Secondary Structure Prediction of Protein using Resilient Back Propagation Learning Algorithm

Jyotshna Dongardive
Dept. of Computer Science, University of Mumbai, Mumbai 400098, India
jyotss.d@gmail.com

Siby Abraham
Dept. of Mathematics and Statistics, Guru Nanak Khalsa College, University of Mumbai, Mumbai, India
sibyam@gmail.com

Abstract

The paper proposes a neural network based approach to predict secondary structure of protein. It uses Multilayer Feed Forward Network (MLFN) with resilient back propagation as the learning algorithm. Point Accepted Mutation (PAM) is adopted as the encoding scheme and CB396 data set is used for the training and testing of the network. Overall accuracy of the network has been experimentally calculated with different window sizes for the sliding window scheme and by varying the number of units in the hidden layer. The best results were obtained with eleven as the window size and seven as the number of units in the hidden layer.

Keywords: Resilient back propagation, point accepted mutation, sliding window, Q3, hidden units.

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 are large, complex molecules consisting of long amino acid chains. Though there are four different structural levels of proteins, tertiary structure prediction is of great interest to biologists because proteins perform their functions by coiling their amino acid sequences into specific three-dimensional shape (tertiary structure). It has its importance in medicine (e.g. drug design) and biotechnology.

In order to understand protein function at the molecular level, it is important to study the structure adopted by a particular sequence. The prediction of protein secondary structure is an important step in the prediction of protein tertiary structure. The usual structure prediction techniques such as X-ray (Sunde & Blake, 1997; Drenth, 1999), nuclear magnetic resonance (NMR) (Jaroniec et al., 2004; Wüthrich, 1986) and electron microscopy (EM) are time-consuming and expensive. However, due to increase in computing power and development of new algorithms, much progress has been made to overcome these problems through computational approaches (Hu et al., 2007).

Several computational approaches like meta predictor based and nearest neighbor methods have been developed to make and improve secondary structure prediction of proteins. Some of the computational methods that are used to achieve secondary structure predictions include statistical analysis (Chou & Fasman, 1974; Chou & Fasman, 1978), simple linear statistics, information theory (Garnier et al., 1996; Garnier et al., 1978; Gibrat et al., 1987), artificial neural networks (Jones, 1999; Chandonia & Karplus, 1999), k-way nearest neighbor (Salamov & Solovyev, 1995; Yi & Lander, 1993), linear discrimination (King & Sternberg, 1996), hydrogen bonding propensities (Frishman & Argos, 1997), conservation number weighted prediction (Zvelebil et al., 1987) and hybrid methods (Rost & Sander, 1993a; Rost & Sander, 1993b; Rost & Sander, 1994; Rost, 1996).

Of these, artificial neural network is the most often used method for secondary structure prediction. This work analyses the prediction of secondary structure of proteins from their sequences using a multi layer feed-forward neural network using resilient back propagation learning algorithm.

2. Related Work

A review of literature on computational techniques for secondary structure prediction using neural network indicates that multilayer feed forward neural networks are the most preferred and effective tool. The first attempt to secondary structure prediction using neural networks was made by Qian and Sejnowski (Qian & Sejnowski, 1988) using multilayered feed forward neural network. The proposed model consisted of 13 groups of 21 units, of which 20 were for amino acid and 1 was for the space between the sliding proteins. Two experiments were conducted. In the first, the network was trained using error back propagation algorithm with 40 units in the hidden layer using real set proteins, resulting in performance measure of 62%. In the second experiment, the performance of the network was improved slightly by using 17 groups of 21 units, 40 units in the hidden layer and 3 outputs units. Window size was varied from 1 to 21 and the peak performance was achieved with window size as 13.

Holley and Karplus (Holley & Karplus, 1989) used feed forward network in which input layer consisted of 17 groups of 21 units, the hidden layer with units varied from 0 to 20 and the output unit predicted the secondary structure. The success rate was improved by structured initialization of the synaptic weights and use of an asymmetric input window. However the success rate was still less compared to earlier methods.

Rost *et al.,* (Rost & Sander, 1994; Rost, 1996) used a basic network architecture, which has 40 units in the hidden layer and an input window of 13 amino acids. They addressed two problems- the first one, called over fitting, in which the training was stopped after the training error came below some threshold. The second problem, related to noise suppression, the arithmetic average was computed over predictions from several networks trained independently using different input information and training procedures.

