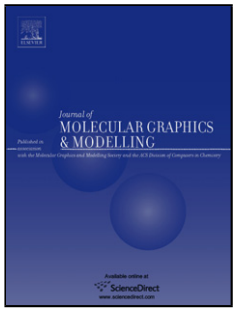


Title: Protein Secondary Structure Prediction: A Survey of the state of the art

Authors: Qian Jiang, Xin Jin, Shin-Jye Lee, Shaowen Yao

PII: S1093-3263(17)30421-7
DOI: http://dx.doi.org/doi:10.1016/j.jmgm.2017.07.015
Reference: JMG 6974

To appear in: *Journal of Molecular Graphics and Modelling*

Received date: 5-6-2017
Revised date: 14-7-2017
Accepted date: 17-7-2017

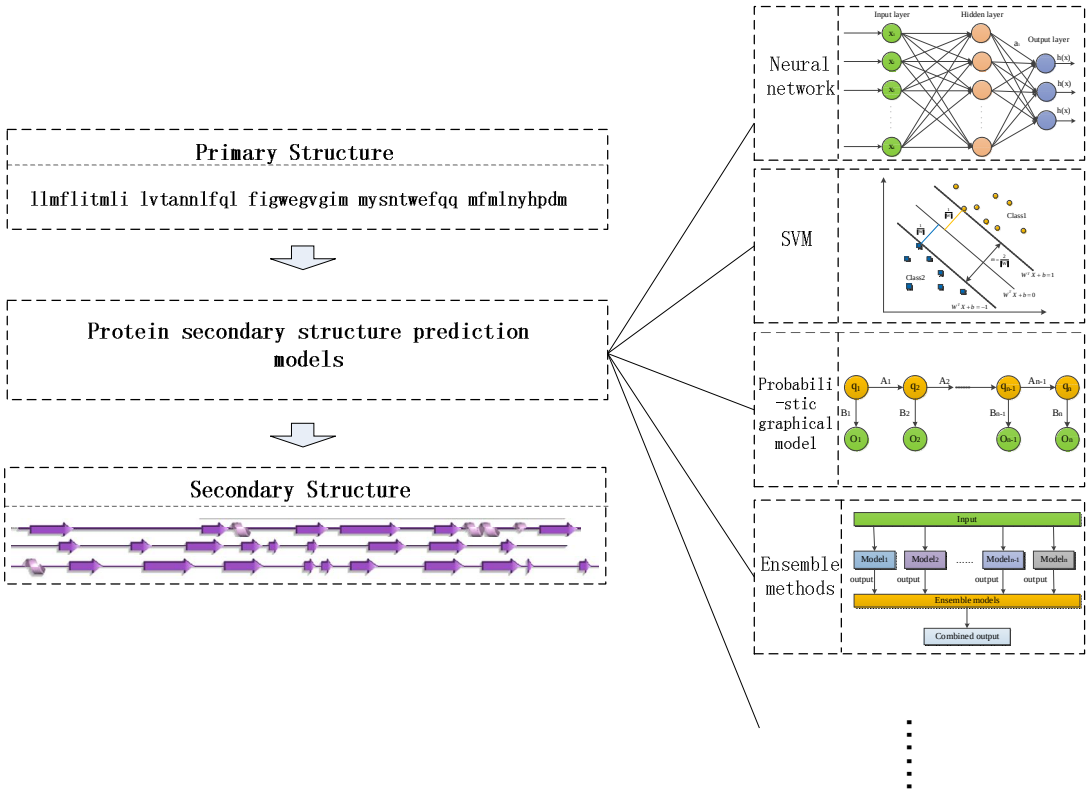
Please cite this article as: Qian Jiang, Xin Jin, Shin-Jye Lee, Shaowen Yao, Protein Secondary Structure Prediction: A Survey of the state of the art, Journal of Molecular Graphics and Modellinghttp://dx.doi.org/10.1016/j.jmgm.2017.07.015

Protein Secondary Structure Prediction: A Survey of the state of the art

Qian Jiang[1], Xin Jin[1], Shin-Jye Lee[2, 3]*, Shaowen Yao[2]*

[1] School of Information, Yunnan University, Kunming, China
[2] School of Software, Yunnan University, Kunming, China
[3] Queens' College, University of Cambridge, Cambridge, England, United Kingdom
* Correspondence: Email: camhero@gmail.com; yaosw@ynu.edu.cn;

Graphical abstract



Highlights:

1. The background and related knowledge of protein secondary structure prediction is introduced.
2. The often-used prediction accuracy assessments methods are described.
3. The recent algorithmic developments of protein secondary structure prediction are reported in detail.
4. The corresponding tendencies and challenges are summarizes.
5. We conclude there are still further improvements and extensions in this field.

Abstract: Protein secondary structure prediction (PSSP) is a fundamental task in protein science and computational biology, and it can be used to understand protein 3-dimensional (3-D) structures, further, to learn their biological functions. In the past decade, a large number of methods have been proposed

for PSSP. In order to learn the latest progress of PSSP, this paper provides a survey on the development of this field. It first introduces the background and related knowledge of PSSP, including basic concepts, data sets, input data features and prediction accuracy assessment. Then, it reviews the recent algorithmic developments of PSSP, which mainly focus on the latest decade. Finally, it summarizes the corresponding tendencies and challenges in this field. This survey concludes that although various PSSP methods have been proposed, there still exist several further improvements or potential research directions. We hope that the presented guidelines will help nonspecialists and specialists to learn the critical progress in PSSP in recent years.

Keywords: Protein secondary structure prediction, Machine learning, Neural networks, Classification algorithm, Feature extraction

1. Introduction

Protein is one of the most important molecules in any living cells and it also is the material basis of all living organisms, which involved in all the processes of life, such as guiding the catalysis of biochemical reactions, signal transduction and transmission or the correct expression of genetic information [1] [2]. A large volume protein sequence data are available in database with very low cost due to the development of advanced sequencing technologies; however, the amount of the structures of these sequences known by man only is about 0.2% as a whole, and the recognized functions are much less[3] [4] [5] [6]. Therefore, it still is a big challenge for biologists and medical scientists to understand the structures and functions of proteins from such massive sequences data [1] [7]. As a result, by the using advanced computer technologies to learn the structural information of proteins becomes a basic task in protein science and bioinformatics, and it can be used to understand how proteins to exercise its biological function and the relationship between protein and protein molecules.

The function of a protein is closely related to its structure; therefore, it is clear that the prediction of unknown protein structures of a proteome could have a strong positive impact on any attempt to discover its function. Generally, protein structures hierarchies are classified into four distinct levels: the primary, secondary, tertiary and quaternary. Protein tertiary structure and quaternary structure determines the 3-D structure of a protein and further determines its functional characteristics. The secondary structure is a bridge between the primary and tertiary structure, and it is the early folding stage of protein molecule as the foundation of protein 3-D structure. As a result, the research of protein secondary structures is indispensable as the first and the most important step in 3-D structure studies, which can help to understand the relationship between the function and primary structure of proteins [8]. Except to learn the 3-D spatial structure of protein, it can also be used in many protein science fields [9], such as the prediction of native tertiary structure [10], prediction of transition-state position [11], real value prediction of solvent accessibility [12], prediction of protein-protein interactions [13], prediction of protein structural classes [14] [15] [16] [17], prediction of protein domains [18] [19], prediction of π-turns in proteins [20] and so on.

Since the rapid development of genomics and proteomics; especially, the DNA sequencing technologies and protein sequencing technologies resulted in enormous accumulation of protein sequence data. It is one of the biggest challenges in bioinformatics to learn the secondary structures and tertiary structures of protein from its sequence data. Generally, protein structure can be obtained by X-ray crystallography and multi-dimensional magnetic resonance in laboratory, and these experimental methods can get accurate protein structure information with high precision; however, these method have the disadvantages of extremely difficult, cost prohibitive, time consuming, limited molecular weight. Consequently, the experimental methods apparently do not cope with the challenge of the rapidly growing protein sequences data [3] [4] [5] [21]. Computational methods for protein structure prediction is one of the most important and effective technologies in the emerging interdisciplinary field of bioinformatics due to the rapid advances of protein sciences, and it has the characteristics of simple, low cost and fast speed, which can overcome the disadvantage of experimental methods [3] [5] [22].

The principle of computational methods for PSSP is that the methods can learn some rules based on the analysis of known protein sequences and their secondary structures to predict the secondary

structures of unknown protein sequences. Generally, it should consider the following questions in PSSP methods: (i) How to construct or select a valid and reliable benchmark dataset for training and testing the predictor [23]; (ii) How to effectively obtain structural information form protein sequences by feature extraction methods; (iii) Which classification algorithm could be used to accurately predict the secondary structures.

The first PSSP method was presented by Chothla and Levitt in 1976, and then the prediction technologies could be regarded as three generations [8] [24]. The first generation was emerging before the 80s, which generally utilized statistical probability of the individual residue which was assigned to different secondary structures; however, the overall accuracy of these methods were less than 60%, which did not satisfy the requirement of protein 3-D structure analysis and prediction, and the representative was Chou-Fasman's method. The second generation emerged about from 1980 to 1992; the sophisticated statistical methods were utilized for PSSP; and the neighboring residues information was took into consideration by a sliding window, besides, other characteristic information of protein (such as physico-chemical) also was considered. These methods could improve the prediction accuracy to some extent, but the overall accuracy was still less than 65%, the typical method was GORIII. The third generation emerged after 1992; these methods generally used multiple sequence alignment (MSA) profile (such as position specific scoring matrices) as the input of advanced machine learning model to predict protein secondary structure. These methods not only considered the compositions and interactions of amino acid, but also considered more other features, such as long-range correlation. The representative methods were PHD and PSIPRED, and the overall precision of this generation was about 76%-80% [25] [26].

Table 1. The development of protein secondary structure prediction methods

Description	First generation	Second generation	Third generation	Current
Time	before the 80s	1980 to 1992	after 1992	the last decade(2006-)
Accuracy	less than 60%	less than 65%	about 76%-80%	about 76%-85%
Method	statistical probability	sophisticated statistical methods	advanced machine learning method	improved machine learning method; hybrid model
Utilized Information	individual residue	neighboring residues information; physico-chemical information	multiple sequence alignment profile; homogeneous information; long-range correlation	multiple sequence alignment profile; homogeneous Information; long-range correlation; other protein natural properties
Representative method	Chou-Fasman's method	GORIII	PHD and PSIPRED	deep leaning; compound pyramid model

Afterwards, we find that the PSSP techniques received more widely attention and emerged lots of new nature and tendencies in the last decade, and the development of PSSP methods are shown in Table 1. The accuracy of PSSP has been further improved due to the new advances of hybrid models and machine learning tools, such as upgraded versions of neural network, SVM, probabilistic graphical model, fuzzy theory and son on [27] [26]; in addition, optimization based methods and ensemble learning methods are also widely used in this field. Besides, after analyzing the most of papers in recent years, this work considers the features of these PSSP methods are that they take into account the evolutionary information, amino acid composition and other protein natural properties, simultaneously.

In addition to the individually modified model based PSSP schemes, we find more and more scholars recognized that the performance of individual model often have some limitations due to the inherent weakness. Different models could extract corresponding protein features according to their different characteristics; therefore, these models may complement the weakness of each other, they could be integrated as a hybrid model which was a new trend in recent years, such as the combination of multiple models and ensemble methods. After statistical analysis of recent papers, we find that many

excellent methods have been proposed over the last decade and the prediction accuracy has nearly researched to 85%; however, further improvements of PSSP methods are needed for more precise 3-D structure prediction [3].

In order to make an overview on recent advances in PSSP, this paper provides a survey on the developments of this field in the last decade. In this paper, we provide the detailed background and introduction of PSSP in the first section. Secondly, we make a survey on the recent advances of PSSP to report its research status, which mainly focus on the latest trends. Finally, the corresponding tendencies and limitations are discussed. We conclude with a brief discussion of the remaining challenges in PSSP, although various prediction methods have been proposed. Using the presented guidelines will help non-specialists and specialists to understand the critical progress of PSSP in recent years.

This rest of paper is arranged as follows. In Section 2, we introduce the related knowledge of PSSP; Section 3 provides a taxonomical survey of the different families of PSSP technique to describe its development. Section 4 presents its future trends and limitations. Section 5 concludes this paper.

2. Related knowledge of protein secondary structure prediction

The protein structure hierarchies can be divided into four levels: primary, secondary, tertiary and quaternary, the four diagrammatic drawings of protein structure as shown in Fig. 1. The primary structure only is a linear sequence of protein polypeptide chain. The secondary structure refers to the periodic structure fragment of polypeptide chain; it is generated by the effect of hydrogen bonds, and along the direction of the polypeptide chain in one-dimensional space. The tertiary structure is a whole polypeptide chain which generated by further combination and fold of multiple secondary structures in 3-D space, and it could already represent the primary biological function of those proteins which only has one polypeptide chain. The quaternary structure is protein complexes, which consists of several polypeptide chains with multiple tertiary structures, and it could completely represent the characteristics of its biological function.

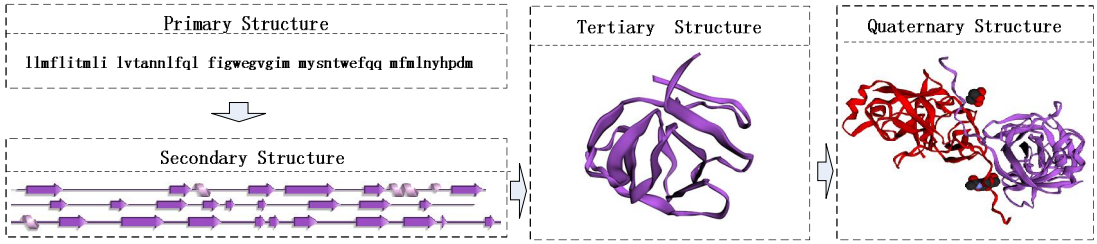


Figure 1. Diagrammatic drawing of protein structure [28]

2.1 The description of protein secondary structure

Among these four structural hierarchies, protein secondary structure plays an important role in protein science, and it is the primary fold of polypeptide chain and the basis of spatial structure for a protein. The secondary structures present different states by the impact of internal hydrogen bond in polypeptide chain. At first, the researchers thought there were only three secondary structures for amino acids in a protein: Helix (H), Strand (E) and Coil (C). These secondary structures could reflect the local spatial arrangement of amino acids: H is a helical configuration which are strengthened by the hydrogen bonds between every fourth amino acid; E is a strand segments structure of linked parallel or antiparallel which generated by hydrogen bonds among interacting amino acids; C is a default class for those amino acids that do not belong to H or E classes [84].

Whereafter, the 3-state is extended to 8-state. Specifically, there are several classification ways for protein secondary structure, including definition of secondary structure of proteins (DSSP) which is based on the repetitive patterns of hydrogen bonding in 3-D space, protein secondary structure assignment from atomic coordinates (STRIDE) which is according to the statistical distribution of hydrogen bonding and dihedral angle, and other classification strategies. Different classification ways for these eight classes will have a big influence on predicted results [29], but DSSP is the most often-

used method in PSSP, which divide residues into eight different secondary structures: H (α-helix), G (310-helix), I (π-helix), E (β-strand), B (isolated β-bridge), T (turn), S (bend), and C (others).

8-state structure based PSSP methods could provide more detailed local structure information [75], but its prediction is going to be more difficult, directly causing its accuracy is lower than 3-state predication about 12% [52]. So far, almost all of the methods were 3-state based PSSP. There are five popular ways to transform 8-state secondary structure into 3-state, including method 1: H {H, G}, E {E, B}, C {S, T, I, C}; method 2: H {H}, E {E}, C {G, S, T, B, I, C}; method 3: H {H, G, I}, E {E, B}, C {S, T, C}; method 4: H {H, G}, E {E}, C {S, T, B, I, C}; method 5: H {H, G, I}, E {E }, C {S, T, B, C} [59], which are shown in Table 2. Therefore, it should be pointed out that the mapping from 3-state to 8-state is obviously vague due to their different classification criterions, and the main reason for this is that there are no very clear definitions or boundaries for these secondary structures in chemistry and biology.

Table 2. 8-to-3 state reduction methods

Class name	3-state	8-state				
		method 1	method 2	method 3	method 4	method 5
Helix	H	H, G	H	H, G, I	H, G	H, G, I
Strand	E	E, B	E	E, B	E	E
Coil	C	S, T, I, C	G, S, T, B, I, C	S, T, C	S, T, B, I, C	S, T, B, C

2.2 Datasets

In PSSP studies, the choice of data sets is a crucial component because it has a seriously impact for the training effect of PSSP model. In general, the data sets themselves do not represent any features, but merely secondary structure assigned to protein sequences [67]. Among these data sets, CB513, CB396, RS126, EVA and PSIPRED are often used.

- CB513 and CB396

CB513 dataset has 513 sequences and comprises 84,107 residues, which is developed by Cuff and Barton [30]. It is a non-homologous and well-established benchmark data set, and the sequence similarity of all 513 proteins are less than 25% to ensure that there is very little homology in the training set [78]. It is one of the most often-used independent dataset including the CB396 dataset and 117 sequences from RS126 [61] [31].

CB396 is a non-redundant dataset and has 396 proteins from CB513. The percentage sequence identity is less than 34% and the average sequence length is 157 residues [32].

- RS126

RS126 has 126 protein sequences and comprises 26,846 residues, which also is one of the most frequently-used non-homologous dataset developed by Rost and Sandar [33]. The average sequence identity is less than 31% and the average sequence length is 185 residues [61].

- EVA

EVA data set is provided by a web-based server, named evaluation of automatic protein structure prediction. It is one of the largest sequence unique subset of the protein data bank (PDB), and there is no pair in the subset has more than 33% identical residues over the more than 100 residues aligned.

