Journal Pre-proof

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

Wafaa Wardah M.G.M. Khan Alok Sharma Mahmood A. Rashid

PII: S1476-9271(18)30501-2

DOI: <https://doi.org/doi:10.1016/j.compbiolchem.2019.107093>

Reference: CBAC 107093

To appear in: *Computational Biology and Chemistry*

Received Date: 12 July 2018

Revised Date: 28 December 2018

Accepted Date: 10 July 2019

Please cite this article as: Wafaa Wardah, M.G.M. Khan, Alok Sharma, Mahmood A. Rashid, Protein secondary structure prediction using neural networks and deep learning: a review, *Computational Biology and Chemistry* (2019), doi: <https://doi.org/10.1016/j.compbiolchem.2019.107093>

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2019 Published by Elsevier.

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*

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.

Keywords: Protein Secondary Structure Prediction; Machine Learning;

*Corresponding author

Email address: mahmood.rashid@griffith.edu.au (Mahmood A. Rashid)

Neural Network; Deep Learning; Feature Selection

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.

Background

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

Protein preliminaries

Proteins are macromolecules that carry out indispensable functions in essentially all biological processes that occur in the human body [1], 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 amino acids, only 20 of them are encoded by the genome for building the proteins required by the human body [2, 3]. 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 [4]. 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 [3]. 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 [5]. Another example is studying structure of Gamma B Crystallin proteins found in the eye lens to better understand the development of cataracts [6]. 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, Alzheimers, Parkinsons, Huntingtons, and amyotrophic lateral sclerosis (ALS) [7], 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 [8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19].

- *Primary Structure* of a protein refers to the linear sequence of amino acid residues that make up the protein [20]. 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 in [21] 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 [8, 22].
- *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)).

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 [23].

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 [24]. Due to these disadvantages, innovative approaches to predict protein structures, such as machine learning, have become the panacea. In the early years, Lim [25] proposed a method that utilized the physicochemical

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

characteristics of amino acids to predict protein structure. Later, a similar approach was also proposed in [26]. 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.

Neural network and deep learning preliminaries

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

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
[27]	1986	[28]	1973
[29]	1986	[30]	1974
[31]	1987	[32]	1977
[33]	1983	[34]	1979
[35]	1983	[36]	1987
[37]	1989	[38]	1973
[39]	1991	[40]	1975
[41]	1990	[42]	1979
[43]	1986	[44]	1988
[45]	1989	[46]	1988
[47]	1991	[48]	1988
[49]	1991	[50]	1989
[39]	1991	[51]	1991
[52]	1992	[53]	1992

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 [54, 55, 56, 57]. Recent popularization of brain-inspired architectures exhibits an acceptance and encouragement to continue further research and application of these methods in various industries [58].

Figure 2: The biological and artificial neuron are shown.

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 [59].

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 $x_0, x_1, x_2, \dots, x_n$). The selection of these features is challenging and further understanding on this can be obtained from [60]. 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 [61, 62]. 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.* [63, 64] and Perutz *et al.* [65] managed to establish the structure of proteins using x-ray crystallography around 1960 (for which Kendrew and Perutz later received the shared Nobel Prize in 1962 [66]), 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 [67, 68]. 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.

Figure 3: A shallow network is compared to a deep network configuration.

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.

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 [69] 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 in [70] 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

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

mutation matrix as the forth encoding option. The remaining 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. The authors [69] 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 Sejnowskis work in 1988 [71], where they adjusted the hidden layer, window size and encoding scheme. This strategy continued to be employed later in early 1990s as found in [72, 73, 74] and towards the end of the decade as mentioned in [75, 76]. 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 to 19.

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.

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 [77]. 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 [78]. 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 its improved version achieving 78% accuracy in 2002 by incorporating ensembles of the SSpro model [79]. In the next few years, Pollastri and McLysaght [79, 80] reached 79% accuracy using Porter - a further improved version of SSpro. Mirabello and Pollastri [77] created a further improved version of Porter [80] called Porter 4.0, which was based on ensembles of BRNN [78] 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 [81] in 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 [82], which is also the result of gradual improvements made to an initial work. The first method, SPIDER [83], was a NN model for predicting θ and τ angles in the protein backbone. SPIDER2 [84], 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 [85]. Finally, SPIDER3 was developed to consider the multiple features and predict secondary structure along with the other mentioned properties, 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%.

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 [86] created a CNN based method for PSSP and managed to achieve up to 84% accuracy. This method, DeepCNF (Deep Convolutional Neural Fields) consisted of two modules: Conditional Random Fields (CRF) module that was initially developed in [87] and deep convolutional NN (DCNN) module covering the input to the CRF [88]. 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.

