A PROGRESS ON PROTEIN STRUCTURE PREDICTION USING VARIOUS SOFT COMPUTING TECHNIQUES

Niharika Chaudhary and Sanjay Saini

Department of Physics and Computer Science, Faculty of Science,
Dayalbagh Educational Institute (Deemed to be University),
Dayalbagh, Agra, India

ABSTRACT

In molecular and computational biology, predicting the three-dimensional structure of a protein from its amino acid sequence has long been an outstanding goal. Soft computing techniques for solving protein structure prediction problems have been gaining the attention of researchers because of their capacity to accommodate imprecision and uncertainty in vast and complicated search spaces. This paper provides a comprehensive overview of recent protein structure prediction efforts and progress using various soft computing techniques. This paper summarises key research in the field of protein structure prediction that has been published in the recent decade. Despite significant research efforts in recent decades, there is still a lot of room for improvement in this field.

KEYWORDS

Nature Inspired Computing, Swarm Intelligence, Deep Learning, Protein folding, Protein structure prediction.

1. INTRODUCTION

Proteins play a vital role in all living organisms. Proteins are complex macromolecules comprised of one or more long chains of amino acid residues connected by peptide bonds. Proteins perform various functions within organisms, including catalysing metabolic reactions, DNA replication, responding to stimuli, providing structure to cells, and transporting molecules from one place to another [1]. The function of a protein is directly linked with its native structure. The protein sequence is composed of 20 different amino acids that are aligned in linear chains via peptide bonds [2]. The classification techniques of popular protein sequences involve the extraction of particular characteristics of the sequences; these characteristics depend on the structural and functional properties of the amino acids [3]. During synthesis, an individual protein is folded into a three-dimensional structure under the influence of various chemical and physical factors, providing essential biological functions and properties that play an essential role in biological science, medicine, drug design, and disease prediction [4]. Sometimes, proteins fold into an incorrect structure (known as misfolding); a single missing or incorrect amino acid could cause such a misfold, which leads to a protein with different properties which can be harmful to the organisms [5]. Therefore, knowledge of protein structure is very crucial in protein research.

Biologists describe a hierarchy of protein structures with four levels to better characterize the structural properties. The levels are organised in steps, with each lower level influencing the construction of the next. A linear sequence of amino acid residues is linked together in long

David C. Wyld et al. (Eds): NLPA, AIS, BDAP, SOENG, IPPR, CSIT - 2022
pp. 113-130, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121410

chains by peptide bonds to form a primary structure of a protein. The secondary structure refers to the periodic structure fragment of the polypeptide chain. The impact of hydrogen bonding between amide hydrogen and carbonyl oxygen in the peptide chain produces it. The fundamental elements of the secondary structure of the proteins are the α-helix and β-sheets [6]. The tertiary structure is a whole polypeptide chain generated by further combination and folding of multiple secondary structures in 3-D space. It could already represent the primary biological function of those proteins that only have one polypeptide chain [7]. The quaternary structure is a protein complex consisting of several polypeptide chains with multiple tertiary structures, and it could completely represent the characteristics of its biological function [8]. As a result, protein structure prediction (PSP) is the technique of determining a protein's three-dimensional structure from its amino acid sequence.

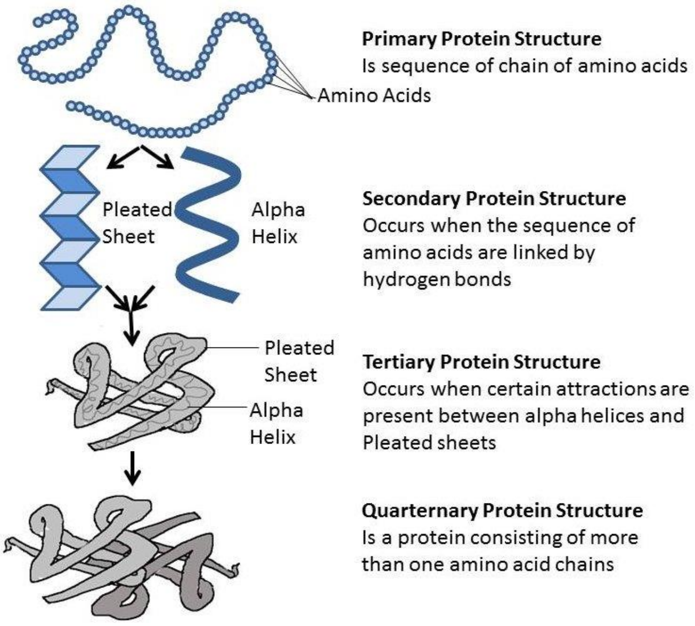
Primary Protein Structure
Is sequence of chain of amino acids
Amino Acids
Pleated Sheet
Alpha Helix
Secondary Protein Structure
Occurs when the sequence of amino acids are linked by hydrogen bonds
Pleated Sheet
Alpha Helix
Tertiary Protein Structure
Occurs when certain attractions are present between alpha helices and Pleated sheets
Quarternary Protein Structure
Is a protein consisting of more than one amino acid chains

Fig.1. Levels of Protein Structure

Protein structure prediction (PSP) is a global optimization problem and today’s most important and stimulating problem of structural bioinformatics, due to the fact that it determines the biological functions and the properties of a protein. PSP is a multimodal optimization problem classified as NP-hard [9]. Anfinsen's thermodynamic hypothesis [10] states that the native structure of a protein corresponds to the global minimum of the free energy surface of the protein, therefore, the protein structure prediction (PSP) problem can be translated into a global optimization problem. Hence, predicting the native structure of a protein is an important and challenging task in computational biology [11].

In general, experimental techniques like X-ray crystallography (XRC), Nuclear Magnetic Resonance (NMR), and cryo-electron microscopy (cryoEM) have been used to determine the native structure of protein, that are not always feasible and are very expensive and time-consuming[12]. More recently, Cryo-EM has fostered an acceleration of the protein structure determination process. This method's secret is in using photographs of frozen molecules to ascertain their structure. Despite this, the method typically produces structures with lesser resolution than those measured by other experimental techniques [13]. Though there has been a continuous improvement in the experimental techniques, the gap between the number of protein sequences and the known structures is increasing rapidly. As of June 2021, the

UniProtKB/TrEMBL [14] database contains 21,91,74,961 protein sequences, whereas in the Protein Data Bank (PDB) there are 1,80,419 protein structures stored.

To bridge this sequence-structure gap, computational approaches are preferred to overcome the difficulties associated with the experimental approaches [15]. The three basic computational approaches for PSP are homology or comparative modelling, threading or fold recognition, and ab-initio or denovo technique. Homology modelling is based on comparing the sequence's similarity, but the threading approach is a method for connecting probably brief sub-conformations of the relevant sub-sequence [16]. On the other hand, ab-initio methods try to find a 3D model of the protein exclusively using the amino acid sequence, according to the laws of physics and chemistry [17]. Critical Assessment of protein Structure Prediction (CASP), a well-known biennial competition of protein structure prediction adopted this classification. Results from this competition serve as a benchmark of computational biology's progress [18].

