Progress in Biophysics and Molecular Biology 173 (2022) 72–82

Contents lists available at ScienceDirect

Progress in Biophysics and Molecular Biology

journal homepage: www.elsevier.com/locate/pbiomolbio

# Different methods, techniques and their limitations in protein structure prediction: A review

Vrushali Bongirwar [a,*], A.S. Mokhade [b]

[a] Shri Ramdeobaba College of Engineering and Management, Nagpur, India
[b] Visvesvaraya National Institute of Technology, Nagpur, India

ARTICLE INFO

*Keywords:*
Conformational sampling
Protein secondary structure prediction
Protein tertiary structure prediction

ABSTRACT

Because of the increase in different types of diseases in human habitats, demands for designing various types of drugs are also increasing. Protein and its structure play a very important role in drug design. Therefore researchers from different areas like mathematics, medicines, and computer science are teaming up for getting better solutions in the said field. In this paper, we have discussed different methods of secondary and tertiary protein structure prediction (PSP), along with the limitations of different approaches. Different types of datasets used in PSP are also discussed here. This paper also tells about different performance measures to evaluate the prediction accuracy of PSP methods. Different software's/servers are available for download, which are used to find the protein structures for the input protein sequence. These softwares will also help to compare the performance of any new algorithm with other available methods. Details of those softwares are also mentioned in this paper.

## 1. Introduction

Protein is an important part of the human body and is a means to perform all biological activities. Proteins are chains of twenty different types of amino acids, and these chains are considered as a protein primary sequence [Christian and Anfinsen, (1973) and Rost et al., 1994]. The length of the string and the order of amino acids play a vital role in the proper functioning of the protein and its structure decides their biological functions that are primarily used in drug design. Therefore, researchers are working on finding the secondary and tertiary structure of a protein from the sequence of amino acids. Since proteins generally fold to their lowest free energy states, the problem can be stated very simply: identify the lowest free energy state of the protein chain. According to Christian and Anfinsen, (1973), amongst all the conformations of the protein, the native state in which it performs its best biological function is having its lowest free energy state.

Proteins are also called polymers; which is a chain of monomers also called amino acids. Generally, 20 different types of amino acids are in existence, given in Table 1.

Each amino acid is represented either by its name (ex. Arginine), or three-letter word (ex. Arg), or one letter word (ex. R). These amino acids are joined together by a peptide bond to form a protein. The structure of a protein is characterized in the following four different ways. The sequence of proteins is called the primary structure of proteins. Repeated regular conformations on the polypeptide chain are called the secondary structures of proteins. A 3D structure of a protein is called the tertiary structure of a protein, which is obtained from a secondary structure. The quaternary structure of a protein is the collection of several polypeptide chains, as well as the addition of the non-protein element. Finding tertiary and quaternary structures is still a very difficult problem because of its structural complexity. Fig. 1[a-d] shows different types of structures of proteins.

Fig. 1(a) shows the protein primary sequence. Any sequence of polypeptides will have a single amine (N-terminus) at one end and carboxylic acid (C-terminus) at the other end [X. H. Hao (2017)]. In Fig. 1(b), the amide plane is shown, also called a secondary protein structure. Each peptide bond shown in RED forms an amide plane also called as polypeptide backbone, shown in yellow due to side chains R1, R2, and R3. The angle between consecutive amide planes is called the torsion angle. The torsion angle is used to describe geometric relations between two molecules. Fig. 1 (c) is a 3D representation of protein structure, which is generally formed by the hydrophobic effect. The Quaternary structure of a protein is a more condensed form of 3D protein structure shown in Fig. 1(d).

* Corresponding author.
*E-mail address:* vrushalibongirwar@gmail.com (V. Bongirwar).

https://doi.org/10.1016/j.pbiomolbio.2022.05.002
Received 23 August 2021; Received in revised form 16 April 2022; Accepted 11 May 2022
Available online 16 May 2022

**Table 1**
Types of amino acids and their representations.

| SN | Name of Amino acid | Three letters word | One letter word |
|---|---|---|---|
| 1 | Alanine | Ala | A |
| 2 | Cysteine | Cys | C |
| 3 | Leucine | Leu | L |
| 4 | Isoleucine | Ile | I |
| 5 | Lysine | Lys | K |
| 6 | Proline | Pro | P |
| 7 | Methionine | Met | M |
| 8 | Histidine | His | H |
| 9 | Phenylalanine | Phe | F |
| 10 | Tyrosine | Tyr | Y |
| 11 | Tryptophan | Trp | W |
| 12 | Asparagine | Asn | N |
| 13 | Valine | Val | V |
| 14 | Glutamine | Gln | Q |
| 15 | Serine | Ser | S |
| 16 | Threonine | Thr | T |
| 17 | Glycine | Gly | G |
| 18 | Aspartic acid | Asp | D |
| 19 | Glutamic acid | Glu | E |
| 20 | Arginine | Arg | R |

Experimental-based methods like X-ray crystallography, nuclear magnetic resonance, cryo-electron microscopy contribute protein structures to the datasets. These datasets consist of protein sequences as shown in Fig. 1(a) and are further used by the researchers for evaluating their algorithms. The resolution of the experimentally determined structure is the smallest distance between two distinguishable features. It means, if a structure has a resolution of 4 Å (0.4 nm), it is possible to differentiate two structures only when two atoms are 4 Å apart or more. Finding structures for higher resolution are more complex. Table 2 shows the features revealed at different resolutions.

Accurate protein structure prediction influences the design of the structure of the drugs. Drugs are biochemical or physiological substances that affect the operations in the body. It may be a single or mixture of compounds. It interacts with specific targets in the body to modify its structure. These targets are generally proteins. The development of a new drug is a time consuming and costly process. The steps involved in drug design are drug discovery and development, preclinical research, clinical development, FDA review and FDA postmarketing safety monitoring. Traditional drug design approaches are structure-based and ligand-based. If the three-dimensional structure of the receptor is known, then a structure-based design is used. These structures are available through X-ray crystallography, NMR, or electron microscopy [Berman et al., 2000, Burley et al., 2019]. The ligand-based design is applied when no structural data is available for biological targets. But, the cost associated with developing drugs is high with experimental methods mentioned above. But computer scientists are using software to identify novel molecules and proteins to speed the drug discovery process along with reduced cost. Use of machine learning for drug discovery, it is possible to develop cheaper, safer medicines in a shorter time. Mostly, machine learning techniques are used in the drug discovery process. Thus, we have so far seen the types of protein

