PREDICTING THE SECONDARY STRUCTURE OF PROTEINS BY CASCADING NEURAL NETWORKS

Maryam Alirezaee[1], Abdollah Dehzangi[2,3] and Eghbal Mansoori[1]

[1] School of Electrical & Computer Engineering, Shiraz University, Shiraz, Iran
alirezaiee@cse.shirazu.ac.ir
mansoori@shirazu.ac.ir
[2]Institute for Integrated and Intelligent Systems (IIIS), Griffith University, Brisbane, Australia
[3]National ICT Australia (NICTA), Brisbane, Australia
Abdollah.dehzangi@griffithuni.edu.au

ABSTRACT

Protein Secondary Structure Prediction (PSSP) is considered as a challenging task in bioinformatics and so many approaches have been proposed in the literature to solve this problem via achieving more accurate prediction results. Accurate prediction of secondary structure is a critical role in deducing tertiary structure of proteins and their functions. Among the proposed approaches to tackle this problem, Artificial Neural Networks (ANNs) are considered as one of the successful tools that are widely used in this field. Recently, many efforts have been devoted to modify, improve and combine this methodology with other machine learning methods in order to get better results. In this work, we have proposed a two-stage feed forward neural network for prediction of protein secondary structures. To evaluate our approach, it is applied on RS126 dataset and its results are compared with some other NN-based methods.

KEYWORDS

Protein Secondary Structure Prediction (PSSP), Artificial Neural Network (ANN), α-helix, β-sheet, coil, Position-Specific Scoring Matrix (PSSM) profile.

1. INTRODUCTION

Proteins are the main building blocks and functional molecules of the cell and play a key role in almost all biological processes. They take part in maintaining the structural integrity of the cell, transport and storage of small molecules, catalysis, regulation, signaling and the immune system [1]. Proteins are large, complex molecules consist of long amino acid chains. Different chemical properties of 20 amino acids cause the protein chains to fold up into complex shapes, and the shape of a protein determines its function.The secondary structure prediction is an intermediate step of deducing the 3D structure of proteins. Secondary structure is the local spatial arrangment of its polypeptide backbone ignoring the conformation of the individual sidechains (R groups). Secondary structures are held together by hydrogen bonds. The PSSP problem aims at predicting each amino acid in a protein sequence as α-helix (H), β-sheet (E) or neither (C) from its primary structure or indeed, the linear sequence of its amino acid structural units. X-ray crystallography, nuclear magnetic resonance (NMR) spectography and electron microscopy are in vitro methods, used to determine the 3D structure of proteins. Though they increasingly being applied in a high-troughput manner, but these methods are time consuming, expensive and not applicable to all proteins. Instead, comutational methods are widly used to predict the secondary structures as a step toward the prediction of 3D structure of proteins. During the last decades, many

DOI : 10.5121/ijaia.2012.3605

computational techniques have been proposed in the litrerature to tackle this problem. The earliest approaches for secondary structure prediction considered just single amino acid statistics and properties, and were limited to a small number of proteins with solved structures. While these early methods are not state-of-the-art, they are the basis of many subsequent approaches [1]. some of the most well-known early secondary structure prediction methods are being the Chou-Fasman method which uses a combination of statistical and heuristic rules [2], GOR method on the basis of information theory framework [3] and Lim method [4] as a stereochemical rule-based approach for predicting secondary structure in globular proteins. Since the secondary structure of each amino acid is affected by its neighbors through the interactions between the constituent amino acids along a protein chain, other methods try to consider higher-order intractions between residues. Information theoretic approaches such as an extension of original GOR method is proposed in [5] which considers the high-order residue interactions. Nearest neighbor approaches which incorporate local dependencies for predicting secondary structure of a residue [6,7], consider a window of residues surrounding it and classify each test residue according to the classification of neighbor residues in training samples. Neural network is another method to consider higher-order interaction between residues [8], [9], [10]. The first attempts to use neural networks for PSSP were made by Qian and Sejnowski [8]. They used a fully connected perceptron with at most one hidden layer and gained the accuracy of 64.3% that was more accurate than other priveous methods. Their work was followed by others in various ways in order to improve the prediction accuracy by applying sophisticated network structures [11], [12], [13], [14], [15]. Further improvement in preformance of PSSP were also acheived by exploiting evolutioary information via multiple sequence alignments (MSAs) profiles [10], Position-Specific Score Matrices (PSSM) [16], homology detection using hidden markov models [17] and PSI-BLAST [18]. Other approaches have been applied to PSS prediction are Support Vector Machines (SVM) [19], [20], [21] and ensemble methods that combine some machine learning methods [22] via the majority voting or weighted majority voting techniques. These methods could give up to a 3% improvement in Q_3 accuracy over the best individual method. In this work, we trained two sequential feed-forward neural networks in order to solve PSSP problem. The results of first ANN shows that predictions for secondary structure of amino acids near the boundary of going from one structure to another are usually unsuccessful, so we used another network to revise the outputs of first network and improve the prediction results. This network is not a structure to structure network have been used by Rost and Sander [10].

