Computational Biology and Chemistry 81 (2019) 1–8

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Review Article

Protein secondary structure prediction using neural networks and deep learning: A review

Wafaa Wardah[a], M.G.M. Khan[a], Alok Sharma[b,c], Mahmood A. Rashid[e,d,*]

[a] School of Computing, Information and Mathematical Sciences, The University of the South Pacific, Suva, Fiji
[b] School of Engineering and Physics, The University of the South Pacific, Suva, Fiji
[c] RIKEN Center for Integrative Medical Sciences, Yokohama, Japan
[d] Institute for Integrated and Intelligent Systems, Griffith University, Queensland, Australia
[e] Institute for Sustainable Industries and Liveable Cities, Victoria University Melbourne, Victoria, Australia

ARTICLE INFO

Keywords:
Protein secondary structure prediction
Machine learning
Neural network
Deep learning
Feature selection

ABSTRACT

Literature contains over fifty years of accumulated methods proposed by researchers for predicting the secondary structures of proteins in silico. A large part of this collection is comprised of artificial neural network-based approaches, a field of artificial intelligence and machine learning that is gaining increasing popularity in various application areas. The primary objective of this paper is to put together the summary of works that are important but sparse in time, to help new researchers have a clear view of the domain in a single place. An informative introduction to protein secondary structure and artificial neural networks is also included for context. This review will be valuable in designing future methods to improve protein secondary structure prediction accuracy. The various neural network methods found in this problem domain employ varying architectures and feature spaces, and a handful stand out due to significant improvements in prediction. Neural networks with larger feature scope and higher architecture complexity have been found to produce better protein secondary structure prediction. The current prediction accuracy lies around the 84% marks, leaving much room for further improvement in the prediction of secondary structures in silico. It was found that the estimated limit of 88% prediction accuracy has not been reached yet, hence further research is a timely demand.

1. Introduction

This study explores the usage of artificial neural networks (ANN) in protein secondary structure prediction (PSSP) – a problem that has engaged scientists and researchers for over 3 decades. ANN, or simply neural networks (NN), have recently gained a lot of popularity in the realm of computational intelligence, and have been observed to be a boost in the rather stagnant field of PSSP.

The uniqueness of the PSSP problem has paved the emergence of an entirely new field in science, one which did not exist a few years ago. Since the 1950s, numerous methods have been devised for predicting secondary structure from amino acid sequences. Methods range from disciplines like biology and physical chemistry to statistics and computer science. In this paper, the use of NN, a recently popular branch of machine learning, and its relevance in the improvement of prediction accuracy for the protein secondary structure problem is explored.

The following section, provides a background to protein and its structure, and an introduction to artificial neural networks (ANN). Then follows a discussion on parameter optimization for PSSP, the different NN architectures being used in PSSP, and features being considered in this domain. Finally, a discussion and conclusion provide insight into where the research stands currently and what can be expected for the future of this field.

2. Background

The following sections provide background information on protein structure, followed by an introduction to NN and deep learning.

2.1. Protein preliminaries

Proteins are macromolecules that carry out indispensable functions in essentially all biological processes that occur in the human body (Berg and Tymoczko, 2002), example, metabolism and homeostasis. They are manufactured in the body within cells using amino acid molecules that act as protein building blocks. Although there are about 500

* Corresponding author at: Institute for Sustainable Industries and Liveable Cities, Victoria University Melbourne, Victoria, Australia.
E-mail address: mahmood.rashid@griffith.edu.au (M.A. Rashid).

https://doi.org/10.1016/j.compbiolchem.2019.107093
Received 12 July 2018; Received in revised form 28 December 2018; Accepted 10 July 2019
Available online 12 August 2019
1476-9271/

amino acids, only 20 of them are encoded by the genome for building the proteins required by the human body (Wagner and Musso, 1983; Sanger, 1959). The amino acids are most likely ingested as part of one's diet, or recycled within the body. Within cells, protein synthesis takes place in two complex stages. First, messenger RNA (mRNA) is transcribed from DNA into templates through a process called transcription. The mRNA templates are then translated into protein chains by organelles called Ribosomes through a process called translation.

The number of possible amino acid configurations is infinite, considering that their length and frequency are highly variable. For a set of 100 amino acids, 20^{100} proteins are possible. However, the human genome codes approximately 3.5×10^4 proteins. Therefore, it can be assumed that relatively less real proteins get synthesized compared to the theoretical possibility (Otaki et al., 2005). Despite, the number of biological proteins that exist is still significantly large, and predicting their structures from amino acid sequences is challenging.

The function of a protein is directly related to its native structure. Often, distinct amino acid sequences embody a similar structure, and the resulting structures exhibit similar functionality solely due to the similarity in their conformations (Sanger, 1959). Many medicinal fields anticipate the proper prediction tools so that studying certain proteins from the component amino acid sequences can be easily realized. One example is exploring encoded proteins based on their three-dimensional spatial relationships in local concentrations of human cancers (Niu et al., 2016). Another example is studying structure of Gamma B Crystallin proteins found in the eye lens to better understand the development of cataracts (Umphred-Wilson et al., 2017). Protein misfolding has been seen to be the major contributor in development of many diseases, like type 2 diabetes as well as neurodegenerative diseases such as, Alzheimer's, Parkinson's, Huntington's, and amyotrophic lateral sclerosis (ALS) (Ken and Justin, 2012), hence, understanding the folded structures is extremely significant for disease prevention and treatment.

Biologists have devised four levels of amino acid organization in the plight to understand protein structures as described below. These are categorized as primary, secondary, tertiary and quaternary structures (see Fig. 1). The levels occur in stages, where each lower level is necessary for the formation of the next level (Pauling et al., 1951; Pauling and Corey, 1951; Levitt and Chothia, 1976; Garnier et al., 1978; Anfinsen et al., 1961; Ewbank and Creighton, 1992; Bradley et al., 1990; Marqusee and Baldwin, 1987; Marqusee et al., 1989; Oas and Kim, 1988; Roder et al., 1988; Udgaonkar and Baldwin, 1988).

- *Primary structure* of a protein refers to the linear sequence of amino acid residues that make up the protein (Sanger, 1952). Amino acid residues are joined together in long chains by peptide bonds, where each residue has 2 neighboring residues. The structure forms a 'backbone' that runs along the entire peptide chain (see Fig. 1(a)).
- *Secondary structure* refers to the 3 dimensional local segments of the protein macromolecule that form after the amino acid residues join in a sequence and before the protein folds into its tertiary structure. The secondary structure involves hydrogen bonds along the backbone that cause the long chain to fold into local shapes, mainly helices, strands and coils (see Fig. 1(b)). The standard defining convention adopted was created by Kabsch and Sander (1983) as the Dictionary of Protein Secondary Structure (DSSP). This convention defines 8 groups, however, it is common to group them into 3 general groups of α helix, β sheet, and random coil (Pauling et al., 1951; Eisenberg, 2003).
- *Tertiary structure* is the 3 dimensional structure of a protein. It consists of one long 'backbone' consisting of the various secondary structures and thus further folds in consequence to the amino acid side chains' interactions. For example, some area may be hydrophobic, which causes it to fold tightly inwards to hide from water molecules, contributing to a globular conformation. Salt bridges, hydrogen bonds, and the tight packing of side chains and disulfide bonds are some forces that contribute to the formation of the tertiary structure (see Fig. 1(c)).
- *Quaternary structure* refers to the further stabilization of the protein molecule by bonding with one or more similar tertiary structures via further non-covalent interactions and disulfide bonding. The final complex achieves stability and functions as one unit (see Fig. 1(d)).