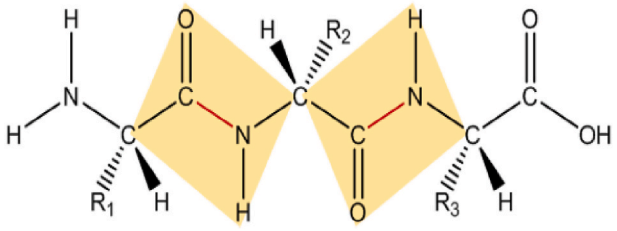


**Fig. 1b.** Secondary Structure of Protein [The amide plane of tri-peptide] [EzMol_Data (2019)].

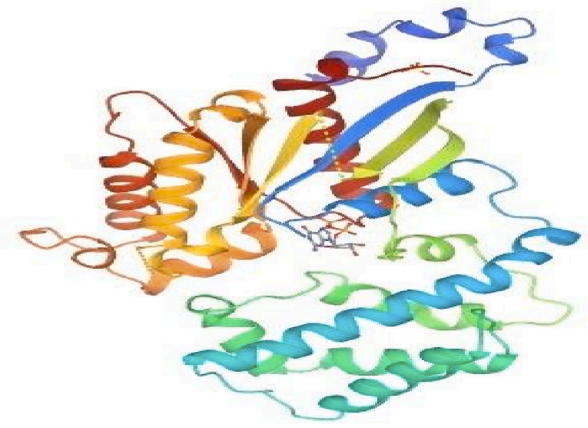

**Fig. 1c.** Tertiary structure of protein [Mixon M.B. et al., 1995].

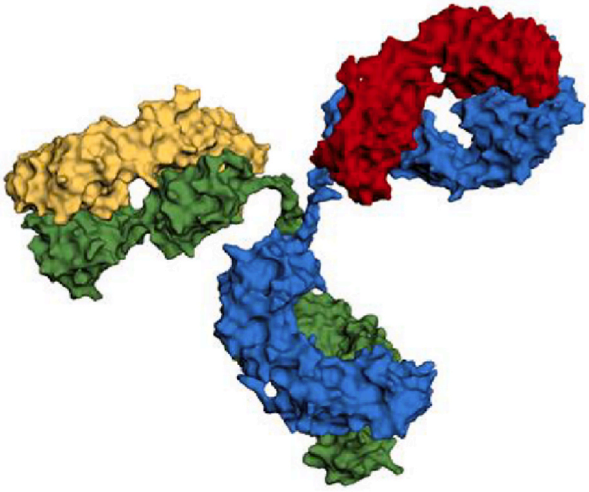

**Fig. 1d.** Quaternary structure of Protein [Mixon M.B. et al., 1995].

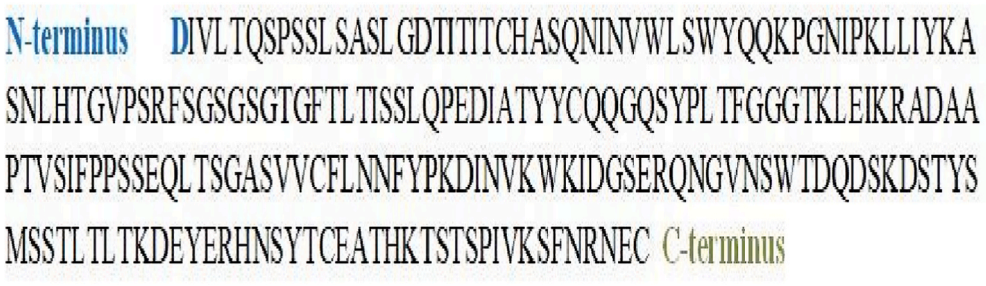


**Fig. 1a.** Primary structure of protein [EzMol_Data (2019)].

**Table 2**
Features at different resolutions.

| Resolution | Features |
|---|---|
| 6 Å | General shape of proteins |
| 4 Å | Backbone of PSS |
| 3.5 Å | Side chain visibility |
| 2.7 Å | Side chain visibility and start to see water molecules |
| 1.5 Å | Start reaching atomic resolution |
| 1.2 Å | Can distinguished any two covalently linked atoms |

structure, their resolution, and the role of protein structure prediction in drug design. The next paragraph gives the outline of this paper.

Section 2 discusses conformational sampling, Section 3 and 4 discuss secondary and tertiary structure prediction methods respectively. Different datasets and available softwares/servers for secondary structure prediction are discussed in section 5. Section 6 discusses various performance measures followed by results and comparison of various previously proposed algorithms in section 7. Lastly, we have concluding remarks on the said work and directions for future work on protein structure prediction.

## 2. Conformational sampling

A conformer is a kind of arrangement of a polypeptide in such a way that total force acting on it is compensated by each other because of pairing of two polypeptides. Each conformer is one saddle point on the potential energy surface and sampling is a method to find that saddle point. Obtaining global minimum-energy conformations of polypeptides is a very hard optimization problem. This problem is difficult to solve because of the presence of many local minima in the conformational space and the ruggedness of the energy landscape in the conformational space. This section discusses various conformational sampling methods.

Monte-Carlo minimization called a global minimization approach is used in protein folding by Z. Li and Scheraga (1987). This method is generally used in finding multiple minima in multimodal problems. The hybrid model of conformation space annealing and the genetic algorithm is proposed to find the global minimum energy conformation of polypeptides. This method finds not only the global minimum but also finds other local minima [J. Lee et al. (1997)]. D. E. Kim et al. (2009) proposed a method for determining the computer power required to accurately predict the structure of the protein. Here, the conformational sampling problem is reformulated as a combinatorial sampling problem in discrete space for easier processing. The transformation from conformational space to a discrete space is achieved by projecting it into torsion bins, secondary structures, and beta-contacts. It also uses Rosetta trajectories for the reverse transformation. Another hybrid model of dynamic fragment assembly (DFA) and conformational space annealing (CSA) are proposed to predict the Ab initio protein structure [Lee et al., 2011]. The weights of the parameters used in the energy function were optimized by CSA. The work aims to develop a force field for the protein structure prediction and this force field is readily available to the community, which can be used for protein structure prediction. B. Olson and Shehu (2014) has identified the problem of template-free protein structure prediction as a multi-objective optimization problem. A rosetta-based memetic algorithm is proposed in larger protein structure prediction, where fragment assembly fails to solve it [M. Garza-Fabre et al. (2016)]. In this method, new genetic operators along with new search schemes were designed for exploration and retention of the diverse conformation having low energies. An abstract convex underestimation method is proposed, where higher dimensional conformation space is dimensionally reduced to a feature space by the feature extraction technique [X. H. Hao et al. (2016)]. In this method, the tight lower bound estimation of searching space is obtained for better-searching direction and to define valid searching space. Followed by dimensionality reduction, a combination of fragment assembly and Monte-Carlo is used to generate a series of meta-stable conformations by sampling in the conformation space. Table 3 shows the limitations of different methods used in conformational sampling. The use of "_____" in all the below-mentioned tables denotes that we cannot be able to draw any clear inferences on the given approach.

Most of the conformational sampling strategies discussed above fail to find global minimum energy conformations because of the stagnation, lesser input features, and lesser parameter tuning. Different variants of evolutionary algorithms, swarm intelligence-based methods or nature-inspired methods can solve this problem effectively. Next section discusses different methods implemented for protein secondary structure prediction.

## 3. Secondary protein structure prediction

The protein secondary structure (PSS) problem is generally considered an intermediate problem in primary sequence and tertiary structures. Typically, PSS has two common elements of protein, namely α-helix and β-sheets, as shown in Fig. 2. In α-helix, the right-handed helical structure of polypeptide chains is formed. The hydrogen bond is formed in the same helical structure. The core of the helix is tightly packed and all the side chains are projected outward. In β-sheets, the hydrogen bond is formed between different chains. The non-repetitive protein structure called coil or loops are also part of the protein. PSS is traditionally characterized as three general states: helix (H), strand (E), and coil (C). Further, these three states are extended into eight states: helix (G), α-helix (H), π-helix (I), β-stand (E), bridge (B), turn (T), bend (S), and others (C) [Zhang et al., 2018]. Different techniques of PSS prediction are proposed in the following subsections.

**Table 3**
Conformational sampling methods.

| SN | Reference | Methodology/Methods | Limitation/Future Scope/ Findings |
|---|---|---|---|
| 1 | Z. Li and Scheraga (1987) | Monte-Carlo minimization | Stagnation at local minima, while finding the best solution |
| 2 | Lee et al. (1997) | The hybrid model of conformation space annealing and the genetic algorithm finds global as well as local minima, slowly decreasing the value of $D_{cut}$ leads to the conformational space annealing and change in seed conformation | No strategy is defined for defining a single initial conformation. |
| 3 | Kim et al. (2009) | The conformational space is projected into torsion bins, secondary structures, and beta-contacts, also uses Rosetta trajectories | Only a few important linchpin features are used, backbone torsion angle reduces the amount of sampling required |
| 4 | Lee et al., 2011 | DFA with CHARMM was used to define suitable energy functions. | The parameter optimization used in this study is neither complete nor rigorous. |
| 5 | B. Olson and Shehu (2014) | It is finding out that evolutionary algorithms can enhance exploration capability | How to balance energetic diversification is not studied here |
| 6 | Garza-Fabre et al. (2016) | It uses Rosetta as a local search strategy | – |
| 7 | Hao et al. (2016) | Abstract convex underestimation method: a higher dimensional conformation space is dimensionally reduced to feature space, followed by dimensionality reduction | The setting of the parameter 'M' should be researched in detail. |