Hybrid neural networks

A few recent PSSP techniques are comprised of multiple architectures intertwined together to improve the overall network prediction. Li and Yu [89] 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 [90] 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.

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 [91] 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 [82] 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. [92] provides an informative background on the various databases and their usage in PSSP. The BLAST service [93, 94] allows identification of similar sequences and structures that evidently enrich the network feature values by taking advantage of homologous proteins. Successful methods like SPIDER3 [82] and JPRED4 [95, 96, 96, 97, 98] 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 [99]. 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).

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 [71]	64.3%	✓		
PHD 1994 [100]	71.4%	✓	✓	
PSIPRED 1997 [76]	76.5%	✓	✓	
JPRED3 2008 [96]	81.5%	✓	✓	
SPIDER3 2017 [82]	84%	✓	✓	✓

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 [92], 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 Figure 5. A collective initiative was made with EVA, a web-based server, to evaluate automatic structure prediction servers continuously and objectively [101]. 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 [102] 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 the growing databases to produce improving prediction accuracies. DCRNN [89] and SPIDER3 [82] are relatively accurate PSSP methods and can be useful for designing improved new methods.

Figure 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. [71, 70] before 1990, [103, 104, 105, 106, 73, 100, 107] between 1990 and 1995, [108, 109, 110, 98, 111, 78, 76] between 1996 and 2000, [97, 112, 113, 79, 114, 115, 116, 117, 80] between 2001 and 2005, [118, 119, 96, 120, 121, 122, 123] between 2006 and 2010, [87, 124, 91, 77, 125, 126, 127, 128, 95, 86] between 2011 and 2015, and [89, 90, 69, 82] post 2015.

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 [129]. 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% [92] which has still not been achieved, however, current progress in NN technology appears to be very promising.

References

- [1] J. M. Berg, J. L. Tymoczko, S. L., Biochemistry, 5th Edition, W H Freeman, 2002.
URL <http://ncbi.nlm.nih.gov/books/NBK21177/>
- [2] I. Wagner, H. Musso, New naturally occurring amino acids, *Angewandte Chemie International Edition in English* 22 (11) (1983) 816–828.
doi:10.1002/anie.198308161.
URL <http://dx.doi.org/10.1002/anie.198308161>
- [3] F. Sanger, Chemistry of insulin, *Science* 129 (3359) (1959) 1340–1344.
doi:10.1126/science.129.3359.1340.
- [4] J. M. Otaki, S. Ienaka, T. Gotoh, H. Yamamoto, Availability of short amino acid sequences in proteins, *Protein Science NCBI* 14 (2005) 617625.
- [5] B. Niu, A. D. Scott, S. Sengupta, M. H. Bailey, P. Batra, J. Ning, M. A. Wyczalkowski, W. Liang, Q. Zhang, M. D. McLellan, S. Q. Sun, P. Tripathi, C. Lou, K. Ye, R. J. Mashl, J. Wallis, M. C. Wendl, F. Chen, L. Ding, Protein-structure-guided discovery of functional mutations across 19 cancer types, *Nature Genetics* 48 (2016) 827837.

- [6] K. Umphred-Wilson, A. Fadden, J. Zanet, K. Mathews, G. Thurston, J. Mills, L. V. Michel, Thermodynamics of the gamma b crystallin protein demonstrated by t1/t2 nmr experiments, *The FASEB Journal* 31 (1 Supplement) (2017) 603.11.
URL http://fasebj.org/content/31/1_Supplement/603.11.abstract
- [7] A. D. Ken, L. M. Justin, Predicting the secondary structure of globular proteins using neural network models, *Science* 1042 (338) (2012) 865–884. doi:10.1126/science.1219021.
- [8] L. Pauling, R. B. Corey, H. R. Branson, The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain, *Proc. Natl. Acad. Sci.* 214 (1951) 205–211.
- [9] L. Pauling, R. B. Corey, Configurations of polypeptide chains with favored orientations around single bonds: Two new pleated sheets., *Proceedings of the National Academy of Sciences of the United States of America* 37 (1951) 729740.
- [10] M. Levitt, C. Chothia, Structural patterns in globular proteins., *Nature* 261 (1976) 552–8.
- [11] J. Garnier, D. J. Osguthorpe, B. Robson, Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins, *Journal of Molecular Biology* 120 (1) (1978) 97–120. doi:www.doi.org/10.1016/0022-2836(78)90297-8.
URL <http://sciencedirect.com/science/article/pii/0022283678902978>
- [12] C. B. Anfinsen, E. Haber, F. H. White, The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain, *Proceedings of the National Academy of Sciences of the United States of America* 47 (9) (1961) 1309–14.
- [13] J. J. Ewbank, T. E. Creighton, Protein folding by stages, *Cell* 2 (1992) 347–349. doi:www.dx.doi.org/10.1016/0960-9822(92)90051-B.
- [14] E. K. Bradley, J. F. Thomason, F. E. Cohen, P. A. Kosen, I. D. Kuntz, Studies of synthetic helical peptides using circular dichroism and nuclear magnetic resonance, *Journal of Molecular Biology* 215 (4) (1990) 607–622.