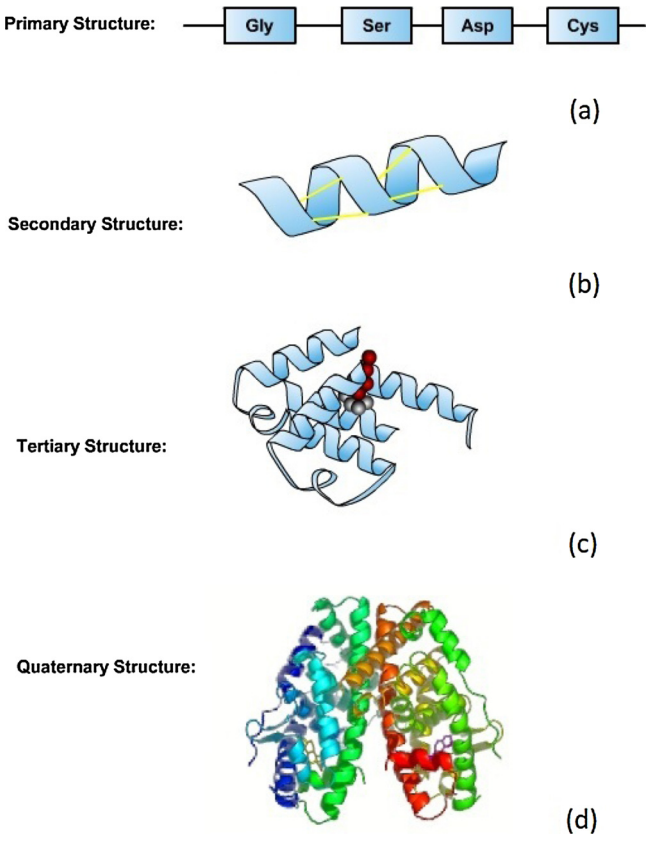

Fig. 1. The four levels of organization are shown, namely, primary, secondary, tertiary and quaternary structures.

Apart from the above mentioned four levels of structure, there are additional classification schemes of protein building blocks. Small groups of secondary structure units that occur commonly are known as super-secondary structures, or motifs. Several motifs pack together to form local semi-independent units called domains. These are helpful in obtaining evolutionary information about the proteins, as it is seen that proteins with similar structure do not always share similar sequence, and domains aid in structural comparison (Richardson, 1981).

In vitro methods of obtaining the detailed structure of proteins include X-ray crystallography, nuclear magnetic resonance spectroscopy and electron micrography. Although these methods are relatively accurate, they are time-consuming and costly (Moraes et al., 2014). Due to these disadvantages, innovative approaches to predict protein structures, such as machine learning, have become the panacea. In the early years, Lim (1974) proposed a method that utilized the physicochemical characteristics of amino acids to predict protein structure. Later, a similar approach was also proposed in Ptitsyn and Finkelstein (1983). Additionally, prediction attempts using sequence patterns and statistical analysis have also been thoroughly investigated in the early years of PSSP as listed in Table 1.

2.2. Neural network and deep learning preliminaries

Traditional computing involves human-written instructions in a computer program. On the contrary, artificial intelligence allows a

Table 1
List of early methods that were developed for the PSSP problem. The works are discriminated according to their approach as either being based on the protein sequence information or statistical analysis.

Sequence patterns		Statistical analysis	
Reference	Year	Reference	Year
Levin et al. (1986)	1986	Wu and Kabat (1973)	1973
Nakashima et al. (1986)	1986	Chou and Fasman (1974)	1974
Zvelebil et al. (1987)	1987	Kozo (1977)	1977
Cohen et al. (1983)	1983	Maxfield and Scheraga (1979)	1979
Taylor and Thornton (1983)	1983	Gibrat et al. (1987)	1987
Rooman et al. (1989)	1989	Nagano (1973)	1973
Rooman et al. (1991)	1991	Nagano and Hasegawa (1975)	1975
King and Sternberg (1990)	1990	Schulz and Schirmer (1979)	1979
Cohen et al. (1986)	1986	Biou et al. (1988)	1988
Taylor and Orengo (1989)	1989	Kanehisa (1988)	1988
Sander and Schneider (1991)	1991	Levin and Garnier (1988)	1988
Vriend and Sander (1991)	1991	Fasman (1989)	1989
Rooman et al. (1991)	1991	Garratt et al. (1991)	1991
Presnell et al. (1992)	1992	Muggleton et al. (1992)	1992

system to modify or write new instructions for itself. One approach of this latter style is through the use of ANNs. This concept is derived from the working patterns of the biological neurons in the brain (see Fig. 2). Just as the millions of neurons in the brain collectively execute the cognitive processes, ANNs are fashioned in a similar way to carry out intelligent computation (McCulloch and Pitts, 1943; Heeb, 1949; Minsky, 1954; Rosenblatt, 1962). Recent popularization of brain-inspired architectures exhibits an acceptance and encouragement to continue further research and application of these methods in various industries (Otoom, 2016).

An ANN is a network created by at least 2 layers of neuron-like processing units. The initial layer is called the input layer as it introduces input variables into the network. The final layer is the output layer, which may contain units for carrying out output classification. For networks that contain more than 2 layers, the remaining inner layers are called the hidden layers. A shallow network is one that ideally contains none or one hidden layer. On the other hand, deep network refers to a network of artificial neurons comprising many hidden layers (refer to Fig. 3). Evidently, deep NNs have been highly successful in solving complex problems (Bianchini and Scarselli, 2014).

Inside an ANN, complex matrix computations take place throughout the inner layers. A standard ANN generally accepts a set of input values in the form of vectors containing feature-values (example x_0, x_1, x_2, …, x_n). The selection of these features is challenging and further understanding on this can be obtained from Kwak and Choi (2002). Each unit (neuron) that is part of the following layer assigns a designated weight (and other parameters such as bias) to the input, which produces some output. In supervised learning, the real corresponding output is also supplied to the algorithm during training. If the produced output does not match the real output for that particular input, the weights get adjusted automatically through an algorithm of choice. In large networks with high dimensionality like those for the protein structure prediction problems, backpropagation is often used for adjusting weights. Backpropagation refers to the method of revisiting the previous layers and adjusting weights so that the calculated output is closer to the actual expected output (Rumelhart et al., 1986a,b). This flow continues until the desired accuracy is obtained or until the specified number of epochs is reached. The assumed optimal weights and biases are achieved once training is complete and the network can apply these parameters to the test inputs for producing predictions as the network outputs. Once a satisfactory model is achieved, the parameters are frozen so that predictions can be made using new input data.

The standard NN has evolved extensively over the years, resulting in a variety of architectural configurations. Their application to the PSSP problem will be explored in this paper.