2. OVERVIEW OF TECHNIQUES USED IN PSP

Although PSP is NP-hard problem, no precise and effective method can provide the best solution in a polynomial amount of time. For decades, a lot of researchers have offered several approaches to solve the PSP. But several methods have limitations when the complexity of a problem increases. In contrast to other available methods, Nature Inspired Computing (NIC) based techniques can be employed to tackle PSP and provide solutions that is close to optimal in terms of complexity [19]. Emerging technology called Nature Inspired Techniques (NIC) intends to create new computational methods by drawing inspiration from how nature behaves in diverse scenarios while tackling challenging issues. A subset of Nature Inspired Computing (NIC) methods is Swarm Intelligence (SI). Swarm intelligence (SI) algorithms are primarily stochastic search and optimization methods that are inspired by the collective behaviour and self-organization of insect swarms. They are effective, flexible, and reliable research methodologies with several latent parallelisms that yield nearly ideal results [20].

SI is a promising stream to develop powerful solutions for optimization problems. Also, Metaheuristics have shown promising outcomes in the solution of difficult optimization problems [21]. For high-dimensional issues, they become constrained in terms of effectiveness and runtime. GPU-based parallel metaheuristics have been implemented in a variety of ways, and they appear to be a viable option for reducing execution time and improving solution quality. GPUs are now an effective tool for high-performance computing. They utilise the single-instruction, multiple-data (SIMD) architecture, which enables the concurrent parallel execution of hundreds of threads [22]. NIC techniques have successfully been experimented with and applied to machine learning and advanced artificial intelligence. Most of the current research is based on methods encouraged by these concepts [23].

Artificial Intelligence (AI) has been a thriving field, with a lot of applications. In recent times AI shows great promise in protein structure prediction. AI can accurately and efficiently predict thousands of possible protein structures [24]. In order to compare the prediction models' output to the known crystal structures and evaluate the quality of the models, artificial intelligence programmes are trained using a variety of numerically represented atomic features from the models (such as bond lengths, bond angles, residue-residue interactions, physio-chemical properties, and potential energy properties) [25].

Machine learning is a technology that allows computers to learn from their experiences and do tasks that are comparable to those performed by humans. ML algorithms formulate rules that link inputs to predicted results. Later, these rules are then applied to unknown data to provide physiochemically plausible accurate solutions. Machine learning techniques in AI have been successfully applied to automate routine labour, understand speech or images perform medical diagnoses, and predict protein structures [26]. Recently, researchers have also proposed machine learning algorithms that are different from the previously dominating deep learning approach, such as the deep forest [27]. Even though machine learning has been seen to be very successful in a lot of fields, there is still a significant gap between methodology and practical applications.

Deep Learning is a sub-field of Machine Learning based on artificial neural networks, usually used to learn the patterns from the input data, and in turn, make accurate predictions from it [28]. DL emphasizes the use of multiple layers to better represent data in different abstractions and has drastically improved the state-of-the-art methods in PSP. The use of DL approaches in structural bioinformatics improves the prediction model performance to a new level [29]. Deep learning has already been applied in a variety of protein processing tasks, including protein structure prediction [30], protein interaction prediction [31], protein secondary structure prediction [32], and protein sub cellular localization prediction [33]. Since the third generation of predictors, a variety of Deep Learning algorithms, as well as more traditional Machine Learning methods like k-Nearest Neighbours, Linear Regression, Hidden Markov Models, Support Vector Machines (SVM), and Support Vector Regression, have been routinely used for PSA prediction.

Every year, a significant gathering known as the Critical Assessment of Protein Structure Prediction (CASP) compares models for predictions and prediction analysis. Since 1994, the CASP has been a community-wide, worldwide project for protein structure prediction that takes place biannually [34]. CASP aims to benchmark the protein-structure prediction methods and stimulate the advancement of the field. Fully blinded testing of structure prediction algorithms is the core principle of CASP [35].

3. SOLVING PSP USING VARIOUS TECHNIQUES (LITERATURE SURVEY)

In this section, we present a literature survey on relevant methods of protein structure prediction based on soft computing techniques. Soft computing is built on both organic and artificial concepts. It is referred to as computational intelligence. The major concepts related to the various successful soft computing techniques are discussed in the following section.

3.1. Computational Approaches to PSP

These can broadly be categorized into Template-Based Modelling (TBM) and Template-Free Modelling (FM). TBM makes use of a template to predict 3D models. The proteins with the same sequences tend to bend in the same structures form the basis of this method. It can be seen that even proteins with a 30% sequence identity bend in the same structures [36].

Free-modelling, ab initio modelling, and de novo modelling are used interchangeably to discuss template-free modelling approaches. Ab initio protein structure prediction or Free modelling (FM) attempts to build a 3D structure without using homologous proteins as templates [37]. FM methods rely on creating algorithms that can quickly discover global energy minima and a scoring system that can choose the best conformation from among the several created models. FM aims to predict the most stable protein spatial arrangement with the lowest free energy [38]. Template-based approaches have a greater prediction accuracy than other methods, especially when the target protein and template are relatively homogeneous, making them useful in practical applications.

SWISS-MODEL [39], developed by Torsten Schwede's structural bioinformatics group, is a well-known online tool for homology modelling of protein structures. In the last two decades, it has become one of the most extensively used homology modelling tools. Template identification, target-template alignment, model construction, and model evaluation are all part of the prediction process. HHblits [40] is employed for template recognition and target-template alignment. Modeller [41] is another homologous modelling tool developed, that uses the target-template alignment to implement structure modelling by satisfying spatial restraints. I-TASSER [42] [43] is a comprehensive structure prediction method based on the threading method developed by the Yang Zhang Lab.

The Yang Zhang Lab has created another good fragment-assembly approach, QUARK [44]. QUARK employs replica-exchange Monte Carlo simulations guided by a knowledge-based force field to build structural fragments ranging from 1 to 20 residues. Bhageerath [45] is one such ab initio-homology hybrid method. It is accessible in the form of a web server called Bhageerath-H [46]. The main focus of the protocol is on the reduction of conformational search space. It is based on a template-based modelling and energy-based function that generates several structures according to a systematic sampling of loop dihedral conformational space. The performance of this software was tested by using CASP10 targets with promising prediction results. After the assessment of its shortcomings, an updated version was introduced in the CASP12 meeting as BhageerathH+ [47]. The David Baker Lab created Rosetta [48], a template-free approach that assembles a full-length structure from fragments of 3-9 residues from PDB Structures. To predict globally minimized protein models it follows a Monte Carlo simulation-based strategy. Its greatest success was recorded in CASP11, when it accurately predicted the structure of a sequence of 256 amino acids [49].

CASP (Critical Assessment of Protein Structure Prediction) determines the current state of the art of protein structure prediction. The first success in protein tertiary structure prediction was observed in CASP4, but primarily for small proteins (≤ 120 residues). In later years, better contact prediction approaches were introduced in CASP11 [50] competing pipelines with promising improvements in prediction accuracy. A similar pattern was observed in CASP12 with the incorporation of alignment-based contact prediction methods [51]. The use of deep learning techniques for structure prediction in CASP13 demonstrated considerable improvement above the average GDT_TS score. That gives an encouragement to delve further into deep learning-based approaches to solve the protein folding problem [52]. Experimental results on CASP11, CASP12 refinement targets, and blind tests in CASP 13 turn out to be promising. Residue-residue interactions, disorder region prediction, model quality assessment (MQA) approaches, tertiary structure refinement, and other structure-related modelling categories are all evaluated by CASP.

For protein sequences with 150 residues or fewer, free-modelling has so far proven to be a good fit. Few instances have been reported when algorithms went too far and attempted to predict the structure of longer proteins. CASP11 witnessed major success in ab initio protein structure prediction for a length of 256 amino acids [53]. Template-Based modelling approaches still have few limitations whereas ab initio approaches can move a step ahead and might help to understand the basic principles of protein folding [54].

