novel point mutations in sulfadoxine resistance genes of Plasmodium falciparum from India. *Acta Tropica, 110*(1), 75–79.

Garnier, J., Osguthorpe, D. J., & Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of Molecular Biology, 120*(1), 97–120.

Geourjon, C., & Deléage, G. (1995). SOPMA: Significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Computer applications in the biosciences: CABIOS, 11*(6), 681–684.

Ginalski, K., Elofsson, A., Fischer, D., & Rychlewski, L. (2003). 3D-Jury: A simple approach to improve protein structure predictions. *Bioinformatics, 19*(8), 1015–1018.

Gong, H., & Rose, G. D. (2005). Does secondary structure determine tertiary structure in proteins? *Proteins: Structure, Function, and Bioinformatics, 61*(2), 338–343.

Guex, N., & Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-Pdb viewer: An environment for comparative protein modeling. *Electrophoresis, 18*(15), 2714–2723.

Heinig, M., & Frishman, D. (2004). STRIDE: A web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Research, 32*(Web Server issue), W500–W502.

Huang, Y. J., Mao, B., Aramini, J. M., & Montelione, G. T. (2014). Assessment of template-based protein structure predictions in CASP10. *Proteins: Structure, Function, and Bioinformatics, 82* (2), 43–56.

Hung, L.- H., Ngan, S.- C., Liu, T., & Samudrala, R. (2005). PROTINFO: New algorithms for enhanced protein structure predictions. *Nucleic Acids Research, 33*(Web Server issue), W77–W80.

Jauch, R., Yeo, H. C., Kolatkar, P. R., & Clarke, N. D. (2007). Assessment of CASP7 structure predictions for template free targets. *Proteins: Structure, Function, and Bioinformatics, 69*(8), 57–67.

Jayaram, B., Bhushan, K., Shenoy, S. R., Narang, P., Bose, S., Agrawal, P., et al. (2006). Bhageerath: An energy based web enabled computer software suite for limiting the search space of tertiary structures of small globular proteins. *Nucleic Acids Research, 34*(21), 6195–6204.

Jones, D. T. (1999a). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology, 292*(2), 195–202.

Jones, D. T. (1999b). GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *Journal of Molecular Biology, 287*(4), 797–815.

Jones, D. T., Taylort, W. R., & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature, 358*, 86–89.

Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., et al. (2012). Template-based protein structure modeling using the RaptorX web server. *Nature Protocols, 7*(8), 1511–1522.

Karplus, K., Barrett, C., & Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics, 14*(10), 846–856.

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research, 30*(14), 3059–3066.

Kelley, L. A., & Sternberg, M. J. E. (2009). Protein structure prediction on the Web: A case study using the Phyre server. *Nature Protocols, 4*(3), 363–371.

Kesheri, M., Richa, & Sinha, R. P. (2011). Antioxidants as natural arsenal against multiple stresses in cyanobacteria. *International Journal of Pharma and Bio Sciences, 2*(2), B168–B187.

Kim, D. E., Chivian, D., & Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research, 32*(Web Server issue), W526–W531.

Kim, H., & Park, H. (2003). Protein secondary structure prediction based on an improved support vector machines approach. *Protein Engineering, 16*(8), 553–560.

Klepeis, J. L., & Floudas, C. A. (2003). ASTRO-FOLD: A combinatorial and global optimization framework for ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. *Biophysical Journal, 85*(4), 2119–2146.

Klepeis, J. L., Wei, Y., Hecht M. H., & Floudas, C. A. (2005). Ab initio prediction of the three-dimensional structure of a de novo designed protein: A double-blind case study. *Proteins: Structure, Function, and Bioinformatics*, *58*(3), 560–570.

Kryshtafovych, A., Fidelis, K., & Moult, J. (2014). CASP10 results compared to those of previous CASP experiments. *Proteins: Structure, Function, and Bioinformatics, 82*(2), 164–174.

Kurowski, M. A., & Bujnicki, J. M. (2003). GeneSilico protein structure prediction meta-server. *Nucleic Acids Research, 31*(13), 3305–3307.

Lambert, C., Léonard, N., De, B. X., & Depiereux, E. (2002). ESyPred3D: Prediction of proteins 3D structures. *Bioinformatics, 18*(9), 1250–1256.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, *23*(21), 2947–2948.

Laskowski, R. A., Macarthur, M. W., Moss, D. S., & Thornton, J. M. (1993). PROCHECK: A program to check the stereochemical quality of protein structures. *Journal of Applied Crystalography, 26*, 283–291.

Lassmann, T., & Sonnhammer, E. (2005). Kalign—An accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics, 6*(1), 298.

Lisewski, A. M., & Lichtarge, O. (2006). Rapid detection of similarity in protein structure and function through contact metric distances. *Nucleic Acids Research, 34*(22), e152.