GSW25 and CASP data set including its reduced datasets: CASP394 and RCASP256 (256 proteins) could be used for blind testing, but not suitable for model development [34] [68] [31]. In addition to the above mentioned data set, PSIPRED, SCOP and SPINE also are used by some researchers [55] [59] [27]. Besides, there are some other data sources for PSSP, such as nuclear magnetic resonance spectra [35], vacuum ultraviolet circular dichroism spectroscopy [36] and Fourier transform infrared spectroscopy [37] and so on.

2.3 Input data feature

Input data feature is another key issue of PSSP; it is generated by data sets and can be directly regarded as inputs of train classifiers. The performance of classifier crucially depends on the judicious

choice of the feature vectors that are utilized by classification algorithm to partition the feature space into disjoint regions with decision boundaries [102]. The features could effectively represent the structure information of amino acid sequence and are expressed as a set of numbers, named feature vectors. Appropriate input data features will effectively improve the prediction accuracy of PSSP [64], and there are many input forms for different prediction methods [41] [119], which can be classified into the following groups.

2.3.1 Single sequence

Single sequence based prediction algorithms cannot utilize the homologous information from other proteins, and its prediction accuracy is lower than that additional evolutionary information based methods in general. However, it is important for protein science because most of the proteins identified in genome sequencing projects have no referable sequence similarity to any known protein [87] [128] [129]. Besides, the single sequence based PSSP method is convenient, and it is simple and easy to use in actual applications of protein science.

2.3.2 Multiple sequence alignment profile

Multiple sequence alignment profile of homologous proteins could represent their structural alignment and aligned residues which usually have similar secondary structures [59]. The multiple sequence alignment profile could be generated by many methods, such as PSI-BLAST, PSI-Search, HMMER3, AMPS and CLUSTALW [31]. Multiple sequence alignments can produce position-specific profiles which provide crucial information about structure and could be used as inputs for PSSP methods. Because the homologous information of proteins is very reliable support to predict unknown protein secondary structures. BLOSUM62 matrix and PSSM matrix are frequently-used multiple sequence alignment profiles, especially the later. Besides, some researchers also proposed other methods to get multiple sequence alignment profiles [38] [39] [40] [41].

- Position specific scoring matrix

Position specific scoring matrix (PSSM) is generally obtained by PSI-BLAST algorithm according to the calculation of position-specific scores for each position in the multiple alignments with the highest scoring hits [31]. The profile information of each amino acid residue is estimated as a 20 element feature-vector which record by 20 columns [77]. Highly and weakly conserved positions are represented by high scores and scores near zero, respectively. The PSSM matrix can effectively discover distantly related protein sequences and evolution information due to it can reflect different substitution patterns at different positions [42] [71]. PSSM can greatly improve the precision of PSSP, especially for beta-fold, and it is the most widely used methods.

- BLOSUM62 matrix

The BLOSUM62 matrix is given as a scoring matrix and proposed by Henikoff and Henikoff. It can effectively measure the difference between two proteins specifically for more distantly related proteins. The BLOSUM62 matrix represent the ‘log-odds’ scores of likelihood by numerical value, and the likelihood represent the possibility that a given amino acid pair will interchange. In BLOSUM62 matrix, a pair of amino acids with similar chemical properties is more likely to replace each other which are given a positive score. Conversely, the given amino acid pairs with very different physicochemical properties are given a negative score [41].

2.3.3 Physicochemical properties

The physical-chemical properties of amino acids will have a serious influence for their protein secondary structures and can be used to predict secondary structures according to the character of amino acids and the arrangement between residues. For the analysis of protein sequences, the 20 amino acids can be grouped into 8 types based on their similarity of physicochemical properties, namely, Hydrophobic, Hydrophilic; Polar, Non-polar; Small, Large; Charged, Uncharged [43]. In the study of protein structure, the 8 representative physicochemical properties can be used to encode each residue and correlative information is examined in relation to the formation of protein secondary structure. In these

8 physicochemical properties, hydrophobic, hydrogen bond and charge properties have a large influence on protein secondary structures, therefore hydrophobic and hydrogen are often used properties in PSSP [31]. Similar to BLOSUM62 matrix, hydrophobicity matrix also is one of physicochemical properties based matrix [41].

2.4 Prediction accuracy assessment

In order to evaluate the prediction quality of PSSP methods in objective way, many assessment indexes are widely adopted by most researchers. The prediction quality assessment methods are generally used to measure the prediction accuracy of PSSP algorithms, and found the optimal parameters for a specific algorithm. Besides, the assessment methods can also be adopted to evaluate and intuitive represent the effectiveness of the PSSP approaches.

The most frequently-used assessment methods are Q score and segment overlap (SOV), which are integrative assessment methods. And critical assessment of methods of protein structure prediction (CASP) identifies SOV as being a more appropriate measure of prediction accuracy [50] [44]. Besides, some other measures are also used to assess the performance of PSSP techniques, such as Matthews correlation coefficient, average absolute error, mean absolute error and so on.

2.4.1 Q score

The estimated performance of PSSP method is often assessed by three-state-per-residue accuracy (Q3) or eight-state-per-residue accuracy (Q8) scoring function which are the simplest and most popular measure methods as (1), and Q score calculates the percent of residues for each secondary structure is correctly predicted [45] [50] [78].

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

where m=3 and m=8 is referred as Q_3 and Q_8 accuracy, respectively. N_{res} is the total number of residues, and M_{ii} is correctly predicted number of residues in state i.

The per-state accuracy is the percentage of correctly predicted residues in a particular state, as (2).

$$Q_i = 100 \frac{M_{ii}}{obs^i} \tag{2}$$

where obs^i is the number of residues observed in state i.

2.4.2 Segment Overlap

Segment Overlap (SOV) score takes into account the segments of continuous structure types instead of simple calculation of the number of correct residues, which is based on the average overlap between the observed and the predicted segments instead of the average per-residue accuracy, as (3). SOV could tolerate a small number of mistakes at the ends of secondary structure segment, but it would seriously penalize these mistakes in the middle region of a secondary structure segment [45] [52].

$$SOV = 100 \times \frac{1}{\sum_i N(i)} \sum_i \sum_{S(i)} \frac{\min ov(s_1, s_2) + \delta(s_1, s_2)}{\max ov(s_1, s_2)} \times len(s_1) \tag{3}$$

where $N(\mathrm{i})$ is the number of residues in state i, s_1 and s_2 are the observed and predicted structure segments, $\min ov(\mathrm{s}_1, \mathrm{s}_2)$ is the length of actual overlap of s_1 and s_2, $\max ov(\mathrm{s}_1, \mathrm{s}_2)$ is the length of the total extent for which either of the segments s_1 and s_2 has a residue in state i. $len(\mathrm{s}_1)$ is the number of residues in the segment of s_1, $\delta(\mathrm{s}_1, \mathrm{s}_2)$ is defined as (4) :

$$\delta(s_1, s_2) = \min((\max ov(s_1, s_2) - \min ov(s_1, s_2)); \min ov(s_1, s_2); \mathrm{int}(len(s_1)/2); \mathrm{int}(len(s_2)/2)) \tag{4}$$

2.4.3 Matthews correlation coefficient

Matthews correlation coefficient (MCC) is a more robust measure for prediction quality, and it is often used in machine learning field as a correlation coefficient to measure the quality of binary classifications (two class) [77] [87]. It takes into account both over- predictions and under-predictions, and it is generally regarded as a balanced measure, which still can be used even the classes have different sizes. It returns a value between −1 and +1, the +1 represents a perfect prediction, 0 represents an average random prediction and −1 represents an inverse prediction. MCC is defined for each type of secondary structure, and the corresponding formulation is given as (5).

$$MCC = \frac{TP \times TN - FP \times FN}{[(TN + FN)(TN + FP)(TP + FN)(TP + FP)]^{1/2}} \tag{5}$$

where TP (true positives) is the number of residues correctly predicted for a secondary structure. TN (true negatives) is the number of residues that are not predicted as the secondary structure. FP (false positives) is the number of residues incorrectly predicted for the secondary structure, and FN (false negatives) is the number of residues observed in the secondary structure but predicted to be other [87].

2.4.4 Average absolute error

Average absolute error (δ^{Θ}) of each secondary structural element is calculated for each protein [75], as (6). δ^{Θ} is the average of the absolute deviations between measured value and the mean value of data set, and it can actually reflect prediction error.

$$\delta^{\Theta} = \frac{1}{N_{res}} \sum_{k=1}^{N} \left| \Theta_k - y_k^{\Theta} \right| \tag{6}$$

where Θ_k is the predicted content of the secondary structural element Θ for the k-th protein, and y_k^{Θ} is the content actually observed.

The second criterion is the standard deviation of the average absolute error formulated as follows:

$$\sigma^{\Theta} = sqrt\left(\frac{1}{N-1} \sum_{k=1}^{N} (\delta^{\Theta} - \left| \Theta_k - y_k^{\Theta} \right|)^2 \right) \tag{7}$$

The third criterion is the overall average error given by

$$\langle \delta \rangle = \frac{1}{8} \sum_{\Theta} \delta^{\Theta} \tag{8}$$

2.4.5 Mean absolute error

The mean absolute error (MAE) is the average of the absolute distance between the observed and predicted value. In order to take in account the periodicity of dihedral angles, the MAE is calculated by [78]:

$$MAE = \frac{1}{N_{res}} \sum_{i=1}^{N_{res}} \min(\left| p_i - x_i \right|, \left| 360^{o} - (p_i - x_i) \right|) \tag{9}$$

where, N_{res} is the total number of residues, x represents the observed value, and p represents the predicted value.

Besides, there are some less used indicators for the prediction quality analysis of PSSP, such as cross validation tests, self-consistency, standard error of prediction (SEP), k-state correlation coefficients, fuzzy Overlap (FOV), fuzzy correlation coefficient (Forr) and so on [45] [75].

3. Current approaches

In this section, we will introduce the last developments of PSSP methods according to its classification strategies, including neural network (NN), SVM, probabilistic graphical model, fuzzy

theory and other improved methods; besides, in hybrid models, especially the combination of multiple models and ensemble methods also are introduced.

3.1 Neural networks

Artificial neural network (ANN) is a kind of biomimetic model, which is the simulation of animal neural system regarded as a black box. Generally, neural network consists of three parts, as input layer, hidden layer and output layer, as shown in Fig. 2. In PSSP, NN can automatically seek and establish a mapping relationship between input vector and the secondary structure of central residue by a local window (generally 11-17 residues), and it can effectively capture the local information of protein to predict the structure of central residue [46] [47], the output of feedforward NN as (10) and (11). NN could represent the information and knowledge of the input data by the interconnected relationships of each neuron, and the connecting weights of each neuron would be constantly adjusted by special algorithm to make its output as close as possible to the correct result (structure of central residue) in the training process. Therefore, different NN models would generate different predicted results due to their different ability to find the relationships.

Since 1988, Qian and Sejnowski proposed one of the earliest NN method for PSSP [1] [26], NN gradually became the most widely used model in this field and achieved remarkable achievements. In the last decade, the most frequently-used NN models are deep learning, recurrent neural networks, BP neural network, radial basis function neural network and complex-valued neural network [127], the summary of neural networks for PSSP methods are shown in Table 3. Nowadays, deep learning is becoming the most promising and widely used NN in PSSP.

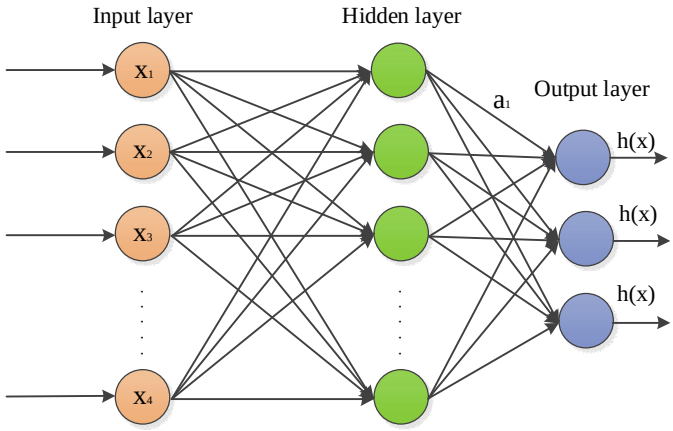


Figure 2. The schematic diagram of neural network

The output function of NN is represented as follows:

$$h(x) = f(\sum_{i=1}^{n} \omega_i a_i + b_i) \tag{10}$$

where $h(x)$ represent the output of PSSP that it have three results for 3-state or eight results for 8-state, $f(z)$ the activation function, a_i is activate value, ω_i is the connecting weights, b_i is intercept, n is the number of nodes of hidden layer for output. The sigmoid function and tanh function is often used: sigmoid function $f(z) = \frac{1}{1+\exp(-z)}$; tanh function $f(z) = \tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$. And a_i is given by:

$$a_i^{(l+1)} = f\left(\sum_{j=1}^{n} \omega_j^{(l)} x_j + b_j^{(l)}\right) \tag{11}$$

where l represent the l-th hidden layer, i represent the i-th node of l+1-th hidden layer, j represent the j-th node of l-th hidden layer that is linked with the i-th node of l+1-th hidden layer.

3.1.1 Deep learning

In recent years, the theories and applications of deep learning get a lot of attentions and achieve good effect in many fields. Meanwhile, it becomes a very powerful artificial intelligence tool in bioinformatics. Compared with other shallow learning models (such as, SVM, traditional and boosting neural network), deep learning has much more multilayer nonlinear operational elements as its hidden layers [48], as shown in Fig. 3. The intent of deep learning is to discover more abstract representation of the features in the higher levels, which is the combination of the features in lower levels [49]. By this way, the model can find more advanced features from raw data which are very suitable for human perception. The pattern recognition of PSSP is highly nonlinear and complex problem, and the distribution of the features in protein sequence is high dimensionality and variability. Therefore, many scholars try to explore the applications of deep learning in PSSP due to its huge potential and good prospect in bio-data processing.

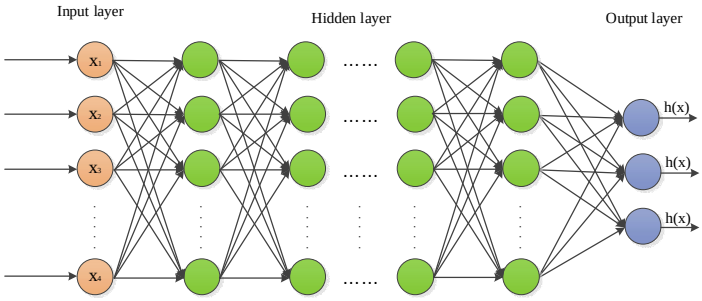


Figure 3. The schematic diagram of deep neural network

In 2014, Spencer et al. proposed the first deep learn based PSSP method, called DNSS, and it was a deep belief network (DBN) model based on restricted Boltzmann machine (RBM) and trained by contrastive divergence46 in an unsupervised manner. The method used PSSM generated by PSI-BLAST to train deep learning network. Due to deep learning network was difficult to train, and the process required a lot of calculation and a large number of training samples, therefore they applied graphical processing units (GPU) and CUDA software to optimize the model [50]. This work firstly introduced deep learning into PSSP and confirmed that deep learning could contribute to the advancement of this field as traditional neural networks which have played an important role in PSSP.

Heffernan et al. considered that more additional features can further improve the prediction accuracy of PSSP; therefore, they designed an iterative features based PSSP method including solvent accessible surface area, backbone angles and dihedrals based on Cα atoms. By considering the multiple features of different physicochemical properties of amino acid and PSSM residues, they proposed a deep learning based PSSP method which showed that the evolutionary information of protein and the excellent performance of deep learning would further improve the prediction accuracy of secondary protein structure [51].

In 2016, Wang et al. proposed a deep convolutional neural fields (DeepCNF) model which was a deep learning extension of conditional neural fields (CNF) for PSSP including 3-state and 8-state. The DeepCNF integrated the advantages of both CNF and deep convolutional neural networks (DCNN), and it took into account the complex sequence-structure relationship and interdependency between adjacent secondary structure labels, and longer-range dependencies information [52]. Afterwards, Wang et al. proposed a deep recurrent encoder-decoder networks for PSSP, named secondary structure recurrent encoder-decoder networks (SSREDNs) which combined deep architecture and recurrent structures to mapping relationship between input protein features and secondary structures, and the mutual interaction among continuous residues of protein [53].

The overall performances of deep learn based PSSP methods were significantly better than the state-of-the-art methods. However, the training process needs a large number of samples due to the characteristics of deep learning, which may lead to the deep learning model have not a good performance

when the training data is limited. And it is normally regarded as a black-box system due to man cannot completely understand its operation mechanism; as a result, when we want to further improve the performance of deep learning based PSSP methods; it could only rely on increasing layer's number, extending computing resources and adjusting parameters. However, it still has a very good prospect in PSSP.

3.1.2 Recurrent neural network

Recurrent neural network (RNN) is powerful connectionist model whose connections among the units form a directed cycle, which is provided with dynamic temporal behavior. It combines the previous step and a hidden representation into the representation of current step, which is operated on the linear progression of time, as shown in Fig. 4, and the output of RNN as (12). These operational modes make RNN have internal memory characteristics which can take into consideration the long-range information in protein, and it overcomes the main disadvantage of feedforward networks that only consider the local information due to the limitation of fixed-size window, and widely used in this field in recent years.