doi:[www.doi.org/10.1016/S0022-2836\(05\)80172-X](http://www.doi.org/10.1016/S0022-2836(05)80172-X).

URL <http://sciencedirect.com/science/article/pii/S002228360580172X>

- [15] S. Marqusee, R. L. Baldwin, Helix stabilization by glu-...lys+ salt bridges in short peptides of de novo design, *Proceedings of the National Academy of Sciences of the United States of America* 84 (1987) 88988902.
- [16] S. Marqusee, V. H. Robbins, R. L. Baldwin, Unusually stable helix formation in short alanine-based peptides, *Proceedings of the National Academy of Sciences of the United States of America* 86 (1989) 52865290.
- [17] G. T. Oas, S. P. Kim, A peptide model of a protein folding intermediate, *Nature* 336 (1988) 42–48. doi:10.1038/336042a0.
- [18] H. Roder, G. A. Elve, S. W. Englander, Structural characterization of folding intermediates in cytochrome c by h-exchange labelling and proton nmr, *Nature* 335 (1988) 700–704. doi:10.1038/335700a0.
- [19] J. B. Udgaonkar, R. L. Baldwin, Nmr evidence for an early framework intermediate on the folding pathway of ribonuclease a, *Nature* 335 (1988) 694–699. doi:10.1038/335694a0.
- [20] F. Sanger, The arrangement of amino acids in proteins, Vol. 7 of *Advances in Protein Chemistry*, Academic Press, 1952, pp. 1–67. doi:[www.doi.org/10.1016/S0065-3233\(08\)60017-0](http://www.doi.org/10.1016/S0065-3233(08)60017-0).
- [21] W. Kabsch, C. Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers* 22 (12) (1983) 2577–2637. doi:10.1002/bip.360221211. URL <http://dx.doi.org/10.1002/bip.360221211>
- [22] D. Eisenberg, The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins., *Proceedings of the National Academy of Sciences of the United States of America* 100 (2003) 11207–10.
- [23] J. S. Richardson, The anatomy and taxonomy of protein structure, *Advances in Protein Chemistry* (1981) 167–339.
- [24] I. Moraes, G. Evans, J. Sanchez-Weatherby, S. Newstead, P. D. S. Stewart, Membrane protein structure determination the next generation, *Biochimica et Biophysica Acta (BBA) - Biomembranes* 1838 (1, Part A)

(2014) 78–87. doi:www.doi.org/10.1016/j.bbamem.2013.07.010.

URL <http://sciencedirect.com/science/article/pii/S0005273613002447>

- [25] V. Lim, Structural principles of the globular organization of protein chains. a stereochemical theory of globular protein secondary structure., J Mol Biol. 88 (1974) 857–872.
- [26] O. B. Ptitsyn, A. V. Finkelstein, Theory of protein secondary structure and algorithm of its prediction., Biopolymers. 22 (1) (1983) 15–25.
- [27] J. M. Levin, B. Robson, J. Garnier, An algorithm for secondary structure determination in proteins based on sequence similarity., FEBS Lett. 205 (1986) 303–8.
- [28] T. T. Wu, E. A. Kabat, 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) (1973) 13–31.
- [29] H. Nakashima, K. Nishikawa, T. Ooi, The folding type of a protein is relevant to the amino acid composition, J Biochem. 99 (1986) 152–62.
- [30] P. Y. Chou, G. D. Fasman, Conformational parameters for amino acids in helical, β -sheet, and random coil regions calculated from proteins, Biochemistry 13 (2) (1974) 211–222. arXiv:www.dx.doi.org/10.1021/bi00699a001, doi:10.1021/bi00699a001. URL <http://dx.doi.org/10.1021/bi00699a001>
- [31] M. J. Zvelebil, G. J. Barton, W. R. Taylor, M. J. Sternberg, Prediction of protein secondary structure and active sites using the alignment of homologous sequences., J Mol. Biol. 195 (1987) 957–61.
- [32] N. Kozo, Triplet information in helix prediction applied to the analysis of super-secondary structures, Journal of Molecular Biology 109 (2) (1977) 251–274. doi:www.doi.org/10.1016/S0022-2836(77)80033-8. URL <http://sciencedirect.com/science/article/pii/S0022283677800338>