3.2. Deep Learning-Based Approaches in PSP

In recent years, deep learning has recently proven extremely effective in tackling challenging problems in several problem domains of protein structure prediction. Their strength comes from their capacity to learn features of data at multiple levels of abstraction, beginning from relatively simple feature sets. Meanwhile, it has become a very powerful artificial intelligence tool in bioinformatics [55]. It is detailed in [55] how the prediction method DeepCov uses fully convolutional neural networks to operate on frequency or covariance data for amino-acid pairs that are directly obtained from sequence alignments. The newly developed DeepMetaPSICOV(DMP) [56] deep learning-based contact prediction tool combines the input feature sets utilised by MetaPSICOV and DeepCov as input to a deep fully convolutional residual neural network. Similarly, MapPred [57] is also a deep learning-based contact prediction method. MapPred consists of two-component methods, DeepMSA and DeepMeta, both trained with the residual neural networks.

Al Quraishi [58] builds a pure deep learning-based prediction approach. It is a one-step algorithm for the prediction of protein structure relying on an end-to-end deferential deep learning strategy. Thus, the RGN (Recurrent Geometric Networks) end-to-end differentiable protein structure predictor was created, allowing for the direct prediction of torsion angles to build the protein backbone. MULTICOM [59] is another protein structure prediction pipeline that has also incorporated recent advances in DL-based approaches to improve the protein structure prediction system. Torrisi et al. [60] recently reviewed DL-based approaches for the evolution of predictive methods for one-dimensional and two-dimensional Protein Structure Annotations, from simple statistical methods. Also looked at how these algorithms database growth has affected our ability to learn about evolution and co-evolution and how that has affected our ability to make better predictions. Moreover, Gao et al. [61] reviewed the recent advance in DL-based approaches for the protein sequence–structure–function paradigm. They analysed the emerging deep learning techniques for protein structural modelling and discussed advances and challenges that must be addressed.

Li et al. [62] proposed a new Rosetta based method trRosetta (transform-restrained Rosetta) for protein structure prediction given a protein sequence developed by Yang's and David Baker's group. In order to predict the inter-residue geometry and orientations, three dihedral and two planar angles between the residues are used to indicate the orientations between two residues. For the first time, RaptorX-Contact [63], a residual neural network (ResNet)-based model, used a deep neural network to predict protein contacts, considerably increasing accuracy on difficult targets with novel folds. Hanson et al. [64] proposed the method called SPOT-Contact (Sequence-based Prediction Online Tools for Contact map prediction) that captures the contextual information from the whole protein 'image' at each layer. SPOT-contact is highly accurate for predicting contacts at all sequence-position separations, and improves over RaptorX-Contact at all Neff values, with the largest improvement for low medium range Neff sequences. TripletRes [65] developed by Zhang's group, is a contact map prediction that ranked first in the Contact Predictions category of CASP 13 for PSP. Co-evolutionary feature extraction, deep neural network modelling, and deep multiple-sequence alignment generation are the components of TripletRes. TripletRes training needed four GPUs operating at the same time using Adam, with an 80% dropout rate. It also correctly predicted both globally and locally multi-domain proteins.

DeepMind's AlphaFold [66] is another tool for protein structure prediction method developed by DeepMind, has achieved the best performance in the template-free modelling domain of CASP13 and has evolved to AlphaFold2 at CASP14 [67]. It uses a two-step process for protein structure determination and also involves the use of co evolutionary profiles to guide model building. This methodology constructed high-accuracy models for 24 out of 43 test proteins achieving a TM-score of 0.7 [68].The protein structure prediction field has seen a lot of progress due to Deep Learning (DL)-based techniques as verified by the success of AlphaFold2 in the most recent CASP14. AlphaFold works on the idea that given a protein sequence, it is feasible to build a learnt, protein-specific potential by training a Deep Neural Network to produce accurate structure predictions and then reducing the potential via gradient descent to predict the structure itself [69].

DL-based Cryo-electron microscopy (Cryo-EM) is a useful method for studying protein structures and determining their dynamics [70]. Cryo-EM is a Nobel Prize-winning technique for creating three-dimensional maps of protein structures. The basic computational step in this method is the analysis of EM data to obtain protein structural information. Esquivel-Rodrguez et al. [71] reviewed recent improvements in computational approaches for modelling protein three-dimensional structures using a 3D EM density map built from 2D maps. For Cryo-EM based protein structure determination, there have been some recent DL-based breakthroughs in several processes such as single-particle picking, backbone prediction, secondary structure prediction, EM density map refining, and all-atom prediction for protein complexes [72]. The authors reviewed some of the recent approaches for single-particle picking, backbone structure prediction, and secondary structure prediction for the construction of high-resolution 3D Cryo-EM maps [73].

The molecular dynamics (MD) simulation is one of the most popular optimization techniques for fine-tuning protein structures. In order to speed up the extraction of the final structure from comparable fragments in the PDB and to improve the convergence of MD-based structural refinement simulations, Raval et al. [74] introduced contact-based restraints into MD simulation. David Shaw's group utilized different sets of restraints to reduce the molecular dynamics simulation runs and prevent the model from getting trapped in a non-native energy state. MELD (Modelling Employing Limited Data) is a newly developed physics-based protein structure prediction approach that uses Bayesian law to use atomic molecular dynamics of proteins for structural modelling. It has proven to be productive in determining high-resolution structures of proteins up to 260 residues long [75]. Replica exchange molecular dynamics (REMD) and the residue-specific force field (RSFF1) in explicit solvent have been demonstrated by another group to shorten simulation times and control energy landscapes [76].

3.3. Support Vector Machines

SVM is a machine learning method used to predict protein structure from its primary sequence. SVM's fundamental approach entails transforming the samples into a high-dimension Hilbert space and searching for a separation plane there [77]. The optimal separation plateau (OSH), also known as the separation hyperplane, is selected to maximize its distance from the nearby training samples. The nine SVM-based contact predictors that Wu et al. [78] created are combined with the sparse template contact restrictions in [79]. They make use of the I-TASSER's energy function as well as contact predictions produced by SVMSEQ's enhanced versions.

Wang et al. [80] proposed a protein secondary structure prediction method based on SVM with position-specific scoring matrix (PSSM) profiles. The PSSM profiles are obtained from commonly used protein CB513 datasets and conclude that accuracy increases by 11.3%. Xie et al. [81] developed a new method for the PSSP based on an improved fuzzy support vector machine (FSVM), in which an approximate optimal separating hyperplane is constructed by iterating the class centres in the feature space, and sample points close to this hyperplane are assigned with large membership values, while outliers with small membership values are assigned with low membership values based on the K-nearest neighbour algorithm. SVMQA [82] is one of the best SVM based single-model quality assessment methods. Using 19 features, including 8 potential energy based and 11 comparing projected and actual models, this method generates a TM Score and GDT_TS score.

Uziela et al. [83] introduced two unique approaches, ProQRosFA and ProQRosCen, modelled after the cutting-edge of ProQ2 technique. ProQ2 [84] is a model quality evaluation algorithm that uses SVM to forecast both the local and overall quality of protein models. While the new predictors are based on Rosetta energies, ProQ2 employs contacts and other features that were computed from a model. ProQ3 inherits all of ProQ2's capabilities and add two new ones based on Rosetta energy: full-atom Rosetta energy terms and coarse-grained centroid Rosetta energy terms. On both the CAMEO and CASP11 datasets, ProQ3 exceeds ProQ2 in correlation and achieves the highest average GDT_TS score [83]. Another secondary structure prediction method, conSSert [85] based on support vector machines (SVM) gives remarkable accuracy for beta-strand prediction, with a QE accuracy of over 0.82 and a Q2-EH of 0.86. Finally, Zhou et al. [86] applied a support vector regression technique that detects the native structures of a protein among decoy sets. This method makes use of the fast Fourier transform, amino acid network score, and contact energy-based score.