The remainder of this paper is organized as follows: section 2 is about related materials which describes dataset, secondary structure assignment, evaluation methods and PSSM generation. The classifier architecture and our proposed method are described in section 3. Section 4, presents and analyzes the exprimental results on RS126 dataset and finally, the conclusions are drawn in section 5.

2. RELATED MATERIALS

In this section we describe RS126 data set, secondary structure assignment methods, Measures of prediction accuracy and PSSM generation details.

2.1. Data Set Description

The dataset we used in this work is the RS126 set which was proposed by Rost and Sander [10]. It contains 126 non-homologous globular proteins which no pairs of proteins in the set have more than 25% similarity over a length of more than 80 residues [10]. RS126 contains 23,346 amino acids with 32% α-helix, 23% β-sheet and 45% coil. Average sequence length of all the proteins in this set is 185. This dataset was used in many researches in protein secondary structure prediction field [14], [23].

2.2. Secondary Structure Assignment

Given the 3D atomic coordinate of a protein structure, there are several methods to assign its secondary structures including dictionary of secondary structure of proteins (DSSP) [24], STRuctural IDEntification (STRIDE) [25] and DEFINE [26]. Since the secondary structure assignment of each residue is not completely well-defined, these methods often disagree on their assignments. For example, DSSP and STRIDE differ on approximately 5% of residues [1]. This inconsistency justifies the need for a certain and standard assignment method that could be used. The method was adopted here is DSSP as the standard algorithm and the most widely used secondary structure definition method. The DSSP specifies eight secondary structure classifications: H (α-helix), E (β- strand), G (3_{10} helix), I (π-helix), B (bridge), T (turn), S (bend) and C (other residues). The eight possible secondary structure specified by DSSP can be reduced to 3 classes [27]:

1. $\{H,I,G\}\rightarrow H(Helix)$, $\{E,B\}\rightarrow B(BetaSheet)$, $\mathrm{Re}st\{S,T,C\}\rightarrow C(Coil)$

2. $\{H,G\}\rightarrow H(Helix)$, $\{E\}\rightarrow B(BetaSheet)$, $\mathrm{Re}st\{S,T,B,I,C\}\rightarrow C(Coil)$

3. $\{H\}\rightarrow H(Helix)$, $\{E\}\rightarrow B(BetaSheet)$, $\mathrm{Re}st\{G,S,T,B,I,C\}\rightarrow C(Coil)$

4. $\{H,G\}\rightarrow H(Helix)$, $\{E,B\}\rightarrow B(BetaSheet)$, $\mathrm{Re}st\{S,T,I,C\}\rightarrow C(Coil)$

In this work we applied method 1.