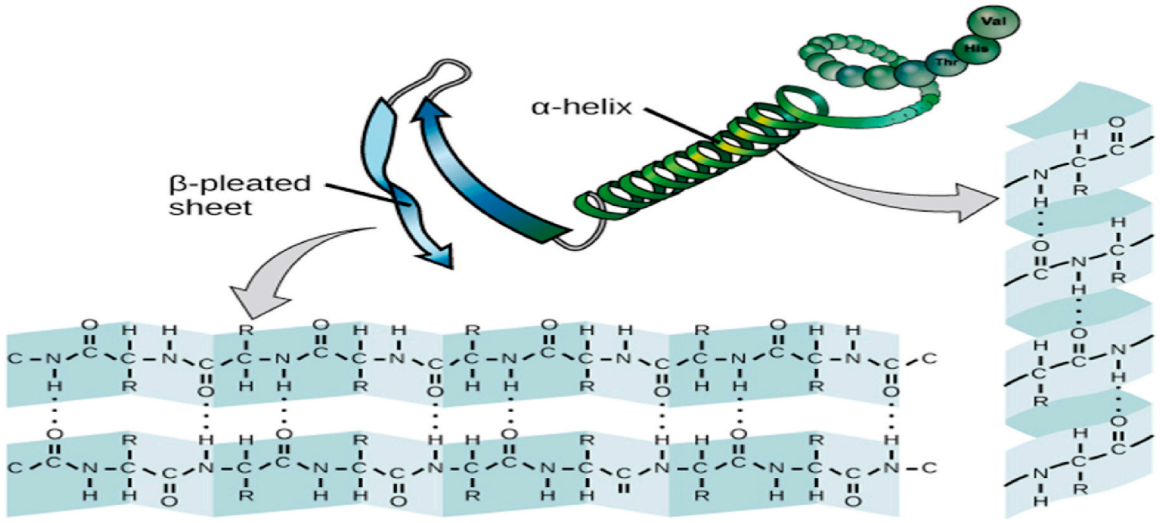


**Fig. 2.** Secondary Structure of Protein [α-helix and β-sheets] [Conway, O.J. et al., (2020)].

### 3.1. Machine learning-based methods

A neural network was primarily applied by L. H. Holley and Karplus (1989) to solve the PSS problem. An ensemble model is constructed, based on different neural networks and their results are combined through majority voting methods for the final solution. The nearest neighbor method was applied for PSS prediction in 1993 [Yi TM and Lander (1993)]. This method is based on probability distributions. Improved accuracy in PSS prediction is achieved through multiple sequence alignments, balanced training, and structure context training [Rost and Sander, 1993]. Proper database search method, alignment method, and scoring mechanism are proposed in two levels neural network-based PSS prediction method [Cuff JA and Barton, 2000]. PSIPRED is a popular protein structure prediction method based on the feed-forward neural network [McGuffin et al., (2000)]. It performs an analysis on the output obtained from PSI-BAST. The average Q3 score for the PSIPRED is 76.5%. The first attempt of applying a support vector for SS prediction is done by J. J. Ward et al. (2003). The author has used a combination of neural networks and support vector machines for PSS prediction. Initially, the marginal probability at each residue is found and then the highest probability is used as predicted SS. Other approaches based on support vectors [Hua S(2001)] and Fuzzy k-nearest [Kim S–Y et al., (2006)] are also proposed for increasing the accuracy of PSS prediction. The iterative hidden semi-Markov model considers patterns of statistically significant amino acid correlation at structural segment borders [Aydin et al., 2006]. Three distance-based classifiers namely minimum distance, K-nearest neighbor, and fuzzy k-nearest neighbor are used for PSS prediction [Ashish Ghosh (2008)]. Dynamic Bayesian network using multivariate Gaussian distribution is proposed for the SS prediction [Yao XQ et al., (2008)]. It considers the profile of the secondary structures as well as dependencies between profiles of neighboring residues for modeling the PSI-Blast profile of protein. P. Ghanty et al. (2013) uses probability-based features for single-sequence SS prediction. Faruk Berat Akcesme (2015) proposed a K-Nearest Neighbor-based PSS prediction model, where physicochemical features and the position-specific scoring matrix is used to solve the PSS problem. A filtering technique is also used here for refining the predicted results. A conditional neural-based method is proposed for 8-class PSS prediction [Wang et al., 2010]. Various neural network-based PSS predictions are also proposed by Tomasz Smolarczyk et al. (2020), Mirko Torrisi et al. (2020), Meng and Kurgan, (2016), Daniel Rademaker et al. (2020).

### 3.2. Deep learning-based methods

Porter, a server specially designed for 3 class prediction is dependent on Recurrent Neural Network (RNN), accurate coding of input profiles, long-range incorporation, large scale ensembles of predictors [Gianluca Pollastri and McL ysaght (2005)]. Ab Initio structure prediction is used for PSS, which is a combination of three deep neural networks (DNN), where the first two DNN predicts the Secondary structure(SS) and the third DNN is used to generate the refined predictions [M. Spencer et al. (2015)]. Wang S (2016) has proposed deep convolutional neural fields (DeepCNF) which are the combination of conditional random field and shallow neural network for secondary structure prediction. Deep Multi-scale convolutional neural networks are proposed to predict eight class PSS [A. Busiay et al. (2016)] and Q8 accuracy is found to be around 77.0% on CB513. After achieving better results in the image processing domain, a multilayer shift-and-stitch convolutional architecture is proposed to solve the PSS problem [Z. Lin et al. (2016)]. Shift-and-stitch is a technique, which is used to compute a convolutional network score for each window in a sequence. This is one of the faster PSS prediction methods. Porter 5, Bidirectional RNN model is proposed in 3 states and 8 states PSS prediction [Torrisi, M. (2018)]. Porter 5 is composed of a bidirectional convolutional and recurrent neural network. But its Q3 accuracy is mentioned to be 84% only. Deeper profiles are cascaded with porter 5, where Q3 accuracy is 2% more as compared to porter 5 [Mirko Torrisi et al., (2020)]. NetSurf-2.0 is a sequence-based PSS prediction tool, which uses the Convolutional and LSTM model for achieving better accuracy and runtime [Michael S. K. (2019)]. Mirko Torrisi et al. (2020) has demonstrated the role of deep learning-based methods on a wider set of databases in PSS prediction.

### 3.3. Hybrid methods

Various hybrid methods [H. Hasic et al. (2017)] involving a combination of nature-inspired and deep learning are proposed for SS prediction [B. Çarkli Yavuz (2018)]. Sixty-five years of secondary protein structure prediction journey is provided by Yuedong Yang et al. (2016). Three states to eight state investigations of proteins, higher theoretical limits of prediction accuracy are discussed here. Iterative use of predicted secondary structure and torsion angle is proposed to improve the secondary structure, backbone torsion angle Cα – atom-based angles and dihedral angles, and solvent accessible surface area [Heffernan et al., 2015].

Table 4 shows a summary of some of the secondary protein structure prediction methods.

**Table 4**
Secondary protein structure prediction methods.

| SN | Reference | Methodology/Methods | Limitation/Future Scope/ Findings |
|---|---|---|---|
| 1 | L. H. Holley and Karplus (1989) | Neural Network | Q3 Prediction accuracy was only 63% |
| 2 | Yi TM and Lander (1993) | Nearest-neighbor model | The selection of the value of n as nearest neighbor is the main challenge |
| 3 | Rost and Sander, 1993 | Sequence profile and neural network | This method is not valid for membrane proteins and other non-globular or non-water-soluble proteins. |
| 4 | Cuff JA and Barton, 2000 | Jnet | – |
| 5 | Ward et al. (2003) | Support vector machine | Works well only on a smaller training set, therefore accuracy needs to be a compromise |
| 6 | Gianluca Pollastri and McLysaght (2005) | Porter: CNN based server | Accuracy affects with training set |
| 7 | Aydin et al., 2006 | Hidden Semi-Markov Model | – |
| 8 | Ghosh and Parai (2008) | Minimum distance, KNN, and fuzzy KNN | Some post-processing techniques might have given better results; an ensemble of classifiers can be created for better results |
| 9 | Yao XQ et al., (2008) | Dynamic Bayesian network (DBN) using multivariate Gaussian distribution; further combine with neural network for better Q3 accuracy | No clear advantage of combining DBN with NN is described here. |
| 10 | Wang et al., 2010 | Conditional neural field; Find complex nonlinear relationship between protein features and SS; Uses interdependence among SS types of adjacent residues. | Long-range inter-residue interactions are not considered; Change in neural network design may retrieve more information from sequence profiles and the chemical property profile of amino acids. |
| 11 | Ghanty et al. (2013) | Hybrid model; position-specific and position-independent features are used by NEUROSVM model | Prediction accuracy is around 67%, which can be increased by Two-stage SVM or any other machine learning methods |
| 12 | Heffernan et al., 2015 | Iterative use of predicted secondary structure and torsion angle | Adding instead of adding only PSS, more features could lead to additional improvement |
| 13 | Faruk Berat Akcesme (2015) | KNN; position-specific scoring matrix (PSSM) profiles, propensity matrix of amino acids in three conformations and three features; hydrophobicity, net charges, and side-chain mass | Hybrid algorithm; More complex algorithm |
| 14 | Spencer et al. (2015) | DNN; Input profiles for DNN are trained using three types of features: the amino acid residues themselves (RES), the Position Specific Scoring Matrix (PSSM) information, and the Atchley factors (FAC). | – |
| 15 | Wang et al., 2016 | Deep convolutional neural fields (DeepCNF); models complex structure relationships of deep hierarchical structure, as well as interdependency between adjacent secondary structure labels. | Finding solutions for 2.7 Å or lower sequence is still a difficult problem |
| 16 | Busiay et al. (2016) | Uses Inception, ResNet, and DenseNet with Batch Normalization; also uses conditioning for past structure labels | This approach suffers less from overfeeding induced by conditioning |
| 17 | Lin et al. (2016) | MUST-CNN | The fully convolutional network can be used for speeding up the process |
| 18 | Hasic et al. (2017) | Window length selection; Binarization of inputs and outputs; Construction of neural network based on MLP | Though this method gives 65% accuracy, still it is not sufficient as compared to the bulkiness of the algorithm. Time consumption is more to carry out all the steps. |
| 19 | Çarkli Yavuz (2018) | Clonal selection (CS) is used to train the model, whereas classification is then performed by a multilayer perceptron (MLP). . MLP is trained with stochastic gradient descent using backpropagation. | Why CSA, why not other nature-inspired methods are used is not provided by the author. But the application of the CS algorithm before MLP gives better results. The CS method can also be replaced with any other nature-inspired methods. |
| 20 | Torrisi, M., Kaleel et al., (2018) | Porter 5 | Q3 accuracy is around 84% |
| 21 | Torrisi et al. (2019) | Deeper profiles with porter 5 | Accuracy is still an issue as compared to the state of the art methods |
| 22 | Michael S. K. (2019) | Net-Surf 2.0; Convolutional and LSTM model | More accuracy and runtime as compared to previously developed servers |

Machine learning (ML) and deep learning (DL) based methods along with various hybrid methods were proposed in recent years. Most of the work is done to improve Q3 accuracy. But it could not reach the theoretical limit as per Yuedong Yang et al. (2016). The accuracy is more on higher resolution and lesser protein sequence length. Very little work is done for Q8 accuracy improvement. The next section discusses different methods in tertiary protein structure prediction.