3.4. Neural Network Techniques

Since Qian and Sejnowski first introduced one of the initial NN approaches for PSSP in 1988, NN has steadily grown to be the most popular model in this area and has accomplished great feats. Various NN architectures have been developed throughout time to produce the best results in certain application areas. Deep learning, recurrent neural networks, BP neural networks, radial basis function neural networks, and complex-valued neural networks have been the most often utilised NN models in the recent decade [87].

Heffernan et al. [88] developed an integrated sequence-based prediction that predicts four different sets of structural properties of proteins by iterative learning in a parallel scheme based on a deep neural network. Local backbone structures that are represented by secondary structure, backbone torsion and Cα angles, and dihedral angles are the structural characteristics. SPIDER2 expands on this technique by predicting solvent accessibility and training an independent set of weights for each iteration. Spencer et al. [89] proposed the first deep learning-based PSSP method, called DNSS, where restricted Boltzmann machine (RBM) based deep belief network (DBN) model was trained by contrastive divergence46 in an unsupervised manner. The method used a position-specific scoring matrix generated by PSI-BLAST to train a deep learning network. GPU and CUDA software optimizes the deep network architecture and efficiently trains the deep networks.

Wang et al. [90] presented Deep Convolutional Neural Fields (DeepCNF) for protein SS prediction which is a conditional neural field deep learning extension for PSSP including Q3 and Q8. The DeepCNF integrated the advantages of both CNF and DCNN and include information about longer-range dependencies as well as the complicated link between neighbouring secondary structure labels and the sequence structure. Also, JPred4 [91] and Porter 5 [92] are the most accurate predictors used in the prediction of 3-class protein secondary structures.

Yaseen et al. [93] proposed a novel method for improving secondary structure prediction accuracy by employing statistical context-based scores as encoded features in neural network training. This approach is being used by a server known as SCORPION (secondary structure prediction). The three common secondary structure states are predicted by SCORPION by grouping (G, H, I) into helices (E, B) into sheets, and (T, S, C) into coils, in accordance with the majority of secondary structure prediction algorithms.

Multiple architectures are interlaced in a few modern PSSP approaches to enhance overall network prediction. Li and Yu [94] employed a deep convolutional recurrent NN architecture (DCRNN) to extract multi-scale local contextual information using CNN with varied kernel sizes. To predict secondary structure given amino acid sequence information, a typical feed-forward NN with one hidden layer is used. Their goal was to identify the ideal parameter configuration for producing the greatest prediction outcomes.

Zhang, B et al. [95] proposed a convolutional residual recurrent neural network (CRRNN) for both Q8 and Q3 secondary structure prediction. The model's parameters are evaluated, and dimensionality reduction is accomplished using a 1D convolutional filter with a single kernel. For PSSP, a recurrent neural network (RNN) is extensively used to solve sequence-based issues. Porter 4.0 [96] and SPIDER3 [97] are successful RNN methods for protein secondary structure prediction where physicochemical parameters are effectively included in the NN model to improve SS prediction and obtain up to 84 per cent accuracy.

The Deep inception-inside-inception (Deep3I) network is presented as a novel deep neural network design for protein secondary structure prediction and developed as a software application. MUFOLD-SS. It is made up of a series of layered inception modules that transfer the input matrix to either eight or three secondary structure states [98]. Zhang and Shen [99] proposed a method called ThreaderAI that applies a deep residual neural network for prediction. By integrating sequence profile, predicted sequential structural features, and predicted residue-residue contacts, ThreaderAI first uses deep learning to predict residue-residue aligning probability matrix. Then, by applying a dynamic programming algorithm to the probability matrix, it builds template-query alignment.

3.5. Evolutionary Computation (EC) Techniques

Based on the Darwinian concepts of evolution, Evolutionary Algorithms (EAs) and Genetic algorithms (GAs) represent two subtypes of Evolutionary Computation (EC). Different protein structural representations, such as dihedral or torsion angles and lattice models, may be used in EC-based methods (direction vector representation and hydrophobic–polar model). The 2D-HP triangular model is addressed in [100], where Evolutionary Programming is used as a search algorithm by applying a hybrid algorithm combining genetic algorithm, tabu search strategy, and local search procedure.

The hybridization of global and local search techniques is commonly known as Memetic Algorithms (MAs) or Hybrid Genetic Algorithms. The method utilizes a structured population using a local search strategy based on the Simulated Annealing (SA) method, ad-hoc crossover, and mutation operators to deal with the problem. By utilising an Angle Probability List (APL), which helps in reducing the search space and providing guidance for the search strategy, it retrieves the structural knowledge recorded in the PDB [101].

Dorn et al. [102] proposed a knowledge-based Genetic Algorithm (GA), aiming to reduce the size of the conformational search space considering the previous occurrences of amino acid residues in experimentally determined proteins. Similar to the APL concept, the technique used appropriate torsion angle intervals for the amino acid targets. The NIAS server was made available by Borguesan et al. [103] to compute ad-hoc APLs to benefit from prediction techniques or in any other application that can use the conformational preferences of amino acids. A hybrid algorithm that combined GA and tabu search (TS) algorithms to solve PSP problems were used, where authors utilized TS instead of the mutation operator in GA for preventing premature convergence and providing a faster convergence rate [104].

Garza-Fabre et al. [105] applied a memetic algorithm (MA), based on fragment assembly technique identified as a search heuristic of the Rosetta ab initio protocol. The phenotypic crowding technique is used as a similarity criterion for the selection of individuals in the multi-objective GA developed in [106]. Based on this approach, two solutions with the most similar ones are selected according to their structural differences, which inferred the template-based delay of the population convergence. Gao et al. [107] attempted a multi-objective evolutionary algorithm for protein structure prediction. They divide the force fields of the Chemistry at

Harvard Macromolecular Mechanics (CHARMM) protein-energy function into bond and non-bond energies as their first and second objectives. Considering the effect of solvent innovatively acquires a solvent-accessible surface area as the third objective, and sixty-six benchmark proteins are used to verify the proposed method.

3.6. Swarm Intelligence (SI) Techniques

The collective behaviour of decentralised, self-organized systems, which can be either natural or artificial, is known as swarm intelligence (SI). Generally, SI systems contain a population of simple agents interacting locally with one another and with the environment. The inspiration usually comes from nature, especially biological systems. Recently, SI techniques have been applied to solve various PSP problems.

In order to establish the protein structure, Cheng-yuan et al. [108] created a quantum-behaved PSO (QPSO), in which the population is split into an elite sub-population, an exploitation sub-population, and an exploration sub-population. Fine-tuning and exploration strategies are used to facilitate faster convergence, bond angles are used to represent the protein conformation, and the AB off-lattice model reveals the protein-energy function. Zhou et al. [109] proposed an improved PSO for protein folding prediction. Levy probability distribution was utilised to update locations, and the velocities of the chosen worst particles were used to accelerate convergence and break out of local optima. For protein structure optimization, an enhanced Artificial Bee Colony (ABC) [110] algorithm was put forth and tested on comparatively less synthetic and natural protein sequences. The performance of the fundamental ABC algorithm is improved by an internal feedback method, which also produces results that are marginally superior to those of other algorithms.