2.3. Measures of Prediction Accuracy

We have used two measures to evaluate the prediction accuracy of the methods. The three state accuracy (Q_3) is defined as the percent of residues that have been predicted correctly:

$$Q_3 = \frac{n_H + n_E + n_C}{N_T} \tag{1}$$

Where n_H, n_E, n_C are the number of correctly predicted residues of type H, E and C, respectively and N_T is the total number of residues in dataset. Although Q_3 is a concise and useful measure to compare different methods, but for known protein structures it has shown that residues are approximately 30% in helices, 20% in sheets and 50% in coil. Because of this imbalanced characteristic of PSS datasets, Q_3 does not convey useful types of information. For example, a classifier that always predicts C has a Q_3 accuracy of 50%. It does not also indicate whether one type of structure is predicted more successfully than another [1], so we have also used Matthews correlation coefficient (MCC) [28] for each of the three secondary structures to judge the quality of prediction. This measure would be a value between −1 and +1 where +1 represents a perfect prediction, 0 no better than random prediction and −1 shows total disagreement between prediction and observation. So, coefficients closer to +1 represent better prediction. The Matthews correlation for a particular state is defined as:

$$c_i = \frac{TP_i \times TN_i - FP_i \times FN_i}{\sqrt{(TP_i + FP_i)\times(TP_i + FN_i)\times(TN_i + FP_i)\times(TN_i + FN_i)}} \tag{2}$$

$$i \in \{H,E,C\}$$

Where TP_i, TN_i, FP_i and FN_i are the number of correctly predicted (true positives), correctly rejected (true negatives), incorrectly predicted (false positives), incorrectly rejected (false negatives) residues, respectively and C_i indicates Matthews correlation coefficient for each three classes H, E and C.

2.4. PSSM Generation

To obtain Position-Specific Scoring Matrix (PSSM) profile for each protein sequence, we performed PSI-BLAST (Position-Specific Iterative Basic Local Alignment Search Tool) [18] against a non-redundant sequence (NR) data base. BLAST is the most widely used sequence similarity tool. PSI-BLAST is an iterative database searching method that uses homologous proteins found in an iteration to build a profile to be used for searching in the next iteration. This profile incorporates sequence weighting so that several closely-related homologs detected in the database do not overwhelm the contribution of more remote homologs [1]. The detected homologous proteins are then used as input into the neural network via the PSSM profile provided by PSI-BLAST. E-value threshold for inclusion is set to 0.001 and number of iteration to 3. The PSSM for each protein sequence has 20×L elements where L is the length of target sequence and each element represents the log-likelihood of particular residue substitution based on a weighted average of BLOSUM62 [29].

3. CLASSIFIER ARCHITECTURE AND CASCADING METHOD

In present work, we have used two sequential neural networks. Here, we briefly describe structure of each network. The first one is a feed-forward neural network with one hidden layer by including PSSM profiles as its inputs. Since the optimum length of window according to the RS126 set is 13 [10], [30], we used a sliding window of size 13 which moves through the protein sequence and the output of the network is attained for the residue in the middle of the window. As a result, the input layer includes 13×20=260 neurons (13 rows of PSSM including 20 elements are concatenated). We assumed two bipolar values to encode secondary structure classes:

H= [-1 1], E= [1 -1] and C= [-1 -1].