- [33] F. E. Cohen, R. M. Abarbanel, I. D. Kuntz, R. J. Fletterick, Secondary structure assignment for alpha/beta proteins by a combinatorial approach., *Biochemistry*. 22 (21) (1983) 4894–904.
- [34] F. R. Maxfield, H. A. Scheraga, Improvements in the prediction of protein backbone topography by reduction of statistical errors, *Biochemistry* 18 (4) (1979) 697–704, pMID: 217422. arXiv:www.dx.doi.org/10.1021/bi00571a023, doi:10.1021/bi00571a023. URL <http://dx.doi.org/10.1021/bi00571a023>
- [35] W. R. Taylor, J. M. Thornton, Prediction of super-secondary structure in proteins., *Nature* 301 (5900) (1983) 540–2.
- [36] J. F. Gibrat, J. Garnier, B. Robson, Further developments of protein secondary structure prediction using information theory. new parameters and consideration of residue pairs., *J Mol Biol.* 198 (1987) 423–43.
- [37] M. J. Rooman, S. J. Wodak, J. M. Thornton, Amino acid sequence templates derived from recurrent turn motifs in proteins: critical evaluation of their predictive power., *Protein Eng.* 3 (1989) 23–7.
- [38] K. Nagano, Logical analysis of the mechanism of protein folding. i. predictions of helices, loops and beta-structures from primary structure., *J Mol Biol.* 75 (1973) 401–20.
- [39] M. J. Rooman, J. P. Kocher, S. J. Wodak, Prediction of protein backbone conformation based on seven structure assignments. influence of local interactions., *J Mol Biol.* 221 (1991) 961–79.
- [40] K. Nagano, K. Hasegawa, Logical analysis of the mechanism of protein folding iii. prediction of the strong long-range interactions., *J Mol Biol.* 94 (1975) 257–81.
- [41] M. R. D. King, J. E. Sternberg, Machine learning approach for the prediction of protein secondary structure, *J Mol Biol.* 216 (1990) 441–457.
- [42] G. E. Schulz, R. H. Schirmer, *Principles of Protein Structure*, Springer, 1979.
- [43] F. E. Cohen, R. Abarbanel, I. D. Kuntz, R. J. Fletterick, Turn prediction in proteins using a pattern-matching approach., *Biochemistry* 25 (1986) 266–75.

- [44] V. Biou, J. F. Gibrat, J. M. Levin, B. Robson, J. Garnier, Secondary structure prediction: combination of three different methods., *Protein Eng.* 2 (1988) 185–91.
- [45] W. R. Taylor, C. A. Orengo, Protein structure alignment, *J Mol Biol.* 208 (1989) 1–22.
- [46] M. Kanehisa, A multivariate analysis method for discriminating protein secondary structural segments., *Protein Eng.* 2 (1988) 87–92.
- [47] C. Sander, R. Schneider, Database of homology-derived protein structures and the structural meaning of sequence alignment., *Proteins* 9 (1991) 56–68.
- [48] J. M. Levin, J. Garnier, 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 (1988) 283–95.
- [49] G. Vriend, C. Sander, Detection of common three-dimensional substructures in proteins, *Proteins: Structure, Function, and Bioinformatics* 11 (1) (1991) 52–58. doi:10.1002/prot.340110107.
URL <http://dx.doi.org/10.1002/prot.340110107>
- [50] G. D. Fasman, Protein conformational prediction, *Trends in Biochemical Sciences* 14 (7) (1989) 295–299. doi:www.doi.org/10.1016/0968-0004(89)90068-6.
URL <http://sciencedirect.com/science/article/pii/0968000489900686>
- [51] R. C. Garratt, J. M. Thornton, W. Taylor, An extension of secondary structure prediction towards the prediction of tertiary structure, *FEBS Letters* 280 (2) (1991) 401. doi:www.doi.org/10.1016/0014-5793(91)80344-3.
URL <http://sciencedirect.com/science/article/pii/0014579391803443>
- [52] S. R. Presnell, B. I. Cohen, F. E. Cohen, A segment-based approach to protein secondary structure prediction., *Biochemistry* 31 (1992) 983–93.
- [53] S. Muggleton, R. D. King, M. J. Stenberg, Protein secondary structure prediction using logic-based machine learning, *Protein Engineering, De-*