## 4. Tertiary protein structure prediction

The function of protein is clearly understood by its 3D structure. The 3D structure of protein helps in understanding biochemical processes in the human body. Mis-folding of these structures leads to various diseases, such as cataracts, mad cow, and Alzheimer's, etc. Therefore, the information about the high-resolution structure of proteins helps to understand the molecular details of it and further helps in the drug design. Determining 3D structure experimentally is costly and time-consuming. Therefore, De novo protein structure prediction methods are used to predict the 3D structure of proteins computationally.

### 4.1. Machine learning-based methods

The first neural network model, the BigNet model [George L. Wilcox (1990)] based on a backpropagation neural network is proposed to solve the protein tertiary structure prediction problem. K. T. Simons et al. (1997) uses simulated annealing and Bayesian scoring function to generate tertiary protein structure based on fragments of unrelated proteins. Protein structure prediction using Rosetta [C. A. Rohl et al.

(2004)] is a widely used method since its inception. Rosetta method uses only sequence information for the generation of models with 3–6 Å $C_\alpha$ root mean square deviation (RMSD). It uses a torsion space representation, where the protein backbone conformation is specified by different torsion angles. In Rosetta strategy, several fluctuating local structures come together to build a compact conformation. Compact structures are then combined by using Monte-Carlo(MC) simulated annealing search, for optimizing energy function in protein design. The design of local movements for optimal energy value is proposed in a template-free protein structure prediction strategy (QUARK) [Xu and Zhang, 2012]. Here, the query sequence is broken into 1–20 residue fragments and the full-length model is built, which is based on Mont-Carlo simulations and optimized knowledge-based force fields.

UniCon3D, a probabilistic model for de novo protein structure prediction, is proposed to improve the performance of traditional energy conformation methods [D. Bhattacharya (2016)]. This model tries to minimize the energy function based on simultaneous capture of local structural preference and side-chain conformational space and further performs conformational sampling via stepwise synthesis and assembly of foldons. CONFOLD, a two-stage modeling method uses predicted contacts and secondary structures for tertiary structure prediction [B. Adhikari et al. (2015)]. I-TASSER, an Iterative Threading ASSEmbly Refinement [Y. Zhang (2008)] is a hierarchical protein structure modeling approach, based on a scoring function computed by the relative clustering density and score of multiple threading templates. In a fragment-based protein structure prediction method with an estimation of distribution algorithm, the learning takes place based on previously searched decoys by taking the uniform probability mass function over the fragment libraries [D. Simoncini et al. (2012)]. This method is named as EdaFold. EdaFold is a fragment-based approach that guides the search by periodically updating probability distribution over the fragment libraries, whereas EdaRose combines simulated annealing sampler from Rosetta AbRelax and estimation of distribution algorithm from EdaFold. Based on these hypotheses, two different probability update policies are proposed in EdaFold with Rosetta algorithm namely, Cluster-based variation ($EdaRose_c$) and an energy-based ($EdaRose_{en}$) [D. Simoncini et al. (2017)]. Apart from this, various machine learning models that are based on feedforward neural network [H. Bohr et al. (1990)], two input neurons [Fariselli and Casadio, (1999)], radial basis function neural network [Zhang et al., 2005], two neural networks (SPINE-2D) [B. Xue et al. (2009)], SVM based contact predictors [S. Wu et al. (2011)] is proposed to find protein tertiary structures.

### 4.2. Deep learning-based methods

Bidirectional recurrent neural network [A. Vullo et al. (2006)] classifies components of principal eigenvectors for tertiary structure prediction. It also uses SS information and hydrophobicity scale for the same. AlphaFold, an improved protein structure prediction based on deep learning is proposed by Senior et al., 2020. In this method, a neural network model is designed to make accurate distance predictions between a pair of residues to find structured information. Then a gradient descent algorithm is applied to find the optimized structure of a protein. The use of residue-residue coevolution derived restraints is proposed to improve the de novo protein structure prediction accuracy [Ovchinnikov et al., 2016]. These restraints are used during sampling and refinement toward searching native sampling. To avoid the problem of the false effect of indirect and transitive contacts in the already available network deconvolution methods, a balanced network deconvolution (BND) is proposed [H. P. Sun et al. (2015)]. In this method, co-evolution-based contact maps are generated and optimized using BND to improve the prediction accuracy of protein contacts. Two residual neural networks based ultra-deep learning model is proposed, which combine evolutionary coupling and sequence information [S. Wang et al. (2017)]. This deep neural network can accurately model higher-quality contact prediction structures. A template and ab initio structure prediction-based server, namely RBO Aleph are proposed by M. Mabrouk et al., 2015 for difficult protein targets, where template detection is difficult. The proposed model builds an approximate model of the energy landscape iteratively and uses it to guide to move towards low energy. RosettaFold rapidly generates accurate protein structure model [Baek et al., 2021]. It uses simultaneous processing of 1D sequence, 2D distance and 3D coordinate information to generate accuracy approaching towards theoretical limits. Therefore, it can be considered as a solution for experimental methods. A neural network based model called as Alphafold2 significantly improves the accuracy of protein structure prediction [Jumper et al., 2021]. It consist of novel NN architecture and training methods based on evolutionary, physical and geometric constraints of protein structures. It outperforms state of the art methods.

### 4.3. Evolutionary computation based methods

A genetic algorithm without using a mutation operator was developed for 3-torsion angle representations [Cui et al., 1998]. The fitness function consists of hydrophobic interactions and Van Der Waals contact measures. A piece of residue-residue distance information is used in a two-stage distance feature-based optimization algorithm [TDFO] for optimizing the energy function [G. Zhang et al. (2019)]. Local and global mutation strategy and state selection strategy are proposed to maintain the balance between exploration and exploitation of the conformation space. Due to the greater impact of evolutionary algorithms in protein structure prediction, T. Braun et al., (2015) has combined the advantages of evolutionary algorithms with the resolution adopted the structural recombination approach of Rosetta, called RASREC. An improved differential evolution involving secondary structure and contact information (SCDE) based method is proposed to enhance the accuracy of tertiary structure prediction [G. J. Zhang et al. (2018)]. Two-stage multi-Subpopulation differential evolution (MDE) is proposed for sampling conformation space, which is further used in de novo protein structure prediction [X. H. Hao (2017)]. In this sampling process, clusters of conformations are formed based on energy scores. Representative conformations from each cluster are then selected as best decoys. The tabu search algorithm [X. Zhang et al. (2010)] included in mutation operator is also proposed for torsion space representations.

### 4.4. Hybrid methods

Comparative protein modeling called MODELLER is proposed by Sali and Bundell to find the most relative structure of a given sequence [Sali and Bundell (1993)]. Sequence alignment with template structures, its atomic coordinates and script file are required as inputs to find 3-D modeling. The best probability density functions (pdfs) are derived from the alignments. Finally, a 3-D model is obtained by optimizing molecular pdfs. A new ab initio protein structure prediction model based on lattice representation, consisting of $C_\alpha$, atoms, $C_\beta$ atoms, and side-chain centers of mass, named as TOUCHSTONE II is developed by Zhang et al., 2003. The conformational search is accelerated by using a parallel hyperbolic sampling algorithm. This technique was used for small and large protein prediction. Chemical Shift Restraints, based on NMR chemical shifts, are proposed by A. Cavalli et al. (2007). The entire process of tertiary structure prediction is broken down into three steps. In the first phase, secondary structures are predicted based on experimental chemical shifts. In the second phase, a library of trial conformations for each of these protein fragments is generated by screening a database. This is used to search for similar sequences of secondary structures and chemical shift patterns. In the third phase, these structures are assembled and refined by using the appropriate scoring function. A new method based on pseudo contact shifts (PCS) using Rosetta [C. Schmitz et al. (2012)] is proposed for protein tertiary structure prediction. Pseudocontact shifts are useful for the detection of changes in chemical shifts. TDFO is found to be better than Rosetta and other

distance-based algorithms on 35 benchmark proteins. RBO Aleph is the first public web server implementing this novel ab initio prediction approach. It is one of the top servers in the template-free modeling category. Table 5 shows a summary of tertiary protein structure prediction methods.