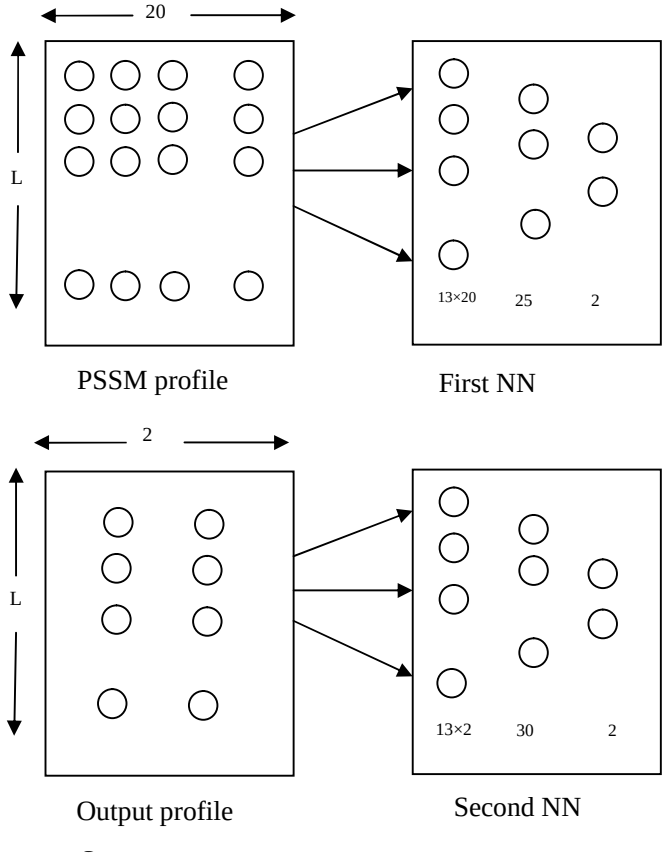
So the output layer has 2 neurons and the number of neurons in hidden layer is set to 25. After training this net, instead of using it as predictor to classify a test residue (by converting negative outputs of network to -1 and positive outputs to 1 and decoding them as predicted secondary structures), these real outputs are injected as inputs to another network. The outputs of first network show that some residues near to the boundary (i.e., where secondary structure changes, from H to C, for example) are influenced by neighbor residues and so wrongly get small positive or negative output values. To solve this problem, we have trained second network to revise the outputs of first network. In second network, in order to incorporate the effects of nearby residues, we have used a sliding window of length 13 to predict the secondary structure of residue in the middle of window. Since we have two output values for each residue from the first network, the input layer of second network would have 13*2=26 neurons. In other words, the output profile for each protein sequence with length L which is obtained from the first network and is used as inputs of second one is a matrix with L×2 elements. The second network has 2 neurons for output layer and 30 neurons are used for hidden layer. Figure 1 shows the proposed cascaded network where the structure of PSSM profile, output profile obtained from first network and the details of both networks are described.

In this figure, the outputs of second network are converted to -1 and +1 and are decoded as secondary structure components. However, since the length of consecutive helices, H, and strands, E, should be at least three and two, respectively, we have used the filtering method

proposed by Salamov and Solovyev [31] to make the predictions more realistic. So the following modifications are applied to the output of network in order to exclude physically unlikely structures:

$$[EHE]\rightarrow[EEE],\ [HEH]\rightarrow[HHH],\ [HCH]\rightarrow[HHH],\ [ECE]\rightarrow[EEE],\ [HEEH]\rightarrow[HHHH]$$

Also, all helices of length one or two and all strands of length 1 are converted to coils, C.



4. EXPERIMENTAL RESULTS

In this section to evaluate our method, we have used seven-fold cross-validation on the RS126 set. In this method the dataset is spitted to seven equal parts A, B, …, G and each part is in turn left out of the training set and used as test set. So, the networks are trained using 108 proteins and the rest 18 proteins are left for test in each turn. We have limited a maximum of 500 epochs for each network and in order to overcome the over fitting tendency of network on training samples, 20% of training residues are considered as validation set and 80% as training set. Table 1 and Table 2 report prediction accuracy for seven test groups when the outputs of first network are used for prediction and the cascaded network is used to refine the outputs. In these tables, Q_H, Q_E and Q_C are the percentage of correctly predicted residues observed in class E, H and C, respectively. The results show that Q_3 is improved in all seven groups using cascaded network. The average value Q_3= 73.28 is attained when the outputs of first network are used as predictions and after revising its outputs by cascaded network, Q_3= 75.22 is obtained. So, the second network can improve the average Q_3 by about 2%.

Table 3 and Table 4 illustrate MCCs for each test group. As results show, C_H, C_E and C_C (which indicate correlation coefficient of H, E and C, respectively), also are improved in each group by using cascaded network and so the second network refinement, makes better predictions for each of three secondary structures. Although the average percentage of correctly predicted residues of type coil decreases by second network, but its correlation coefficient improves such that we can say it makes better prediction for residues of type coil. The main reason for higher accuracy of coil structure by first network is that the network has tendency to predict a residue as coil (as the majority class) to obtain higher overall accuracy. However, the number of residues of type α-helix and β-sheet that are predicted as coil also increases. So, higher accuracy does not necessarily indicate better prediction.