Ever since Kendrew et al. (1958, 1960) and Perutz et al. (1960) managed to establish the structure of proteins using x-ray crystallography around 1960 (for which Kendrew and Perutz later received the shared Nobel Prize (1962)), researchers have been attempting to understand the protein folding problem. By 1988, it was realized that the PSSP problem would require researchers to move away from traditional computing onto newer ways of computation (Rooman and Wodak, 1988; Kneller et al., 1990). Hence, machine learning techniques such as ANN were explored. Fig. 4 is a graph that represents the accumulation of the efforts made in improving PSSP with NN over the past 3 decades.

3. Review of NN in PSSP problems

Exploration of the various methods applied to the PSSP problem reveals the diversity in approaches undertaken by researches over the past few decades. Within the NN field of machine learning, the options are endless when selecting a model for the task of PSSP. The following sections discuss major concepts relating to the various successful NN models.

3.1. Parameter optimization

As described earlier, standard feed forward back propagation networks have helped give insight into the promising potential of this machine learning technique. For a very long time, the standard NN was heavily experimented with in the PSSP domain. Today, there are multiple architectures to choose from when designing a NN. In 2017, Dongardive and Abraham (2017) used a standard feed forward NN with one hidden layer to predict secondary structure given the amino acid sequence information. Their objective was to find an optimal parameter set that would produce the best prediction results. The parameters involved were encoding scheme (ES), window size (WS), number of neurons in the hidden layer (HN) and the type of learning algorithm (LA). The first parameter ES is chosen from a set of 8 options. Firstly, the orthogonal encoding scheme is a popular one-hot convention suggested by Holley and Karplus (1989) and has been used in many early works. Secondly, the hydrophobicity encoding scheme involves creating a matrix from the hydrophobicity index of each amino acid in the given sequence. Thirdly, BLOSUM62, a substitution matrix that includes evolutionary based information, along with the similar PAM250 mutation matrix as the forth encoding option. The remaining

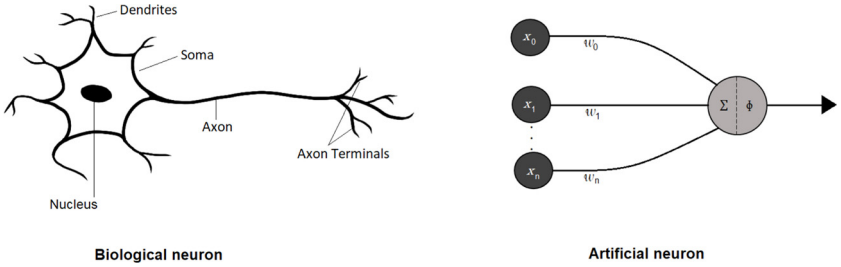

Fig. 2. The biological and artificial neuron are shown.

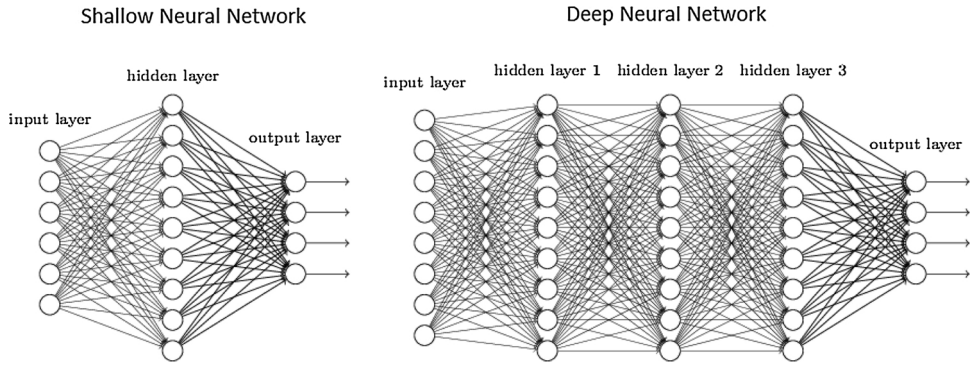

Fig. 3. A shallow network is compared to a deep network configuration.

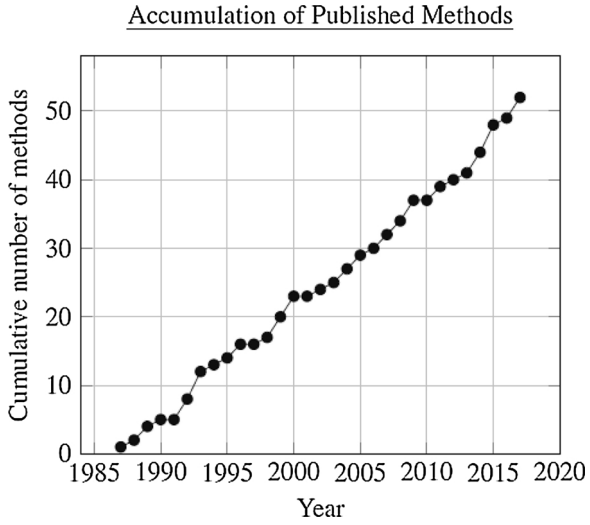

Fig. 4. Accumulation of PSSP methods using ANN over the past 3 decades.

four encoding schemes were hybrids of the earlier mentioned schemes. The parameter WS is chosen from a range of odd numbered window sizes between 3 and 19. The sliding window protocol involves selecting a variable length of the amino acid sequence, which would correspond to the rows in the input matrix of the network. The parameter NS referring to the number of neurons in the hidden layer is picked from the range 1 to 20 and would determine the overall network topology. The final parameter LA involved 8 distinct learning algorithms that would contribute to the best prediction accuracy of the network. These were gradient descent with momentum and adaptive learning rate, resilient back propagation algorithm, scaled conjugate gradient algorithm, conjugate gradient back propagation with Fletcher-Reeves updates, conjugate gradient back propagation with Polak-Rabiere updates, conjugate gradient back propagation with Powell-Beale restarts, quasi-Newton (Broyden-Fletcher-Goldfarb-Shanno) algorithm, and finally quasi-Newton one-step secant learning algorithm. Dongardive and Abraham (2017) found that the optimal parameter set in their experiment consisted of BLOSUM62 as the best encoding scheme, window size of 19, 19 neurons in the hidden layer and one-step secant as the optimal learning algorithm, which altogether produced a prediction accuracy of 78%. Changing parameters to find the optimal set for a NN model in PSSP has been a central task since the beginning of the field itself, starting with Qian and Sejnowski's work in 1988 (Qian and Sejnowski, 1988), where they adjusted the hidden layer, window size and encoding scheme. This strategy continued to be employed later in early 1990s as found in Sasagawa and Tajima (1993), Rost and Sander (1993a,b) and towards the end of the decade as mentioned in Chandonia and Karplus (1995), Jones (1999). The most common parameter that is seen to be adjusted is the window size, and the optimal number suggested in the mentioned works is window size of 13–19.

3.2. Review of different NN architectures

Over the years, various architectures of NN have been devised to produce best results in certain application areas. While a few of these architectures are readily applicable to the sequence-based nature of the PSSP problem, researchers have sought improvement in PSSP by experimenting with almost all major techniques of NN, as discussed below.