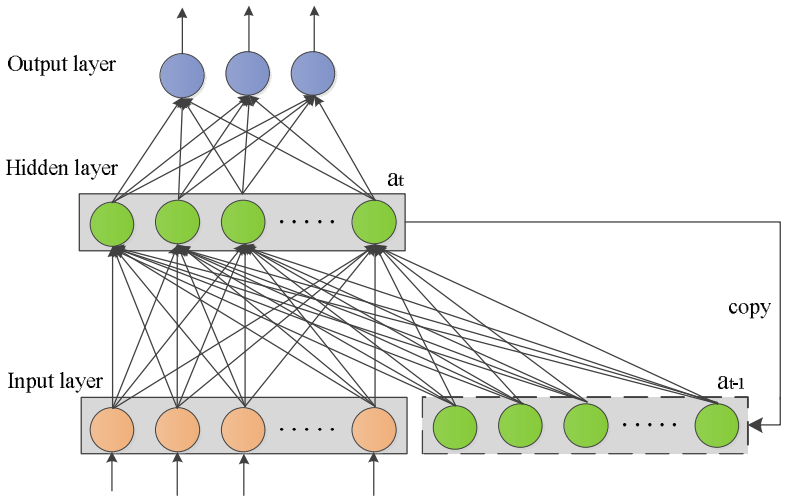


Figure 4. The schematic diagram of RNN

The activate value a_i is represented as follows:

$$a_i^{(l+1)}(t) = f\left(\sum_{j=1}^{n} \omega_j^{(l)} x_j^{(l)}(t) + u_i a_i^{(l+1)}(t-1)\right) \tag{12}$$

where t is current time step, u_i are the weights which connect the previous state and the current state.

But RNN also have restrictions as the future input information cannot be reached from the current state. Pierre et al. introduced bidirectional recurrent neural network (BRNN) into PSSP, which connect two non-causal hidden layers of opposite directions to capture both upstream and downstream information [54], and it could overcome the shortcoming of RNN and further improve the prediction accuracy of PSSP. Since then, scholars proposed some improved model on the basis of BRNN to get more accurate results for PSSP. Kountouris et al. established several prediction methods based on an ensemble of six BRNN architectures and filtering methods, which would be introduced in Section 3.8 [134]. Chen et al. developed a cascaded BRNN which used the results of the first BRNN (sequence-to-structure BRNN) as the input of second BRNN (structure-to-structure BRNN) to get final prediction results [55]. In 2013, Mirabello et al. provided another cascaded BRNN based PSSP method, the first BRNN was used to predict secondary structure from primary sequence and multiple sequence alignments (MSA), and the second BRNN was used to filter the predictions of the first stage [56]. The advantage of the cascaded architecture is that it can take into account the nonlocal information and secondary structure-structure correlations of protein, the cascaded architecture.

However, RNN is also have some limitations to learn the long-term dependencies of protein by its gradient descent algorithms in its training process due to the problem of vanishing gradients [57]. And the error propagation in both forward and backward chains is also subject to exponential decay which prevents BRNN from learning the remote information of protein sequence effectively [59].

For the disadvantages of BRNN, Babaei et al. proposed a modular prediction system for PSSP. The method integrated a multilayer bidirectional recurrent neural network (MBR-NN) which considered the strong correlations of nearby secondary structure elements and multilayer reciprocal recurrent neural networks (MRR-NN) that was to capture the long-range intramolecular interactions of amino acids. MBR-NN was a RNN model with double context layers including the extensive correlations between the secondary structure elements. And MRR-NN dispatched the past and future outputs of both hidden layers to the present state in recurrent links. The characteristic of the combined two models was to arbitrarily engage the neighboring effects of the secondary structure types and memorize the sequential dependencies of amino acids along the protein chain [58].

Inspire by human memory type, J. Chen et al. proposed a segmented-memory recurrent neural network (SMRNN), it would memorize each segment which obtained by breaking the entire sequence, and then each segment would be cascaded to get the final sequence. Their SMRNN based PSSP method could improve the effectiveness for capturing the long-term dependencies in proteins [59]. In recent, Heffernan et al. proposed a long short term memory BRNN (LSTM-BRNN) method for PSSP. BRNNs could capture long range interactions without using a window, which could avoid the shortcomings of sliding window based methods and extract the long-range features in proteins [60]. Both the two improved RNN versions focused on the simulation of human memory form to more effectively capture the long-term dependencies of protein.

Adapting variable width temporal dependencies is the advantage of RNN (a specific memory mode), which could store contextual information by applying dynamic temporal states. RNN relaxed the limit of fixed-size window based approaches to take into account not only the strong correlation between nearby secondary structure elements, but also long-range interaction of amino acids. But it still is a problem that RNN's limited memory lead to it can't effectively learn the long-range interaction information of proteins due to its vanishing gradient phenomenon, which become the main restriction for improving the accuracy of RNN based PSSP methods.

3.1.3 Feedforward neural network

Feedforward neural network (also known as backpropagation neural network, BPNN) is a kind of multilayer neural network which is trained by backpropagation algorithm. BPNN uses gradient descent method to constantly adjust its weights and threshold to minimize the error sum of squares of the network by backpropagation as shown in Fig. 5, and the formula as (13) to (16). The characteristics of BPNN are forward propagation of signal and backpropagation of error.

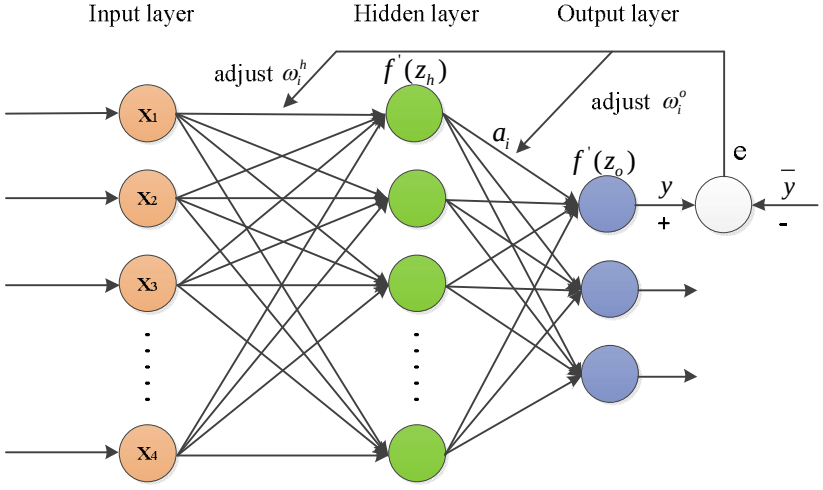


Figure 5. The schematic diagram of BPNN

Generally, error function is expressed as $e=\frac{1}{2}\sum_{i=1}^{n}(\bar{y}-y_i)^2$; and the formula of adjusting weights as $\omega=\omega-\eta\frac{\partial e}{\partial \omega}$.

The weight of output layer:

$$\omega_i^o=\omega_i^o+\eta\delta^o a_i \tag{13}$$

$$\delta^o=-(\bar{y}_i-y_i)f'(z^o) \tag{14}$$

The weight of hidden layer:

$$\omega_i^h=\omega_i^h+\eta\delta^h x_i \tag{15}$$

$$\delta^h=-(\sum_{o=1}^{m}\delta^o\omega^o)f'(z_h) \tag{16}$$

where, y is output and $\bar{y}$ is the expected outputs, δ_o is the partial derivative of error function, η is step size. m represent total number of type for output of PSSP, o and h represent output layer and hidden layer, respectively.

In the past decade, BPNN were generally combined with other models for PSSP, such as, Bidargaddi et al integrated the Bayesian segmentation and a feed-forward BPNN for PSSP [46], and Qu et al proposed several compound pyramid models (CPM) which were a combined architecture of multi-modal BPNN, mixed-modal SVM and knowledge discovery in databases for PSSP [31]. Similarly, Patel et al coupled BPNN with knowledge base model for PSSP [61]. These works will be introduced in section 3.8.

The advantage of BPNN is that its backpropagation of error which has also been increased to other neural network to improve the performance, and it also has strong nonlinear mapping ability. Thus, it can effectively accomplish identification and classification tasks for nonlinear data by approximating a nonlinear function with arbitrary precision. However, BPNN also have two obvious shortcomings. Firstly, it is easy to trap in local optimum so that it may not get the global optimal. Secondly, its convergence process usually takes lots of time with lower learning efficiency mainly due to the using of gradient descent method, therefore many improved BPNN models are proposed for this problem, which is mostly employed momentum method and adaptive adjustment method to replace the gradient descent method to improve its learning speed and reliability.

3.1.4 Radial basis function neural network

Radial basis function neural network (RBFNN) is also a kind of non-linear multilayer feedforward neural network that based on function approximation theory. Based on the principle of Cover's theorem, RBFNN can find the best-fitting plane in high-dimensional space (hidden layer) by the mapping relationship of low dimensional space (input layer) in its training process. Its activation function of hidden layer is radial basis function, where the most typical one is Gaussian form, as shown in Fig. 6, and the activation value of hidden units are activated according to the distance between the input vector and the center of radial basis functions, and the output of RBFNN is linear weighted sum of hidden unit due to the linear mapping from hidden layer to output layer, as (17) and (18). Those are the difference between with other feedforward neural networks. What needs to be pointed out is that it can approximate any continuous function with arbitrary precision, so it is suitable for solving the classification problem of PSSP.

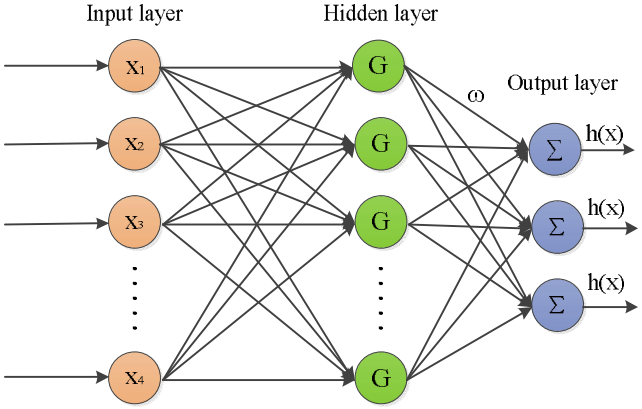


Figure 6. The schematic diagram of RBFNN

Gaussian form is given as follows:

$$f(z) = G_i(z) = \exp\left(\frac{\|z - c_i\|^2}{2\sigma_i^2}\right) \tag{17}$$

$$h(x) = \sum_{i=1}^{l} \omega_i G_i(x) \tag{18}$$

where c_i and σ_i is the center and width of Gaussian function of ith neuron, $\|\ \|^2$ is the Euclidean distance.

Based on 340 protein sequences and their corresponding secondary structures from PDB, Zhang et al. divided 20 amino acids into three groups, as f (Former), b (Breaker) and n (Neutral). The classification information could reflect the inherent attribute of different residues and be used to reconstruct the input vectors of RBFNN to improve the accuracy of PSSP [25]. This method relied on the improvement of input vector to further improve its prediction accuracy.

Some scholars improved traditional RBFNN structure, such as, Jing et al. proposed a two level RBFNN method for PSSP by the model of sequence-to-structure-to-structure. It utilized the output of first layer as the input of second layer and also took into count the evolutionary information of proteins [62] [63].

The RBF neural network can overcome many shortcomings of traditional feedforward neural network (such as BPNN). It can avoid the problem of local minimum due to its only best approximation characteristics. RBFNN is superior to BPNN on the characteristics of good classification ability and approximation performance. And its input-output mapping function and learning speed is better than other feedforward neural networks. But the center of the hidden nodes is difficult to find, this is the key problem of such NN.

3.1.5 Other neural network

Except the above mentioned neural network models, there are some other neural networks based PSSP methods, such as complex-valued neural network(CVNN), denoeux belief neural network (DBNN) and so on [64] [65] [66].

CVNN could process complex data with complex-valued parameters and variables. On the basis of CVNN, Rashid et al introduced two methods for PSSP based on the fully complex-valued relaxation network (FCRN) which was a single hidden layer neural network classifier [67] [68]. For constant random input parameters, FCRN utilized a logarithmic error function to estimate the output weights which are corresponded to the minimum energy point that represents both the magnitude and phase error explicitly [67][68]. A computed energy potentials based feature encoding was used to represent protein residues as its features which were used to train FCRN classifier for PSSP. The characteristics of the two

methods were that the computed energy potentials features could reduce the dependent of the predictions method on training set. When compared to real-valued neurons, the CVNN used a complex plane as its decision boundary which can result in more computational power. [67] [68] [69].

In PSSP study of Zhong et al., they proposed a tertiary classifier based on DBNN which was based on Dempster-Shafer theory and was a multistep neural network [70]. Pthread and OpenMP were employed to parallelize DBNN in the hyper-threading enabled Intel architecture to speed up the training process [71]. Besides, Faraggi et al. proposed another multistep neural network model, called SPINE X, it coupled the prediction of real-value residue solvent accessibility, backbone torsion angles and secondary structure in six steps of iterative way [72]. Every module of the method used the same NN whose weights were proportional to the sequence distance between their corresponding residues in the sliding window. SPINE X had the ability of modeling native distribution of three secondary structures. This work indicated that a more consistent consensus assignment would improve the accuracy of PSSP, which was similar to Ref. [51].

546

Table 3 Summary of neural networks for proteins secondary structure prediction

Method	Ref.	Year	Acc (%)	Dataset	Description
DL	[50]	2015	Q3=80.7%, SOV=74.2%	CASP (CASP9: 105 proteins, CASP10: 93 proteins)	deep learning (belief) network; PSSM by PSI-BLAST
	[51]	2015	Q3=81.8%	CASP11	local backbone angles; PSSM ;physical chemical properties; deep learning neural network
	[52]	2016	Q3=84.7%, SOV=86.5%, Q8=72.3%; Q3=82.3%, SOV=84.8%, Q8=68.3%	CASP; CB513	deep convolutional neural fields; conditional neural fields (CNF); PSSM; http://raptorx.uchicago.edu/download/;
	[53]	2016	Q3=82.9%, Q8= 68.20% Q3= 84.2%,Q8= 73.1%	CB513; CullPDB	PSSM, recurrent neural networks, encoder–decoder networks, bidirectional gated recurrent units
RNN	[59]	2006	Q3=73.1%, SOV= 63.0%	CB396 (RS126 set for training)	bidirectional segmented-memory recurrent neural network; dynamics; multiple alignment profile generated from BLAST
	[55]	2007	Q3=74.38%, SOV= 66.05%	PSIPRED (training dataset :EVA)	cascaded bidirectional recurrent neural networks, long-range interactions; strong correlation; PSSM
	[58]	2010	Q3 = 79.36%, SOV= 70.09%.	PSIPRED	bidirectional recurrent neural network; reciprocal recurrent neural network, long-range interactions; strong correlations; PSSM;
	[56]	2013	Q3=82.2%	1630 proteins of lower quality form PDB	relative solvent accessibility; cascaded architecture; http://distillf.ucd.ie/porterpaleale/
	[60]	2017	Q3=83.9%	TS115	BRNN, PSSM, DSSP
RBF	[25]	2005	Q3=77.4%	340 protein from PDB	conformational classification; structure transition
	[63]	2008	Q3=71.3%	RS126	two level RBFNN; evolutionary information
other neural networks	[71]	2007	Q3=72.01%	CB513 data	PSSM by PSI-BLAST; parallelization; Pthread and OpenMP; BLOSUM62; tertiary classifier
	[64]	2008	Q3=73.4%	RS126	two-stage architecture; fully connected multilayer perceptrons (MLP) neural network; backpropagation algorithm; Sequence Profiles; http://raptorx.uchicago.edu/download/
	[72]	2011	Q3=81.8%; Q3=82.0%	CASP9; SPINE (Sub-dataset of PISCES)	multistep NN model; torsion angle prediction; solvent accessible surface area;
	[67]	2013	Q3=82.14%	CB513	three layered of CVNN; circular transformation; energy potentials; CABS algorithm;

	[68]	2016	Q3=81.72%	CB513	heuristics; Complex-valued relaxation network; inhibitor peptides; compact model; energies computed

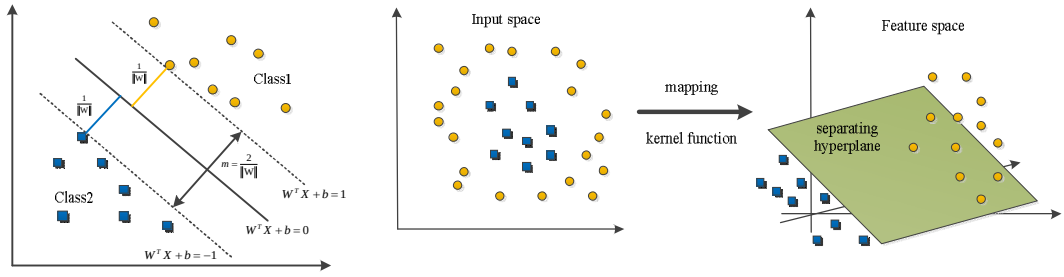
547

NN can not only consider multiple sequence alignment profile as its input vector, but also the local arrangement of adjacent amino acids in protein sequence. The significant advantage of NN for PSSP is its nonlinear mapping ability which can learn and store a lot of mapping relationships between input and output vector by different construction rules in its training process; it can fully approximate the complex nonlinear relationships between protein data and secondary structure. Therefore, it has the advantages of strong learning ability, good flexibility, fault tolerance ability and good self-adaptive. Besides, it can perform large-scale parallel processing and easy to design.

Most NN could not effectively capture the long-range interaction of amino acids in proteins. And almost all NN are regarded as black box model due to the lack of theoretical description; it does not explain the process of how a predict result is reached and why a decision is being made [119]; besides, the selection of hidden layer and hidden layer nodes is difficult and easily cause overtraining or inadequate training. It is so extremely depends upon the training data lead to its poor predict effect for untrained new proteins, and it is also sensitive to input data encoding method lead to its different predict effect. In general, its training process usually takes a great deal of time and samples as well.