Various methods based on machine and deep learning as well as on evolutionary computing are proposed to predict protein tertiary structures. However, it is hard to find protein tertiary structures because PSS prediction methods do not give 100% accuracy. The accuracy of protein tertiary structure can be improved by using parameters tuning in ML and DL techniques, properly generating and using conformational information, attending greater accuracy in PSS prediction, etc. These methods are evaluated on different types of datasets. Different types of servers are also available for comparing our results. These datasets and servers are discussed in the next section.

## 5. Datasets and softwares

Various datasets are available for researching the field of Bioinformatics. Table 6 is the list of mostly used/basic datasets for PSS prediction.

There are other datasets, which are formed by extracting some of the sequence-structure information from the datasets mentioned in Table 4. These datasets are CullPDB [Wang and Dunbrack (2003,2005)], CB513 [Cuff JA and Barton (1999, 2000)] (created from RS126 and CB396), TS115 [Hefferman R. (2017)], CAMEO [Wang S. et al., (2016)], ASTRAL40 [Fox NK et al., (2014)], Carugo-338 [Carugo (2000)], Manesh215 [Ahmad et al., 2003] etc.

Various servers are available, where anyone can find the protein structure by providing input sequences. Table 7 shows the different servers/software available for determining protein structures.

## 6. Performance measures

The performance measure is an important parameter in any research. This is because it helps us to evaluate and compare the algorithm with other algorithms. Different types of measures are available in different domains. This section discusses different performance measures used in protein structure prediction. Protein tertiary structure prediction uses RMSD, GDT_TS, MaxSub, TM-Score to evaluate the performance of an algorithm.

### 6.1. Root-mean-square deviation (RMSD)

RMSD is a Structural quality measure where the structural quality is usually assessed concerning another reference structure in terms of individual Cα atoms. The aim is to minimize RMSD. Lower values for this measure, expressed in Ångströms (Å), indicate better correspondence to the native structure [Kabsch, 1976].

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left| r_i^{model} - r_i^{real}\right|^2}$$

where $r_i^{model}$ and $r_i^{real}$ are the positions of ith Cα atoms in the model and the real protein, respectively.

### 6.2. Global distance test-total score (GDT_TS)

This method considers both local and global structures to rank the level of similarity between two structures and find 3D similarities between protein structures. Its value lies in [0,1]. An online LGA service is accessible at http://PredictionCenter.llnl.gov/local/lga [Zemla A. (2003)].

**Table 5**
Tertiary protein structure prediction methods.

| SN | Reference | Methodology/Methods | Limitation/Future Scope/ Findings |
|---|---|---|---|
| 1 | Simons et al. (1997) | The bayesian scoring function is useful for series expansion in the residue of the protein database, whereas simulated annealing is to rapidly develop small protein structures. | The scoring function defined here could not consistently distinguish native-like structures from non-native structures. It requires much improvement in its scoring function formation. |
| 2 | Zhang et al., 2003 | It is based on lattice representation consisting of $C_{\alpha,}$ atoms, $C_{\beta}$ atoms, and side-chain centers of mass; Focused on energy terms is optimization | However, this method is used for both small and large protein prediction; efficiency and accuracy are major issues. |
| 3 | Rohl et al. (2004) | It uses only sequence information for the generation of models with 3–6 Å $C_{\alpha}$ RMSD. It also uses Monte-Carlo(MC) simulated annealing search | This is the base model de novo protein structure prediction and is widely used for the Protein structure prediction |
| 4 | Cavalli et al. (2007) | Chemical Shift Restraints based on NMR chemical shifts | The method must have been tested on a higher resolution of proteins. The quantitative structural analysis needs to be carried out |
| 5 | Zhang (2008) | A target sequence is first threaded to search for the possible folds by using the hidden Markov model, PSI-Blast, Needleman-Wunsch, and Smith-Waterman algorithm for the same. Also uses Monte-Carlo for conformational space searching, SPICKER for clustering structure trajectories, followed by fragment assembly simulation to refine the model. | Very small Conformational space is used by the ITASSER. Easy decoys were easily clustered into the first cluster, but lowered rank clusters were difficult to find |
| 6 | Xu and Zhang, 2012 | It is based on Mont-Carlo simulations. The energy force field of QUARK covers three levels of structural packing: atom-, residue, and topology | Due to the complexity of β-protein, global energy minima could not be reached. The incorrect trap in local and global optima is due to the simultaneous effect of force field and search engine. |
| 7 | Schmitz et al. (2012) | Pseudocontact shifts (PCS) using Rosetta called PCS-ROSETTA | It is necessary to know chemical shifts and PCSs data of proteins. |
| 8 | Simoncini et al. (2012) | EdaFold; build protein models by assembling short fragments; Fragment-based method | Initially random orders of fragments are selected randomly for putting them into the solution. Providing a fixed method instead of random insertion may result in better performance. Simulated annealing and hill-climbing are used as a sample and minimization function. There are much better techniques like swarm optimization, nature-inspired algorithms for optimization. |
| 9 | Adhikari et al. (2015) | CONFOLD, a two-stage modeling method; In the first stage contacts and secondary structures translates into distance, | Repetitive task of generating contacts; required the certain value (minimum number) of contacts for the best results |

(continued on next page)

**Table 5** (*continued*)

| SN | Reference | Methodology/Methods | Limitation/Future Scope/ Findings |
|---|---|---|---|
| | | dihedral angle, and hydrogen bond restraints for tertiary structures; In the second stage, these structures are used to regenerate the contacts for final prediction | |
| 10 | T. Braun et al., (2015) | RASREC, A hybrid model of evolutionary algorithms and resolution adapted structural recombination approach of Rosetta | The refinement step runs only once. Prediction accuracy can be tested by running the refinement step iteratively. |
| 11 | Sun et al. (2015) | Co-evolution based contact maps are generated and optimized using balanced network deconvolution | – |
| 12 | Mabrouk et al., (2015) | RBO Aleph; uses evolutionary and physicochemical information for residue-residue contact prediction and uses this information for conformational sampling. | – |
| 13 | Ovchinnikov et al., 2016 | Residue-residue coevolution derived restraints are used | – |
| 14 | Bhattacharya et al. (2016) | UniCon3D model to improve the performance of traditional energy conformation methods | This method is purely based on Coarse-grain sampling and scoring function. Change or refinement in scoring function and sampling may give better results |
| 15 | Hao et al. (2017) | MDE; In the first stage, MDE finds all the models with the ultra shape recognition method, whereas in the second stage, differential evolution preserves these modals in the further evolution process. | The inaccuracy of the energy model results in the lowest energy may misguide the MDE towards wrong conformation sampling. To solve this problem more accurate energy functions need to be designed. Another work is to balance the low energy and high reasonability of conformations. A model can be designed to ensure the diversity of the population. A solution for the inaccuracy of the energy model may be the distance profile and dihedral angle profile. |
| 16 | Simoncini et al. (2017) | Two different probability update policies are proposed in EdaFold with the Rosetta algorithm namely, Cluster-based variation ($EdaRose_c$) and energy-based ($EdaRose_{en}$) | Any alternate probability update policies and generation of new search dynamics can be tested |
| 17 | Wang et al. (2017) | An ultra-deep learning model is proposed which combine evolutionary coupling and sequence information | This method is inspired by the ultra-deep learning model proposed in 2015 for the image processing problem. The author tries to map image-processing problems with protein contact prediction, but find those input parameters are more complex as compared to the image processing domain. |
| 18 | Zhang et al. (2018) | SCDE; uses two different selection strategies, the first is the conformation generation based on Secondary structure; the contact-based strategy for generating compact and low-energy conformations. | Accuracy of the system decreases with an increase in protein fold complexity and sequence length of the protein. However, the proposed method provides highly accurate results, but could always result in a high-quality model. Future work is to provide more attention to utilizing the conformation information for high-quality models of protein. |
| 19 | Zhang et al. (2019) | Bisecting k-means is used to extract feature information from the distance profiles, which are further used in the similarity model. Evolutionary algorithm-based strategy is also established | Stagnation criteria are not considered while designing an algorithm. Multimodal and multi-objective techniques can be used to achieve greater accuracy. |
| 20 | Senior et al., 2020 | AlphaFold, protein structure prediction based on deep learning | – |

**Table 6**
Datasets.

| SN | Dataset | Details |
|---|---|---|
| 1 | CASP dataset [CASP 1- CASP 14] from 1994 to 2020 | Critical Assessment of Structure Prediction (CASP) helps to advance the methods of identifying protein structure from sequence. CASP experiments are used for detecting the current state of protein structure prediction, highlighting progress and future efforts in the said direction. Total 14 experiments are available from 1994 to 2020 namely CASP 1-CASP 14. The numbers of targets are different in different CASP experiments. |
| 2 | Protein Data Bank (PDB) | Protein Data Bank consists of 3D shapes of proteins, nucleic acids, and complex assemblies. It is a collection of 1,72,175 protein structures. |
| 3 | UniProt Knowledge Bases | The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent, and rich annotation. Most of the protein sequences are derived from the translation of the coding sequences (CDS). It is a collection of 184,998,855 protein structures |

$$GDT_{TS} = 100 \times \frac{\sum_{d_i} \frac{GDT_i}{NT}}{4}$$

where *GDTi* is the number of α-carbons of a prediction not deviating from more than an established cutoff di (in Å) from the α -carbons of the targets, after optimal superimposition. NT is the number of amino acids of the protein and $d_i \in \{1, 2, 4, 8\}$ expressedin Å.