3.2.1. Recurrent neural network

Recurrent neural network (RNN) is widely known for processing sequence based problems, and this makes it the first option when considering NN techniques for PSSP. A successful RNN method for PSSP is Porter 4.0 (Mirabello and Pollastri, 2013). This method has evolved over several years as RNN techniques got further refined. The project was initiated by Baldi, Pollastri and colleagues, where they explored RNN for PSSP in late 1990s and published their method SSpro in 1999 (Baldi et al., 1999). This method consisted of a bidirectional RNN (BRNN) that was able to consider past and future elements of the amino acid sequence when predicting secondary structure. SSpro managed to reach 76% accuracy, which was exceeded by it's improved version achieving 78% accuracy in 2002 by incorporating ensembles of the SSpro model (Pollastri et al., 2002). In the next few years, Pollastri et al. (2002) and Pollastri and McLysaght (2005) reached 79% accuracy using Porter – a further improved version of SSpro. Mirabello and Pollastri (Mirabello and Pollastri, 2013) created a further improved version of Porter (Pollastri and McLysaght, 2005) called Porter 4.0, which was based on ensembles of BRNN (Baldi et al., 1999) and achieved 82.2% accuracy. They followed a cascaded architecture where a first BRNN predicted secondary structure from the primary sequence and multiple sequence alignments (MSA), and a second BRNN filtered the predictions of the first stage. The initial SSpro method was also the basis for a multi-class variant named SSpro8 created by Magnan and Baldi (2014). This method includes both sequence similarity and sequence-based structural similarity which they claimed to achieve up to 92.91% accuracy for proteins with homologs found in the Protein Data Bank.

Another successful RNN based method is SPIDER3 (Heffernan et al., 2017), which is also the result of gradual improvements made to an initial work. The first method, SPIDER (Lyons et al., 2014), was a NN model for predicting θ and τ angles in the protein backbone. SPIDER2 (Yang et al., 2014), a NN with three hidden layers, each with 150 units, was used in three iterations that predicted four different sets of structural properties: secondary structure, torsion angles, C_α atom based angles and dihedral angles, and solvent accessible surface area. This method achieved 82% three state accuracy (Heffernan et al., 2015). Finally, SPIDER3 was developed to consider the multiple features and predict secondary structure along with the other mentioned properties,

Table 2
The major periodically relevant state-of-the-art methods are shown along with the types of feature values they employed in their networks.

Neural network method	Accuracy (Q3)	Seq info	Evo info	Physico chem info
Qian &Sejnowski 1988 (Qian and Sejnowski, 1988)	64.3%	✓		
PHD 1994 (Rost et al., 1994)	71.4%	✓	✓	
PSIPRED 1997 (Jones, 1999)	76.5%	✓	✓	
JPRED3 2008 (Cole et al., 2007)	81.5%	✓	✓	
SPIDER3 2017 (Heffernan et al., 2017)	84%	✓	✓	✓

and was a BRNN model that contained long short-term memory (LSTM) cells. The LSTM cells aid in capturing both local and non-local intra-sequence relationships efficiently, and the authors hold these responsible for the improvement in prediction to 84%.

3.2.2. Convolutional neural network

Another variant of ANN is convolutional neural network (CNN) and is mostly known for its success in image recognition applications. Wang et al. (2015) created a CNN based method for PSSP and managed to achieve up to 84% accuracy. This method, deep convolutional neural fields (DeepCNF) consisted of two modules: conditional random fields (CRF) module that was initially developed in Wang et al. (2011) and deep convolutional NN (DCNN) module covering the input to the CRF (Wang et al., 2016). The researchers emphasized the contribution of the deep layers in the network to the improvement achieved by the method, in contrast to the contribution of the network parameters such as window size. They ran several tests that clearly showed the effects of increasing the number of hidden layers within the network and the number of neurons per layer towards the incremental improvement in SS prediction.

3.2.3. Hybrid neural networks

A few recent PSSP techniques are comprised of multiple architectures intertwined together to improve the overall network prediction. Li and Yu (2016) used a deep convolutional recurrent NN architecture (DCRNN) that leveraged CNN with different kernel sizes to extract multi-scale local contextual features. Due to long-range dependencies existing in amino acid sequences, a BRNN consisting of gated recurrent unit (GRU) was set up to capture global contextual features. They achieved up to 85.3% three state accuracy with this hybrid architecture.

Wang et al. (2017) achieved up to 84.2% accuracy, and also used DCRNN with GRU, consisting of a feature embedding layer, multiscale CNN layers for local context extraction, stacked bidirectional RNN layers for global context extraction, and softmax layers for secondary structure and solvent accessibility classification.

3.3. Importance of feature selection in NN

Early methods commonly used sliding windows of residue sequence information as feature values for network inputs. Gradually, addition of physicochemical properties (PP) and evolutionary based structural information were also included in the model inputs as they showed to increase SS prediction accuracy. Faraggi et al. (2012) achieved 83.8% accuracy by including physicochemical properties like a steric parameter (graph shape index), hydrophobicity, volume, polarizability, isoelectric point, helix probability, and sheet probability, into the NN model SPINE-X. Their standard NNs were made of two hidden layers with 101 hidden units, and they also carried out 6 steps of iterative prediction of secondary structure, real-value residue solvent accessibility (RSA), and torsion angles. SPIDER3 (Heffernan et al., 2017) also successfully incorporates such physicochemical properties in the NN model to improve SS prediction. Another useful feature that successful methods employ is evolutionary based information. These are obtained from large databases through automatic searches, and include profiles such as position scoring matrices and hidden Markov models. Rost (2001) provides an informative background on the various databases and their usage in PSSP. The BLAST service (Altschul et al., 1997; BLAST, 2017) allows identification of similar sequences and structures that evidently enrich the network feature values by taking advantage of homologous proteins. Successful methods like SPIDER3 (Heffernan et al., 2017) and JPRED4 (Drozdetskiy et al., 2015; Cole et al., 2007, 2007; Cuff and Barton, 2000; Cuff et al., 1998) also include hidden Markov model (HMM) profiles to increase the feature space and improve SS prediction. The use of evolutionary information in PSSP has been comprehensively explored in Heringa (2000). The study shows that including extensively descriptive features in the NN improves prediction accuracy as shown in Table 2 (accuracies are obtained from the cited publications).

4. Discussion

The advancement in technology and growing activity in the area of bioinformatics research suggests that there is potential for improvement in the prediction of protein secondary structure in silico. Machine learning has recently leveraged areas such as online commerce, autonomous cars and speech and image recognition, so much so that most predictive instances achieve prediction accuracies of up to 99%. Additionally, the ab initio methods for obtaining protein data that accumulate into the protein databases are also expected to improve in reliability and timeliness as protein ab initio technology improves. As estimated at the beginning of the century (Rost, 2001), the 88% ceiling has not been reached and there is ample room for improving PSS prediction accuracy, especially given the recent advancements in machine learning techniques and technology.