Table 1. The accuracy of first network on seven test groups

Folds / Accuracies	A	B	C	D	E	F	G	AVG
Q_H	61.47	71.67	74.18	70.34	67.37	66.64	62.84	**67.79**
Q_E	54.29	49.54	51.95	59.53	45.99	57.60	47.59	**52.35**
Q_C	90.11	88.07	86.23	85.27	84.83	87.08	88.92	**87.21**
Q_3	72.92	74.44	75.14	74.58	70.51	73.84	71.51	**73.28**

Table 2. The accuracy of cascaded network on seven test groups

Folds / Accuracies	A	B	C	D	E	F	G	AVG
Q_H	69.27	76.61	77.79	75.86	72.40	73.14	70.14	**73.60**
Q_E	60.45	55.03	61.34	66.19	52.09	63.40	56.39	**59.27**
Q_C	87.73	86.11	83.59	80.73	82.65	82.92	85.22	**84.13**
Q_3	75.33	76.23	77.09	75.59	72.61	75.40	74.30	**75.22**

Table 3. The MCCs of first network on seven test groups

Folds / MCCs	A	B	C	D	E	F	G	AVG
C_H	0.64	0.68	0.68	0.68	0.63	0.64	0.61	**0.65**
C_E	0.57	0.53	0.56	0.56	0.46	0.57	0.49	**0.53**
C_C	0.50	0.53	0.54	0.52	0.46	0.53	0.50	**0.51**

Table 4. The MCCS of cascaded network on seven test groups

Folds / MCCs	A	B	C	D	E	F	G	AVG
C_H	0.70	0.71	0.70	0.70	0.66	0.66	0.65	**0.68**
C_E	0.59	0.57	0.61	0.60	0.50	0.60	0.53	**0.57**
C_C	0.54	0.56	0.57	0.53	0.50	0.55	0.54	**0.54**

Figure 2 visualizes comparison between single and cascaded network. This figure compares the average of Q_H, Q_E, Q_C and Q_3 of these two networks and shows that in single network there is high difference between prediction accuracy of beta-sheet and coil structure. This is because of class imbalance problem in protein secondary structure datasets which causes that the classifiers give more importance to majority class. But cascaded network have more balanced prediction of three secondary structures and less difference between prediction accuracy of H,E and C.

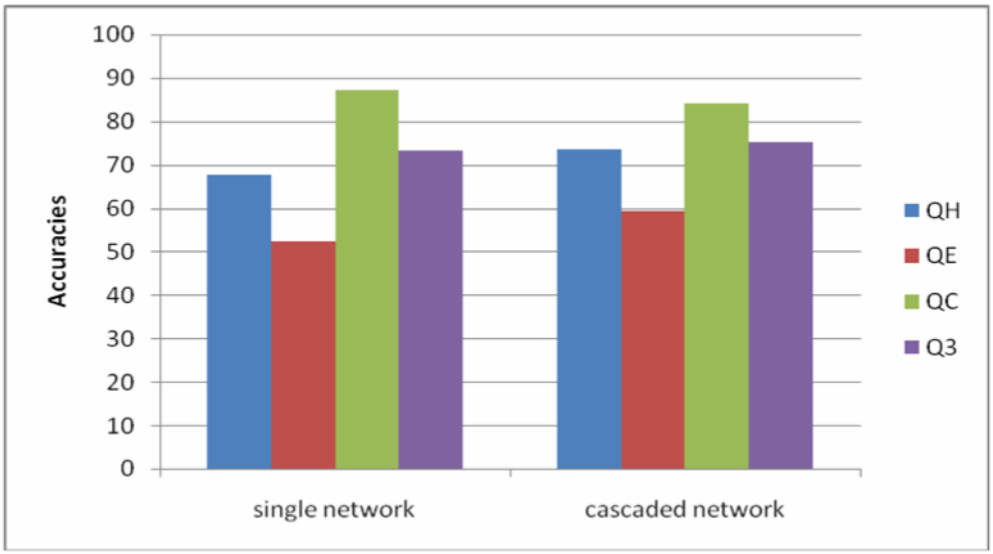


Figure 2. Q_H,Q_E and Q_C of single and cascaded network

In Table 5 we have compared the performance of cascaded network with some other NN-based methods such as Multi Layer Perceptron (MLP), Rost and Sander (PHD) [10], Bidirectional Recurrent Neural Network (BRNN) [32] and Bidirectional Segmented-Memory Recurrent Neural Network (BSMRNN) [14]. In this table, both prediction accuracy and correlation coefficient for helices (H), sheets (E) and coils (C) are included. As the results show, our proposed cascaded network performs better in most of the criteria; though its sheet accuracy is the worst.