Another ABC variant for 3D PSP is presented in [111], where structure quality was improved by using the convergence information of the ABC algorithm during the execution process. The new algorithm performed better than previous cutting-edge methods on some cases of artificial and actual proteins, according to the authors' demonstrations. By applying the same technique as in [103] to thirteen genuine protein sequences, Li et al. [112] expanded their previous work and compared the outcomes with those found in the literature. Additionally, the chaotic ABC method has a high computing cost. Swarm intelligence systems were compared in a study of protein folding issues based on the 3D AB off-lattice model. Only a small number of artificial protein sequences with short lengths were used by the authors to analyse the performance of the standard versions of the algorithms [113]. When comparing actual protein sequences to artificial protein sequences, the comparison might be more accurate.

The Particle Swarm Optimization search algorithm is used to discover the ideal pair of dihedral angles of a structure utilising a profile-level knowledge-based force field. Detail the population-based harmony search as the optimizer as well as the 2D-AB off-lattice model [114]. In an effort to improve the artificial intelligence-based protein structure refinement method known as AIR, the authors of Wang et al. [115] developed a unique multi-objective particle swarm optimization (PSO) structure refinement protocol. The basic objective is to use multiple energy functions as multi-objective for correcting the potential bias problem caused by minimizing only a single energy function. When large conformation spaces are required, it can achieve optimal search with swarm intelligence and the information-sharing method amongst the particles in PSO.

In order to propose a neural network model for protein secondary structure prediction (PSSP), PSO has been investigated [116]. Six common datasets, including PSS504, RS126, EVA6, CB396 and Manesh, have been used for the neural network's training and testing. The suggested model is validated using cross-validation with 10, 20, 30, and 40 folds, as well as sensitivity analysis. Self-organizing map (SOM) based PSO approach with the efficient classification of secondary and tertiary proteins explored in [117]. They grouped 2D and 3D proteins, where 2D proteins contain fewer hydro-carbons than 3D proteins. The angles of the proteins are considered by evaluating the SOMs with the Bounding Box technique for a quicker analysis. An effective protein structure prediction method that combines template-based and template-free techniques was used by Gao et al. [118]. The initial protein conformations, in particular, can be constructed using a non-redundant protein database and a random selection approach with secondary structure limitations. To update the protein structures while keeping the secondary structures the same, three alternative structure evolution approaches are used: enhanced particle swarm optimization (PSO) algorithm, random perturbation, and fragment substitution.

4. SUMMARY

Protein structure prediction (PSP) is an important field of research in computational biology and protein science. Due to growing demand in the last decade, a large number of PSP methods have been proposed, but still, there is considerable scope for further improvement. This paper tried to cover the recent works published from 2012. In this review, we have outlined the most cutting-edge soft computing methods used to solve the protein structure prediction problem.

This paper first provides an introduction and the related knowledge of PSP; then, the overview of most relevant methods used in PSP are reported; finally, the literature review on various soft computing techniques like TBM (template-based modelling), TFM (template-free modelling), deep learning, support vector machines, neural networks, evolutionary computations, and swarm intelligence for solving PSP problem is provided. Various soft computing techniques and their tools that have been applied to solve the PSP problem are summarized in Table1.

Table.1. Summary of protein structure prediction methods

Method/Tools	Dataset Size	Architecture	Year	Reference
ProQ2	CASP7-9	Support Vector Machine (SVM)	2012	[84]
Porter 4.0	TS115, PDB	2 stages with Bidrectional Recurrent Neural Network (BRNN)	2013	[96]
Bhageerath-H	CASP10	Template-Free based modelling	2014	[46]
SCORPION	Cull7987	3 stages with Feed-forward Neural Network (FNN)	2014	[93]
PSI-BLAST	CASP9-10	Position-specific scoring matrix (PSSM)	2014	[89]
SPIDER 2	TR4590,CASP11	Deep artificial neural network (DeepANN)	2015	[88]
Rosetta	CASP11	Template-Free modelling	2016	[48]
ProQ3	CAMEO, CASP8-10	Support Vector Machine (SVM)	2016	[83]
conSSert	PDBSelect25, PISCES	Support Vector Machine (SVM)	2016	[85]
DeepCNF	CASP,CB513	Deep Neural Network with 5 hidden layers	2016	[90]
SVMQA	CASP8-12	Support Vector Machine (SVM)	2017	[82]
SPIDER 3	TR4590, TS115	Long Short Term Memory (LSTM) Bidirectional Recurrent Neural Networks (BRNNs)	2017	[97]
RaptorX-Contact	PDB25	Residual network (ResNet)	2017	[63]
Porter 5	PDB	Two-stage ensemble of BRNN and CNN	2018	[92]
MUFOLD-SS	CullPDB, CASP10-12	Deep Neural Network (inception-inside-inception)	2018	[98]
CRRNN	CB513, CASP10-12	Deep Neural Network with 5 hidden layers	2018	[95]
DeepCov	PDB, CASP12	2D Convolutional Neural Network (CNN)	2018	[55]
SPOT	PDB, UniProt	Residual-bidirectional-LSTM	2018	[64]
QUARK	CASP13	Deep Residual CNN	2019	[44]
C-I-TASSER	CASP13	2D Convolutional Neural Network (CNN)	2019	[42]
RGN	ProteinNet12	bi-Long Short Term Memory (LSTM)	2019	[58]
DeepMetaPSICOV	PDB, CASP13	Residual network (ResNet)	2019	[56]
AlphaFold	PDB, CASP13	Deep Neural Network	2020	[66]
trRosetta	CAMEO, CASP13	Residual network (ResNet)	2020	[62]
MapPred	PISCES	Residual network (ResNet)	2020	[57]
ThreaderAI	SCOPe40, CASP13	Deep Residual Neural Network (RNN)	2020	[99]
AIR	CASP11-12	Artificial intelligence-based multi-objective optimization protocol (MOOP)	2020	[115]
TripletRes	CASP11-13	Deep Neural Network	2021	[65]
MULTICOM	CASP8-11	Deep Convolutional Neural Network (CNN)	2021	[59]

REFERENCES

[1] Lesk, A. (2019). Introduction to bioinformatics. Oxford university press.

[2] Wang, Y., Mao, H., & Yi, Z. (2017). Protein secondary structure prediction by using deep learning method. Knowledge-Based Systems, 118, 115-123.

[3] Hendy, H., Khalifa, W., Roushdy, M., & Salem, A. B. (2015). A study of intelligent techniques for protein secondary structure prediction. International Journal of Information Models and Analyses, 4(1).

[4] Neudecker, P., Robustelli, P., Cavalli, A., Walsh, P., Lundström, P., Zarrine-Afsar, A & Kay, L. E. (2012). Structure of an intermediate state in protein folding and aggregation. Science, 336(6079), 362-366.

[5] Argyrou, A. (2020). The Misfolding of Proteins. In GeNeDis 2018 (pp. 249-254). Springer, Cham.

[6] Jiang, Q., Jin, X., Lee, S. J., & Yao, S. (2017). Protein secondary structure prediction: A survey of the state of the art. Journal of Molecular Graphics and Modelling, 76, 379-402.

[7] Voet, D., Voet, J. G., & Pratt, C. W. (2016). Fundamentals of biochemistry: life at the molecular level. John Wiley & Sons.