3.2 Support vector machines

Support vector machines (SVM) is a kind of representative machine learning algorithm that is proposed on the basis of statistical learning theory and structural risk minimization principle, and it is widely used in pattern recognition field including bioinformatics [18] [112]. It can create an optimal separating hyperplane that separates the data into two classes with maximum margin in the case of linear separable problem, and it also can find out the boundary between non-linear separable classes by using a kernel function which maps the input data from its low-dimensional space into a high-dimensional feature space, as shown in Fig. 7, the formula of linear and non-linear separable as (19) to (24). It makes SVM perform well in PSSP due to the high nonlinearity and complexity characteristics of its structure pattern classification. As a result, SVM becomes one of the most common prediction methods, second only to the neural network. The summary of SVM for PSSP methods are shown in Table 4.



(**a**) (**b**)

Figure 7. The schematic diagram of SVM: (**a**) linear separable; (**b**) non-linear separable

(**a**) Linear separable:

The classification function as $f(x)=W^T x_i+b$.

The forms of optimal solution as follows:

$$\min \frac{1}{2}\|W\|^2 \tag{19}$$

$$\text{Subject to: } y_i(W^T x_i+b)\geq 1,\ y_i=\begin{cases}+1, \text{class1}\\ -1, \text{class2}\end{cases} \tag{20}$$

It can be converted into as follows by introducing Lagrange multiplier:

$$\max_{\alpha} \sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x \tag{21}$$

Subject to:

$$0 \le \alpha_i \le C, i = 1.....n, \sum_{i=1}^{n} \alpha_i y_i = 0 \quad (22)$$

(b) Non-linear separable: the classification function as $f(x) = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b$

The formula of optimal solution as follows:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x) \quad (23)$$

$$\text{Subject to: } 0 \le \alpha_i \le C, \ i = 1.....n, \sum_{i=1}^{n} \alpha_i y_i = 0 \quad (24)$$

where W^T is the weight vector, b_i is bias, α is Lagrange multiplier, $K(x_i, x)$ is the kernel function.

Based on frequent pattern phenomenon of consecutive amino acids in protein, Birzele et al. proposed a representation method to show these frequent patterns which were used by SVM to predict protein secondary structure. The authors considered that these frequent patterns could be located anywhere in protein sequence and may be of arbitrary length [73]. In another research, Shoyaib et al. reported a similar method based on the frequent words in protein sequence and utilized SVM as classifier for PSSP [74]. The advantage of using frequency patterns as the input vector of SVM was that the generated descriptors of arbitrarily parameterized. Chen et al. took into account the pair-coupled amino acid composition (pair occurrence frequency) with its corresponding content of protein secondary structural elements to train a SVM regressing system [75], and the pattern is similar to in Ref. [73][74]. More frequent pattern of consecutive amino acids based methods would be introduced in Section 3.4. But it still had some drawbacks because the frequency of the patterns must be greater than a specific threshold in this method. As a result, it is obvious that the key point of these methods is how to solve the problem of threshold setting by manual and make them be more effectively and rationality.

Karypis et al. presented a method of PSSP by using an exponential kernel function which could capture sequence conservation signals around the local window of each residue and a combined coding scheme which was the integration of position-specific and nonposition-specific information generated by PSI-BLAST and BLOSUM62, named YASSPP, and it was a cascaded models constructed by two stages of SVM-based models whose second stage incorporated both the output structure of the first phase and information about the original sequence to obtained the final result, the second stage was a sequence+structure-to-structure model [76]. Afterwards, Chatterjee et al. used a similar scheme by cascading two multiclass SVM classifiers which was sequence-to-structure-to-structure model, and the results of the first stage SVM were used as input data of the second stage SVM to get the final predictions [77]. The strategy of this method could be considered as a sequence-to-structure-to-structure model, which allowed SVM to explicitly capture the dependencies between amino acid compositions and predicted secondary structure of different positions, and it was able to achieve measurable prediction improvements. But these cascaded modes may have the problem of cumulative error, and it needed a special treat.

In order to improve the accuracy of PSSP, some researchers tried apply more additional features to SVM. Kountouris et al. considered that the backbone dihedral angles of protein had a highly correlated with its secondary structures, which could provide crucial information about local 3-D structure of protein, so this work combined the independently results of dihedral angles and secondary structure prediction in a loop by SVM to improve the predictive accuracy [78]. Besides, some other SVM based PSSP methods were proposed by researchers [41] [79] [80]. Besides, there are some other SVM based PSSP method proposed by researchers, which were introduced in Section 3.8 [21] [112] [118] [129].

Table 4. Summary of SVM for proteins secondary structure prediction

Ref.	Year	Acc. (%)	Dataset	Description
[73]	2006	Q3=75.34%-77%, SOV=70.29%-74.13%	EVA150	frequent amino acid patterns to features; structure to structure layer; level-wise search
[76]	2006	Q3=77.91%, SOV=72.81%; Q3=77.83%, SOV=75.05%; Q3=79.34%, SOV=78.65%	RS126; CB513; EVA(common subset 4)	cascaded models; exponential kernel function; input sequence coding;
[74]	2007	Q3=80.79%; Q3=89.67%	RS126 CB513	frequent patterns; frequent words; 3 amino acid together; radial basis function
[75]	2007	δΘ=0.018 (self-consistency test), 0.050 (seven-fold cross validation)	CB513	pair-coupled amino acid composition; 8-state;
[78]	2009	Q3= 80%; Q3= 79.7%; Q3= 81.7%	CB513; PDB-Select25; EVA	backbone dihedral angles; PSSM; http://comp.chem.nottingham.ac.uk/disspred/
[77]	2011	Q3=75%, SOV=71%	CASP9	physicochemical properties; PSSM; two-stage cascaded classifiers
[41]	2012	Q3= 70.4% SOV =69.5%	RS126	codon encoding; 16-dimensional binary vector

SVM is widely used in PSSP due to its outstanding pattern recognition capacity, which is very suitable for the high-dimensional and nonlinear characteristics of PSSP task. SVM has various kernel functions which can map nonlinear protein data into high-dimensional space, and this operation is performed in low-dimensional space to avoid it suffering from the "curse of dimensionality". Therefore, SVM is suitable for the characteristic of large feature spaces of PSSP; it can find an optimal separating hyperplane for protein structure classification. It also has many attractive features for bioinformatics, including effective avoidance of over-fitting which is a common drawback of supervised learning techniques. Need to point that SVM is originally designed for binary classification, even there are some multiclass SVM based PSSP method, but how to more effectively improve it for multiclass classification of PSSP is still an ongoing research issue [127].

3.3 Probabilistic graphical model

Probability graph model is the combination of probability theory and graph theory, and it uses graph to represent joint probability distribution of variables. The Bayesian networks and hidden Markov models are both directed graph of probability graph model, which are often used in PSSP. The summary of probabilistic graphical model for PSSP methods are shown in. Table 5.

- Bayesian network

Bayesian network is a directed acyclic graph and reflects a series of probabilistic dependency relationships among different variables without considering time factors to the variables. When time is considered as an additional factor of Bayesian network, it would become dynamic Bayesian network (DBN) which can reflect time-varying and probabilistic dependency relations of variables through its topology structure.

Because most probabilistic methods utilize single sequence for PSSP, therefore, Li et al. presented a Bayesian model based on knob-socket model of protein packing in secondary structure for PSSP that considered the packing influence of residues on secondary structure, including those packed which are close in space but distant in sequence [81]. And knob-socket model was employed to provide constructs for the direct inclusion and prediction of the secondary states of coil and turn. This research showed that the incorporation of multiple sequence alignment data would improve the accuracy of PSSP. Besides, Yao et al. also proposed a method based on DBN; it used a multivariate Gaussian distribution to generate PSI-BLAST profile of a protein sequence, which takes into account the correlation between entries of the PSSM [26].

- Hidden Markov Models

As a special Bayesian networks, Hidden Markov models (HMM) is a statistical Markov model, which is a practical tool for sequence analysis and have been successfully implemented in PSSP. In a HMM, the states cannot be observed directly, but each state of this model emits a single observation from a set of observations by a corresponding probability density distribution, as shown in Fig. 8; these states are usually three secondary structures in protein sequences. In general, the observations of HMMs are assumed to be amino acids [82] [84].

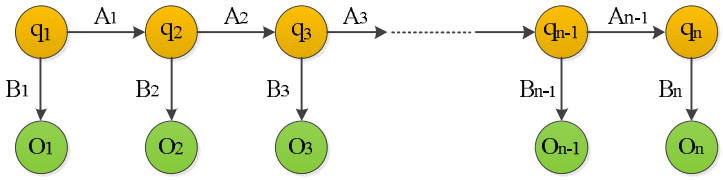


Figure 8. The schematic diagram of HMM

HMM is determined by the pair (π, A, B), the basic algorithm of HMM is forward-backward algorithm, Viterbi algorithm, Baum-welch algorithm respectively according to the different situation.

$$\pi = (\pi_i); \ \pi_i = P(q_0 = S_i), \ i = 1.....\mathrm{N} \tag{25}$$

where π is the initial distribution of the state:

$A = \{a_{ij}\}$ is the transition probabilities matrix:

$$a_{ij} = P\left[q_{t+1} = S_j \mid q_t = S_i\right] \tag{26}$$

$B = \{b_i(k)\}$ is the Emission probability matrix:

$$b_i(k) = P\left[O_t = o_k \mid q_t = S_j\right] \tag{27}$$

where $S = \{S_1, S_2, ..., \mathrm{S}_N\}$ is the set of states, and q_t represent as the state of time t; $O = \{o_1, o_2, ..., o_m\}$ is the set of limited number of output symbols.

Segmental semi Markov model (SSMM) is a generalization of HMM where a hidden state generates segments of various length and secondary structure types. In this model, a segment of observations is emitted from a single state. One advantage of SSMM is that it could incorporate varied sources of protein sequence information by using a joint sequence-structure probability distribution based on structural segments [83]. By the incorporation of segmental semi-Markov model (SSMM) and multiple sequence alignment profiles, Chu et al. proposed a PSSP method which based the notion that unrelated proteins may share the similar structure in a collection of local structural fragments or segments. They proposed a parameterized model for the likelihood function which could explicitly represent multiple sequence alignment profiles to capture segmental conformation, and presented a definition of beta sheet contact space which can effectively capture the long-range information under the realistic constraint [83].

Many implementations of SSMMs only consider the information on the left hand side of an amino acid, and the right side always was neglected. Therefore, Malekpour et al presented a SSMM based PSSP method which took into account the information of both sides of amino acids. The method could capture the correlation between the observations in dependency windows by broke it into shorter dependencies to estimate the probability of the segments in structural classes, and used a weighted model which was based on these conditional probabilities to calculate the probability of each segment in a structure [84]. The advantage of this method was that the parameters number do not exponentially increase when the length of the dependency window increase. In other paper of Malekpour et al., they proposed a two-sided modified (Modified Bidirectional) HMM to consider the dependencies among the emissions in the protein sequence; besides, a modified Viterbi algorithm and a posterior decoding algorithm were used to find the most probable state sequence [85]. Their researches proved that the information on both sides of a residue in a protein sequence could improve the accuracy of HMM based PSSP methods.

Hidden semi-Markov model (HSMM) was initially considered in Bayesian segmentation of protein secondary structure (BSPSS) algorithm [86], afterwards, Aydin et al. introduced a PSSP model for single sequence by refining and extending HSSM [87]. They proposed an improved residue dependency model by considering the statistical patterns of amino acid segment, and performed a statistical analysis identifying the correlations between sequence and structure to build the architecture of HSMM. This model specifically considered the correlations at proximal positions of structural segments and dependencies to upstream and downstream residues. Besides, they implemented an iterative training method to refine estimates of HSMM parameters. On the basis of Ref. [87], Aydin et al. proposed another PSSP method based on the N-best paradigm where a set of suboptimal segmentations (N-best list) was computed as an alternative to the most likely segmentation by modified stack decoder and N-best Viterbi algorithms [88]. Both the two algorithms employed a HSMM [87], and used Viterbi scoring to compute the N-best list. And the final prediction results could be computed by a weighted voting scheme which applied to a selected set of top scoring results. This method not only extracted the local correlation information of proteins in a relatively simple way, but also contained some long-range interaction information. However, N-best method did not so effectively extract long-range interaction information, which may be further improved.

Table 5. Summary of probabilistic graphical model for proteins secondary structure prediction

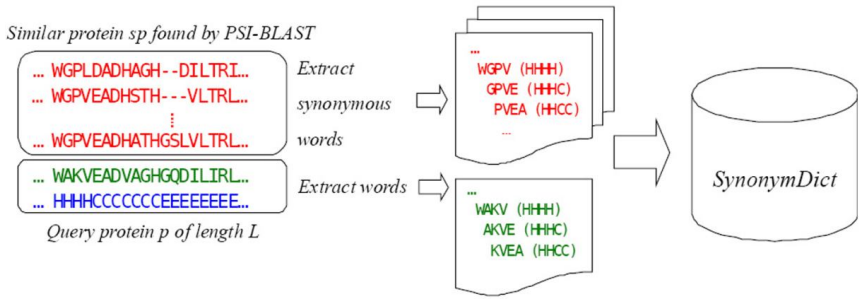
Ref.	Year	Acc. (%)	Dataset	Description
[83]	2006	Q3=72.23%	CB513	Bayesian segmental semi-Markov models; multiple sequence alignment profiles; parametric models; long range correlation; http://public.kgi.edu/~wild/bsm.html
[87]	2006	Q3=70.300%; Q3=67.899%	EVA; CASP6	single-sequence algorithms; modified bidirectional hidden Markov model
[88]	2007	Q3=63.71%; Q3=65.80%	CB513; EVA	single-sequence algorithms; HSMM, BSPSS algorithm; N-best list; suboptimal segmentations
[84]	2009	Q3=66.48%	CB513	left-to-right and right-to-left dependencies; segmentation; HMM
[85]	2010	Q3=62%	1342 protein from DSSP	posterior decoding; Viterbi algorithm; left-to-right and right-to-left dependency model
[81]	2014	Q3=74%; Q3=88%	CASP9; ASTRAL30	knob-socket model; long range correlation; multiple sequence alignment; http://bamboo.byu.edu

In contrast to NN and SVM, probabilistic graphical methods has more explicit theoretical basis and computational process, so that their mechanisms is more easy to understand, such as specific correlation structure between neighboring residues [26]. And it could be used to capture the distant relations of protein sequence by a joint sequence-structure probability distribution based on structural segments [46]. However, most of probabilistic graphical methods are designed to deal with single sequence without homologous information [87] [26], and it is very difficult to incorporate the evolutionary information of different protein sequences lead to their limitation on prediction precision of PSSP.

3.4 Amino acid composition statistics (statistical dictionaries) based methods

It is well known that secondary structure continuously appear in protein sequences, therefore amino acid composition statistics (statistical dictionaries) based methods mainly focus on the statistical characteristic of the certain length of continuous amino acid compositions; furtherly, the features of these amino acid compositions would be found by someway; and then protein secondary structures would be predicted according to the found features. Some frequent pattern of consecutive amino acids and SVM based methods have been introduced in Section 3.2. The summary of statistical dictionaries model for PSSP is shown in Table 6.

Lin et al. presented an improved dictionary-based PSSP method on the basis of PROSP [89] called SymPred; and a meta-predictor called SymPsiPred also was proposed by combing SymPred and PSIPRED [90]. In their methods, the synonymous words of natural language processing field were adopted to capture local sequence similarities in a group of similar proteins. For PSSP, a protein-dependent synonymous dictionary was generated [91], the procedure as shown in Fig. 9.



738

739 **Figure 9.** The procedure used to extract protein words and synonymous words for a query protein [91]

740 Based on the frequency dictionaries of amino acid, Popov et al. first calculated the total observed
741 frequency of every amino acid and determined the most probable secondary structure at every position
742 by calculating a score for every potential secondary structure; then six separate predictions for each
743 sequence was created by six structural class dictionaries; at last, the obviously wrong or very low
744 probability elements were cleared [92]. The advantage of this method was that it took into account six
745 possible readings of amino acid sequence which were common protein secondary structure. And the
746 method maintained a low computational profile and independence from sequence data bases.
747 In the study of Feng et al., the frequency features of the tetrapeptide were calculated by a sliding
748 window of four residues along protein sequences; and then a binomial distribution based method was
749 used to filter noise and unnecessary features to find tetrapeptide structural words with high confidence
750 level; at last, the predictions were performed by the increment of diversity combined with quadratic
751 discriminant analysis method, named IDQD [23]. In another paper, Feng proposed a similar PSSP
752 method based on the frequency features of tetrapeptide structural words combined with long-range
753 contact number that assumed two residues to be in long-range contact if the separation of two residues
754 exceeded more than 10 residues in protein sequences [93].

755 **Table 6.** Summary of statistical dictionaries model for proteins secondary structure prediction

Ref.	Year	Acc. (%)	Dataset	Description
[91]	2010	Q3=81.0% (SymPred) Q3=83.9% (SymPsiPred)	DsspNr-25	dictionary-based; language processing techniques; synonymous words; http://bio-cluster.iis.sinica.edu.tw/SymPred/
[92]	2010	Q3=56.72%	EVA	amino acid frequencies; six structural class dictionaries
[23]	2014	Q3=87.8 %	PISCES	tetrapeptide structural words, tetrapeptide frequency, feature selection, increment of diversity combined with quadratic discriminant analysis
[93]	2014	Q3=83%	194 proteins from CB513	long-range contact number, tetrapeptide structural words, quadratic discriminant analysis, increment of diversity

756 The above mentioned works proved that the frequency dictionaries based approaches were
757 applicable for PSSP, and it was useful for novel sequences where homology based approaches have less
758 or nothing to work with. The performance of this kind of methods obviously rely on the extractive
759 features of the small peptides fragments or small continuous amino acid fragments; therefore, these
760 methods could effectively find the local features of proteins; however, it would be difficult to find out
761 long-range interaction information. And the fixed length composition statistical method will lose the
762 information of very short composition or long composition, because the length of the continuous
763 secondary structure of the amino acid are unfixed and polytropic. The key problems of these methods are
764 how to effectively set threshold value and extract local secondary structure information for PSSP.