Mejia and Fogelman-Soulie (Mejia & Fogelman-Soulie, 1990) improved the neural network performance by eliminating the hidden layer and synaptic weights during training. They also stated that accuracies of Holley and Karplus were due to the presence of homologies in the training set and that of Qian and Sejnowski was due to the size of the database they used.

Nanda *et al.,* (Sathya et al., 2001) used a first level secondary structure prediction network based on 'sliding window' approach to iteratively predict the secondary structure of each residue in the protein. The window size was 15 to 27, 22 units in input layer, output layer consisted of two units, H (Helix) and E (Strand). The output was compared to a cutoff value. If both H and E values were greater than the cutoff, coil was predicted as the secondary structure. Otherwise, the secondary structure corresponding to the larger of the two values was predicted. Also a second-level network was used to refine the results produced by the primary network.

Mottalib *et al.,* (Mottalib et al., 2010) obtained a prediction only for Helix (H) and Sheet (S). They used the feed-forward network architecture, which was built in java named Java Object Oriented Neural Engine (JOONE). The network consisted of 2 nodes in input layer, 3 nodes in hidden layer and 1 node in output layer. The learning rate was 0.9, momentum was 0.1 and 10000 epochs were considered for training of around 20 proteins. The network was tested for helix and sheet prediction and the accuracy was 71% and 65% respectively.

Agarwal *et al.,* (Pankaj, 2010) developed two learning rules for predicting the secondary structure of proteins. The first rule used feed forward back propagation method in which entire primary sequence was divided into patterns with window size changed from 5 and 11. The learning algorithm was then applied on these patterns, which after training were stored within the database of learnt patterns. The second rule used feed forward back propagation network using delta rule which provided a significant improvement in the number of iterations required to train the patterns when compared with first learning rule.

3. Materials and Methods

The work proposes to train the neural network to respond to the primary sequence of proteins whose secondary structures are known.

A. Dataset

The dataset used for this work is CB396. This dataset contains 396 non-redundant sequences derived from the 3Dee database created by Cuff and Barton (Cuff & Barton, 1999). It contains 396 proteins with their respective secondary structure as shown in Figure 1.

Figure 1: Snapshot of the CB396 dataset

The primary structure is a sequence of amino acids, which are represented by a one letter code. The secondary structures are made of 3 classes: H, E, C and rest marked a dash (−).

B. Encoding Scheme: PAM250

Feature extraction (Cuff & Barton, 1999) is a form of pre-processing in which the original variables are transformed into new inputs for classification. This initial process is important in protein structure prediction as the amino acids in the primary sequences are represented as single letter codes. It is therefore important to transform them into numbers. Different procedures can be adopted for this purpose. However, for the purpose of the present study, Point Accepted Mutations (PAM) coding is used to convert the letters into numbers.

The PAM matrix (Dayhoff et al., 1978) describes the probability that original amino acid will be replaced by another amino acid over a defined evolutionary interval. The unit of evolutionary divergence is defined as the interval in which 1% of the amino acids have been changed between two sequences. The work uses PAM250, which assumes the occurrence of 250-point mutations per 100 amino acids.

So, for the given the protein sequence GIVEQCCASVCSLYQLENYCN, A will be replaced by 1 -3 0 1 -3 -1 0 5 -2 -3 -4 -2 -3 -5 0 1 0 -7 -5 -1 as shown in Figure 2.

2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-3	1	1	1	-6	-3	0
-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2
0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2
0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2
-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2
0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2
0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2
1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1
-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2
-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4
-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2
-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2
-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2
-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1
1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6	1	0	-6	-5	-1
1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1
1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	-5	-3	0
-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6
-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2
0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4
0	-1	2	3	-4	1	3	0	1	-2	-3	1	-2	-4	-1	0	0	-5	-3	-2
0	0	1	3	-5	3	3	0	2	-2	-3	0	-2	-5	0	0	-1	-6	-4	-2
0	-1	0	-1	-3	-1	-1	-1	-1	-1	-1	-1	-1	-2	-1	0	0	-4	-2	-1
-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8

Figure 2. PAM250 matrix for the encoded sequence

C. Sliding Window

The input is controlled by window size, which determines how much local context information we want to consider in the prediction. The window size usually takes an odd length so that the amino acid at the centre of the window is predicted. Ideally, one may expect that the larger the window size, the more information given to the predictor, hence the performance should increase. Unfortunately, the increase of window size also means the increase of possible noises. It is observed that beyond some threshold size, the signal to noise ratio would decrease. Typical window sizes range from 9 to 25 residues (Vullo, 2002).