[8] Jana, N. D., Das, S., & Sil, J. (2018). A Metaheuristic Approach to Protein Structure Prediction: Algorithms and Insights from Fitness Landscape Analysis (Vol. 31). Springer.

[9] Kuhlman, B., & Bradley, P. (2019). Advances in protein structure prediction and design. Nature Reviews Molecular Cell Biology, 20(11), 681-697.

[10] Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. Science, 181(4096), 223-230.

[11] Zhou, C., Hou, C., Wei, X., & Zhang, Q. (2014). Improved hybrid optimization algorithm for 3D protein structure prediction. Journal of molecular modelling, 20(7), 2289.

[12] Boiani, M., &Parpinelli, R. S. (2020). A GPU-based hybrid jDE algorithm applied to the 3D-AB protein structure prediction. Swarm and Evolutionary Computation, 58, 100711.

[13] Lawson, C. L., Berman, H. M., & Chiu, W. (2020). Evolving data standards for cryo-EM structures. Structural Dynamics, 7(1), 014701.

[14] UniProt: the universal protein knowledge base in 2021." Nucleic Acids Research 49, no. D1 (2021): D480-D489.

[15] Hanson, J., Paliwal, K. K., Litfin, T., Yang, Y., & Zhou, Y. (2020). Getting to know your neighbor: protein structure prediction comes of age with contextual machine learning. Journal of Computational Biology, 27(5), 796-814.

[16] Feig, M. (2017). Computational protein structure refinement: almost there, yet still so far to go. Wiley Interdisciplinary Reviews: Computational Molecular Science, 7(3), e1307.

[17] Márquez-Chamorro, A. E., Asencio-Cortés, G., Santiesteban-Toca, C. E., & Aguilar-Ruiz, J. S. (2015). Soft computing methods for the prediction of protein tertiary structures: A survey. Applied Soft Computing, 35, 398-410.

[18] Kinch, L. N., Li, W., Schaeffer, R. D., Dunbrack, R. L., Monastyrskyy, B., Kryshtafovych, A., & Grishin, N. V. (2016). CASP 11 target classification. Proteins: Structure, Function, and Bioinformatics, 84, 20-33.

[19] Krishnaveni, A., Shankar, R., &Duraisamy, S. (2019). A Survey on Nature Inspired Computing (NIC): Algorithms and Challenges. Global journal of computer science and technology: D Neural & Artificial Intelligence, 19(3).

[20] Al Kawam, A., Sen, A., Datta, A., & Dickey, N. (2017). Understanding the bioinformatics challenges of integrating genomics into healthcare. IEEE journal of biomedical and health informatics, 22(5), 1672-1683.

[21] Rajakumar, R., Dhavachelvan, P., &Vengattaraman, T. (2016, October). A survey on nature inspired meta-heuristic algorithms with its domain specifications. In 2016 international conference on communication and electronics systems (ICCES) (pp. 1-6). IEEE.

[22] Essaid, M., Idoumghar, L., Lepagnot, J., &Brévilliers, M. (2019). GPU parallelization strategies for metaheuristics: a survey. International Journal of Parallel, Emergent and Distributed Systems, 34(5), 497-522.

[23] Shandilya, S. K., Shandilya, S., & Nagar, A. K. (Eds.). (2019). Advances in nature-inspired computing and applications (Vol. 1). Switzerland: Springer International Publishing.

[24] Hessler, G., &Baringhaus, K. H. (2018). Artificial intelligence in drug design. Molecules, 23(10), 2520.

[25] Xu, J., & Wang, S. (2019). Analysis of distance-based protein structure prediction by deep learning in CASP13. Proteins: Structure, Function, and Bioinformatics, 87(12), 1069-1081.

[26] Gawehn, E., Hiss, J. A., & Schneider, G. (2016). Deep learning in drug discovery. Molecular informatics, 35(1), 3-14.

[27] Dorn, M., e Silva, M. B., Buriol, L. S., & Lamb, L. C. (2014). Three-dimensional protein structure prediction: Methods and computational strategies. Computational biology and chemistry, 53, 251-276.

[28] Torrisi, M., Kaleel, M., &Pollastri, G. (2019). Deeper profiles and cascaded recurrent and convolutional neural networks for state-of-the-art protein secondary structure prediction. Scientific reports, 9(1), 1-12.

[29] Goodfellow, I.,`Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

[30] Mirabello, C., &Wallner, B. (2019). RAWMSA: End-to-end deep learning using raw multiple sequence alignments. PloS one, 14(8), e0220182.

[31] Xu, J. (2019). Distance-based protein folding powered by deep learning. Proceedings of the National Academy of Sciences, 116(34), 16856-16865.

[32] Smolarczyk, T., Roterman-Konieczna, I., &Stapor, K. (2020). Protein secondary structure prediction: a review of progress and directions. Current Bioinformatics, 15(2), 90-107.

[33] Barberis, E., Marengo, E., & Manfredi, M. (2021). Protein Subcellular Localization Prediction. In Proteomics Data Analysis (pp. 197-212). Humana, New York, NY.

[34] Moult, J., Pedersen, J. T., Judson, R., & Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods.

[35] Fleishman, S. J., Whitehead, T. A., Strauch, E. M., Corn, J. E., Qin, S., Zhou, H. X., ...& Baker, D. (2011). Community-wide assessment of protein-interface modelling suggests improvements to design methodology. Journal of molecular biology, 414(2), 289-302.

[36] Yang, J., Zhang, W., He, B., Walker, S. E., Zhang, H., Govindarajoo, B., ...& Zhang, Y. (2016). Template-based protein structure prediction in CASP11 and retrospect of I-TASSER in the last decade. Proteins: Structure, Function, and Bioinformatics, 84, 233-246.

[37] Dhingra, S., Sowdhamini, R., Cadet, F., &Offmann, B. (2020). A glance into the evolution of template-free protein structure prediction methodologies. Biochimie, 175, 85-92.

[38] Bhattacharya, D., Cao, R., & Cheng, J. (2016). UniCon3D: de novo protein structure prediction using united-residue conformational search via stepwise, probabilistic sampling. Bioinformatics, 32(18), 2791-2799.

[39] Bienert, S., Waterhouse, A., de Beer, T. A., Tauriello, G., Studer, G., Bordoli, L., &Schwede, T. (2017). The SWISS-MODEL Repository—new features and functionality. Nucleic acids research, 45(D1), D313-D319.

[40] Remmert, M., Biegert, A., Hauser, A., &Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature methods, 9(2), 173-175.

[41] Webb, B., &Sali, A. (2016). Comparative protein structure modelling using MODELLER. Current protocols in bioinformatics, 54(1), 5-6.

[42] Zheng, W., Zhang, C., Bell, E. W., & Zhang, Y. (2019). I-TASSER gateway: A protein structure and function prediction server powered by XSEDE. Future Generation Computer Systems, 99, 73-85.

[43] Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. Nature methods, 12(1), 7-8.

[44] Zheng, W., Li, Y., Zhang, C., Pearce, R., Mortuza, S. M., & Zhang, Y. (2019). Deep-learning contact-map guided protein structure prediction in CASP13. Proteins: Structure, Function, and Bioinformatics, 87(12), 1149-1164.

[45] Jayaram, B., Dhingra, P., Lakhani, B., & Shekhar, S. (2012). Bhageerath—Targeting the near impossible: Pushing the frontiers of atomic models for protein tertiary structure prediction. Journal of Chemical Sciences, 124(1), 83-91.