Table 5. Performance comparison of some NN-based methods

Criterion / NN-based method	Q_H	Q_E	Q_C	Q_3	C_H	C_E	C_C
MLP	73.25	61.24	81.47	74.42	0.66	0.55	**0.56**
PHD	72.00	66.00	72.00	70.80	0.60	0.52	0.51
BRNN	70.50	59.70	77.50	71.40	0.62	0.50	0.52
BSMRNN	70.90	**67.50**	76.00	72.30	0.67	0.53	0.53
Cascaded NN	**73.60**	59.27	**84.13**	**75.22**	**0.68**	**0.57**	0.54

5. CONCLUSION

In this work, we proposed a two-stage feed-forward neural network for prediction of protein secondary structures. The results of neural networks on PSSP problem show unsuccessful predictions for secondary structure of amino acids near the boundary of going from one structure to another. So instead of converting the real outputs of this network to bipolar values which can be decoded to secondary structures, output profiles obtained from this network are fed to another network in order to revise the outputs and improve the prediction accuracy. The results show that this cascaded network can achieve better performance and can improve in prediction accuracy when compared to a single-stage neural network. Also The comparison between the performance of cascaded network with some other NN-based methods showed that our proposed cascaded network performs better in most of the criteria and has higher prediction accuracy.

Because of class imbalance problem in protein secondary structure datasets, the network tends to be biased towards the coil (majority class) and beta-sheet (minority class) samples are more likely to be misclassified. The comparison between the results of single and cascaded network showed that using cascaded network can decrease the difference between prediction accuracy of beta-sheet and coil structure and have more balanced prediction of three secondary structures.

REFERENCES

[1] M. Singh, "Predicting Protein Secondary and Supersecondary Structure," in Handbook of Computational Molecular Biology by S. Aluru, Chapman & Hall/CRC, 2006.

[2] P. Chou and G. Fasman, "Prediction of protein conformation," Biopolymers, vol. 13, no. 2, pp. 211-215, 1974.

[3] J. Garnier, D. Osguthorpe, and B. Robson, "Analysis and implications of simple methods for predicting the secondary structure of globular proteins," Journal of Molecular Biology, vol. 120, pp. 97-120, 1978.

[4] V. I. Lim, "Algorithms for prediction of alpha helices and structural regions in globular proteins," Journal of Molecular Biology, vol. 88, pp. 873-894, 1974.

[5] J. Gibrat, J. Garnier, and B. Robson, "Further developments of protein secondary structure prediction using information theory: new parameters and consideration of residue pairs," Journal of Molecular Biology, vol. 198, pp. 425-443, 1987.

[6] K. Nishikawa and T. Ooi, "Amino acid sequence homology applied to prediction of protein secondary structure and joint prediction with existing methods," Biochimica et Biophysica Acta, vol. 871, no. 1, pp. 45-54, 1986.

[7] J. Levin, B. Robson, and J. Garnier, "An algorithm for secondary structure determination in proteins based on sequence similarity," FEBS Letters, vol. 205, no. 2, pp. 303-308, 1986.

[8] N. Qian and T. Sejnowski, "Predicting the secondary structure of globular proteins using neural network models," Journal of Molecular Biology, vol. 202, no. 4, pp. 865-884, 1988.

[9] L. H. Holley and M. Karplus, "Protein secondary structure prediction with a neural net," Proceedings of the National Academy of Sciences (USA), vol. 86, pp. 152-156, 1989.

[10] B. Rost and C. Sander, "Prediction of protein secondary structure at better than 70%," Journal of Molecular Biology, vol. 232, pp. 584-599, 1993.