765 *3.5 Fuzzy logic*

Fuzzy logic systems could be used to encode human reasoning into a program to make decisions, which contains five functional components: fuzzifier, inference, defuzzifier, fuzzy set, and fuzzy rule, as shown in Fig. 10 [94]. It can be used to represent affiliation relationship with indefinite boundaries between element and set by using membership function of fuzzy set that can quantify the element attributes by a scalar in range of [0, 1] [95], as (28) and (29) [96] [97] [98]. Fuzzy inference applies the satisfied fuzzy rules to map the primary fuzzified features to other secondary fuzzy features, and it can get the fuzzy results by definitively mathematical computation process according to fuzzy rules. Fuzzy logic can be widely used in information incomplete or imprecise situation so that it also can effectively process the data in PSSP [45] [99] [100]. The summary of fuzzy logic model for PSSP is shown in Table 7.

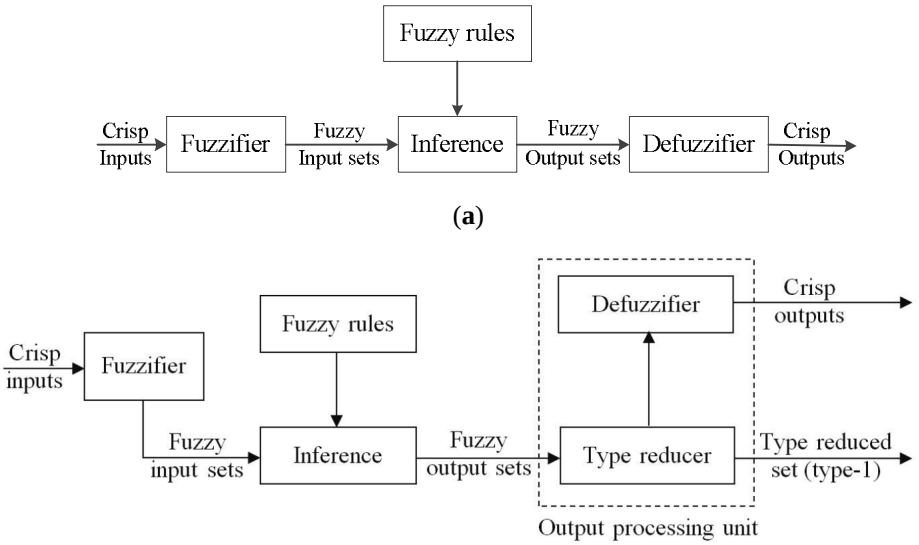


Figure 10. The structure of Fuzzy logic systems: (**a**) type-1 FLS; (**b**) type-2 FLS [101]

A fuzzy set usually can be defined as

$$A = \{(x, \mu A(x) \mid x \in U\} \tag{28}$$

where $U = \{x_1, x_2, ..., x_n\}$, $0 \le \mu A(x) \le 1$, $\mu A(x)$ is the membership function of x in the fuzzy set A; and the sum of membership functions for x is always 1, which can be defined as follows

$$\sum_{i=1}^{n} uA_i(x) = 1 \tag{29}$$

where n represent the number of the fuzzy set which x belonging.

The type-1 fuzzy logic system (Type-1 FLS) utilized membership to represent the uncertainty of affiliation of an element and a set. Krishnaji et al. applied Type-1 FLS combined genetic algorithms and neural networks for PSSP. It utilized fuzzy C-Means clustering to reduce the inputs number of attributes which was protein structural class, solvent accessibility and physicochemical properties. And a fuzzy inference engine was used to convert fuzzy predictions to actions according to fuzzy rule base [94]. The type-2 fuzzy logic system (Type-2 FLS) was the extension of Type-1 FLS, which described the uncertainty of the membership function in Type-1 FLS by three dimensional membership functions, and it can more efficiently handle uncertainties. Nguyen et al. proposed a multi-output interval type-2 fuzzy logic system (MOIT2FLS) for PSSP and this was the first time that Type-2 FLS was used in this field. Quantitative properties of amino acids were first employed to characterize twenty amino acids, which could be used as the input of MOIT2FLS.Three clustered outputs of the MOIT2FLS were assigned to

three protein secondary structures by its adaptive vector quantization (AVQ) method. Genetic algorithm was applied to optimize parameters of MOIT2FLS [101].

The k-nearest neighbor (K-NN) method was the simplest algorithm for pattern classification and was easily adapted for parallel computation; however, it has some restrictions in PSSP. Fuzzy theory was combined with K-NN method to provide more classified information than a definite prediction of the query secondary structure. Kim proposed a parallel algorithm for PSSP based on fuzzy k-nearest neighbor (FK-NN) method, which used the evolutionary profile obtained from PSI-BLAST as its input feature vectors [102]. Bondugula et al. also combined FK-NN with neural network for PSSP, which would be introduced in section 3.8 [114]. Afterwards, Ghosh et al. made an attempt in PSSP by using three low cost pattern classification techniques as minimum distance, K-NN and fuzzy k-nearest neighbor classifiers, and the FK-NN can achieve better performance with window size 3 [103].

Table 7. Summary of fuzzy logic model for proteins secondary structure prediction

Ref.	Year	Acc. (%)	Dataset	Description
[102]	2006	Q3=71.8%, SOV= 67.9%; Q3=89%, SOV= 84%	EVA; RS126	parallel algorithm; fuzzy k-nearest neighbor method
[103]	2008	MCC=0.0753	RS126	fuzzy K-nearest neighbor;
[94]	2012	-	-	artificial neural networks; genetic algorithms; Type-1 FLS
[101]	2015	Q3= 68.15%	62 proteins	Type-2 FLS; adaptive vector quantization; genetic algorithm;

Fuzzy logic has the ability to handle the imprecise and vague data, and this characteristic make it be superior to other model for PSSP due to the complexity of protein structure predication. It also can estimate the probability of the affiliation between the amino acid and secondary structural class, which is the lack of other models.

3.7 Other methods

In addition to the above mentioned models and their improved versions, there also are some methods were used for PSSP, such as RICO, Prote2S, K-nearest and LMNN. The summary of other methods for PSSP is shown in Table 8.

Based on rule induction from coverings (RICO) method [104], Lee et al. proposed a relaxed threshold RICO (RT-RICO) model by using a rule-based method [105]. RICO utilized some of concepts form rough sets, and it was a classification scheme based on partitions of entities in a protein data set. RT-RICO could identify the dependencies between amino acids in protein sequences and generate rules for PSSP. But the two algorithms had the shortages, including high computational complexity and long program running time. In another paper of the authors, a slightly improved RT-RICO approach was proposed by parallelization in order to acquire faster running speed [106].

In order to more effectively exploit the information in protein databases and deliver ever-improving prediction accuracy as the databases expand rapidly, Chang et al. proposed a Prote2S predictor with a kernel density estimation algorithm [107], and this method had a unique advantage with its low time complexity in the training process.

Yüksektepe et al. proposed a two-stage method to predict the location of secondary structure elements in a protein using primary structure. The first stage employed hyper-boxes approach to determine the protein folding type, and the second stage utilized the conditional probabilities that were calculated for each folding type separately to predict three-state probabilities [108]. In this method, the hyper-boxes approach can define the boundaries of protein sets to improve the performance of multi-class secondary structure classification in folding type prediction of protein.

Yang et al. proposed a K-nearest neighbor model which based large margin nearest neighbour (LMNN). This model utilized supervised distance metric learning (SDML) technique to obtain a linear transform matrix by LMNN, and energy-based classification rule was used to alleviate the influence of unbalanced class distribution. It was also the first time to apply the SDML into the field of PSSP [109].

Table 8. Summary of other methods for proteins secondary structure prediction

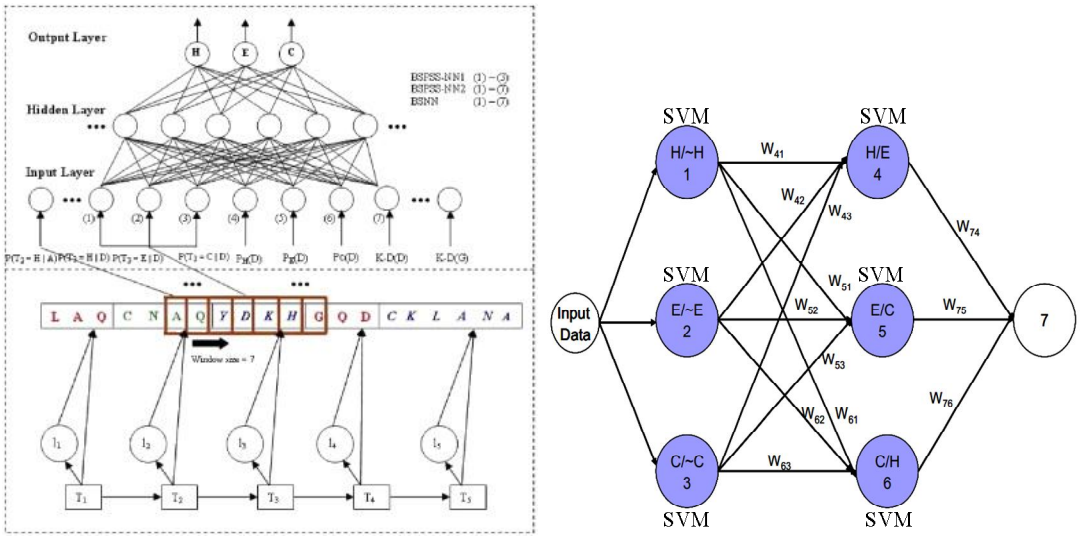
Ref.	Year	Acc. (%)	Dataset	Description
[107]	2008	Q3=80.3%, SOV= 76.9%	EVA	kernel density estimation algorithm; Prote2S predictor
[108]	2008	Q3=74.1%	PDB	SCOP, two-stage algorithm; mixed-integer linear program; folding type; probabilistic search algorithm
[105]	2009	Q3=80.3%	SCOP	RT-RICO; data mining
[106]	2009	Q3=74.6%	CB396	RT-RICO; data mining; a rule-based method; parallelized implementation
[109]	2013	Q3=75.09%, SOV=72.18%; Q3=75.44%, SOV= 75.66%	RS126; CB513	K-nearest neighbor; large margin nearest neighbor; energy-based rule; supervised distance metric learning

3.8 Hybrid method

In the last decade, the prediction accuracy of protein secondary structure has gained some improvements which largely due to the new advances of machine learning tools and the hybrid models in different prediction methods [26]. Because more and more scholars recognize that the performance of single model often have its limitations; therefore, the hybrid model which is the combined predictors of different machine learning tools are becoming a new trend in recent years due to the shortcomings of single model would be remedied by other models that can extract protein features from different perspectives [125]. Especially, the introduction of ensemble prediction model further improves the accuracy of PSSP by combining the results from different predictors [118]. Another significant trend is that multi-step process based hybrid methods for PSSP. The prediction process often composes of protein structural features extraction and secondary structure prediction; the former mainly focus on effectively extracting features of protein data by advanced and suitable methods; and the latter adopts appropriate method to predict secondary structure based on advanced machine learning models and classifiers (clusters) [137].

3.8.1 The combination of multiple methods

Different models have different natures. Each of them has some disadvantages, such as probabilistic graphical model is difficult to consider evolutionary information of proteins, and neural network is not good at capture distant interactions among residues and has unapprehensive mechanism [46]. Thus, the combination of these two kinds of models can complement the shortcomings of each other, and provide possibilities to get better prediction accuracy than any of individual [26]. The hybrid form is divided into two kinds: one is cascaded architecture that the output of first stage is the input of second stage, such as Ref. [46]; the other one is cross structure, such as Ref. [112], SVM as the neurons in neural network, as shown in Fig 11. In these multiple methods, the neural network is the most popular model.



(**a**) (**b**)

Figure 11. The structure of multiple methods: (**a**) cascaded architecture [46]; (**b**) cross structure [112]

1. Neural networks-probabilistic graphical models

Bidargaddi et al. proposed three variants of hybrid model for PSSP, which combined the Bayesian segmentation based on SSMM as the first stage to capture non-local interactions through a joint sequence-structure probability distribution, and neural networks was used as the second stage to capture local correlation by a local window of residues. A sliding window of residues with segmentation scores and physical-chemical properties was provided as inputs of neural network. At last, neural network ensembles based on the posterior probability scores of the structural states of each residue sample were used for training. The method was suitable for single sequences without homogeneous information of multiple sequences [46].

It was difficulty for probability graph model to consider evolutionary information; therefore, probability graph model usually was combined with neural networks which often used as input layer for considering evolutionary information. Malekpour et al. proposed a hybrid model based on three neural networks as first stage and SSMM as second stage. The multiple sequence alignment profile was used as the inputs of the three neural networks to get four types of outputs, and SSMM was used as a decision function for final prediction results [110]. Zhang et al. proposed method for PSSP, called MD-OAO+Bayesian, which were consisted of three neural networks as first stage and a Bayesian model as second stage. And the output of the three neural networks were used as the input of Bayesian model to obtain the final prediction results [111]. Yao et al. also proposed a DBNN model which was combined by DBN and neural networks (the typical three-layered feed-forward back-propagation) for PSSP [26].

2. Neural networks-SVM

Zhang et al. proposed a hybrid PSSP method which utilized sliding multi-window scheme to form input data for SVM. And the method included two different tertiary classifiers; the first classifier employed SVM as the neurons in neural network architecture which optimized by genetic algorithms; the other was a granular decision tree based on granular computing, decision tree and SVM [112]. Afterwards, Ghanty et al. described two methods for PSSP, NSVMps and NSVM, both them were based on Neuro-SVM (two cascaded modules: MLP neural networks for feature extraction module and SVM for classification module). The two methods utilized position-specific probability-based features and position-independent probability-based features, respectively. The proposed method employed single-sequence as its turning data and did not rely on any sequence alignment profile information [113].

3. Neural networks-other

In addition to above mentioned neural network based hybrid models, neural network also is combined with other methods, such as fuzzy k-nearest neighbor algorithm, knowledge base, cascaded nonlinear components analysis and conditional random fields.

- Neural network - fuzzy k-nearest neighbor algorithm

Nearest neighbor method could predict a target protein secondary structure using local sequence similarity of the segments in known proteins. It would be successful when the similar sequences could be found in PDB, but it would be limited if there were no well-defined sequence profile for the query protein. Fortunately, sequence profile based methods performed well if the query protein had many similar sequences in sequence database by building a good PSSM, such as neural networks. Therefore, Bondugula et al presented a MUPRED system that used two types of features: the first was membership values of each residue in the 3-state structure classes generated by fuzzy k-nearest neighbor algorithm, and the second was normalized PSSM. These all were the input of feed-forward neural network which is trained by standard back-propagation learning method [114]. This method combined the advantages of the two methods and had a better potential to use the information in both the sequence and structure databases.

- Neural network-conditional random fields

Compared to 3-states PSSP, 8-states prediction received fewer attentions; however, it was also a challenged problem, especially when there were few homologous sequences. Based on probabilistic model, Wang et al. proposed conditional neural fields (CNFs) based method for 8-states PSSP. This model was a combination of conditional random fields (CRF) and NNs [115]. Similar to HMM, CRFs could extract the features of the interdependency among adjacent secondary structures; and NNs could extract the nonlinear features between protein sequence and its corresponding secondary structure. By combing the advantages of both NNs and CRFs, the proposed method could take into account the complex relationship between sequence features and secondary structures, and the interdependency among adjacent residues. Besides, the method could also provide a probability distribution over all the possible secondary structures types and non-evolutionary information of proteins. Recently, two deep learning based models also were used to predict protein 8-state secondary structure, which have been introduced in section 3.1 [52] [53].

- Neural network-knowledge base

A hybrid algorithm for PSSP was proposed by Patel et al., named KB-PROSSP-NN which combined knowledge based method (KB-PROSSP) and neural network. KB-PROSSP-NN had two phases: the first phase was KB-PROSSP that was a hierarchical lateral-validation technique built from the knowledge base which consisted of the statistics of association between the 5-residue words and corresponding secondary structures; the second phase was a pre-trained feedforward neural network (BPNN) which was used to correct the discrepancies of the knowledge base, and the neural network was trained by the output of KB-PROSSP and the actual secondary structure to predict protein secondary structure [61].

- Neural network-cascaded nonlinear components analysis

Due to the high complexity of PSSP, the effective pre-processing can contribute enough for the improvement of accuracy for predictors systems. Principal components analysis (PCA) was an algorithm that can effectively provide to statistical analysis for data set. So Botelho et al utilized cascaded nonlinear components analysis (C-NLPCA) method which was concerned with the utilization of the nonlinear potential of neural networks to reduce dimensionality of protein data and acquire useful information for classification phase. The pre-processing protein date was used as input of three neural networks with different topologies. The neural networks were trained by resilient propagation method and independently found classification results which were combined for the attainment of better results [116].

4. Other hybrid method

Except the above mentioned hybrid models, there are some other integrated methods, such as k-mers and CRF, SVM-decision tree and knowledge discovery in databases model. The summary of the combination of multiple methods for PSSP is shown in Table 9.

- Knowledge discovery in databases model

For non-trivial problems, such as PSSP, the general single-method models and simple combinations of prediction models may not obtain satisfactory prediction results. Fortunately, the data mining technologies receive a good development in recent years and have a good application prospect in PSSP. Therefore, Yang et al proposed several compound pyramid models (CPM) for PSSP based on data mining

and machine learning approaches. The CPM is composed of several layers which work in close coordination, and it adopts a gradually refining, multi-hierarchical configuration, in which the layers focus on independent functions [118], one of the CPM is shown in Fig. 12. Knowledge discovery in databases (KDD*) process model is the core technology in CPM, and its most essential mechanisms are heuristic coordinator and maintaining coordinator [117]. The heuristic coordinator simulates the "intention creation" in cognitive psychology, so that the shortage of knowledge could be detected by the system itself [118].