Liwo, A., Lee, J., Ripoll, D. R., Pillardy, J., & Scheraga, H. A. (1999). Protein structure prediction by global optimization of a potential energy function. *Proceedings of the National Academy of Sciences, USA, 96*(10), 5482–5485.

Lundström, J., Rychlewski, L., Bujnicki, J., & Elofsson, A. (2001). Pcons: A neural-network–based consensus predictor that improves fold recognition. *Protein Science, 10*(11), 2354–2362.

Luthy, R., Bowie, J. U., & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature, 356*, 83–85.

Manepalli, S., Surratt, C., Madura, J., & Nolan, T. (2012). Monoamine transporter structure, function, dynamics, and drug discovery: A computational perspective. *American Association of Pharmaceutical Scientists Journal, 14*(4), 820–831.

McGuffin, L. J., Bryson, K., & Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics, 16*(4), 404–405.

Morris, G. M., & Lim-Wilby, M. (2008). Molecular docking. *Methods in Molecular Biology*, *443*, 365–382.

Moult, J. (2005). A decade of CASP: Progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structural Biology, 15*(3), 285–289.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., & Tramontano, A. (2014). Critical assessment of methods of protein structure prediction (CASP)-round x. *Proteins: Structure, Function, and Bioinformatics, 82*(2), 1–6.

Moult, J., & Melamud, E. (2000). From fold to function. *Current Opinion in Structural Biology, 10*(3), 384–389.

Moult, J., Pedersen, J. T., Judson, R., & Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*, *23*(3), ii–iv.

Nair, R., & Rost, B. (2003). Better prediction of sub-cellular localization by combining evolutionary and structural information. *Proteins: Structure, Function, and Bioinformatics, 53* (4), 917–930.

Nair, R., & Rost, B. (2005). Mimicking cellular sorting improves prediction of subcellular localization. *Journal of Molecular Biology, 348*(1), 85–100.

Nanias, M., Chinchio, M., Ołdziej, S., Czaplewski, C., & Scheraga, H. A. (2005). Protein structure prediction with the UNRES force-field using replica-exchange Monte Carlo-with-minimization; comparison with MCM, CSA, and CFMC. *Journal of Computational Chemistry, 26*(14), 1472–1486.

Nielsen, M., Lundegaard, C., Lund, O., & Petersen, T. N. (2010). CPHmodels-3.0—Remote homology modeling using structure-guided sequence profiles. *Nucleic Acids Research, 38*(Web Server issue), W576–W581.

Norel, R., Petrey, D., & Honig, B. (2010). PUDGE: A flexible, interactive server for protein structure prediction. *Nucleic Acids Research, 38*(Web Server issue), W550–W554.

Peng, J., & Xu, J. (2009). *Boosting protein threading accuracy* (Vol. 5541, pp. 31–45). Lecture Notes in Computer Science.

Peng, J., & Xu, J. (2010). Low-homology protein threading. *Bioinformatics, 26*(12), i294–i300.

Pieper, U., Webb, B. M., Barkan, D. T., Schneidman-Duhovny, D., Schlessinger, A., Braberg, H., et al. (2011). MODBASE, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Research*, *39*(Database issue), D465–D474.

Pollastri, G., & McLysaght, A. (2005). Porter: A new, accurate server for protein secondary structure prediction. *Bioinformatics, 21*(8), 1719–1720.

Przybylski, D., & Rost, B. (2002). Alignments grow, secondary structure prediction improves. *Proteins: Structure, Function, and Bioinformatics, 46*(2), 197–205.

Przytycka, T., Aurora, R., & Rose, G. D. (1999). A protein taxonomy based on secondary structure. *Nature Structural & Molecular Biology, 6*(7), 672–682.

Rost, B., Sander, C., & Schneider, R. (1994). PHD-an automatic mail server for protein secondary structure prediction. *Computer Applications in the Biosciences: CABIOS, 10*(1), 53–60.

Runthala, A., & Chowdhury, S. (2013). Protein structure prediction: Are we there yet?. In D. P. Tuan, & L. C. Jain (Eds.), *Knowledge-based systems in biomedicine and computational life science* (Vol. 450, pp. 9–115). Berlin, Heidelberg: Springer.

Šali, A., & Blundell, T. L. (1993). Comparative Protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology, 234*(3), 779–815.

Sali, A., Matsumoto, R., McNeil, H. P., Karplus, M., & Stevens, R. L. (1993). Three-dimensional models of four mouse mast cell chymases. Identification of proteoglycan binding regions and protease-specific antigenic epitopes. *Journal of Biological Chemistry, 268*(12), 9023–9034.

Sen, T. Z., Jernigan, R. L., Garnier, J., & Kloczkowski, A. (2005). GOR V server for protein secondary structure prediction. *Bioinformatics, 21*(11), 2787–2788.