[46] Jayaram, B., Dhingra, P., Mishra, A., Kaushik, R., Mukherjee, G., Singh, A., & Shekhar, S. (2014). Bhageerath-H: a homology/ab initio hybrid server for predicting tertiary structures of monomeric soluble proteins. BMC bioinformatics, 15(16), 1-12.

[47] Rahul, K., Ankita, S., Debarati, D., Amita, P., Shashank, S., and B. Jayaram. (2016). Bhageerath H+ : A hybrid methodology based software suite for protein tertiary structure prediction. In CASP12 Proceedings, pages 25–26, 2016.

[48] Park, H., DiMaio, F., & Baker, D. (2016). CASP 11 refinement experiments with ROSETTA. Proteins: Structure, Function, and Bioinformatics, 84, 314-322.

[49] Kinch, L. N., Li, W., Monastyrskyy, B., Kryshtafovych, A., & Grishin, N. V. (2016). Evaluation of free modelling targets in CASP11 and ROLL. Proteins: Structure, Function, and Bioinformatics, 84, 51-66.

[50] Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., &Tramontano, A. (2016). Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. Proteins: Structure, Function, and Bioinformatics, 84, 4-14.

[51] Abriata, L. A., Tamò, G. E., Monastyrskyy, B., Kryshtafovych, A., & Dal Peraro, M. (2018). Assessment of hard target modelling in CASP12 reveals an emerging role of alignment-based contact prediction methods. Proteins: Structure, Function, and Bioinformatics, 86, 97-112.

[52] Hou, J., Wu, T., Cao, R., & Cheng, J. (2019). Protein tertiary structure modelling driven by deep learning and contact distance prediction in CASP13. Proteins: Structure, Function, and Bioinformatics, 87(12), 1165-1178.

[53] Lee, J., Freddolino, P. L., & Zhang, Y. (2017). Ab initio protein structure prediction. In From protein structure to function with bioinformatics (pp. 3-35). Springer, Dordrecht.

[54] Cao, R., Bhattacharya, D., Adhikari, B., Li, J., & Cheng, J. (2016). Massive integration of diverse protein quality assessment methods to improve template-based modelling in CASP11. Proteins: Structure, Function, and Bioinformatics, 84, 247-259.

[55] Jones, D. T., &Kandathil, S. M. (2018). High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. Bioinformatics, 34(19), 3308-3315.

[56] Kandathil, S. M., Greener, J. G., & Jones, D. T. (2019). Prediction of interresidue contacts with DeepMetaPSICOV in CASP13. Proteins: Structure, Function, and Bioinformatics, 87(12), 1092-1099.

[57] Wu, Q., Peng, Z., Anishchenko, I., Cong, Q., Baker, D., & Yang, J. (2020). Protein contact prediction using metagenome sequence data and residual neural networks. Bioinformatics, 36(1), 41-48.

[58] AlQuraishi, M. (2019). End-to-end differentiable learning of protein structure. Cell systems, 8(4), 292-301.

[59] Liu, J., Wu, T., Guo, Z., Hou, J., & Cheng, J. (2021). Improving protein tertiary structure prediction by deep learning and distance prediction in CASP14. bioRxiv.

[60] Torrisi, M., Pollastri, G., & Le, Q. (2020). Deep learning methods in protein structure prediction. Computational and Structural Biotechnology Journal, 18, 1301-1310.

[61] Gao, W., Mahajan, S. P., Sulam, J., & Gray, J. J. (2020). Deep learning in protein structural modelling and design. Patterns, 100142.

[62] Li, Y., Hu, J., Zhang, C., Yu, D. J., & Zhang, Y. (2019). ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. Bioinformatics, 35(22), 4647-4655.

[63] Wang, S., Sun, S., Li, Z., Zhang, R., & Xu, J. (2017). Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS computational biology, 13(1), e1005324.

[64] Hanson, J., Paliwal, K., Litfin, T., Yang, Y., & Zhou, Y. (2018). Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. Bioinformatics, 34(23), 4039-4045.

[65] Chen, C., Wu, T., Guo, Z., & Cheng, J. (2021). Combination of deep neural network with attention mechanism enhances the explainability of protein contact prediction. Proteins: Structure, Function, and Bioinformatics, 89(6), 697-707.

[66] Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., ...& Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. Nature, 577(7792), 706-710.

[67] Kinch, L. N., Schaeffer, D. R., Kryshtafovych, A., & Grishin, N. V. (2021). Target Classification in the 14th Round of the Critical Assessment of Protein Structure Prediction (CASP14). Proteins: Structure, Function, and Bioinformatics

[68] Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., & Moult, J. (2019). Critical assessment of methods of protein structure prediction (CASP)—Round XIII. Proteins: Structure, Function, and Bioinformatics, 87(12), 1011-1020.

[69] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ...& Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 1-11.

[70] Murata, K., & Wolf, M. (2018). Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. BiochimicaetBiophysicaActa (BBA)-General Subjects, 1862(2), 324-334.

[71] Esquivel-Rodríguez, J., &Kihara, D. (2013). Computational methods for constructing protein structure models from 3D electron microscopy maps. Journal of structural biology, 184(1), 93-102.

[72] Pakhrin, S. C., Shrestha, B., Adhikari, B., & Kc, D. B. (2021). Deep Learning-Based Advances in Protein Structure Prediction. International Journal of Molecular Sciences, 22(11), 5553

[73] Alnabati, E., &Kihara, D. (2020). Advances in structure modelling methods for cryo-electron microscopy maps. Molecules, 25(1), 82.

[74] Raval, A., Piana, S., Eastwood, M. P., & Shaw, D. E. (2016). Assessment of the utility of contact-based restraints in accelerating the prediction of protein structure using molecular dynamics simulations. Protein Science, 25(1), 19-29.

[75] Morrone, J. A., Perez, A., MacCallum, J., & Dill, K. A. (2017). Computed binding of peptides to proteins with MELD-accelerated molecular dynamics. Journal of chemical theory and computation, 13(2), 870-876.

[76] Jiang, F., & Wu, Y. D. (2014). Folding of fourteen small proteins with a residue-specific force field and replica-exchange molecular dynamics. Journal of the American Chemical Society, 136(27), 9536-9539.

[77] Li, D., Ju, Y., & Zou, Q. (2016). Protein folds prediction with hierarchical structured SVM. Current Proteomics, 13(2), 79-85.

[78] Wu, S., Szilagyi, A., & Zhang, Y. (2011). Improving protein structure prediction using multiple sequence-based contact predictions. Structure, 19(8), 1182-1191.

[79] Savojardo, C., Fariselli, P., Martelli, P. L., &Casadio, R. (2013). Prediction of disulfide connectivity in proteins with machine-learning methods and correlated mutations. BMC bioinformatics, 14(1), 1-8.

[80] Wang, Y., Cheng, J., Liu, Y., & Chen, Y. (2016, May). Prediction of protein secondary structure using support vector machine with PSSM profiles. In 2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference (pp. 502-505). IEEE.

[81] Xie, S., Li, Z., & Hu, H. (2018). Protein secondary structure prediction based on the fuzzy support vector machine with the hyperplane optimization. Gene, 642, 74-83.

[82] Manavalan, B., & Lee, J. (2017). SVMQA: support–vector-machine-based protein single-model quality assessment. Bioinformatics, 33(16), 2496-2503.