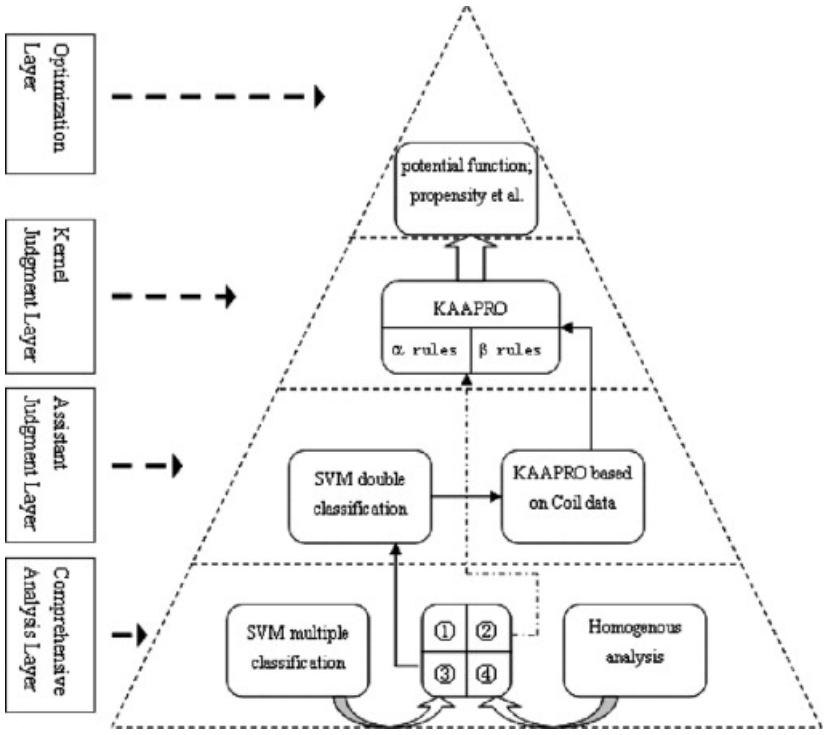


Figure 12. The compound pyramid model [118]

In 2009, Yang et al. proposed a CPM which was a four layers systematic prediction model for PSSP. In this CPM, KDD* association analysis protein secondary structure prediction (KAAPRO) was proposed based on Maradbcm algorithm and D-CBA algorithm; the former was induced by KDD* model; the latter was derived from a classical CBA algorithm that was based on complex measure. KAAPRO worked as the kernel in CMP, and SVM was used in two layers of CMP [118].

One year after Ref. [118], Yang et al. presented another CPM model for PSSP based on the concept of associate classification algorithm (SAC). The knowledge discovery theory based on inner cognitive mechanism (KDTICM) theory and Maradbcm algorithm were used in association rules mining process. The knowledge base of protein secondary structure was established with three kinds of protein secondary structure. The classification process applied a classic classification that was based on multiple association rules (CMAR) to determine which secondary structure a given window was more likely to be through the adjacent information on amino acid sequential window and screening of three different rule sets. The bottom level of CPM was homologous analysis and multi-SVM, and the higher level was SAC that used to pass judgments about the windows over the bottom level and then put the results into application [21].

Afterwards, Yang et al. proposed another CPM based PSSP method by multi-modal backpropagation neural network (MMBP). CPM was constructed by using a KDTICM method, and this CPM was composed of three layers of intelligent interface, as MMBP, mixed-modal SVM (MMS), KDD* process and so on. And four most commonly used multiple sequence alignment methods (PSI-BLAST, PSI-Search, HMMER3 and AMPS) were used to construct four profiles together which was used as input data for MMBP module. In this CPM, MMS and MMBP were used in the bottom layer which as comprehensive analysis layer, and the SAC module was used in second layer which as kernel

judgment layer, attribute association classifier (AAC) module was used in top layer as assistant judgment layer[31].

These methods could be regarded as the complicated combinations of knowledge discovery techniques and machine learning models to perform more predicted accuracy than single-method models and simple combinations of models. These methods combined the advantages of several approaches by the concept of data mining, so that they got the higher prediction accuracy comparatively. Even these methods are more complex and need more computation power relatively, they still may become a trend due to its innovative thinking of data mining and good performance.

- SVM-decision tree

SVM have shown strong power for PSSP; however, they did not produce comprehensible models that account for the predictions they made, which affected the prediction accuracy of SVM. Complementary to SVM, decision tree had good comprehensibility. As a result, He et al. presented an approach, named SVM_DT which combined the strong generalization ability of SVM and the strong comprehensibility of rule induction of decision tree. This method employed SVM as a preprocess of decision tree and used a encoding scheme by combining orthogonal matrix and BLOSUM62 matrix to train SVM, and the output of SVM to train decision tree learning system, the corresponding rule sets was also extracted [119]. The comprehensibility of SVM_DT was better than that of the SVM.

- k-mers - conditional random field (CRF)

Madera et al. proposed a general framework based on k-mers and CRF which could take into account short range, medium range and longer range interactions. The occurred frequency of each individual amino acid (1-mer), each possible pair (2-mer) or every combination of up to k amino acids (k-mer) could be measured from a real sequence. Markov chains and k-mer model were used to the problem of generating realistic emissions from secondary structure profiles. Especially, the longer range interactions in the pterion were described as a k-mer order model. For each protein sequences, the method would generate two local structure profiles by PREDICT-2ND neural networks [120]: one described the amino acid at each position of the target sequence, and another from the alignment; these local structure profiles were regarded as single-sequence, and alignment were input to the k-mer model for PSSP, respectively [121].

1009 **Table 9.** Summary of the combination of multiple methods for proteins secondary structure prediction

Method	Ref.	Year	Acc. (%)	Dataset	Description
Neural networks-probabilistic graphical model	[111]	2005	Q3=75.8%; Q3=74.8%	PDB; CB513	neural networks; Bayesian; one-against-one modeling;
	[110]	2009	Q3=75.35%	EVA	SSMM; neural networks; multi-class pattern classification
	[26]	2008	Q3= 78.1%, SOV = 74.0%; Q3= 78.8%, SOV = 74.8%; Q3= 80%, SOV = 78.1%	CB513; EVAc6; SD576	DBN; multivariate Gaussian distribution; segment length distribution; neural networks; PSI-BLAST profile
	[46]	2009	Q3=71.23%; Q3=70.89%	PDB_SELECT; CB513	SSMM, single sequence; feed-forward backpropagation neural network; graphical models
Neural networks-SVM	[112]	2009	Q3= 66.7%	RS126	tertiary classifier; granular decision tree; neural networks; SVM
	[113]	2013	Q3=71.5%; 68.3%; 68.1%	RS126; CB513; CASP9	probability-based features, SVM, neural networks, single-sequence
SVM-decision tree	[119]	2006	overall average error 3.3%	RS126	decision tree, rule extraction, SVM
Neural network and fuzzy k-nearest neighbor algorithm	[114]	2007	Q3=80.14%	March 2006 release of PDBSelect17 database	fuzzy nearest neighbor; neural network; hybrid prediction system; sequence profile; template; PSSM; http://digbio.missouri.edu/mupred.
Neural network-knowledge base	[61]	2014	Q3=90.16%; Q3=82.28%	RS126; CB396	5-residue words; knowledge base; lateral association and validation; hierarchical validation; feedforward neural network (BPNN)
Neural network and cascaded nonlinear components analysis	[116]	2006	Q3=76.1%	CB396	cascaded nonlinear components analysis ; dimensional reductio ; matrix of scores PSSM; principal components analysis
Neural network and conditional random fields	[115]	2011	Q8=64.9; Q8=64.7%, Q8=81.17%	CB513; RS126; CullPDB;	conditional random field; neural network; 8-state, PSSM; PSIPRED package
k-mers and conditional random field	[121]	2010	SOV=83% Q3=77.4%	1763 protein from PISCES server	neural network; dynamic conditional random field (CRF). k-mers; longer range interactions; http://supfam.cs.bris.ac.uk/kmer
Knowledge discovery in databases model	[118]	2009	Q3=86.18%, 94.34%, 92.68%, 79.25%	256B, 351C, 9PAP, 1BP2.	association analysis; KDD*, CBA, CMP, ensemble prediction model; PSSP

	[21]	2010	Q3=83.06%; Q3=80.49%	RS126; CB513	KDD*; SAC; CMAR; multi-SVM; CMP ; KDTICM; Association analysis;
	[31]	2011	Q3=86.13%, SOV99=84.66% Q3=83.99%, SOV99=80.25% Q3=85.58%, SOV99=81.15%	256 proteins from RCASP256; RS126; CB513	multi-modal BP; CMP; physicochemical properties; multiple sequence alignment profiles ; KDTICM http://kdd.ustb.edu.cn/protein_Web/.

1010

3.8.2 Optimization based methods

In the last decade, with the development of optimization theories, it has been used to solve the nonlinear and complex optimum problem of PSSP as well. Optimization algorithms are suitable for global optimization which usually maximize or minimize a multivariable function by satisfying some constraints of equation or inequality to find some optimal solutions or approximate optimal solutions in theory. At present some popular optimization algorithms are used in PSSP, such as genetic algorithm (GA) [94] [101] [112] [125], genetic programming (GP) [122], bee colony algorithm [123], and the most widely used method is GA. GA belongs to the larger class of evolutionary algorithms (EA), which is inspired by the process of natural selection. It commonly used to generate high-quality solutions by relying on bio-inspired operators such as mutation, crossover and selection [124], an example of optimization based method is shown in Fig. 13. GA often used to optimize the parameters of models to achieve better predicted results in PSSP. The summary of optimization based methods for PSSP is shown in Table 10.

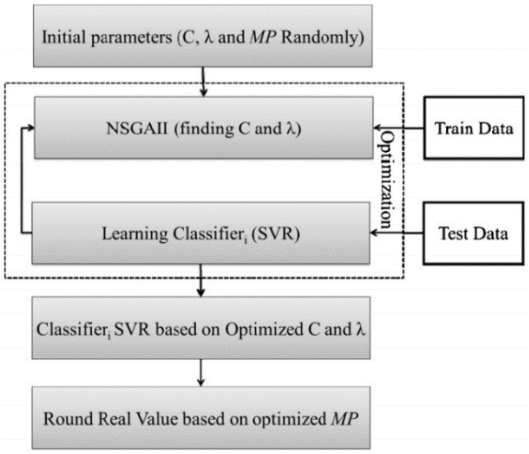


Figure 13. The block diagram in Ref. [127]

Won et al. developed a PSSP method based on HMMs whose structure automatically optimized by GA [125]. During the evolutionary optimization process, the HMM's structure was assembled from biologically meaningful building blocks. GA and the standard HMM estimation algorithm was used to choose model parameters of such Block-HMMs. The optimized HMM would capture the features of protein sequence and structure. The advantages of this method was that the structure of HMM was automatically evolved without prior knowledge and could get good results with a smaller number of states.

Green et al proposed a PSSP method based on parallel cascade identification (PCI) which was a dynamic nonlinear system identification system that was created by a black-box approach to model the process of protein folding. The method was two layers of sequence-to-structure-to-structure PCI classifier which was built by three binary ways in each layer, and GA was used to optimize the parameters of the PCI models. The PSSM form PSI-BLAST as input data [126].

Based on PSSM profiles, Zangooei et al. proposed a PSSP method based on SVR and non-dominated sorting genetic algorithm-II (NSGAII). Since the major challenge for applying SVR was how to tune and set its parameters for a given dataset. NSGAII was used to find mapping points for rounding a real-value to an integer one and also used to optimize SVR kernel parameters to achieve better results. Due to there may be more than one kernel that can get acceptable predict output, they also proposed a dynamic weighted kernel fusion method for fusing of three SVR kernels to get good performance [127].

The local interaction of individual residues define the global protein conformation; therefore, Chopra et al. utilized cellular automata (CA) to simulate global phenomena by localized interactions for PSSP. In this study, protein sequence was used as the input data of CA, and the rules for updating states were optimized by GA [128].

Table 10. Summary of optimization based methods for proteins secondary structure prediction

Ref.	Year	Acc. (%)	Dataset	Description
[128]	2007	Q3=58.21%; Q3=56.51%	RS126; CB513	protein folding；evolved cellular automata；genetic algorithms；cellular automata
[125]	2007	SOV=70.5% Q3=75.0%	SABMark Twilight Zone data set(consists of 2230 high quality structures)	GA; Block-HMM; homologous sequences by PSI-BLAST; http://www.binf.ku.dk/~won/pshmm.htm
[126]	2009	Q3=73%	EVA test set	GA; parallel cascade identification; evolutionary PSI-BLAST; http://bioinf.sce.carleton.ca/PCISS
[127]	2012	Q3=85.79% Q3=84.94%	RS126; CB513	machine learning approach; Support vector regression; multi-objective genetic algorithm
[122]	2015	Q3=65.1%, SOV=64.8%; Q3=66.4%, SOV=67.1%	RS126; CB513	genetic programming ;k-means clustering method; IF rules; codon encoding

Optimization could automatically optimize model's parameters or structures to achieve better performance and reduce the dependence on labour. However, its biggest disadvantage is that optimization based method need more computation power to find the best solution; besides, it should be pointed out that this best solution may be not the real optimal solution, just is a relatively optimal solution.

3.8.3 Ensemble methods and multi-models integration

Ensemble learning methods have been extensively studied in the last decade [133], and they have been an important topic in machine learning, pattern recognition and data mining fields, where it has shown great promise to improve the performance of diverse applications [129]. Ensemble learning methods take the advantages of various classifiers to build an improved classifier which integrates the results from multiple machine learning algorithms by some certain rules to obtain better predictive performance than individual [130] [131] [132]. As a result, ensemble learning can improve prediction accuracy for the problems of high complexity by different characteristic of classifiers, especially when there are significant complementarities among the ensemble members [129], such as PSSP problem [134]. The general structure of ensemble learning method is shown in Fig. 14. The summary of ensemble methods for PSSP is shown in Table 11.

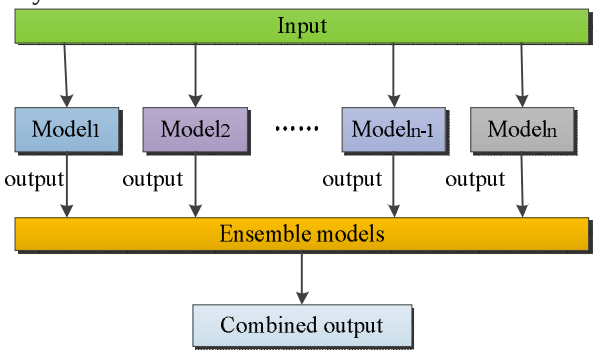


Figure 14. The structure of ensemble learning methods

Bouziane et al. evaluated the performance of two ensemble methods (simple rule-based ensemble method and weighted pooling) for PSSP, and the ensemble members were multi-class SVM and artificial neural networks [129]. But they did not consider the influence of the long-range interactions between amino acids in protein. In other paper of Bouziane et al., they reported an ensemble method by combining the outputs of two feed-forward ANNs, k-nearest neighbor (k-NN) and three multi-class SVM classifiers, and incorporating PSSM profiles to improve secondary structure prediction of globular proteins, simultaneously. Ensemble members were combined by using two variants of majority voting rules which were simple majority voting and weighted majority voting; then a heuristic based filter also was applied to refine the prediction [133]. Besides, Kountouris et al. established a prediction method by combining several neural networks and filter techniques [134]. The method utilized an ensemble of six BRNN architectures and a local window of five residues to get preliminary

output, and then the outputs were used for filtering to improve prediction accuracy. Besides, various filtering methods were used in this model to comparative predictive effect and analyze the filters performance.

Montgomerie et al. proposed an ensemble model, called PROTEUS, which integrated several structural alignment approaches with sequence-based PSSP methods (JNET, PSIPRED and TRANSSEC) according to a neural network classifier. The method combined the results with a jury-of-experts consensus tool and a robust PDB-based structure alignment process to generate prediction results [135]. The main idea of this model was to use sequence alignments as part of the secondary structure prediction process which mainly rely on the predicted results from ensemble models.

Palopoli et al. proposed a server to improve the accuracy of PSSP by selecting and integrating the prediction results computed by some available on interne (Porter, Psipred, PSA, Jufo, Prof, SAM, HMMSTR/Rosetta, Yaspin, Nn-predict, Gor IV, Hnn). This server had two main steps to obtain a final prediction for the input protein: (i1) select a team of prediction tools; (ii) fuse their prediction results [136]. Besides, Wei et al. proposed another ensemble model for PSSP, called CONCORD, which was a mixed integer linear optimization (MILP)-based consensus method that combined the results of seven selected PSSP methods (SSpro, DSC, PROF, PROFphd, PSIPRED, Predator and GorIV) to achieve better prediction effect [137].

Table 11. Summary of ensemble methods for proteins secondary structure prediction

Ref.	Year	Acc.(%)	Dataset	Description
[135]	2006	Q3=81.3%	EVA	Homologue sequence; sequence-based; http://wishart.biology.ualberta.ca/proteus
[136]	2009	other	180 proteins from CASP	Data integration; http://polifemo.deis.unical.it/jsspred.
[133]	2011	Q3 =78.24%; Q3 =76.34%	RS126; CB513	PSI-BLAST; PSSM; k-NNs, ANNs and Multi-class SVM; weighted majority voting (WMV)
[134]	2012	Q3=77.04%, SOV=72.83%	CB513	bidirectional recurrent neural networks; filtering; cascaded architecture; PSSM
[137]	2012	Q3=83.04%; Q3=82.3%	PDBselect25; CASP9 targets	mixed integer linear optimization; consensus method; http://helios.princeton.edu/CONCORD.
[129]	2015	Q3 =78.50%; Q3 =76.65%	RS126; CB513	PSSM form PSI-BLAST derived profiles; multi-class SVM; Artificial neural networks; ensemble methods

Ensemble learning boast robustness and spearheaded in the application of PSSP, and it has potential possibilities to increase the classification accuracy beyond the level reached by an individual classifier alone; however, it generally requires more computation than a single classifier, which can be regarded as a way to compensate for accuracy by performing a lot of extra computation. Besides, ensemble learning merely relies on the prediction effect of the existing methods, and its progress seriously depends on the progress of other models because it only could optimize decision results.