Comparing the performance of methods is challenging as so many of them have been published in the past 3 decades and they differ significantly in their approaches and overall style of reporting. An overview of the methods and their performance is represented by Fig. 5. A collective initiative was made with EVA, a web-based server, to evaluate automatic structure prediction servers continuously and objectively (Koh et al., 2003). Critical Assessment of Techniques for Protein Structure Prediction (CASP), also a benchmark determining community-wide project, tested protein structure prediction methods including methods for PSSP. However, while CASP continues to assess protein structure prediction tools every two years, both of these initiatives abandoned PSSP-specific evaluation around the year 2004. Additionally, a comprehensive review has been carried out in Zhang et al. (2011) where methods using NN and other machine learning techniques have been evaluated. Challenges of carrying out a review of prediction methods include variation in size and content of datasets used by the methods, validation methods used, and convention chosen for denoting validation. The ideal configuration for comparing the various methods would be to test them all with the same test dataset and carry out the same validation techniques. It would take significantly long to complete the training and testing of each method unless powerful computational resources are available. For this reason, the accuracy levels claimed by the researchers have been assumed as accurate in this review. It can be seen through this study that NN has been explored as a promising strategy for PSSP since the identification of the PSSP problem itself, however, accuracy had plateaued around the 65% mark. It is the hybridization of the various NN techniques that have efficiently utilized

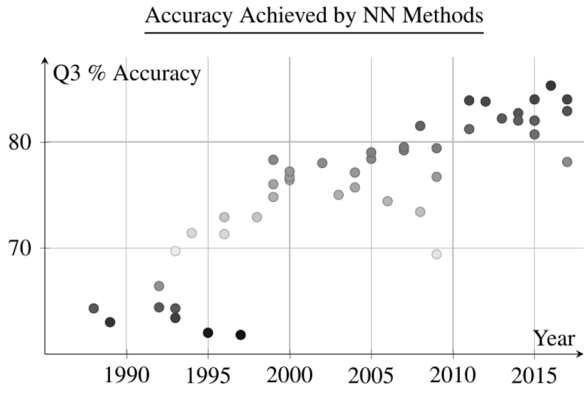

Fig. 5. The graph shows a timeline of methods for PSSP using NNs that were published since 1985, along with the three state accuracy they achieved. The methods have been extracted from literature based on PSSP and NN. (Qian and Sejnowski, 1988; Holley and Karplus, 1989) before 1990, (Stolorz et al., 1992; Zhang et al., 1992; Maclin and Shavlik, 1993; Reczko, 1993; Rost and Sander, 1993a; Rost et al., 1994; Vivarelli et al., 1995) between 1990 and 1995, (Riis and Krogh, 1996; Chandonia and Karplus, 1996, 1999; Vivarelli et al., 1997; Cuff et al., 1998; Baldi et al., 1999; Jones, 1999) between 1996 and 2000, (Cuff and Barton, 2000; Ouali and King, 2000; Petersen et al., 2000; Pollastri et al., 2002; Meiler and Baker, 2003; Wood and Hirst, 2004; Lin et al., 2005; Adamczak et al., 2005; Pollastri and McLysaght, 2005) between 2001 and 2005, (Bondugula and Xu, 2007; Dor and Zhou, 2007; Cole et al., 2007; Kakumani et al., 2008; Malekpour et al., 2009; Lakizadeh and Marashi, 2009; Babaei et al., 2010) between 2006 and 2010, (Wang et al., 2011, 2015; Qu et al., 2011; Faraggi et al., 2012; Mirabello and Pollastri, 2013; Yang et al., 2012; Yaseen and Li, 2014; Spencer et al., 2015; Patel and Mazumdar, 2014; Drozdetskiy et al., 2015) between 2011 and 2015, and (Li and Yu, 2016; Wang et al., 2017; Dongardive and Abraham, 2017; Heffernan et al., 2017) post 2015.

the growing databases to produce improving prediction accuracies. DCRNN (Li and Yu, 2016) and SPIDER3 (Heffernan et al., 2017) are relatively accurate PSSP methods and can be useful for designing improved new methods.

5. Conclusion

This review work has analyzed the progress of PSSP from the early ages of discovering 3 dimensional structures of protein molecules around the 1950s, to today's era of sophisticated machine learning. The emergence of the field has been explored as the first generation of single-residue statistics evolved into the second generation of segment statistics, which later evolved into multi-featured third generation of PSSP methods that utilize evolutionary information along with physicochemical and structural based information (Rost and Sander, 2000). It can be seen that improvement can be attributed to the advancement in the computer hardware, growth in algorithm efficiency, and expansion in the range of inputs that describe the polypeptide properties. The most recent improved methods lie in the 80-85% accuracy benchmark and owe the success to the large databases that provide large training datasets and also support the inclusion of expanding input features. The limit of the prediction accuracy has been estimated to be around 88% (Rost, 2001) which has still not been achieved, however, current progress in NN technology appears to be very promising.

References

Adamczak, R., Porollo, A., Meller, J., 2005. Combining prediction of secondary structure and solvent accessibility in proteins. Proteins 59 (3), 467–475. https://doi.org/10.1002/prot.20441.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Anfinsen, C.B., Haber, E., White, F.H., 1961. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc. Natl. Acad. Sci. USA 47 (9), 1309–1314.

Babaei, S., Geranmayeh, A., Seyyedsalehi, S.A., 2010. Protein secondary structure prediction using modular reciprocal bidirectional recurrent neural networks. Comput. Methods Progr. Biomed. 100 (3), 237–247. https://doi.org/10.1016/j.cmpb.2010.04.005.

Baldi, P., Brunak, S., Frasconi, P., Soda, G., Pollastri, G., 1999. Exploiting the past and the future in protein secondary structure prediction. Bioinformatics 15, 937–946.

Berg, J.M., Tymoczko, J.L., Stryer, L., 2002. Biochemistry, fifth edition. W. H. Freeman. http://ncbi.nlm.nih.gov/books/NBK21177/.

Bianchini, M., Scarselli, F., 2014. On the complexity of neural network classifiers: a comparison between shallow and deep architectures. IEEE Trans. Neural Netw. Learn. Syst. 25 (8), 1553–1565. https://doi.org/10.1109/TNNLS.2013.2293637.

Biou, V., Gibrat, J.F., Levin, J.M., Robson, B., Garnier, J., 1988. Secondary structure prediction: combination of three different methods. Protein Eng. 2, 185–191.

BLAST, 2017. http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE=Proteins&PROGRAM=blastp&RUN_PSIBLAST=on.

Bondugula, R., Xu, D., 2007. MUPRED: a tool for bridging the gap between template based methods and sequence profile based methods for protein secondary structure prediction. Proteins 66 (3), 664–670. https://doi.org/10.1002/prot.21177.

Bradley, E.K., Thomason, J.F., Cohen, F.E., Kosen, P.A., Kuntz, I.D., 1990. Studies of synthetic helical peptides using circular dichroism and nuclear magnetic resonance. J. Mol. Biol. 215 (4), 607–622. https://doi.org/10.1016/S0022-2836(05)80172-X.

Chandonia, J.M., Karplus, M., 1995. Neural networks for secondary structure and structural class predictions. Protein Sci. 4, 275–285.

Chandonia, J.M., Karplus, M., 1996. The importance of larger data sets for protein secondary structure prediction with neural networks. Protein Sci. 5, 768–774.

Chandonia, J.M., Karplus, M., 1999. New methods for accurate prediction of protein secondary structure. Proteins 35, 293–306.

Chou, P.Y., Fasman, G.D., 1974. Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. Biochemistry 13 (2), 211–222. https://doi.org/10.1021/bi00699a001.