[83] Uziela, K., Shu, N., Wallner, B., &Elofsson, A. (2016). ProQ3: Improved model quality assessments using Rosetta energy terms. Scientific reports, 6(1), 1-10.

[84] Ray, A., Lindahl, E., &Wallner, B. (2012). Improved model quality assessment using ProQ2. BMC bioinformatics, 13(1), 1-12.

[85] Kieslich, C. A., Smadbeck, J., Khoury, G. A., &Floudas, C. A. (2016). conSSert: consensus SVM model for accurate prediction of ordered secondary structure.

[86] Zhou, J., Yan, W., Hu, G., & Shen, B. (2014). SVR_CAF: an integrated score function for detecting native protein structures among decoys. Proteins: Structure, Function, and Bioinformatics, 82(4), 556-564.

[87] Wardah, W., Khan, M. G., Sharma, A., & Rashid, M. A. (2019). Protein secondary structure prediction using neural networks and deep learning: A review. Computational biology and chemistry, 81, 1-8.

[88] Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., Sattar, A., Yang,Y., Zhou, Y., (2015). Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. Sci. Rep. 5, 11476.

[89] Spencer, M., Eickholt, J., & Cheng, J. (2014). A deep learning network approach to ab initio protein secondary structure prediction. IEEE/ACM transactions on computational biology and bioinformatics, 12(1), 103-112.

[90] Wang, S., Peng, J., Ma, J., & Xu, J. (2016). Protein secondary structure prediction using deep convolutional neural fields. Scientific reports, 6(1), 1-11.

[91] Drozdetskiy, A., Cole, C., Procter, J., & Barton, G. J. (2015). JPred4: a protein secondary structure prediction server. Nucleic acids research, 43(W1), W389-W394.

[92] 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.

[93] Yaseen, A., & Li, Y. (2014). Context-based features enhance protein secondary structure prediction accuracy. Journal of chemical information and modelling, 54(3), 992-1002.

[94] Li, Z., & Yu, Y. (2016). Protein secondary structure prediction using cascaded convolutional and recurrent neural networks. arXiv preprint arXiv:1604.07176.

[95] Zhang, B., Li, J., & Lü, Q. (2018). Prediction of 8-state protein secondary structures by a novel deep learning architecture. BMC bioinformatics, 19(1), 1-13.

[96] Mirabello, C., & Pollastri, G. (2013). Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility. Bioinformatics, 29(16), 2056-2058.

[97] Heffernan, R., Yang, Y., Paliwal, K., Zhou, Y., (2017). Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers, and solvent accessibility. Bioinformatics 33 (18), 2842–2849.

[98] Fang, C., Shang, Y., & Xu, D. (2018). MUFOLD-SS: New deep inception-inside-inception networks for protein secondary structure prediction. Proteins: Structure, Function, and Bioinformatics, 86(5), 592-598.

[99] Zhang, H., & Shen, Y. (2020). Template-based prediction of protein structure with deep learning. BMC genomics, 21(11), 1-9.

[100] Boumedine, N., & Bouroubi, S. (2019). A new hybrid genetic algorithm for protein structure prediction on the 2D triangular lattice. arXiv preprint arXiv:1907.04190.

[101] Correa, L., Borguesan, B., Farfan, C., Inostroza-Ponta, M., & Dorn, M. (2016). A memetic algorithm for 3D protein structure prediction problem. IEEE/ACM transactions on computational biology and bioinformatics, 15(3), 690-704.

[102] Dorn, M., Inostroza-Ponta, M., Buriol, L. S., &Verli, H. (2013, June). A knowledge-based genetic algorithm to predict three-dimensional structures of polypeptides. In 2013 IEEE Congress on Evolutionary Computation (pp. 1233-1240). IEEE.

[103] Borguesan, B., e Silva, M. B., Grisci, B., Inostroza-Ponta, M., & Dorn, M. (2015). APL: An angle probability list to improve knowledge-based metaheuristics for the three-dimensional protein structure prediction. Computational biology and chemistry, 59, 142-157.

[104] Borguesan, B., Inostroza-Ponta, M., & Dorn, M. (2017). Nias-server: Neighbors influence of amino acids and secondary structures in proteins. Journal of Computational Biology, 24(3), 255-265.

[105] 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. Evolutionary computation, 24(4), 577-607.

[106] Rocha, G. K., Custódio, F. L., Barbosa, H. J., & Dardenne, L. E. (2016, July). Using crowding-distance in a multiobjective genetic algorithm for protein structure prediction. In Proceedings of the 2016 on Genetic and Evolutionary Computation Conference Companion (pp. 1285-1292).

[107] Gao, S., Song, S., Cheng, J., Todo, Y., & Zhou, M. (2017). Incorporation of solvent effect into multi-objective evolutionary algorithm for improved protein structure prediction. IEEE/ACM transactions on computational biology and bioinformatics, 15(4), 1365-1378.

[108] Cheng-yuan, L., Yan-rui, D., & Wen-bo, X. (2010, August). Multiple-layer quantum-behaved particle swarm optimization and toy model for protein structure prediction. In 2010 Ninth International Symposium on Distributed Computing and Applications to Business, Engineering and Science (pp. 92-96). IEEE.

[109] Zhou, C., Hou, C., Wei, X., & Zhang, Q. (2014). Improved hybrid optimization algorithm for 3D protein structure prediction. Journal of molecular modelling, 20(7), 1-12.

[110] 110. Li, B., Li, Y., & Gong, L. (2014). Protein secondary structure optimization using an improved artificial bee colony algorithm based on AB off-lattice model. Engineering Applications of Artificial Intelligence, 27, 70-79.

[111] Li, B., Chiong, R., & Lin, M. (2015). A balance-evolution artificial bee colony algorithm for protein structure optimization based on a three-dimensional AB off-lattice model. Computational biology and chemistry, 54, 1-12.

[112] Li, B., Lin, M., Liu, Q., Li, Y., & Zhou, C. (2015). Protein folding optimization based on 3D off-lattice model via an improved artificial bee colony algorithm. Journal of molecular modelling, 21(10), 1-15.

[113] Parpinelli, R. S., Benitiez, C. M., Cordeiro, J., & Lopes, H. S. (2014). Performance Analysis of Swarm Intelligence Algorithms for the 3D-AB off-lattice Protein Folding Problem. J. Multiple Valued Log. Soft Comput., 22(3), 267-286.

[114] Khakzad, H., Karami, Y., & Arab, S. S. (2015). Accelerating protein structure prediction using particle swarm optimization on GPU. BioRxiv, 022434.

[115] Wang, D., Geng, L., Zhao, Y. J., Yang, Y., Huang, Y., Zhang, Y., & Shen, H. B. (2020). Artificial intelligence-based multi-objective optimization protocol for protein structure refinement. Bioinformatics, 36(2), 437-448.

[116] Akbar, S., Pardasani, K. R., & Khan, F. (2021). Swarm optimization-based neural network model for secondary structure prediction of proteins. Network Modelling Analysis in Health Informatics and Bioinformatics, 10(1), 1-9.

[117] Kamal, M. S., Chowdhury, L., Khan, M. I., Ashour, A. S., Tavares, J. M. R., & Dey, N. (2017). Hidden Markov model and Chapman Kolmogrov for protein structures prediction from images. Computational biology and chemistry, 68, 231-244.

[118] Gao, P., Wang, S., Lv, J., Wang, Y., & Ma, Y. (2017). A database-assisted protein structure prediction method via a swarm intelligence algorithm. RSC advances, 7(63), 39869-39876.