The steps used in the sliding window protocol steps used in the present study is as follows

- All the windows are extracted using linear indexing.
- They are loaded into a bigger array.
- They are then processed using vector operations.
- The indices of all the sliding windows of the matrix are obtained.
- Index, which is the centre most cell of that window, guarantees that the length of the windows be an odd number.

We have considered the sliding windows of the sizes from 11 to 19 over the entire dataset.

D. Feed Forward Neural Network using Resilient Back propagation Learning Algorithm

Neural networks have been trained to perform complex functions in various fields, including pattern recognition, identification, classification, and speech, vision and control systems. Most of the neural network models considered for protein predictions are of the feed forward type which has two passes through the network-the forward pass and the backward pass. For the forward pass, during training, a sample is presented to the network as input. For each layer, the output from the previous layer is used as an input to the next hidden layer until the output layer is reached and the output is produced. The output response is then compared to the known target output. Based on the value of the error, the connection weights are adjusted. In the backward pass, weights are adapted to ensure that the minimum error between the targets and the actual outputs is achieved (Haykin, 1994).

In this study, multilayer feed forward network with resilient back propagation (RPROP) (Riedmiller & Braun, 1993) learning algorithm is used. The algorithm performs a direct adaptation of the weight step based on local gradient information. In this, the effort of adaptation is not blurred by gradient behavior whatsoever, it only depends on the sign of the derivative not its value.

Therefore it will converge from ten to one hundred times faster than the simple back propagation algorithms.

E. Performance Measure Q_3

The performance measure Q_3 measures the expected accuracy of an unknown residue against the number of residues correctly predicted divided by the total number of residues.
The Q_3 is expressed as:

$$Q_3 = \frac{Q_H + Q_E + Q_C}{\text{Total number of residues}} * 100$$

Where Q_H, Q_E, and Q_C are defined as the total number of α-helix, β-strands and C-coil correctly predicted respectively.

F. Software

The software used for the experiments is Matlab Version 8.2.0.701 (R2013b). The Neural Network Toolbox Version 8.1 (R2013b) is used for the implementation of neural networks. The computer that was used to perform the experiments for model selection was an Intel(R) Core(TM) 2CPU6300@1.86GHz.

4. Results

The exhaustive experiments were conducted with window size ranges from 11 to 19 and number of neurons in the hidden layer ranges from 1 to 7. The detail analysis of the results is provided in the following subsections.

A. Dependence of testing success on window size

The best window size for the sliding window scheme is obtained by testing different window lengths from 11 to 19 as discussed in 3(C).

Table1. Dependence of testing success on window size

No	Window Size	Q_3
1	**11**	**64.0497**
2	13	62.8519
3	15	63.9995
4	17	63.0777
5	19	63.2094

Table 1 shows the dependence of testing accuracy rate on the size of the input window. This indicates the highest accuracy of the system is with window size as 11.

B. Dependence on the number of hidden units

The best number of hidden units is obtained by testing different hidden units from 1 to 7. Table 2 shows the peak performance on the testing set depending on the number of hidden units. The highest accuracy of the system was with 7 hidden units.

Table 2. Dependence on the number of hidden units

Hidden Neurons	Q_3
1	55.06
2	60.58
3	61.10
4	63.81
5	62.73
6	62.25
7	**64.04**

C. Prediction Accuracy (Q_3, Q_C, Q_E, Q_H)

The estimated accuracy for the α- helices (Q_H), β- strands (Q_E), C-coil states (Q_C), and three state together (Q_3) for the system is shown in Figure 3.