Cohen, F.E., Abarbanel, R.M., Kuntz, I.D., Fletterick, R.J., 1983. Secondary structure assignment for alpha/beta proteins by a combinatorial approach. Biochemistry 22 (21), 4894–4904.

Cohen, F.E., Abarbanel, R., Kuntz, I.D., Fletterick, R.J., 1986. Turn prediction in proteins using a pattern-matching approach. Biochemistry 25, 266–275.

Cole, C., Barber, J.D., Barton, G.J., 2007. The Jpred 3 secondary structure prediction server. Nucleic Acids Res. 36, W197–W201. https://doi.org/10.1093/nar/gkn238.

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

Cuff, J.A., Clamp, M.E., Siddiqui, A.S., Finlay, M., Barton, G.J., 1998. JPred: a consensus secondary structure prediction server. Bioinformatics 14, 892–893.

Dongardive, J., Abraham, S., 2017. Reaching optimized parameter set: protein secondary structure prediction using neural network. Neural Comput. Appl. 28, 1947. https://doi.org/10.1007/s00521-015-2150-2.

Dor, O., Zhou, Y., 2007. Achieving 80 ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. Proteins 66 (4), 838–845. https://doi.org/10.1002/prot.21298.

Drozdetskiy, A., Cole, C., Procter, J., Barton, G.J., 2015. Jpred4: a protein secondary structure prediction server. Nucleic Acids Res. 43 (W1), W389–W394. https://doi.org/10.1093/nar/gkv332.

Eisenberg, D., 2003. The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins. Proc. Natl. Acad. Sci. USA 100, 11207–11210.

Ewbank, J.J., Creighton, T.E., 1992. Protein folding by stages. Cell 2, 347–349. https://doi.org/10.1016/0960-9822(92)90051-B.

Faraggi, E., Zhang, T., Yang, Y., Kurgan, L., Zhou, Y., 2012. SPINE X: improving protein secondary structure prediction by multi-step learning coupled with prediction of solvent accessible surface area and backbone torsion angles. J. Comput. Chem. 33, 259–267. https://doi.org/10.1002/jcc.21968.

Fasman, G.D., 1989. Protein conformational prediction. Trends Biochem. Sci. 14 (7), 295–299. https://doi.org/10.1016/0968-0004(89)90068-6.

Garnier, J., Osguthorpe, D.J., Robson, B., 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol. 120 (1), 97–120. https://doi.org/10.1016/0022-2836(78)90297-8.

Garratt, R.C., Thornton, J.M., Taylor, W., 1991. An extension of secondary structure prediction towards the prediction of tertiary structure. FEBS Lett. 280 (2), 401. https://doi.org/10.1016/0014-5793(91)80344-3.

Gibrat, J.F., Garnier, J., Robson, B., 1987. Further developments of protein secondary structure prediction using information theory. new parameters and consideration of residue pairs. J. Mol. Biol. 198, 423–443.

Heeb, D., 1949. The Organization of Behavior. Wiley.

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

Heffernan, R., Yang, Y., Paliwal, K., Zhou, Y., 2017. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. Bioinformatics 33 (18), 2842–2849. https://doi.org/10.1093/bioinformatics/btx218.

Heringa, J., 2000. Computational methods for protein secondary structure prediction using multiple sequence alignments. Curr. Protein Pept. Sci. 273–301.

Holley, L.H., Karplus, M., 1989. Protein secondary structure prediction with a neural network. Proc. Natl. Acad. Sci. USA 86, 152–156.

Jones, D.T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292 (2), 195–202. https://doi.org/10.1006/jmbi.1999.3091.