- sign and Selection 5 (7) (1992) 647–657. doi:10.1093/protein/5.7.647.
URL <http://dx.doi.org/10.1093/protein/5.7.647>
- [54] W. S. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity, The bulletin of mathematical biophysics 5 (4) (1943) 115–133. doi:10.1007/BF02478259.
URL <http://doi.org/10.1007/BF02478259>
- [55] D. Heeb, The Organization of Behavior, Wiley, 1949.
- [56] M. Minsky, Neural-analog networks and the brain-model problem - phd thesis (1954).
- [57] F. Rosenblatt, Principles of Neurodynamics, Spartan, 1962.
- [58] M. Ootom, Beyond von neumann: Brain-computer structural metaphor, in: 2016 Third International Conference on Electrical, Electronics, Computer Engineering and their Applications (EECEA), 2016, pp. 46–51. doi:10.1109/EECEA.2016.7470764.
- [59] M. Bianchini, F. Scarselli, On the complexity of neural network classifiers: A comparison between shallow and deep architectures, IEEE Transactions on Neural Networks and Learning Systems 25 (8) (2014) 1553–1565. doi:10.1109/TNNLS.2013.2293637.
- [60] N. Kwak, C. H. Choi, Input feature selection for classification problems, IEEE Transactions on Neural Networks 13 (1) (2002) 143–159. doi:10.1109/72.977291.
- [61] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors, Nature 323 (1986) 533–536. doi:10.1038/323533a0.
URL <http://dx.doi.org/10.1038/323533a0>
- [62] D. E. Rumelhart, G. Hinton, R. Williams, Parrellal Distributed Processing, Vol. 1, MIT Press, 1986.
- [63] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, D. C. Phillips, A three-dimensional model of the myoglobin molecule obtained by x-ray analysis., Nature 181 (4610) (1958) 662–666. doi:10.1038/181662a0.

- [64] J. C. KENDREW, R. E. DICKERSON, B. E. STRANDBERG, R. G. HART, D. R. DAVIES, D. C. PHILLIPS, V. C. SHORE, Structure of myoglobin: A three-dimensional fourier synthesis at 2 . resolution, *Nature* 185 (1960) 422427. doi:10.1038/185422a0.
- [65] M. F. PERUTZ, M. G. ROSSMANN, A. F. CULLIS, H. MUIRHEAD, G. WILL, A. C. T. NORTH, Structure of hmoglobin: A three-dimensional fourier synthesis at 5.5-. resolution, obtained by x-ray analysis, *Nature* 185 (1960) 416422. doi:10.1038/185416a0.
- [66] The nobel prize in chemistry 1962.
- [67] M. J. Rooman, S. J. Wodak, Identification of predictive sequence motifs limited by protein structure data base size, *Nature* 335 (1988) 45–49. doi:10.1038/335045a0.
- [68] D. Kneller, F. Cohen, R. Langridge, Improvements in protein secondary structure prediction by an enhanced neural network, *Journal of Molecular Biology* 214 (1) (1990) 171–182. doi:www.doi.org/10.1016/0022-2836(90)90154-E.
URL <http://sciencedirect.com/science/article/pii/002228369090154E>
- [69] J. Dongardive, S. Abraham, Reaching optimized parameter set: protein secondary structure prediction using neural network, *Neural Comput & Applic* 28 (2017) 1947. doi:www.doi.org/10.1007/s00521-015-2150-2.
- [70] L. H. Holley, M. Karplus, Protein secondary structure prediction with a neural network, *Proc. Natl. Acad. Sci.* 86 (1989) 152–156.
- [71] N. Qian, T. J. Sejnowski, Predicting the secondary structure of globular proteins using neural network models, *Journal of Molecular Biology* 202 (4) (1988) 865–884. doi:www.doi.org/10.1016/0022-2836(88)90564-5.
URL <http://sciencedirect.com/science/article/pii/0022283688905645>
- [72] F. Sasagawa, K. Tajima, Prediction of protein secondary structures by a neural network, *Bioinformatics* 9 (2) (1993) 147–152. doi:10.1093/bioinformatics/9.2.147.
URL <http://dx.doi.org/10.1093/bioinformatics/9.2.147>