4. Future trends and limitations

PSSP is highly nonlinear and complex task, which can be regarded as a synthesis problem of features extraction and pattern classification. In this survey we find that the prediction performances of PSSP method mainly depend on the development of machine learning techniques and the consideration of multiple natural properties of proteins in the last decade; the former mainly relied on the theoretical and technological advance in computer science, such as deep learning and data mining, the later mainly relied on the progress of protein data sets and biological knowledge. At present, most of PSSP methods were based on common machine learning techniques and their improved versions, therefore maybe it should design some special techniques for PSSP by focusing on protein natural properties. The progress of PSSP will help to contribute to the prediction of 3-D structure and function prediction, and provide effective protein structure information for biologists and medical scientists.

4.1 Hybrid method is a trendy

Hybrid method is a trend in PSSP in recent years, such as the combination of multiple models, optimization based hybrid method and ensemble methods. The advanced machine learning techniques and their improved versions is a significant impetus for the increasing performances of PSSP. However, the single-method models often have some limitations due to their unique characteristics. Different models can extract unique protein

structure features according to their build-in attribute, which provides a basis for complementing their weakness by integrating different individual methods into a hybrid model.

Most hybrid methods adopt multi-step process for PSSP and normally composed of two processes: protein structural features extraction and secondary structure prediction; the former mainly focus on effectively extracted features of protein data by effective and suitable feature extraction methods; and the latter adopted appropriate method to predict secondary structure based on advanced machine learning models and classifiers (clusters).

Neural network is the most popular techniques in hybrid methods. It has outstanding nonlinear mapping ability to build the relationship between input and output data by different construction rules in its training process. Neural network can take both protein sequences and multiple sequence alignment profile as its input to consider protein's sequence information and evolutionary information for PSSP. However, neural network has some shortcomings, such as it normally is used as a black-box due to the lack of theoretical description, and it is not very effectively to consider the long-range dependencies of proteins. Fortunately, there are many method can remedy the shortcomings of neural network, such as SVM, probabilistic graphical model, KDD*, fuzzy k-nearest neighbor algorithm and so on, which can be used to enhance the long-range dependencies capture capability to and classification capacity of PSSP method and improve prediction accuracy.

In addition to neural network, optimization methods and ensemble learning methods are also widely used in hybrid methods for PSSP. Optimization methods are often utilized to optimize the parameters or algorithm structure of PSSP models to get better predicted results. And ensemble learning methods are extensively used to integrate different prediction results from various classifiers to obtain better predictive performance by building an ensemble classifier model. In recent years, the concept of data mining also is used in PSSP, which can be regarded as complicated combinations knowledge discovery techniques and machine learning models to perform better predicted effect. These hybrid methods are more complex and require more computation, and can be regarded as a way to compensate for accuracy by performing a lot of extra computation, however, they still is a trend due to their excellent performance and bright future.

4.2 The integration of different protein natural properties

The protein natural properties, including protein sequence information, evolutionary information, local dependencies, long-range information, physicochemical properties and biological properties, should be effectively and reasonably combined by PSSP methods. The above mentioned characteristics of proteins properties are often used as the input feature which is an extremely key component of PSSP. The performance of classifier for PSSP crucially depends on the input feature vectors. More and more researchers recognize that the individually propertied features of proteins cannot provide enough information for further improving its prediction accuracy. As a result, many scholars try to combine the features of different protein properties as the input of PSSP to achieve better predicted effect.

The physical-chemical properties of amino acids, local spatial arrangement of amino acids and long-range interdependency in protein will have a serious influence on protein secondary structures, which reflect complex sequence-structure relationship and interdependency. These protein natural properties can be used to encode each residue and correlative information is examined in relation to the formation of protein secondary structure. Evolution information is represented by multiple sequence alignment profile, which can provide additional structure information from similar protein sequences. In addition to these frequently-used protein properties, some researchers adopt other protein properties as its additional features for PSSP, such as backbone dihedral (torsion) angles, solvent accessible surface area, and dihedrals based on Cα atoms, which also have a highly correlated with protein secondary structures and provide crucial information about local 3-D structure of protein. The methods by considering the multiple features of different protein properties is adopt by many researchers to further improve the prediction accuracy of protein secondary structure. However, how to more effectively generate the input feature from different protein natural properties to represent its secondary structure information still is a challenge, which also is a subject worth being studied in the future, because it will provide many possibilities to further improve PSSP performance.

4.3 Training sample set and input data

Choosing effective training sample sets are a significant basis of PSSP research and have a serious influence on the performance of PSSP due to all prediction methods need to train. Therefore, the quality and quantity of training samples will play an important role in prediction models. How to construct or select a valid

and reliable benchmark dataset for training and testing the predictor is a key problem in PSSP. Especially, some PSSP methods are very sensitive to training data set that should be carefully choose, which also lead to its limitations, such as poor robustness, poor extensibility and difficult to train. As a result, researchers provide many sub-datasets for their PSSP methods by compressing or combining some often-used data sets. However, the training set must be big enough and include all kinds of structures in proportion to achieve higher prediction accuracy.

Besides, some machine learning algorithms need a large number of samples due to their inherent properties, which would lead to these models have unsatisfactory performances when training data is limited, such as deep learning and ensemble learning model. Even, some small samples trained models are proposed by researchers to overcome this drawback, there are still some queries for their performances in actual PSSP applications due to their finite knowledge for such complex question.

In general, the homology or non-homology and sequence identity are often used as criterion to choose training samples. At present, most methods adopt multiple sequence alignment profile as the input to train and to predict. However, there are some limitations for the situation of homologous or similar protein sequence deficiency in data sets; because the accuracy of multiple sequence alignment based PSSP method drops significantly when low homology sequences are considered. Single sequence based prediction algorithms cannot utilize the evolution information from other homologous proteins lead to its prediction accuracy is lower; besides, most of the proteins identified in genome sequencing projects have no referable sequence similarity to any known protein. As a result, new methods should be able to use either multiple sequence alignment profiles or protein sequences as their train samples or predication inputs, simultaneously.

4.4 Long-range correlation

Long-range correlation has two implications; the first is the mentioned in “The integration of different protein natural properties”, it is represented by data formatting and can be used as the input features for PSSP according to multiple sequence alignment profile; the second is the ability of PSSP methods, it refers to the prediction method could effectively extract the long-range dependence information in proteins. PSSP will be able to achieve higher prediction precision if it could consider long-range correlation information due to that is vitally important for the formation of 3-D structure in proteins. As a result, how to more effective find the long-range correlation information of proteins for PSSP still is a problem needed to be advanced.

At present, the methods with high predication accuracy usually have powerful ability for knowledge discovery in protein data; especially, they can effective extract long-range dependence information, such as deep learning and knowledge discovery based methods. Deep learning can discover more abstract representations by the combination of the features in lower levels by their numerous nonlinear operational elements. The essential mechanisms of knowledge discovery process models are heuristic coordinator and maintaining coordinator, which can be used to efficaciously find the abstract features in protein. These two kinds of methods are potentially promising method for PSSP.

4.5 Robustness and adaptability

PSSP is a high dynamic and complexity problem including feature extraction and pattern classification, which provides a big challenge for the robustness and adaptability of PSPP methods. Even many machine learning techniques based methods could get a good prediction results in PSSP, but there are few methods get more than 85% prediction accuracy. Therefore, the boost of robustness and adaptability for PSSP is a key point to improve its prediction accuracy. Researchers mainly rely on hybrid methods to improve secondary structure prediction performance in recent years, but there are a few limitations due to their dependence on individual machine learning models.

In recent years, the volume of protein data in database is keeping growing with the development of advanced sequencing technologies and protein science; therefore the dynamics of protein sequence data would be higher and higher. At present, most PSSP methods cannot cope with the incremental characteristic of protein data. Especially, the most widely used neural networks based PSSP method has a bad adaptability in newly added data. More unfortunately, only a few researchers mention their methods can cope with the incremental study task of PSSP. Therefore, the PSSP methods should be actively pursued with better robustness and adaptability to achieve better prediction performance. The boast robustness and adaptability of PSSP will spearhead their application in 3-D structure prediction and many other fields in protein science, even bioinformatics and biology.

5 Conclusion

PSSP is important research field in computational biology and protein science, and it is a fundamental task to learn protein 3-D structures and biological function. Although, there have been proposed a large number of new PSSP methods due to the ever-growing demands in the last decade. However, it still cannot meet the needs of protein 3-D structure and function prediction, and provide enough protein structure information for biologists and medical scientists. This paper provides a survey on this field to learn the latest progress and try to promote the development of PSSP. In this survey, we first provide an introduction and the relate knowledge of PSSP; then, the recent algorithmic advances of PSSP are reported to show the research status; finally, the corresponding tendencies and challenges is discussed. We consider that the remaining challenges and demands would further promote the development of this kind of techniques, although many prediction methods have been proposed. This work considers there are some possibilities to improve the performance of PSSP: (i) the hybrid methods are the trend in PSSP due to they could integrate the different advantages of single model; (ii) the integration of different protein natural properties would provide more biological information of protein; (iii) the carefully selected training sets and the consideration of long-range correlation in protein are capable of constructing an effective PSSP model; (IV) the utilization of multiple sequence alignment profile could provide more biological evolution information from other known proteins; (V) the employment of powerful machine learning techniques would enhance the reliability of PSSP method, such as deep learning. As a result, we consider that highly effective PSSP methods can potentially provide more accurate information about protein structure for biologists and medical scientists.

Acknowledgements: This study is supported by the National Natural Science Foundation of China (No.61640306), and Key Laboratory of Software Engineering of Yunnan Province (No.2017SE202). We also thank to the supporhoughtt of Scientific Research Fund of Education Department of Yunnan Province (No. 2017YJS108) and Doctoral Candidate Academic Award of Yunnan Province.

Author Contributions: Qian Jiang and Xin Jin analyzed the relevant articles. Qian Jiang, Xin Jin and Shin-Jye Lee prepared the manuscript. Shin-Jye Lee and Shaowen Yao made a critical revision of the paper.

Conflict of Interests: The authors declare that there is no conflict of interests regarding the publication of this manuscript. This article does not contain any studies with human participants performed by any of the authors. Informed consent was obtained from all individual participants included in the study.

References

[1] Liu Lin Tang, Wen Dong, Shaowen Yao et al. An overview of topic modeling and its current applications in bioinformatics. Springerplus, 2016, 5(1):1608.

[2] Saini A, Hou J. Progressive Clustering Based Method for Protein Function Prediction. Bulletin of Mathematical Biology, 2013, 75(2):331-350.

[3] Li D, Li T, Cong P, et al. A novel structural position-specific scoring matrix for the prediction of protein secondary structures. Bioinformatics, 2012, 28(1):32-9.

[4] Blazewicz J, Lukasiak P, Wilk S. New machine learning methods for prediction of protein secondary structure. Control & Cybernetics, 2007(36):183-201.

[5] Hui K H, Lei Z, Ramamohanarao K, et al. A Survey of Machine Learning Methods for Secondary and Super secondary Protein Structure Prediction. Methods in Molecular Biology, 2013, 932:87-106.

[6] Jin X, Nie R, Zhou D, et al. A novel DNA sequence similarity calculation based on simplified pulse-coupled neural network and Huffman coding. Physica A Statistical Mechanics & Its Applications, 2016, 461:325-338.

[7] Lin Liu, Lin Tang, Libo He, Shaowen Yao, Wei Zhou. Predicting protein function via multi-label supervised topic model on gene ontology. Biotechnology & Biotechnological Equipment, 2017, 31(3):630-638. DOI:

10.1080/13102818.2017.1307697

[8] Yoo P D, Zhou B B, Zomaya A Y. Machine Learning Techniques for Protein Secondary Structure Prediction: An Overview and Evaluation. Current Bioinformatics, 2008, 3(2):74-86.

[9] Lee J. Measures for the assessment of fuzzy predictions of protein secondary structure. Proteins Structure Function & Bioinformatics, 2006, 65(2):453.

[10] Furuta T, Shimizu K, Terada T. Accurate prediction of native tertiary structure of protein using molecular dynamics simulation with the aid of the knowledge of secondary structures. Chemical Physics Letters, 2009, 472(1-3):134-139.

[11] Huang J T, Cheng J P. Prediction of folding transition-state position (betaT) of small, two-state proteins from local secondary structure content. Proteins-structure Function & Bioinformatics, 2007, 68(1):218-222.

[12] Garg A, Kaur H, Raghava G P. Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. Proteins Structure Function & Bioinformatics, 2005, 61(2):318–324.

[13] Yu J T, Guo M Z. Prediction of Protein-Protein Interactions from Secondary Structures in Binding Motifs Using the Statistic Method. International Conference on Natural Computation. IEEE, 2008:100-103.

[14] Yang J Y, Peng Z L, Chen X. Prediction of protein structural classes for low-homology sequences based on predicted secondary structure. BMC Bioinformatics, 2010, 11 (1):S9.

[15] Zhang S, Ding S, Wang T. High-accuracy prediction of protein structural class for low-similarity sequences based on predicted secondary structure. Biochimie, 2011, 93(4):710-714.

[16] Liu T, Jia C. A high-accuracy protein structural class prediction algorithm using predicted secondary structural information. Journal of Theoretical Biology, 2010, 267(3):272-275.

[17] Zhang L, Liang K, Han X, et al. Structural class prediction of protein using novel feature extraction method from chaos game representation of predicted secondary structure. Journal of Theoretical Biology, 2016, 400:1.

[18] Cheng J, Sweredoski M J, Baldi P. DOMpro: Protein Domain Prediction Using Profiles, Secondary Structure, Relative Solvent Accessibility, and Recursive Neural Networks. Data Mining and Knowledge Discovery, 2006, 13(1):1-10.

[19] Gewehr J E, Zimmer R. SSEP-Domain: protein domain prediction by alignment of secondary structure elements and profiles. Bioinformatics, 2006, 22(2):181-187.

[20] Yan Wanga, Zhi-Dong Xueb, Xiao-Hong Shia, Jin Xua, Prediction of π-turns in proteins using PSI-BLAST profiles and secondary structure information. Biochemical and Biophysical Research Communications, 2006, 347(3): 574-580.

[21]Zhou Z, Yang B, Hou W. Association classification algorithm based on structure sequence in protein secondary structure prediction. Expert Systems with Applications, 2010, 37(9):6381-6389.

[22] Meng F, Kurgan L. Computational Prediction of Protein Secondary Structure from Sequence, Current Protocols in Protein Science. 2016:2.3.1-2.3.10.

[23] Feng Y, Lin H, Luo L. Prediction of protein secondary structure using feature selection and analysis approach. Acta Biotheoretica, 2014, 62(1):1-14.

[24] Levitt M, Chothia C. Structural patterns in globular proteins. Nature, 1976, 261(5561):552.

[25] Zhang G Z, Huang D S, Zhu Y P, et al. Improving protein secondary structure prediction by using the residue conformational classes. Pattern Recognition Letters, 2005, 26(15):2346-2352.

[26] Yao X Q, Zhu H, She Z S. A dynamic Bayesian network approach to protein secondary structure prediction. BMC Bioinformatics, 2008, 9(1):49.

[27] Lee S, Lee B C, Kim D. Prediction of protein secondary structure content using amino acid composition and evolutionary information. Proteins-structure Function & Bioinformatics, 2006, 62(4):1107-1114.

[28] http://www.ebi.ac.uk/

[29] Yuan Z, Sagui C. Secondary structure assignment for conformationally irregular peptides: Comparison between DSSP, STRIDE and KAKSI. Journal of Molecular Graphics & Modelling, 2015, 55:72-84.

[30] J.A.Cuff,G.J.Barton, Evaluation and improvement of multiple sequence methods for protein secondary structure prediction, Proteins, 1999, 34(4): 508-519.

[31] Qu W, Sui H, Yang B, et al. Improving protein secondary structure prediction using a multi-modal BP method. Computers in Biology & Medicine, 2011, 41(10):946.

[32] Cuff J A, Barton G J. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins Structure Function & Bioinformatics, 1999, 34(4):508.

[33] Rost B, Sander C, Schneider R. PHD-an automatic mail server for protein secondary structure prediction. Bioinformatics, 1994, 10(1):53.

[34]Saraswathi S, Fernández-Martínez J L, Kolinski A. Fast learning optimized prediction methodology (FLOPRED) for protein secondary structure prediction. Journal of Molecular Modeling, 2012, 18(9):4275.

[35] Moreau V H, Valente A P, Almeida F C L. Prediction of the amount of secondary structure of proteins using unassigned NMR spectra: a tool for target selection in structural proteomics. Genetics & Molecular Biology, 2006, 29(4):762-770.

[36] Pukáncsik M, Ágnes Orbán, Nagy K, et al. Secondary Structure Prediction of Protein Constructs Using Random Incremental Truncation and Vacuum-Ultraviolet CD Spectroscopy. Plos One, 2014, 11(6).

[37] Hering J A, Innocent P R, Haris P I. Neuro-fuzzy structural classification of proteins for improved protein secondary structure prediction. Proteomics, 2003, 3(8):1464-1475.

[38] Li D, Li T, Cong P, et al. A novel structural position-specific scoring matrix for the prediction of protein secondary structures. Bioinformatics, 2012, 28(1):32-39.

[39] Dai Q, Li Y, Liu X, et al. Comparison study on statistical features of predicted secondary structures for protein structural class prediction: From content to position. BMC Bioinformatics, 2013, 14(1):1-14.