Kabsch, W., Sander, C., 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22 (12), 2577–2637. https://doi.org/10.1002/bip.360221211.
Kakumani, R., Devabhaktuni, V., Ahmad, M.O., 2008. A two-stage neural network based technique for protein secondary structure prediction. 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. https://doi.org/10.1109/IEMBS.2008.4649416.
Kanehisa, M., 1988. A multivariate analysis method for discriminating protein secondary structural segments. Protein Eng. 2, 87–92.
Ken, A.D., Justin, L.M., 2012. Predicting the secondary structure of globular proteins using neural network models. Science 1042 (338), 865–884. https://doi.org/10.1126/science.1219021.
Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H., Phillips, D.C., 1958. A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. Nature 181 (4610), 662–666. https://doi.org/10.1038/181662a0.
Kendrew, J.C., Dickerson, R.E., Strandberg, B.E., Hart, R.G., Davies, D.R., Phillips, D.C., Shore, V.C., 1960. Structure of myoglobin: a three-dimensional Fourier synthesis at 2 Å. resolution. Nature 185, 422–427. https://doi.org/10.1038/185422a0.
King, M.R.D., Sternberg, J.E., 1990. Machine learning approach for the prediction of protein secondary structure. J. Mol. Biol. 216, 441–457.
Kneller, D., Cohen, F., Langridge, R., 1990. Improvements in protein secondary structure prediction by an enhanced neural network. J. Mol. Biol. 214 (1), 171–182. https://doi.org/10.1016/0022-2836(90)90154-E.
Koh, I.Y.Y., Eyrich, V.A., Marti-Renom, M.A., Przybylski, D., Madhusudhan, M.S., Eswar, N., Graña, O., Pazos, F., Valencia, A., Sali, A., Rost, B., 2003. EVA: evaluation of protein structure prediction servers. Nucleic Acids Res. 31, 3311–3315.
Kozo, N., 1977. Triplet information in helix prediction applied to the analysis of super-secondary structures. J. Mol. Biol. 109 (2), 251–274. https://doi.org/10.1016/S0022-2836(77)80033-8.
Kwak, N., Choi, C.H., 2002. Input feature selection for classification problems. IEEE Trans. Neural Netw. 13 (1), 143–159. https://doi.org/10.1109/72.977291.
Lakizadeh, A., Marashi, S.-A., 2009. Addition of Contact Number Information can Improve Protein Secondary Structure Prediction by Neural Networks.
Levin, J.M., Garnier, J., 1988. Improvements in a secondary structure prediction method based on a search for local sequence homologies and its use as a model building tool. Biochim. Biophys. Acta 955, 283–295.
Levin, J.M., Robson, B., Garnier, J., 1986. An algorithm for secondary structure determination in proteins based on sequence similarity. FEBS Lett. 205, 303–308.
Levitt, M., Chothia, C., 1976. Structural patterns in globular proteins. Nature 261, 552–558.
Li, Z., Yu, Y., 2016. Protein secondary structure prediction using cascaded convolutional and recurrent neural networks. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI'16. AAAI Press. pp. 2560–2567.
Lim, V., 1974. Structural principles of the globular organization of protein chains. a stereochemical theory of globular protein secondary structure. J. Mol. Biol. 88, 857–872.
Lin, K., Simossis, V.A., Taylor, W.R., Heringa, J., 2005. A simple and fast secondary structure prediction method using hidden neural networks. Bioinformatics 21 (2), 152–159. https://doi.org/10.1093/bioinformatics/bth487.
Lyons, J., Dehzangi, A., Heffernan, R., Sharma, A., Paliwal, K., Sattar, A., Zhou, Y., Yang, Y., 2014. Predicting backbone alpha c angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. J. Comput. Chem. 35 (28), 2040–2046. https://doi.org/10.1002/jcc.23718.
Maclin, R., Shavlik, J.W., 1993. Using knowledge-based neural networks to improve algorithms: refining the Chou-Fasman algorithm for protein folding. Mach. Learn. 11 (2), 195–215. https://doi.org/10.1023/A:1022609403428.
Magnan, C.N., Baldi, P., 2014. SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. Bioinformatics 30 (18), 2592–2597. https://doi.org/10.1093/bioinformatics/btu352.
Malekpour, S.A., Naghizadeh, S., Pezeshk, H., Sadeghi, M., Eslahchi, C., 2009. Protein secondary structure prediction using three neural networks and a segmental semi Markov model. Math. Biosci. 217 (2), 145–150. https://doi.org/10.1016/j.mbs.2008.11.001.
Marqusee, S., Baldwin, R.L., 1987. Helix stabilization by Glu-… Lys+ salt bridges in short peptides of de novo design. Proc. Natl. Acad. Sci. USA 84, 8898–8902.
Marqusee, S., Robbins, V.H., Baldwin, R.L., 1989. Unusually stable helix formation in short alanine-based peptides. Proc. Natl. Acad. Sci. USA 86, 5286–5290.
Maxfield, F.R., Scheraga, H.A., 1979. Improvements in the prediction of protein backbone topography by reduction of statistical errors. Biochemistry 18 (4), 697–704. https://doi.org/10.1021/bi00571a023.
McCulloch, W.S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. 5 (4), 115–133. https://doi.org/10.1007/BF02478259.
Meiler, J., Baker, D., 2003. Coupled prediction of protein secondary and tertiary structure. Proc. Natl. Acad. Sci. USA 100 (21), 12105–12110. https://doi.org/10.1073/pnas.1831973100.
Minsky, M., 1954. Neural-Analog Networks and the Brain-Model Problem (Ph.D. thesis).
Mirabello, C., Pollastri, G., 2013. Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility. Bioinformatics 29 (16), 2056–2058. https://doi.org/10.1093/bioinformatics/btt344.
Moraes, I., Evans, G., Sanchez-Weatherby, J., Newstead, S., Stewart, P.D.S., 2014. Membrane protein structure determination – the next generation. Biochim. Biophys. Acta 1838 (1 Part A), 78–87. https://doi.org/10.1016/j.bbamem.2013.07.010.
Muggleton, S., King, R.D., Stenberg, M.J., 1992. Protein secondary structure prediction using logic-based machine learning. Protein Eng. Des. Sel. 5 (7), 647–657. https://doi.org/10.1093/protein/5.7.647.
Nagano, K., Hasegawa, K., 1975. Logical analysis of the mechanism of protein folding: III. Prediction of the strong long-range interactions. J. Mol. Biol. 94, 257–281.
Nagano, K., 1973. Logical analysis of the mechanism of protein folding. I. Predictions of helices, loops and beta-structures from primary structure. J. Mol. Biol. 75, 401–420.
Nakashima, H., Nishikawa, K., Ooi, T., 1986. The folding type of a protein is relevant to the amino acid composition. J. Biochem. 99, 152–162.
Niu, B., Scott, A.D., Sengupta, S., Bailey, M.H., Batra, P., Ning, J., Wyczalkowski, M.A., Liang, W., Zhang, Q., McLellan, M.D., Sun, S.Q., Tripathi, P., Lou, C., Ye, K., Mashl, R.J., Wallis, J., Wendl, M.C., Chen, F., Ding, L., 2016. Protein-structure-guided discovery of functional mutations across 19 cancer types. Nat. Genet. 48, 827–837.
The Nobel Prize in Chemistry 1962.
Oas, G.T., Kim, S.P., 1988. A peptide model of a protein folding intermediate. Nature 336, 42–48. https://doi.org/10.1038/336042a0.
Otaki, J.M., Ienaka, S., Gotoh, T., Yamamoto, H., 2005. Availability of short amino acid sequences in proteins. Protein Sci. 14, 617–625.
Otoom, M., 2016. Beyond von neumann: Brain-computer structural metaphor. 2016 Third International Conference on Electrical, Electronics, Computer Engineering and their Applications (EECEA) 46–51. https://doi.org/10.1109/EECEA.2016.7470764.
Ouali, M., King, R.D., 2000. Cascaded multiple classifiers for secondary structure prediction. Protein Sci. 9, 1162–1176. https://doi.org/10.1110/ps.9.6.1162.
Patel, M.S., Mazumdar, H.S., 2014. Knowledge base and neural network approach for protein secondary structure prediction. J. Theor. Biol. 361 (Suppl. C), 182–189. https://doi.org/10.1016/j.jtbi.2014.08.005.
Pauling, L., Corey, R.B., 1951. Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. Proc. Natl. Acad. Sci. USA 37, 729–740.
Pauling, L., Corey, R.B., Branson, H.R., 1951. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Natl. Acad. Sci. USA 214, 205–211.
Perutz, M.F., Rossmann, M.G., Cullis, A.F., Muirhead, H., Will, G., North, A.C.T., 1960. Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-Å. resolution, obtained by X-ray analysis. Nature 185, 416–422. https://doi.org/10.1038/185416a0.
Petersen, T.N., Lundegaard, C., Nielsen, M., Bohr, H., Bohr, J., Brunak, S., Gippert, G.P., Lund, O., 2000. Prediction of protein secondary structure at 80% accuracy. Proteins 41, 17–20.
Pollastri, G., McLysaght, A., 2005. Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics 21 (8), 1719–1720. https://doi.org/10.1093/bioinformatics/bti203.
Pollastri, G., Przybylski, D., Rost, B., Baldi, P., 2002. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins 47, 228–235.
Presnell, S.R., Cohen, B.I., Cohen, F.E., 1992. A segment-based approach to protein secondary structure prediction. Biochemistry 31, 983–993.
Ptitsyn, O.B., Finkelstein, A.V., 1983. Theory of protein secondary structure and algorithm of its prediction. Biopolymers 22 (1), 15–25.
Qian, N., Sejnowski, T.J., 1988. Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol. 202 (4), 865–884. https://doi.org/10.1016/0022-2836(88)90564-5.
Qu, W., Sui, H., Yang, B., Qian, W., 2011. Improving protein secondary structure prediction using a multi-modal BP method. Comput. Biol. Med. 41 (10), 946–959. https://doi.org/10.1016/j.compbiomed.2011.08.005.
Reczko, M., 1993. Protein secondary structure prediction with partially recurrent neural networks. SAR QSAR Environ. Res. 1 (2–3), 153–159. https://doi.org/10.1080/10629369308028826.
Richardson, J.S., 1981. The anatomy and taxonomy of protein structure. Adv. Protein Chem. 167–339.
Riis, S.K., Krogh, A., 1996. Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. J. Comput. Biol. 3, 163–183.
Roder, H., Elöve, G.A., Englander, S.W., 1988. Structural characterization of folding intermediates in cytochrome c by H-exchange labelling and proton NMR. Nature 335, 700–704. https://doi.org/10.1038/335700a0.
Rooman, M.J., Wodak, S.J., 1988. Identification of predictive sequence motifs limited by protein structure data base size. Nature 335, 45–49. https://doi.org/10.1038/335045a0.
Rooman, M.J., Wodak, S.J., Thornton, J.M., 1989. Amino acid sequence templates derived from recurrent turn motifs in proteins: critical evaluation of their predictive power. Protein Eng. 3, 23–27.
Rooman, M.J., Kocher, J.P., Wodak, S.J., 1991. Prediction of protein backbone conformation based on seven structure assignments. Influence of local interactions. J. Mol. Biol. 221, 961–979.
Rosenblatt, F., 1962. Principles of Neurodynamics. Spartan.
Rost, B., Sander, C., 1993a. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Mach. Learn. 90, 7558–7562.
Rost, B., Sander, C., 1993b. Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. 232 (2), 584–599. https://doi.org/10.1006/jmbi.1993.1413.
Rost, B., Sander, C., 2000. Third Generation Prediction of Secondary Structure. Humana Press.
Rost, B., Sander, C., Schneider, R., 1994. PHD-an automatic mail server for protein secondary structure prediction. Bioinformatics 10, 53–60. https://doi.org/10.1093/bioinformatics/10.1.53.
Rost, B., 2001. Review: protein secondary structure prediction continues to rise. J. Struct. Biol. 134 (2), 204–218. https://doi.org/10.1006/jsbi.2001.4336.
Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986a. Learning representations by back-propagating errors. Nature 323, 533–536. https://doi.org/10.1038/323533a0.