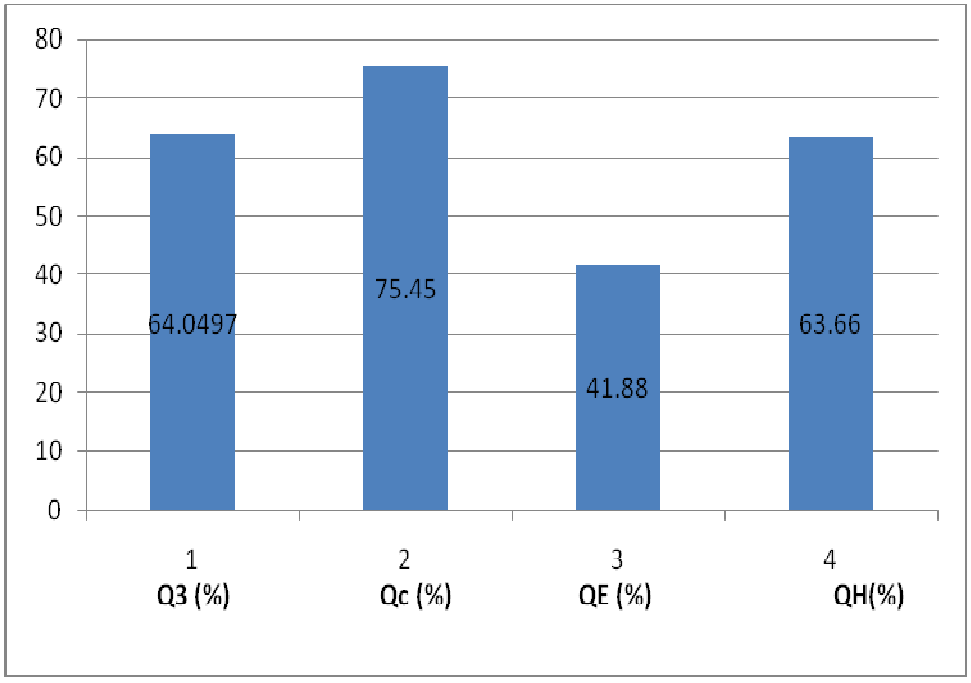

Figure 3. Distribution of Q_3 values

The overall accuracy of the system was 64.04%. The individual accuracy for the α-helices, β-strands, c-coils were 63%, 41% and 75% respectively. It is to be noted that this moderate accuracy is characteristic of secondary structure prediction structure as it is a fact that secondary structure prediction have a proven record of only around 60% accuracy. In that sense, the proposed method offers an incremental improvement with the existing methods. In addition, the method has not used any information concerning long-range interactions (Burgess & Scheraga, 1975) to increase the accuracy.

5. Conclusion and Future work

This paper deals with prediction of protein secondary structure using feed forward neural network using resilient back propagation learning algorithm. The model developed illustrates that window size 11 with 7 units in the hidden layer gives the highest accuracy of 64.04% over the selected dataset. The future work will primarily focus on looking at the feasibility of applying different encoding schemes and various learning algorithms of neural network to increase the overall accuracy.

6. References

Burgess, A.W and Scheraga, H.A. (1975). Assessment of some problems associated with prediction of the three-dimensional structure of a protein from its amino-acid sequence. Proc Natl Acad Sci U S A.72 (4):1221–1225.

Chandonia, J.M. and Karplus, M. (1999). New methods for accurate prediction of protein secondary structure. Proteins: Structure, Function and Genetics. 35: 293-306.

Chou, P. Y and Fasman, G.D. (1974). Prediction of protein conformation. Biochemistry. 13(2): 222-245.

Chou, P.Y and Fasman, G.D. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. Advanced Enzymology. 47: 45-148.

Cuff J. and Barton G. (1999).Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins: Structure, Function and Genetics. 34:508–519.

Dayhoff, M, Schwartz, R.M. and Orcutt, B.C. (1978). A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure. 5(3):345–358.

Drenth, J. (1999). Principles of Protein X-Ray Crystallography. 2nd ed. Springer-Verlag

Frishman, D. and Argos, P. (1997). 75% accuracy in protein secondary structure prediction accuracy. Proteins. 27:329-335

Garnier, J. Gibrat, JF. and Robson, B. (1996). GOR method for predicting protein secondary structure from amino acid sequence. Methods in Enzymology. 266: 540-553.