Shi, J., Blundell, T. L., & Mizuguchi, K. (2001). FUGUE: Sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *Journal of Molecular Biology, 310*(1), 243–257.

Siew, N., Elofsson, A., Rychlewski, L., & Fischer, D. (2000). MaxSub: An automated measure for the assessment of protein structure prediction quality. *Bioinformatics, 16*(9), 776–785.

Skolnick, J., Fetrow, J. S., & Kolinski, A. (2000). Structural genomics and its importance for gene function analysis. *Nature Biotechnology, 18*(3), 283–287.

Skolnick, J., Kihara, D., & Zhang, Y. (2004). Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. *Proteins: Structure, Function, and Bioinformatics, 56*(3), 502–518.

Skolnick, J., & Kolinski, A. (2002). A unified approach to the prediction of protein structure and function. In R. Friesner (Ed.), *A Computational Methods for Protein Folding* (Vol. 120, pp. 131-192). USA: Wiley.

Söding, J., Biegert, A., & Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, *33*(Web Server issue), W244–W248.

Su, E., Chiu, H.- S., Lo, A., Hwang, J.- K., Sung, T.- Y., & Hsu, W.- L. (2007). Protein subcellular localization prediction based on compartment-specific features and structure conservation. *BMC Bioinformatics, 8*(1), 330.

Takeda-Shitaka, M., Takaya, D., Chiba, C., Tanaka, H., & Umeyama, H. (2004). Protein structure prediction in structure based drug design. *Current Medicinal Chemistry, 11*(5), 551–558.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution, 30*(12), 2725–2729.

Teodorescu, O., Galor, T., Pillardy, J., & Elber, R. (2004). Enriching the sequence substitution matrix by structural information. *Proteins: Structure, Function, and Bioinformatics, 54*(1), 41–48.

Vakser, I. A. (1997). Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex. *Proteins: Structure, Function, and Bioinformatics, 29*(1), 226–230.

Wallner, B., Larsson, P., & Elofsson, A. (2007). Pcons.net: Protein structure prediction meta server. *Nucleic Acids Research, 35*(Web Server issue), W369–W374.

Wiederstein, M., & Sippl, M. J. (2007). ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Research, 35*(Web Server issue), W407–W410.

Wolf, E., Vassilev, A., Makino, Y., Sali, A., Nakatani, Y., & Burley, S. K. (1998). Crystal structure of a GCN5-related N-acetyltransferase: Serratia marcescens aminoglycoside 3-N-acetyltransferase. *Cell, 94*(4), 439–449.

Wu, S., & Zhang, Y. (2007). LOMETS: A local meta-threading-server for protein structure prediction. *Nucleic Acids Research, 35*(10), 3375–3382.

Wu, S., & Zhang, Y. (2008). MUSTER: Improving protein sequence profile–profile alignments by using multiple sources of structure information. *Proteins: Structure, Function, and Bioinformatics, 72*(2), 547–556.

Xu, Y., & Xu, D. (2000). Protein threading using PROSPECT: Design and evaluation. *Proteins: Structure, Function, and Bioinformatics, 40*(3), 343–354.

Xu, J., & Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics, 26*(7), 889–895.

Xu, D., & Zhang, Y. (2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Structure, Function, and Bioinformatics, 80*(7), 1715–1735.

Yang, Y., Faraggi, E., Zhao, H., & Zhou, Y. (2011). Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics, 27*(15), 2076–2082.

Yi, T.-M., & Lander, E. S. (1993). Protein secondary structure prediction using nearest-neighbor methods. *Journal of Molecular Biology, 232*(4), 1117–1129.

Zemla, A., Venclovas, Č., Fidelis, K., & Rost, B. (1999). A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins: Structure, Function, and Bioinformatics, 34*(2), 220–223.

Zhang, Y., Arakaki, A. K., & Skolnick, J. (2005). TASSER: An automated method for the prediction of protein tertiary structures in CASP6. *Proteins: Structure, Function, and Bioinformatics, 61*(7), 91–98.

Zhang, Y., Kolinski, A., & Skolnick, J. (2003). TOUCHSTONE II: A new approach to Ab initio protein structure prediction. *Biophysical Journal, 85*(2), 1145–1164.

Zhang, Y., & Skolnick, J. (2004). Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins. *Biophysical Journal, 87*(4), 2647–2655.

Zhou, H., & Zhou, Y. (2005). Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins: Structure, Function, and Bioinformatics, 58*(2), 321–328.

Zuccotto, F. Z. M., Brun, R., Chowdhury, S. F., Di, L. R., Leal, I., Maes, L., et al. (2001). Novel inhibitors of Trypanosoma cruzi dihydrofolate reductase. *European Journal of Medicinal Chemistry, 36*(5), 395–405.