Rumelhart, D.E., Hinton, G., Williams, R., 1986b. Parallel Distributed Processing, vol. 1. MIT Press.
Sander, C., Schneider, R., 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9, 56–68.
Sanger, F., 1952. The Arrangement of Amino Acids in Proteins, Advances in Protein Chemistry, vol. 7. Academic Press, pp. 1–67. https://doi.org/10.1016/S0065-3233(08)60017-0.
Sanger, F., 1959. Chemistry of insulin. Science 129 (3359), 1340–1344. https://doi.org/10.1126/science.129.3359.1340.
Sasagawa, F., Tajima, K., 1993. Prediction of protein secondary structures by a neural network. Bioinformatics 9 (2), 147–152. https://doi.org/10.1093/bioinformatics/9.2.147.
Schulz, G.E., Schirmer, R.H., 1979. Principles of Protein Structure. Springer.
Spencer, M., Eickholt, J., Cheng, J., 2015. A deep learning network approach to ab initio protein secondary structure prediction. IEEE/ACM Trans. Comput. Biol. Bioinform. 12, 103–112. https://doi.org/10.1109/TCBB.2014.23439602.
Stolorz, P., Lapedes, A., Xia, Y., 1992. Predicting protein secondary structure using neural net and statistical methods. J. Mol. Biol. 225 (2), 363–377. https://doi.org/10.1016/0022-2836(92)90927-C.
Taylor, W.R., Orengo, C.A., 1989. Protein structure alignment. J. Mol. Biol. 208, 1–22.
Taylor, W.R., Thornton, J.M., 1983. Prediction of super-secondary structure in proteins. Nature 301 (5900), 540–542.
Udgaonkar, J.B., Baldwin, R.L., 1988. NMR evidence for an early framework intermediate on the folding pathway of ribonuclease A. Nature 335, 694–699. https://doi.org/10.1038/335694a0.
Umphred-Wilson, K., Fadden, A., Zanet, J., Mathews, K., Thurston, G., Mills, J., Michel, L.V., 2017. Thermodynamics of the gamma B crystallin protein demonstrated by T1/T2 NMR experiments. FASEB J. 31 (1 Suppl), 603–611. http://fasebj.org/content/31/1_Supplement/603.11.abstract.
Vivarelli, F., Giusti, G., Villani, M., Campanini, R., Fariselli, P., Compiani, M., Casadio, R., 1995. LGANN: a parallel system combining a local genetic algorithm and neural networks for the prediction of secondary structure of proteins. Comput. Appl. Biosci. 11, 253–260.
Vivarelli, F., Fariselli, P., Casadio, R., 1997. The prediction of protein secondary structure with a cascade correlation learning architecture of neural networks. Neural Comput. Appl. 6, 57–62. https://doi.org/10.1007/BF01670152.
Vriend, G., Sander, C., 1991. Detection of common three-dimensional substructures in proteins. Proteins 11 (1), 52–58. https://doi.org/10.1002/prot.340110107.
Wagner, I., Musso, H., 1983. New naturally occurring amino acids. Angew. Chem. Int. Ed. Engl. 22 (11), 816–828. https://doi.org/10.1002/anie.198308161.
Wang, Z., Zhao, F., Peng, J., Xu, J., 2011. Protein 8-class secondary structure prediction using conditional neural fields. Proteomics 11, 3786–3792. https://doi.org/10.1002/pmic.201100196.
Wang, S., Peng, J., Ma, J., Xu, J., 2015. Protein secondary structure prediction using deep convolutional neural fields. Cornell University Library: Biomolecules. http://arxiv.org/abs/1512.00843.
Wang, S., Li, W., Liu, S., Xu, J., 2016. Raptorx-property: a web server for protein structure property prediction. Nucleic Acids Res. 44 (W1), W430–W435. https://doi.org/10.1093/nar/gkw306.
Wang, Y., Mao, H., Yi, Z., 2017. Protein secondary structure prediction by using deep learning method. Knowl. Based Syst. 118 (Suppl. C), 115–123. https://doi.org/10.1016/j.knosys.2016.11.015.
Wood, M.J., Hirst, J.D., 2004. Predicting protein secondary structure by cascade-correlation neural networks. Bioinformatics 20 (3), 419–420. https://doi.org/10.1093/bioinformatics/btg423.
Wu, T.T., Kabat, E.A., 1973. An attempt to evaluate the influence of neighboring amino acids (n-1) and (n + 1) on the backbone conformation of amino acid (n) in proteins. use in predicting the three-dimensional structure of the polypeptide backbone of other proteins. J. Mol. Biol. 75 (1), 13–31.
Yang, J., Roy, A., Zhang, Y., 2012. Biolip: a semi-manually curated database for biologically relevant ligand-protein interactions. Nucleic Acids Res. 41, 1096–1103.
Yang, Y., Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., Sattar, A., Zhou, Y., 2014. Spider2: a package to predict secondary structure, accessible surface area, and main-chain torsional angles by deep neural networks. Predict. Protein Second. Struct. 55–63.
Yaseen, A., Li, Y., 2014. Context-based features enhance protein secondary structure prediction accuracy. J. Chem. Inf. Model. 54 (3), 992–1002. https://doi.org/10.1021/ci400647u.
Zhang, X., Mesirov, J.P., Waltz, D.L., 1992. Hybrid system for protein secondary structure prediction. J. Mol. Biol. 225 (4), 1049–1063. https://doi.org/10.1016/0022-2836(92)90104-R.
Zhang, H., Zhang, T., Chen, K., Kedarisetti, K.D., Mizianty, M.J., Bao, Q., Stach, W., Kurgan, L., 2011. Critical assessment of high-throughput standalone methods for secondary structure prediction. Brief. Bioinform. 12 (6), 672–688. https://doi.org/10.1093/bib/bbq088.
Zvelebil, M.J., Barton, G.J., Taylor, W.R., Sternberg, M.J., 1987. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. J. Mol. Biol. 195, 957–961.