### 6.3. MaxSub

This is the heuristic-based algorithm proposed for finding the largest 'well predicted' Subset. MaxSub returns the S score based on the user input model and an experimental structure. The value of the score lies in the range of 0 and 1. MaxSub is available at: http://www.cs.bgu.ac.il/~dfischer/MaxSub/MaxSub.html [Siew et al., 2000].

**Table 7**
List of servers.

| SN | Server | Web Link | Year |
|---|---|---|---|
| 1 | PSIPRED | http://bioinfadmin.cs.ucl.ac.uk/index.html | 2000 |
| 2 | APSSP2 | http://crdd.osdd.net/raghava/apssp2/ | 2002 |
| 3 | Robetta | http://robetta.bakerlab.org | 2004 |
| 4 | SCRATCH | http://download.igb.uci.edu/download.html | 2005 |
| 5 | Porter | http://distilldeep.ucd.ie/porter/ | 2005 |
| 6 | I-TASSER | https://zhanglab.ccmb.med.umich.edu/I-TASSER/ | 2010 |
| 7 | SPARKS-X | https://sparks-lab.org/server/sparks-x/ | 2011 |
| 8 | QUARK | https://zhanglab.ccmb.med.umich.edu/QUARK/ | 2012 |
| 9 | CABS-flex | http://biocomp.chem.uw.edu.pl/CABSflex/ | 2013 |
| 10 | Jpred | http://barton.ebi.ac.uk/ | 2015 |
| 11 | BND | http://www.csbio.sjtu.edu.cn/bioinf/BND/ | 2015 |
| 12 | clmDCA | http://protein.ict.ac.cn/clmDCA/ | 2015 |
| 13 | IJCAI2016 | https://github.com/icemansina/IJCAI2016 | 2016 |
| 14 | DeepCNF | http://raptorx.uchicago.edu/download/ | 2016 |
| 15 | SPIDER3 | https://sparks-lab.org/server/spider3/ | 2018 |
| 16 | trRosetta | https://yanglab.nankai.edu.cn/trRosetta/ | 2020 |

### 6.4. Template modeling score (TM-score)

TM-scores were used to assess the quality of protein structure templates. The value always lies between (0, 1]. Higher values represent better templates[Zhang and Skolnick, 2004]. TM-score program is available on http://bioinformatics.buffalo.edu/TM-score.

$$\delta(s1,s2) = \left\{ (maxov(s1,s2) - minov(s1,s2)); minov(s1,s2); int\left(\frac{len(s1)}{2}\right); int\left(\frac{len(s2)}{2}\right) \right\}$$

$$TM_{score} = max\left[\frac{1}{L_N}\sum_{i=1}^{L_T}\frac{1}{1+\left(\frac{d_i}{d_o}\right)^2}\right]$$

Various criteria are also defined for Protein Secondary Structure Prediction. These criteria are summarized as follows.

### 6.5. Q3 success criteria

The percentage of the total number of residues correctly predicted for helices (qH), strands (qE), and coils (qC) assigned to the three secondary structure states.

### 6.6. Q8 success criteria

The Q8 accuracy is the percentage of the amino acid residues for which the predicted secondary structure labels are correct. These structures are helix (G), α-helix (H), π-helix (I), β-stand (E), bridge (B), turn (T), bend (S), and others (C).

In general Q score, which calculates the percentage of residues for which the predicted secondary structures are correct is defined as

$$Q_m = 100\% \times \frac{\sum_{i=1}^{m} M_{ii}}{N_{res}}$$

where m is the number of classes; $N_{res}$ is the total number of residues and $M_{ii}$ is the number of correctly predicted residues in state i.

### 6.7. Matthew's correlation coefficient

Matthew's correlation coefficients are not influenced by the percentage of true positives in a sample and are considered a good way of evaluating different methods. A result is a number between −1 and 1, where 1 represents a perfect coincidence, −1 signifies a total inequality and 0 indicates that the predictions do not correlate with the true values [Matthews B. (1975)].

### 6.8. The segment overlap measure (SOV)

SOV measures the accuracy with different segments (of any length) of a sequence into account. The segment overlaps the quantity to measure for a single conformational state x (H, E, or C).

$$Sov(i) = 100 \times \left[\frac{1}{N}\sum_{i \in \{H,E,C\}}\sum_{S(i)}\frac{minov(s1,s2)+\delta(s1,s2)}{maxov(s1,s2)} \times len(s1)\right]$$

where.

N: sum of $N(i)$ over all three conformational states, where $N(i)$ *is*

$$N(i) = \sum_{S(i)} len(s1) + \sum_{S'(i)} len(s1)$$

$minov(s1,s2)$ : length of the actual overlap of $s1$ and. $s2 maxov(s1,s2)$ : total extent for which either of the segments s1 and s2 has a residue in state i.

$len(s1)$: is the number of residues in segment $s1$

And $\delta(s1,s2)$ is defined as follows,

In addition to these measures, two additional per class measures are precision (proportion of correct positive predictions) and recall (proportion of correctly predicted positive structures) [Zhang B. (2018), Wang S.(2016)]. Fuzzy F score and fizzy overlap (FOV) are less popular measures [Jiang et al., (2017), Lee, 2006].

**Table 8**
Comparative study of various protein secondary structure prediction methods.

| Reference | Method | Test Data Set | Q3 Accuracy (%) | SOV (%) |
|---|---|---|---|---|
| Cuff JA and Barton, 2000 | Jnet | Protein data bank | 76.4 | 82.9 |
| Ward et al. (2003) | SVM | Protein data bank | 75.44 | 70.74 |
| Hua (2001) | SVM_Jury | CB513 | 73.5 | 76.2 |
| | | RS126 | 71.6 | 74.6 |
| Yao XQ et al., (2008) | DBNN | CB513 | 78.1 | 74 |
| Ghanty et al. (2013) | NSVM | CB513 | 68.3 | 66.7 |
| | | RS126 | 71.5 | – |
| Torrisi et al. (2018) | Porter 5 | 3154 proteins from PDB from June 14, 2017 | 84.19 | 81.19 |
| Zhang et al., 2018 | **eCRRNN** | **CB513** | **87.3** | 0.723 |
| | | **CASP10** | **87.8** | 0.740 |
| | | **CASP11** | **85.9** | 0.738 |
| | | **CASP12** | **83.7** | 0.698 |
| Fang et al., 2018 | MUFOLD-SS | CB513 | 70.63 | – |
| | | CASP10 | 76.47 | – |
| | | CASP11 | 74.51 | – |
| | | CASP12 | 72.10 | – |
| Busia and Jaitly, 2017 | Conditioned CNN | CB513 | 70.03 | – |

## 7. Results & comparison

Table 8 shows a comparative study of various protein secondary structure prediction methods based on neural/deep networks. The test was conducted on different datasets mentioned in section 5. Comparison with other methods including evolutionary computation is not quoted here, because most of them have used other measures to evaluate the algorithm. It is also difficult to quote all the methods in one format for the comparison, because of the use of different test sets and accuracy measures. Table 8 shows that deep learning-based eCRRNN methods perform well amongst all other methods. The maximum Q3 accuracy is 87.8 on CASP10 and 87.3 on the CB513 dataset, which is not much good for small-medium length protein sequences. One more difficulty in comparing different methods is a protein sequence length. Therefore, this factor is not considered here while comparing. It is necessary to achieve this much accuracy on a larger protein sequence.

## 8. Conclusion

This paper discusses the review of different approaches used in protein structure prediction. Protein structure prediction is an important problem in Bioinformatics, due to its application in drug design. Researchers have put their efforts into designing different approaches for protein secondary and tertiary structure prediction from the input protein sequence. Due to the complexity in finding tertiary structures, finding the secondary structure of the protein is one of the core problems in this field. But the maximum accuracy for PSS prediction is 88% for small-medium length protein sequences. Achieving better accuracy for larger-length protein sequences is still an issue.

Based on the study, the given areas are identified as a scope of research: Building a model for improved SS prediction for both Q3 and Q8 accuracy by Optimal window size selection, Effective binarization of inputs (for low processing power), and Machine learning techniques for classification (maybe a hybridization of two or more from Convolutional neural networks; Radial Basis Function; Support Vector machines; Ensemble learning; Naïve Bayes classification).

## Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References


Adhikari, B., Bhattacharya, D., Cao, R.Z., Cheng, J.L., 2015. Confold:residue-residue contact-guided ab initio protein folding. Proteins:Struct. Funct. Bioinf. 83 (8), 1436–1449.

Ahmad, S., Gromiha, M.M., Sarai, A., 2003. Real value prediction of solvent accessibility from amino acid sequence. Proteins 50 (4), 629–635. https://doi.org/10.1002/prot.10328 PMID: 12577269.

Aydin, Z., Altunbasak, Y., Borodovsky, M., 2006. Protein secondary structure prediction for a single-sequence using hidden semi-Markov models. BMC Bioinf. 7 (7), 178.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J.H., Rodrigues, A.V., van Dijk, A.A., Ebrecht, A.C., Opperman, D.J., Baker, D., 2021. Accurate prediction of protein structures and interactions using a three-track neural network. Science (New York, N.Y.) 373 (6557), 871–876. https://doi.org/10.1126/science.abj8754.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The protein data bank. Nucleic Acids Res. 28, 235–242.

Bhattacharya, D., Cao, R.Z., Cheng, J.L., 2016. Unicon3D: de novo protein structure prediction using united- residue conformational search via stepwise, probabilistic sampling. Bioinformatics 32 (18), 2791–2799.

Bohr, H., Bohr, J., Brunak, S., Cotterill, R.M.J., Fredholm, H., Lautrupt, B., Petersen, S.B., 1990. A novel approach to prediction of the 3-dimensional structures ofprotein backbones by neural networks. FEBS Lett. 261 (1), 43–46.

Braun, T., Leman, J., Lange, O.F., 2015. Combining evolutionary information and an iterative sampling strategy for accurate protein structure prediction. PLoS Comput. Biol. 11 (12), e1004661.

Burley, S.K., Berman, H.M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., Christie, C., Dalenberg, K., Duarte, J.M., Dutta, S., et al., 2019. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Res. 47, D464–D474.

Busia, A., Jaitly, N., 2017. Next-step Conditioned Deep Convolutional Neural Networks Improve Protein Secondary Structure Prediction, Conference on Intelligent Systems for Molecular Biology & European Conference on Computational Biology 2017.

Busiay, A., Collins, J., Jaitly, N., 2016. Protein secondary structure prediction using deep multi-scale convolutional neural networks and next-step conditioning. RECOMB 2017 arXiv:1611.01503v1 [cs.LG].

Çarkli Yavuz, B., 2018. Prediction of protein secondary structure with CSA and MLP. IEEE Access 6. https://doi.org/10.1109/ACCESS.2018.2864665.

Carugo, O., 2000. Predicting residue solvent accessibility from protein sequence by considering the sequence environment. Protein Eng. 13 (9), 607–609. https://doi.org/10.1093/protein/13.9.607 PMID: 11054454.

Cavalli, A., Salvatella, X., Dobson, C.M., Vendruscolo, M., 2007. Protein structure determination from nmr chemical shifts. Proc. Natl. Acad. Sci. U. S. A 104 (23), 9615–9620.

Christian, B., Anfinsen, 1973, Jul 20. Principles that govern the folding of protein chains. Science 181 (4096), 223–230.

Conway, O.J., An, Y., Bejagam, K.K., Deshmukh, S.A., 2020. Development of Transferable Coarse-Grained Models of Amino Acids.

Cuff, J.A., Barton, G.J., 1999. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins 34 (4), 508–519. https://doi.org/10.1002/(SICI)1097- 0134(19990301)34:4<508::AID-PROT10>3.0.CO;2-4, 1999, PMID: 10081963.

Cuff, J.A., Barton, G.J., 2000. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins: Struct. Funct. Bioinf. 40, 502–511.

Cui, Y., Chen, R.S., Hung, W., 1998. Protein folding simulation with genetic algorithm and super secondary structure constraints. Proteins 31, 247–257.

EzMol_Data, 2019. http://www.sbg.bio.ic.ac.uk/~ezmol/EzMol_Data/images/student/pdf/.

Fang, C., Shang, Y., Xu, D., 2018. MUFOLD-SS: new deep inceptioninside-inception networks for protein secondary structure prediction. Proteins 86 (5), 592–598. https://doi.org/10.1002/prot.25487 PMID: 29492997.

Fariselli, P., Casadio, R., 1999. A neural network based predictor of residue contacts in proteins. Protein Eng. 12, 15–21.

Faruk Berat, Akcesme, 2015. Protein secondary structure prediction based on physicochemical features and PSSM by KNN. Southeast Eur. J. Soft Comput. 4 (1), 37–42.

Fox, N.K., Brenner, S.E., Chandonia, J.-M., 2014. SCOPe: structural classification of proteins–extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res. 42 (Database issue), D304–D309. https://doi.org/10.1093/nar/gkt1240 PMID: 24304899.

Garza-Fabre, M., Kandathil, S.M., Handl, J., Knowles, J., Lovell, S.C., 2016. Generating, maintaining, and exploiting diversity in a memetic algorithm for protein structure prediction. Evol. Comput. 24 (4), 577–607.

Ghanty, P., Pal, N.R., Mudi, R.K., 2013. Prediction of protein secondary structure using probability based features and a hybrid system. J. Bioinf. Comput. Biol. 11 (5), 1350012.

Ghosh, Ashish, Parai, Bijnan, 2008. Protein secondary structure prediction using distance based classifiers. Int. J. Approx. Reason. 47, 37–44.

Hao, X.H., Zhang, G.J., Zhou, X.G., Yu, X.F., 2016. A novel method using abstract convex underestimation in ab-initio protein structure prediction for guiding search in conformational feature space. IEEE ACM Trans. Comput. Biol. Bioinf 13 (5), 887–900.

Hao, X.H., Zhang, G.J., Zhou, X.G., 2017. Conformational space sampling method using multi-subpopulation differential evolution for de novo protein structure prediction. IEEE Trans. NanoBioscience 16 (7), 618–633.

Hasic, H., Buza, E., Akagic, A., May 2017. A hybrid method for prediction of protein secondary structure based on multiple artificial neural networks. In: Proc. 40th Int. Conv. Inf. Commun. Technol., Electron. Microelectron. MIPRO, p. 1195_1200.

Heffernan, R., Paliwal, K., Lyons, J., et al., 2015. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. Sci. Rep. 5, 11476. https://doi.org/10.1038/srep11476.

Holley, L.H., Karplus, M., 1989. Protein secondary structure prediction with a neural network. Proc. Nat. Acad. Sci. USA 86 (1), 152_156.

Hua, S, 2001. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J Mol Biol. 308 (2), 397–407. https://doi.org/10.1006/jmbi.2001.4580. In this issue.

Jiang, Q., Jin, X., Lee, S.-J., Yao, S., 2017. Protein secondary structure prediction: a survey of the state of the art. J. Mol. Graph. Model. 76, 379–402.

Jumper, J., Evans, R., Pritzel, A., et al., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

Kabsch, W., 1976. A solution for the best rotation to relate two sets of vectors. Acta Crystallogr. A 32 (5), 922–923. https://doi.org/10.1107/s0567739476001873.

Kim, S.-Y., Sim, J., Lee, J., 2006. Fuzzy K-Nearest Neighbor Method for Protein Secondary Structure Prediction and its Parallel Implementation Computational Intelligence and Bioinformatics, pp. 444–453.

Kim, D.E., Blum, B., Bradley, P., Baker, D., 2009. Sampling bottlenecks in de novo protein structure prediction. J. Mol. Biol. 393 (1), 249–260.