- [73] B. Rost, C. Sander, Improved prediction of protein secondary structure by use of sequence profiles and neural networks, *Machine Learning* 90 (1993) 75587562.
- [74] B. Rost, C. Sander, Prediction of protein secondary structure at better than 70% accuracy, *Journal of Molecular Biology* 232 (2) (1993) 584–599. doi:www.doi.org/10.1006/jmbi.1993.1413.
URL <http://sciencedirect.com/science/article/pii/S0022283683714130>
- [75] J. M. Chandonia, M. Karplus, Neural networks for secondary structure and structural class predictions, *Protein Science: A Publication of the Protein Society* 4 (1995) 275285.
- [76] D. T. Jones, Protein secondary structure prediction based on position-specific scoring matrices, *Journal of Molecular Biology* 292 (2) (1999) 195–202. doi:www.doi.org/10.1006/jmbi.1999.3091.
URL <http://sciencedirect.com/science/article/pii/S0022283699930917>
- [77] C. Mirabello, G. Pollastri, Porter, paleale 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility, *Bioinformatics* 29 (16) (2013) 2056–2058. doi:10.1093/bioinformatics/btt344.
- [78] P. Baldi, S. Brunak, P. Frasconi, G. Soda, G. Pollastri, Exploiting the past and the future in protein secondary structure prediction, *Bioinformatics* 15 (1999) 937–46.
- [79] G. Pollastri, D. Przybylski, B. Rost, P. Baldi, Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles, *Proteins* 47 (2002) 228–35.
- [80] G. Pollastri, A. McLysaght, Porter: a new, accurate server for protein secondary structure prediction, *Bioinformatics* 21 (8) (2005) 1719–1720. doi:10.1093/bioinformatics/bti203.
URL <http://dx.doi.org/10.1093/bioinformatics/bti203>
- [81] C. N. Magnan, P. Baldi, Sspro/accpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity, *Bioinformatics* 30 (18) (2014) 25922597. doi:www.doi.org/10.1093/bioinformatics/btu352.

- [82] R. Heffernan, Y. Yang, K. Paliwal, Y. Zhou, 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) (2017) 2842–2849. doi:10.1093/bioinformatics/btx218.
URL <http://dx.doi.org/10.1093/bioinformatics/btx218>
- [83] J. Lyons, A. Dehzangi, R. Heffernan, A. Sharma, K. Paliwal, A. Sattar, Y. Zhou, Y. Yang, Predicting backbone alpha c angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network, *Journal of Computational Chemistry* 35 (28) (2014) 2040–2046. doi:10.1002/jcc.23718.
URL <http://dx.doi.org/10.1002/jcc.23718>
- [84] Y. Yang, R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Zhou, Spider2: A package to predict secondary structure, accessible surface area, and main-chain torsional angles by deep neural networks, *Prediction of Protein Secondary Structure* (2014) 55–63.
- [85] R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Yang, Y. Zhou, Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning, *Scientific Reports* 5 (11476).
- [86] S. Wang, J. Peng, J. Ma, J. Xu, Protein secondary structure prediction using deep convolutional neural fields, *Cornell University Library: Biomolecules*.
URL <http://arxiv.org/abs/1512.00843>
- [87] Z. Wang, F. Zhao, J. Peng, J. Xu, Protein 8-class secondary structure prediction using conditional neural fields, *Proteomics* 11 (2011) 37863792. doi:www.doi.org/10.1002/pmic.201100196.
- [88] S. Wang, W. Li, S. Liu, J. Xu, Raptorx-property: a web server for protein structure property prediction, *Nucleic Acids Research* 44 (W1) (2016) W430–W435. doi:10.1093/nar/gkw306.
- [89] Z. Li, Y. Yu, 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, 2016, pp. 2560–2567.

- [90] Y. Wang, H. Mao, Z. Yi, Protein secondary structure prediction by using deep learning method, *Knowledge-Based Systems* 118 (Supplement C) (2017) 115–123. doi:www.doi.org/10.1016/j.knosys.2016.11.015.
URL <http://sciencedirect.com/science/article/pii/S0950705116304713>
- [91] E. Faraggi, T. Zhang, Y. Yang, L. Kurgan, Y. Zhou, Spine x: Improving protein secondary structure prediction by multi-step learning coupled with prediction of solvent accessible surface area and backbone torsion angles, *Journal of Computational Chemistry* 33 (2012) 259267. doi:www.doi.org/10.1002/jcc.21968.
- [92] B. Rost, Review: Protein secondary structure prediction continues to rise, *Journal of Structural Biology* 134 (2) (2001) 204–218. doi:www.doi.org/10.1006/jsbi.2001.4336.
- [93] S. F. Altschul, T. L. Madden, A. A. Schffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped blast and psi-blast: a new generation of protein database search programs, *Nucleic Acids Research* 25 (1997) 33893402.
- [94] Blast.
URL ["http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE=Proteins&PROGRAM=blastp&RUN_PSIBLAST=on"](http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE=Proteins&PROGRAM=blastp&RUN_PSIBLAST=on)
- [95] A. Drozdetskiy, C. Cole, J. Procter, G. J. Barton, Jpred4: a protein secondary structure prediction server, *Nucleic Acids Research* 43 (W1) (2015) W389–W394. doi:10.1093/nar/gkv332.
- [96] C. Cole, J. D. Barber, G. J. Barton, The jpred 3 secondary structure prediction server, *Nucleic Acids Research* 36 (2007) W197W201. doi:www.doi.org/10.1093/nar/gkn238.
- [97] J. A. Cuff, G. J. Barton, Application of multiple sequence alignment profiles to improve protein secondary structure prediction, *Proteins* 40 (2000) 502–11.
- [98] J. A. Cuff, M. E. Clamp, A. S. Siddiqui, M. Finlay, G. J. Barton, Jpred: a consensus secondary structure prediction server, *Bioinformatics* 14 (1998) 892–3.