[40] Bettella F, Rasinski D, Knapp E W. Protein secondary structure prediction with SPARROW. Journal of Chemical Information & Modeling, 2012, 52(2):545-56.

[41] Zamani M, Kremer S C. Protein secondary structure prediction using support vector machines and a codon encoding scheme. IEEE International Conference on Bioinformatics and Biomedicine Workshops. IEEE, 2012:22-27.

[42] Carrascoza F, Zaric S, Silaghidumitrescu R. Computational study of protein secondary structure elements: Ramachandran plots revisited. Journal of Molecular Graphics & Modelling, 2014, 50(5):125-133.

[43] Pok G, Jin C H, Ryu K H. Correlation of Amino Acid Physicochemical Properties with Protein Secondary Structure Conformation. National Congress of Chemical Engineering. 2008:117-121.

[44] Hubbard T. Critical Assessment of Methods of Protein StructurePrediction (CASP). Proteins-structure Function & Bioinformatics, 2003, 53(6):334-339.

[45] Lee J. Measures for the assessment of fuzzy predictions of protein secondary structure. Proteins-structure Function & Bioinformatics, 2006, 65(2):453–462.

[46] Bidargaddi N P, Chetty M, Kamruzzaman J. Combining segmental semi-Markov models with neural networks for protein secondary structure prediction. 2009, 72(16-18):3943-3950.

[47] Huang X, Huang D S, Zhang G Z, et al. Prediction of protein secondary structure using improved two-level neural network architecture. Protein & Peptide Letters, 2005, 12(8):805.

[48] Schmidhuber J. Deep learning in neural networks: an overview. Neural Networks, 2014, 61:85.

[49] Bengio Y, Delalleau O. On the expressive power of deep architectures. International Conference on Discovery Science. Springer-Verlag, 2011:1-1.

[50] Spencer M, Eickholt J, Cheng J. A deep learning network approach to ab initio protein secondary structure prediction. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 2014, 12(1):103-112.

[51] Heffernan R, Paliwal K, Lyons J, et al. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. Scientific Reports, 2015, 5(11476):11476.

[52] Wang S, Peng J, Ma J, et al. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. Scientific Reports, 2016, 6.

[53] Wang Y, Mao H, Yi Z. Protein Secondary Structure Prediction by using Deep Learning Method. Knowledge-Based Systems, 2016.

[54] Baldi P, Brunak S, Frasconi P, et al. Exploiting the past and the future in protein secondary structure prediction. Bioinformatics, 1999, 15(11):937-946.

[55] Chen J, Chaudhari N S. Cascaded Bidirectional Recurrent Neural Networks for Protein Secondary Structure Prediction. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 2007, 4(4):572-582.

[56] Mirabello C, Pollastri G. Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility. Bioinformatics, 2013, 29(16):2056-8.

[57]Ceroni A, Frasconi P, Pollastri G. Learning protein secondary structure from sequential and relational data. Neural Networks, 2005, 18(8):1029-1039.

[58] Babaei S, Geranmayeh A, Seyyedsalehi S A. Protein secondary structure prediction using modular reciprocal bidirectional recurrent neural networks. Computer Methods & Programs in Biomedicine, 2010, 100(3):237-247.

[59] Chen J, Chaudhari N S. Bidirectional segmented-memory recurrent neural network for protein secondary structure prediction. Soft Computing, 2006, 10(4):315-324.

[60] Heffernan R, Yang Y, Paliwal K, Zhou Y., 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. 2017, In press.

[61] Patel M S, Mazumdar H S. Knowledge base and neural network approach for protein secondary structure prediction. Journal of Theoretical Biology, 2014, 361:182-189.

[62] Jing N, Xia B, Zhou C G, et al. Protein secondary structure prediction methods based on RBF neural networks. 1st International Conference on Computational Methods (ICCM04). 2006:1037-1043.

[63] Zhang Z, Jing N. Radial basis function method for prediction of protein secondary structure. International Conference on Machine Learning and Cybernetics. IEEE Xplore, 2008:1379-1383.

[64] Kakumani R, Devabhaktuni V, Ahmad M. A two-stage neural network based technique for protein secondary structure prediction. 2008, 2008(2008):1355-8.

[65] Sakthivel S, Habeeb S K M. NNvPDB: Neural Network based Protein Secondary Structure Prediction with PDB Validation. Bioinformation, 2015, 11(8):416-421.

[66] Liu K H, Xia J F, Li X. Efficient ensemble schemes for protein secondary structure prediction. Protein & Peptide Letters, 2008, 15(5):488-93.

[67] Shamima B, Savitha R, Suresh S, et al. Protein secondary structure prediction using a fully complex-valued relaxation network. International Joint Conference on Neural Networks. IEEE, 2013:1-8.

[68] Rashid S, Saraswathi S, Kloczkowski A, et al. Protein secondary structure prediction using a small training set (compact model) combined with a Complex-valued neural network approach. BMC Bioinformatics, 2016, 17(1):362.

[69] Nitta T. Orthogonality of decision boundaries in complex-valued neural networks. Neural Computation, 2004, 16(1):73.

[70] Denoeux T. A neural network classifier based on Dempster-Shafer theory. IEEE Transactions on Systems Man and Cybernetics-Part A Systems and Humans, 1997, 30(2):131-150.

[71] Zhong W, Altun G, Tian X, et al. Parallel protein secondary structure prediction schemes using Pthread and OpenMP over hyper-threading technology. The Journal of Supercomputing, 2007, 41(1):1-16.

[72] Faraggi E, Al E. SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. Journal of Computational Chemistry, 2012, 33(3):259-67.

[73] Birzele F, Kramer S. A new representation for protein secondary structure prediction based on frequent patterns. Bioinformatics, 2006, 22(21):2628-2634.

[74] Shoyaib M, Baker S M, Jabid T, et al. Protein secondary structure prediction with high accuracy using Support Vector Machine. International Conference on Computer and Information Technology. IEEE Xplore, 2008:1-4.

[75] Chen C, Tian Y, Zou X, et al. Prediction of protein secondary structure content using support vector machine. Talanta, 2007, 71(5):2069-2073.

[76] Karypis G. YASSPP: better kernels and coding schemes lead to improvements in protein secondary structure prediction. Proteins-structure Function & Bioinformatics, 2006, 64(3):575-586.

[77] Chatterjee P, Basu S, Kundu M, et al. PSP_MCSVM: brainstorming consensus prediction of protein secondary structures using two-stage multiclass support vector machines. Journal of Molecular Modeling, 2011, 17(9):2191-2201.

[78] Kountouris P, Hirst J D. Prediction of backbone dihedral angles and protein secondary structure using support vector machines. BMC Bioinformatics, 2009, 10(1):437.

[79] Jin C, Gubbi J, Buyya R, et al. Jeeva: Enterprise Grid Enabled Web Portal for Protein Secondary Structure Prediction. International Conference on Advanced Computing and Communications. IEEE Xplore, 2009:141-147.

[80] Altun G, Hu H J, Brinza D, et al. Hybrid SVM Kernels for Protein Secondary Structure Prediction. IEEE International Conference on Granular Computing, Grc 2006, Atlanta, Georgia, Usa, May. DBLP, 2006:762-765.

[81] Li Q, Dahl D B, Vannucci M, et al. Bayesian Model of Protein Primary Sequence for Secondary Structure Prediction. Plos One, 2014, 9(10):e109832.

[82] Asai K, Hayamizu S, Handa K. Prediction of protein secondary structure by the hidden Markov model. Bioinformatics, 1993, 9(2):141.

[83] Chu W, Ghahramani Z, Podtelezhnikov A, et al. Bayesian segmental models with multiple sequence alignment profiles for protein secondary structure and contact map prediction. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 2006, 3(2):98-113.

[84] Malekpour S A, Naghizadeh S, Pezeshk H, et al. A Segmental Semi Markov Model for protein secondary structure prediction. Mathematical Biosciences, 2009, 221(2):130-135.

[85] Pezeshk H, Naghizadeh S, Malekpour S A, et al. A modified bidirectional hidden Markov model and its application in protein secondary structure prediction. International Conference on Advanced Computer Control. IEEE, 2010:535-538.

[86] Schmidler S C, Liu J S, Brutlag D L. Bayesian segmentation of protein secondary structure. Journal of Computational Biology, 2000, 7(1-2):233-248.

[87] Z Aydin, Y Altunbasak, M Borodovsky. Protein secondary structure prediction for a single-sequence using hidden semi-Markov models . BMC Bioinformatics, 2006, 7(1):178

[88] Z Aydin, Y Altunbasak, H Erdogan, Bayesian Protein Secondary Structure Prediction With Near-Optimal

Segmentations. Signal Processing IEEE Transactions on, 2007, 55(7):3512-3525.

[89] Wu K P, Lin H N, Chang J M, et al. HYPROSP: a hybrid protein secondary structure prediction algorithm--a knowledge-based approach. Nucleic Acids Research, 2004, 32(17):5059-5065.

[90] Jones D T. Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology, 1999, 292(2):195-202.

[91] Lin H N, Sung T Y, Ho S Y, et al. Improving protein secondary structure prediction based on short subsequences with local structure similarity. BMC Genomics, 2010, 11 Suppl 4(4):S4.

[92] Popov I, Vassilev D. A Semi-Automated Structural Class Dependent Method for the Prediction of Protein Secondary Structures. Biotechnology & Biotechnological Equipment, 2010, 24(3):2044-2048.

[93] Yonge, Feng, Liaofu. Using long-range contact number information for protein secondary structure prediction. International Journal of Biomathematics, 2014, 07(5):83-93.

[94] Krishnaji A, Rao A A. Implementation of a hybrid Neuro Fuzzy Genetic System for improving protein secondary structure prediction. Computing and Communication Systems. 2012:1-5.

[95] Mocz G. Fuzzy cluster analysis of simple physicochemical properties of amino acids for recognizing secondary structure in proteins. Protein Science, 1995, 4(6):1178.

[96] Lee S J, Zeng X J. A three-part input-output clustering-based approach to fuzzy system identification. International Conference on Intelligent Systems Design and Applications. IEEE, 2011:55-60.

[97] Lee S J, Zeng X J. A similarity-based learning algorithm for fuzzy system identification with a two-layer optimization scheme. IEEE International Conference on Fuzzy Systems. IEEE, 2012:1-8.

[98] Lee S J, Wang H S. A Dynamic Modular Method for Estimating Null Values in Relational Database Systems. International Journal of Computer Information Systems and Industrial Management Applications. 2009:249-257.

[99] Zhang C T, Chou K C, Maggiora G M. Predicting protein structural classes from amino acid composition: application of fuzzy clustering. Protein Engineering, 1995, 8(5):425-435.

[100] Boberg J, Salakoski T, Vihinen M. Accurate prediction of protein secondary structural class with fuzzy structural vectors. Protein Engineering, 1995, 8(8):505-512.

[101] Thanh Nguyen, Abbas Khosravi, Douglas Creighton, et al. Multi-Output Interval Type-2 Fuzzy Logic System for Protein Secondary Structure Prediction. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2015, 23(05):735-760.

[102] Kim S Y, Sim J, Lee J. Fuzzy k-Nearest Neighbor Method for Protein Secondary Structure Prediction and Its Parallel Implementation. International Conference on Computational Intelligence and Bioinformatics. Springer-Verlag, 2006:444-453.

[103] Ghosh A, Parai B. Protein secondary structure prediction using distance based classifiers. International Journal of Approximate Reasoning, 2008, 47(1):37-44.

[104] Maglia A M, Leopold J L, Ghatti V R. Identifying character non-independence in phylogenetic data using data mining techniques. Asia-Pacific Bioinformatics Conference. DBLP, 2004:181-189.

[105] Lee L, Leopold J L, Frank R L, et al. Protein secondary structure prediction using rule induction from coverings. IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology. IEEE Press, 2009:79-86.

[106] Lee L, Kandoth C, Leopold J L, at el. Protein Secondary Structure Prediction Using Parallelized Rule Induction from Coverings, World Academy of Science, Engineering and Technology, 2009, 60: 388-394.

[107] Chang D T, Ou Y Y, Hung H G, et al. Prediction of protein secondary structures with a novel kernel density estimation based classifier. BMC Research Notes, 2008, 1(1):51

[108] Yüksektepe F Ü, Özlem Yılmaz, Türkay M. Prediction of secondary structures of proteins using a two-stage

method. Computers & Chemical Engineering, 2008, 32(1–2):78-88.

[109] Yang W, Wang K, Zuo W. Prediction of protein secondary structure using Large Margin Nearest Neighbour classification. International Journal of Bioinformatics Research & Applications, 2013, 9(2):207.

[110] Malekpour S A, Naghizadeh S, Pezeshk H, et al. Protein secondary structure prediction using three neural networks and a segmental semi Markov model. Mathematical Biosciences, 2009, 217(2):145-150.

[111] Zhang B, Chen Z, Murphey Y L. Protein secondary structure prediction using machine learning. IEEE International Joint Conference on Neural Networks, 2005. IJCNN '05. Proceedings. IEEE Xplore, 2005:532-537 vol. 1.

[112] Anjum R A, Zhang Y Q, Robert W. Harrison. Granular Decision Tree and Evolutionary Neural SVM for Protein Secondary Structure Prediction. International Journal of Computational Intelligence Systems, 2009, 2(4):343-352.

[113] Pradip G, Nikhi R P, Rajani K M. Prediction of Protein Secondary Structure Using probability based fetures and a hybrid system. Journal of Bioinformatics & Computational Biology, 2013, 11(5):1350012.

[114] Bondugula R, Xu D. MUPRED: a tool for bridging the gap between template based methods and sequence profile based methods for protein secondary structure prediction. Proteins, 2007, 66(3):664.

[115] Wang Z, Zhao F, Peng J, et al. Protein 8-class secondary structure prediction using conditional neural fields. Proteomics, 2010, 11(19):3786-92.

[116] Silvia Botelho, Gisele Simas, and Patricia Silveira, Prediction of Protein Secondary Structure Using Nonlinear Method. Lecture Notes in Computer Science, 2006, 4634:40-47.

[117] Yang B R, Sun H H, Xiong F L. Ming Quantitative association rules with standard SQL queries and its evaluation. Journal of Computer Research and Development, 2002, 39(3): 307–312.

[118] Yang B, Wei H, Zhun Z, et al. KAAPRO: An approach of protein secondary structure prediction based on KDD* in the compound pyramid prediction model. Expert Systems with Applications, 2009, 36(5):9000-9006.

[119] He J, Hu H J, Harrison R, et al. Rule generation for protein secondary structure prediction with support vector machines and decision tree. IEEE Transactions on Nanobioscience, 2006, 5(1):46-53.

[120] Katzman S, Barrett C, Thiltgen G, et al. Predict-2nd: a tool for generalized protein local structure prediction. Bioinformatics, 2008, 24(21):2453-2459.

[121] Madera M, Calmus R, Thiltgen G, et al. Improving protein secondary structure prediction using a simple k-mer model. Bioinformatics, 2010, 26(5):596-602.

[122] Zamani M, Kremer S C. Protein secondary structure prediction using an evolutionary computation method and clustering. IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology. IEEE, 2015:1-6.

[123] Li B, Gong L, Yao Y. On the performance of internal feedback artificial bee colony algorithm (IF-ABC) for protein secondary structure prediction. Sixth International Conference on Advanced Computational Intelligence. IEEE, 2013:33-38.

[124] Mitchell M. An Introduction to Genetic Algorithms. MIT Press, Cambridge. DBLP, 1998.

[125] Won K J, Hamelryck T, Prügelbennett A, et al. An evolutionary method for learning HMM structure: prediction of protein secondary structure. BMC Bioinformatics, 2007, 8(1):357.

[126] Green J R, Korenberg M J, Aboul-Magd M O. PCI-SS: MISO dynamic nonlinear protein secondary structure prediction. BMC Bioinformatics, 2009, 10(1):222.

[127] Zangooei M H, Jalili S. Protein secondary structure prediction using DWKF based on SVR-NSGAII. Neuro computing, 2012, 94(3):87-101.

[128] Chopra P, Bender A. Evolved cellular automata for protein secondary structure prediction imitate the determinants for folding observed in nature. Silico Biology, 2007, 7(1):87-93.

[129] Bouziane H, Messabih B, Chouarfia A. Effect of simple ensemble methods on protein secondary structure prediction. Soft Computing, 2015, 19(6):1663-1678.

[130] Maclin R, Opitz D. Popular Ensemble Methods: An Empirical Study. Journal of Artificial Intelligence Research, 1999, 11:169-198.

[131] Polikar, Robi. Ensemble based systems in decision making. Circuits & Systems Magazine IEEE, 2006, 6(3):21-45.

[132] Rokach L. Ensemble-based classifiers. Artificial Intelligence Review, 2010, 33(1):1-39.

[133] Bouziane H, Messabih B, Chouarfia A. Profiles and Majority Voting-Based Ensemble Method for Protein Secondary Structure Prediction. Evolutionary Bioinformatics Online, 2011, 7(7):171

[134] Kountouris P, Agathocleous M, Promponas V J, et al. A comparative study on filtering protein secondary structure prediction. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 2012, 9(3):731-739.

[135] Montgomerie S, Sundararaj S, Gallin W J, et al. Improving the accuracy of protein secondary structure prediction using structural alignment. BMC Bioinformatics, 2006, 7(1):301.

[136] Palopoli L, Rombo S E, Terracina G, et al. Improving protein secondary structure predictions by prediction fusion. Information Fusion, 2009, 10(3):217-232.

[137] Wei Y, Thompson J, Floudas C A. CONCORD: a consensus method for protein secondary structure prediction via mixed integer linear optimization. Proceedings of the Royal Society A Mathematical Physical & Engineering Sciences, 2012, 468(2139):831-850.