[11] L. Han, 1. Cuil, H. Lin, Z. Ji, Z. Cao,Y. Li, and Y. Chen, "Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity," Proteomics, vol. 6, pp. 4023-4037, 2006.

[12] A. Ceronia, P. Frasconia, and G. Pollastri, "Learning protein secondary structure from sequential and relational data," Neural Networks, vol. 18, pp. 1029-1039, 2005.

[13] J. Chen and N. Chaudhari, "Cascaded bidirectional recurrent neural networks for protein secondary structure prediction," IEEE Trans. Computational Biology and Bioinformatics, vol. 4, no. 4, 2007.

[14] J. Chen and N. Chaudhari, "Bidirectional segmented-memory recurrent neural network for protein secondary structure prediction," Soft Computing, vol. 10, pp. 315-324, 2006.

[15] H. Sui, B. Yang and W. Qian "Improving protein secondary structure prediction using a multi-modal BP method," Computers in Biology and Medicine, pp. 946-955, 2011.

[16] D. Jones, “Protein secondary structure prediction based on position-specific scoring matrices,” Journal of Molecular Biology, vol. 292, 1999.
[17] K. Karplus, C. Barret, and R. Hughey, “Hidden Markov models for detecting remote protein homologies,” Bioinformatics, vol. 14, pp. 846-856, 1998.
[18] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman, “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs,” Nucleic Acids Research, vol. 25, pp. 3389-3402, 1997.
[19] S. Hua and Z. Sun, “A novel method of protein secondary structure prediction with high segment overlap measure: Support vector machine approach,” Journal of Molecular Biology, vol. 308, pp. 397-407, 2001.
[20] P. Kountouris and J. Hirst, “Prediction of backbone dihedral angles and protein secondary structure using support vector machines,” BMC Bioinformatics, 2009.
[21] M. Nguyen, J. Rajapkse, “Two-stage multi-class support vector machines to protein secondary structure prediction,” Pac. Symp. Biocomput., vol. 10, pp. 346–57, 2005.
[22] R. King, M. Ouali, A. Strong, A. Aly, A. Elmaghraby, M. Kantardzic, and D. Page, “Is it better to combine predictions?,” Protein Engineering, vol. 13, pp. 15-19, 2000.
[23] H. Kim and H. Park, “Protein Secondary Structure Prediction Based on an Improved Support Vector Machines Approach,” Protein Engineering, vol. 16, pp. 553-560, 2003.
[24] W. Kabsch and C. Sander, “A dictionary of protein secondary structure,” Biopolymers, vol. 22, pp. 2577-2637, 1983.
[25] D. Frishman and P. Argos. “Knowledge-based secondary structure assignment,” Proteins: Structure, Function and Genetics, vol. 23, pp. 566-579, 1995.
[26] F. M. Richards and C. E. Kundrot, “Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure,” Proteins, vol. 3, pp. 71–84, 1988.
[27] P. Baldi and S. Brunak, Bioinformatics: The machine learning approach, Cambridge, MA, MIT Press, Second Edt., 2001.
[28] B. Matthews, “Comparison of the predicted and observed secondary structure of T4 phage lysozyme,” Biochimica et Biophysica Acta, vol. 405, pp. 442-451, 1975.
[29] S. Heniko and J. Henikoff, “Amino acid substitution matrices from protein blocks,” Proc. Natl. Acad. Sci. USA, vol. 89, pp. 10915–10919, 1992.
[30] M. Mirto, M. Cafaro, S. Luigi Fiore, D. Tartarini, and G. Aloisio, "A grid-enabled protein secondary structure predictor," IEEE Trans. Nanobioscience, vol. 6, no. 2, 2007.
[31] A. A. Salamov and V. V. Solovyev, “Prediction of protein secondary structure by combining nearest neighbor algorithms and multiple sequence alignments,” Journal of Molecular Biology, vol. 247, pp. 11–15, 1995.
[32] G. Pollastri, D. Przybylski, B. Rost and P. Baldi, “Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles,” Proteins, vol. 4, pp. 228–235, 2002.