Lee, J., 2006. Measures for the assessment of fuzzy predictions of protein secondary structure. Proteins 65 (2), 453–462. https://doi.org/10.1002/prot.21164 PMID: 16948155.
Lee, J., Scheraga, H.A., Rackovsky, S., 1997. New optimization method for conformational energy calculations on polypeptides: conformational space annealing. J. Comput. Chem. 18 (9), 1222–1232.
Lee, J., Sasaki, T.N., Sasai, M., Seok, C., Lee, J., 2011. De novo protein structure prediction by dynamic fragment assembly and conformational space annealing. Proteins: Struct. Funct. Bioinf. 79 (8), 2403–2417. https://doi.org/10.1002/prot.23059.
Li, Z., Scheraga, H.A., 1987. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. Proc. Natl. Acad. Sci. U. S. A 84 (19), 6611–6615.
Lin, Z., Lanchantin, J., Qi, Y., 2016. MUST-CNN: a multilayer shiftand-stitch deep convolutional architecture for sequence-based protein structure prediction. AAAI 2016 arXiv:1605.03004v1 [cs.LG].
Mabrouk, M., Putz, I., Werner, T., Schneider, M., Neeb, M., Bartels, P., Brock, O., 2015. Rbo aleph: leveraging novel information sources for protein structure prediction. Nucleic Acids Res. 43 (W1), W343–W348.
McGuffin, L.J., Bryson, K., Jones, D.T., 2000. The PSIPRED protein structure prediction server. Bioinformatics 16 (4), 404–405. https://doi.org/10.1093/bioinformatics/16.4.404.
Meng, F., Kurgan, L., 2016. Computational prediction of protein secondary structure from sequence. Curr. Protein Pept. Sci. 86, 2.3.1–2.3.10. https://doi.org/10.1002/cpps.19.
Mixon, M.B., Lee, E., Coleman, D.E., Berghuis, A.M., Gilman, A.G., Sprang, S.R., 1995. Tertiary and quaternary structural changes in Gi alpha 1 induced by GTP hydrolysis. Science 270, 954–960. https://doi.org/10.1126/science.270.5238.954.
Olson, B., Shehu, A., 2014. Multi-objective optimization techniques for conformational sampling in template-free protein structure prediction. Int. Conf. Bioinf. Computat. Biol. (BICoB) 143–148.
Ovchinnikov, S., Kim, D.E., Wang, R.Y.-R., Liu, Y., DiMaio, F., Baker, D., 2016. Improved de novo structure prediction in CASP11 by incorporating coevolution information into Rosetta. Proteins: Struct. Funct. Bioinf. 84, 67–75. https://doi.org/10.1002/prot.24974.
Pollastri, G., McLysaght, A., 2005. Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics 21, 1719–1720.
Rademaker, Daniel, Jarek van Dijk, Titulaer, Willem, Lange, Joanna, Vriend, Gert, Xue, Li, 2020. The future of protein secondary structure prediction was invented by Oleg Ptitsyn. Biomolecules 10, 910. https://doi.org/10.3390/biom10060910.
Rohl, C.A., Strauss, C.E.M., Misura, K.M.S., Baker, D., 2004. Protein structure prediction using rosetta. Methods Enzymol. 383, 66–93.
Rost, B., Sander, C., 1993. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proc. Natl. Acad. Sci. USA 90 (16), 7558–7562.
Rost, B., Sander, C., Schneider, R., 1994. Redefining the goals of protein secondary structure prediction. J. Mol. Biol. 235 (1), 13–26.
Sali, A., Blundell, T.L., 1993. Comparative protein modeling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779–815. https://doi.org/10.1006/jmbi.1993.1626.
Schmitz, C., Vernon, R., Otting, G., Baker, D., Huber, T., 2012. Protein structure determination from pseudocontact shifts using rosetta. J. Mol. Biol. 416 (5), 668–677.
Senior, A.W., Evans, R., Jumper, J., et al., 2020. Improved protein structure prediction using potentials from deep learning. Nature 577, 706–710. https://doi.org/10.1038/s41586-019-1923-7.
Siew, N., Elofsson, A., Rychlewski, L., Fischer, D., 2000. MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics 16, 776–785.
Simoncini, D., Berenger, F., Shrestha, R., Zhang, K.Y.J., 2012. A probabilistic fragment-based protein structure prediction algorithm. PLoS One 7 (7), e38799.
Simoncini, D., Schiex, T., Zhang, K.Y.J., 2017. Balancing exploration and exploitation in population-based sampling improves fragment-based de novo protein structure prediction. Proteins 85 (5), 852–858.
Simons, K.T., Kooperberg, C., Huang, E., Baker, D., 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. J. Mol. Biol. 268 (1), 209–225.
Smolarczyk, Tomasz, Roterman-Konieczna, Irena, Stapor, Katarzyna, 2020. Protein secondary structure prediction: a review of progress and directions. Curr. Bioinf. 15, 90–107.
Spencer, M., Eickholt, J., Cheng, J., 2015. A deep learning network approach to ab initio protein secondary structure prediction. IEEE ACM Trans. Comput. Biol. Bioinf 12 (1), 103_11.
Sun, H.P., Huang, Y., Wang, X.F., Zhang, Y., Shen, H.B., 2015. Improving accuracy of protein contact prediction using balanced network deconvolution. Proteins 83 (3), 485–496.
Torrisi, M., Kaleel, M., Pollastri, G., 2018. Porter 5: fast, state-of-the-art ab initio prediction of protein secondary structure in 3 and 8 classes. bioRxiv 289033. https://doi.org/10.1101/289033.
Torrisi, Mirko, Kaleel, Manaz, Pollastri, Gianluca, 2019. Deeper profiles and cascaded recurrent and convolutional neural networks for state-of-the art protein secondary structure prediction. Sci. Rep. (9), 12374. https://doi.org/10.1038/s41598-019-48786-x, 2019.
Torrisi, Mirko, Pollastri, Gianluca, Quan, Le, 2020. Deep learning methods in protein structure prediction. Comput. Struct. Biotechnol. J. 18 (2020), 1301–1310.
Vullo, A., Walsh, I., Pollastri, G., 2006. A two-stage approach for improved prediction of residue contact maps. BMC Bioinf. 7 (180), 1–12.
Wang, G., Dunbrack Jr., R.L., 2003. PISCES: a protein sequence culling server. Bioinformatics 19 (12), 1589–1591. https://doi.org/10.1093/bioinformatics/btg224 PMID: 12912846.
Wang, G., Dunbrack Jr., R.L., 2005. PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Res. 33, W94–W98. https://doi.org/10.1093/nar/gki402. Web Server issue).
Wang, Z., Zhao, F., Peng, J., Xu, J., 2010. Protein 8-class secondary structure prediction using Conditional Neural Fields. IEEE Int. Conf. Bioinf. Biomed. (BIBM) 109–114.
Wang, S., Peng, J., Ma, J.Z., 2016. Protein secondary structure prediction using deep convolutional neural fields. Sci. Rep. 6, 18962. https://doi.org/10.1038/srep18962.
Wang, S., Sun, S.Q., Li, Z., Zhang, R.Y., Xu, J.B., 2017. Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS Comput. Biol. 13 (1), e1005324.
Ward, J.J., McGuf n, L.J., Buxton, B.F., Jones, D.T., 2003. Secondary structure prediction with support vector machines. Bioinformatics 19 (13), 1650–1655.
Wu, S., Szilagyi, A., Zhang, Y., 2011. Improving protein structure prediction using multiple sequence-based contact predictions. Structure 19 (8), 1182–1191.
Xu, D., Zhang, Y., 2012. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 80 (7), 1715–1735.
Xue, B., Faraggi, E., Zhou, Y., 2009. Predicting residue–residue contact maps bya two-layer: integrated neural-network method. Proteins 76 (1), 176–183.
Yang, Yuedong, Gao, Jianzhao, Wang, Jihua, Heffernan, Rhys, Hanson, Jack, , Kuldip Paliwal, Zhou, Yaoqi, 2016. Sixty- five years of the long march in protein secondary structure prediction: the final stretch? Briefings Bioinf. 1–13. https://doi.org/10.1093/bib/bbw129.
Yao, X.Q., Zhu, H., She, Z.S., 2008. A dynamic Bayesian network approach to protein secondary structure prediction. BMC Bioinf. 9, 49.
Yi, T.M., Lander, E.S., 1993. Protein secondary structure prediction using nearest neighbor methods. J. Mol. Biol. 232, 1117–1129.
Zemla, A., 2003. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res. 31 (13), 3370–3374. https://doi.org/10.1093/nar/gkg571.
Zhang, Y., 2008. I-tasser server for protein 3d structure prediction. BMC Bioinf. 9, 40.
Zhang, Y., Skolnick, J., 2004. Scoring function for automated assessment of protein structure template quality. Proteins: Struct. Funct. Bioinf. 57 (4), 702–710. https://doi.org/10.1002/prot.20264.
Zhang, Y., Kolinski, A., Skolnick, J., 2003. Touchstone II: a new approach to ab initio protein structure prediction. Biophys. J. 85, 1145–1164.
Zhang, G., Huang, D., Quan, Z., 2005. Combining a binary input encoding scheme withRBFNN for globulin protein inter- residue contact map prediction. Pattern Recogn. Lett. 16 (10), 1543–1553.
Zhang, X., Wang, T., Luo, H., Yang, J., Deng, Y., Tang, J., Yang, M., 2010. 3D Protein struc-ture prediction with genetic tabu search algorithm. BMC Syst. Biol. 4, S6.
Zhang, G.J., Ma, L.F., Q, W.X., G, Z.X., 2018. Secondary structure and contact guided differential evolution for protein structure prediction. IEEE/ACM Trans. Comput. Biol. Bioinf. Tetrahedron Comput. Methodol. 3 (3/4), 191–211.
Zhang, B., Li, J., Lu, Q., 2018. Prediction of 8-state protein secondary structures by a novel deep learning architecture. BMC Bioinf. 19 (293), 1–13.
Zhang, G., Wang, X., Ma, L., Wang, L., Hu, J., Zhou, X., 2019. Two-stage distance feature-based optimization algorithm for de novo protein structure prediction. In: IEEE/ACM Transactions on Computational Biology and Bioinformatics. https://doi.org/10.1109/TCBB.2019.2917452.