Garnier, J. Osguthorpe, D and Robson, B.(1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. Journal of Molecular Biology.120: 97-120.

Gibrat, JF. Garnier, J. and Robson, B. (1987). Further developments of protein secondary structure prediction using information theory, Journal of Molecular Biology. 198(3):425-443.

Haykin, S. (1994).Neural Networks: A Comprehensive Foundation, Macmillan, New York.

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

Hu, H.J. Harrison, R. W. Tai, P. C. and Pan, Y. (2007). Current Methods for Protein Secondary-Structure Prediction Based on Support Vector Machines In: *Knowledge Discovery in Bioinformatics: Techniques, Methods, and Applications,* eds Hu, X and Pan, Y. John Wiley & Sons, Inc., Hoboken, NJ, USA. doi: 10.1002/9780470124642.ch1.

Jaroniec, C. MacPhee, C. Bajaj, V. McMahon, M. Dobson, C and Griffin, R. (2004). High resolution molecular structure of a peptide in an amyloid fibril determined by magic angle spinning NMR spectroscopy. Proceedings of the National Academy of Sciences of the USA.101 (3): 711–716.

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

King, R.D and Sternberg, M.J.E. (1996). Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. Protein Science. 5: 2298-2310.

Mejia. C and Fogelman-Soulie, F. (1990). Incorporating knowledge in multi-layer networks: The example of protein secondary structure prediction In *Neurocomputing: Algorithms,*

Architectures, and Applications, Fogelman-Soulie F and Herault, J eds., (pp. 3-13), Springer-Verlag, Berlin.

Mottalib, M.A. Safiur Rahman Mahdi, Md. Zunaid Haque, A.B.M. Al Mamun, S.M and Hawlader Abdullah Al-Mamun. (2010). Protein Secondary Structure Prediction using Feed-Forward Neural Network. Journal of Cases on Information Technology (JCIT). 1(1):64-68.

Pankaj, A. Sakshi, J. Deepika. Swati, V. (2010). Secondary Structure Prediction Using ANN Learning. International Journal of Computer Science & Engineering Technology (IJCSET). 1(4):2229-3345.

Qian, N and Sejnowski, T. (1988). Predicting the secondary structure of globular proteins using neural network models. Journal of Molecular Biology. 202(4): 865-884.

Riedmiller, M. and Braun, H. (1993). Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm. In: Proceedings of the IEEE International Conference on Neural Networks. 586-591.

Rost, B and Sander, C. (1993a). Improved prediction of protein secondary structure by use of sequence profiles and neuronal networks. Proceedings of the National Academy of Science U.S.A. 90(16):7558-7562.

Rost, B and Sander, C. (1993b). Prediction of protein secondary structure at better than 70%. Journal of Molecular Biology. 232: 584-599.

Rost, B and Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. Proteins. 19: 55-72.

Rost, B. (1996). PHD: Predicting one-dimensional protein structure by profile based neural networks. Methods in Enzymology. 266:525-539.

Salamov, A.A and Solovyev, V.V. (1995) Prediction of protein secondary structure by combining nearest-neighbour algorithms and multiply sequence alignments. Journal of Molecular Biology. 247:11-15.

Sathya nanda, Saraai Deris and Rosli Md Illias. (2001). Prediction of protein secondary structure. Jurnal Teknologi 35(C): 81-90.

Sunde, M. Blake, C. (1997). The structure of amyloid fibrils by electron microscopy and X-ray diffraction. Advances in Protein Chemistry. 50: 123–159.

Vullo (2002). On the role of machine learning in protein structure determination. Journal of the Italian Association for Artificial Intelligence. XV (3):22-30.

Wüthrich K. (1986). NMR of Proteins and Nucleic Acids. John Wiley & Sons, New York.

Yi, T.M and Lander, S. (1993). Protein secondary structure prediction using nearest-neighbour methods. Journal of Molecular Biology.232 (4):1117-1129.

Zvelebil, M. Barton, G. Taylor, W and Sternberg, M. (1987). Prediction of protein secondary structure and active sites using the alignment of homologous sequences. Journal of Molecular Biology. 195(4): 957-961.