- [99] J. Heringa, Computational methods for protein secondary structure prediction using multiple sequence alignments, *Current Protein & Peptide Science* (2000) 273–301.
- [100] B. Rost, C. Sander, R. Schneider, Phd-an automatic mail server for protein secondary structure prediction, *Bioinformatics* 10 (1994) 5360. doi:doi.org/10.1093/bioinformatics/10.1.53.
- [101] I. Y. Y. Koh, V. A. Eylich, M. A. Marti-Renom, D. Przybylski, M. S. Madhusudhan, N. Eswar, O. Graa, F. Pazos, A. Valencia, A. Sali, B. Rost, Eva: evaluation of protein structure prediction servers, *Nucleic Acids Research* 31 (2003) 33113315.
- [102] H. Zhang, T. Zhang, K. Chen, K. D. Kedarisetti, M. J. Mizianty, Q. Bao, W. Stach, L. Kurgan, Critical assessment of high-throughput standalone methods for secondary structure prediction, *Briefings in Bioinformatics* 12 (6) (2011) 672–688. doi:10.1093/bib/bbq088. URL <http://dx.doi.org/10.1093/bib/bbq088>
- [103] P. Stolorz, A. Lapedes, Y. Xia, Predicting protein secondary structure using neural net and statistical methods, *Journal of Molecular Biology* 225 (2) (1992) 363–377. doi:www.doi.org/10.1016/0022-2836(92)90927-C. URL <http://sciencedirect.com/science/article/pii/002228369290927C>
- [104] X. Zhang, J. P. Mesirov, D. L. Waltz, Hybrid system for protein secondary structure prediction, *Journal of Molecular Biology* 225 (4) (1992) 1049–1063. doi:www.doi.org/10.1016/0022-2836(92)90104-R. URL <http://sciencedirect.com/science/article/pii/002228369290104R>
- [105] R. Maclin, J. W. Shavlik, Using knowledge-based neural networks to improve algorithms: Refining the chou-fasman algorithm for protein folding, *Machine Learning* 11 (2) (1993) 195–215. doi:10.1023/A:1022609403428. URL <http://doi.org/10.1023/A:1022609403428>
- [106] M. Reczko, Protein secondary structure prediction with partially recurrent neural networks, *SAR and QSAR in Environmental Research* 1 (2-3) (1993) 153–159. doi:10.1080/10629369308028826. URL <http://doi.org/10.1080/10629369308028826>

- [107] F. Vivarelli, G. Giusti, M. Villani, R. Campanini, P. Fariselli, M. Compiani, R. Casadio, Lgann: a parallel system combining a local genetic algorithm and neural networks for the prediction of secondary structure of proteins, *Comput Appl Biosci*. 11 (1995) 253–60.
- [108] S. K. Riis, A. Krogh, Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments, *Journal of Computational Biology* 3 (1996) 163–183.
- [109] J. M. Chandonia, M. Karplus, The importance of larger data sets for protein secondary structure prediction with neural networks, *Protein Science: A Publication of the Protein Society* 5 (1996) 768774.
- [110] F. Vivarelli, P. Fariselli, R. Casadio, The prediction of protein secondary structure with a cascade correlation learning architecture of neural networks, *Neural Comput & Applic* 6 (1997) 5762. doi:www.doi.org/10.1007/BF01670152.
- [111] J. M. Chandonia, M. Karplus, New methods for accurate prediction of protein secondary structure, *Proteins*. 35 (1999) 293–306.
- [112] M. Ouali, R. D. King, Cascaded multiple classifiers for secondary structure prediction, *Protein Sci* 9 (2000) 1162–76. doi:10.1110/ps.9.6.1162.
- [113] T. N. Petersen, C. Lundegaard, M. Nielsen, H. Bohr, J. Bohr, S. Brunak, G. P. Gippert, O. Lund, Prediction of protein secondary structure at 80% accuracy, *Proteins* 41 (2000) 17–20.
- [114] J. Meiler, D. Baker, Coupled prediction of protein secondary and tertiary structure, *Proceedings of the National Academy of Sciences* 100 (21) (2003) 12105–12110. doi:10.1073/pnas.1831973100.
- [115] M. J. Wood, J. D. Hirst, Predicting protein secondary structure by cascade-correlation neural networks, *Bioinformatics* 20 (3) (2004) 419–420. doi:10.1093/bioinformatics/btg423.
- [116] K. Lin, V. A. Simossis, W. R. Taylor, J. Heringa, A simple and fast secondary structure prediction method using hidden neural networks, *Bioinformatics* 21 (2) (2005) 152–159. doi:10.1093/bioinformatics/bth487. URL <http://dx.doi.org/10.1093/bioinformatics/bth487>

- [117] R. Adamczak, A. Porollo, J. Meller, Combining prediction of secondary structure and solvent accessibility in proteins, *Proteins: Structure, Function, and Bioinformatics* 59 (3) (2005) 467–475. doi:10.1002/prot.20441.
- [118] R. Bondugula, D. Xu, Mupred: A tool for bridging the gap between template based methods and sequence profile based methods for protein secondary structure prediction, *Proteins: Structure, Function, and Bioinformatics* 66 (3) (2007) 664–670. doi:10.1002/prot.21177.
- [119] O. Dor, Y. Zhou, Achieving 80ten-fold cross-validated accuracy for secondary structure prediction by large-scale training, *Proteins: Structure, Function, and Bioinformatics* 66 (4) (2007) 838–845. doi:10.1002/prot.21298.
URL <http://dx.doi.org/10.1002/prot.21298>
- [120] R. Kakumani, V. Devabhaktuni, M. O. Ahmad, A two-stage neural network based technique for protein secondary structure prediction, *Engineering in Medicine and Biology Society. 30th Annual International Conference of the IEEE* doi:10.1109/IEMBS.2008.4649416.
- [121] S. A. Malekpour, S. Naghizadeh, H. Pezeshk, M. Sadeghi, C. Eslahchi, Protein secondary structure prediction using three neural networks and a segmental semi markov model, *Mathematical Biosciences* 217 (2) (2009) 145–150. doi:www.doi.org/10.1016/j.mbs.2008.11.001.
- [122] A. lakizadeh, S.-A. Marashi, Addition of contact number information can improve protein secondary structure prediction by neural networks (2009).
- [123] S. Babaei, A. Geranmayeh, S. A. Seyyedsalehi, Protein secondary structure prediction using modular reciprocal bidirectional recurrent neural networks, *Computer Methods and Programs in Biomedicine* 100 (3) (2010) 237–247. doi:www.doi.org/10.1016/j.cmpb.2010.04.005.
- [124] W. Qu, H. Sui, B. Yang, W. Qian, Improving protein secondary structure prediction using a multi-modal bp method, *Computers in Biology and Medicine* 41 (10) (2011) 946–959. doi:www.doi.org/10.1016/j.compbimed.2011.08.005.
- [125] J. Yang, A. Roy, Y. Zhang, Biolip: a semi-manually curated database for biologically relevant ligand-protein interactions., *Nucleic Acids Res* 41 (2012) 1096–103.

- [126] A. Yaseen, Y. Li, Context-based features enhance protein secondary structure prediction accuracy, *Journal of Chemical Information and Modeling* 54 (3) (2014) 992–1002. doi:10.1021/ci400647u.
- [127] M. Spencer, J. Eickholt, J. Cheng, A deep learning network approach to ab initio protein secondary structure prediction, *IEEE/ACM Transactions on Computational Biology and Bioinformatics / IEEE, ACM* 12 (2015) 103112. doi:www.doi.org/10.1109/TCBB.2014.23439602.
- [128] M. S. Patel, H. S. Mazumdar, Knowledge base and neural network approach for protein secondary structure prediction, *Journal of Theoretical Biology* 361 (Supplement C) (2014) 182–189. doi:www.doi.org/10.1016/j.jtbi.2014.08.005.
- [129] B. Rost, C. Sander, *Third Generation Prediction Of Secondary Structure*, Humana Press, 2000.