Computer and Information Science; Vol. 6, No. 1; 2013
ISSN 1913-8989 E-ISSN 1913-8997
Published by Canadian Center of Science and Education

The Hybrid Method of Fuzzy Feed-Forward Neural Network for Predicting Protein Secondary Structure

Sania Vahedian Movahed[1]

[1] School of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran

Correspondence: Sania Vahedian Movahed, School of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran. E-mail: saniavahedian@yahoo.com

Received: September 19, 2012 Accepted: December 4, 2012 Online Published: January 22, 2013

doi:10.5539/cis.v6n1p133 URL: http://dx.doi.org/10.5539/cis.v6n1p133

Abstract

With respect to the fact that the prediction of Protein secondary structure based on amino acids is very important, therefore, this study tries to present a new method based on the fuzzy combinational structure of a set of feed-forward neural networks so that the prediction accuracy of Protein secondary structure can be improved compared with the existing methods. Neural networks used in this paper are based on time windows; also, different methods have been established and trained to infer the three states of α- helix, β- sheet and coils from DSSP results, and finally, combining the results of the abovementioned networks in a fuzzy manner, the prediction method of Protein secondary structure based on neural network has been improved. It should be noted that in this paper, CB513 and RS126 data sets which are valid data sets in evaluating prediction methods of Protein secondary structure known in research studies in this area have been used to train and evaluate the proposed method.

Keywords: protein secondary structure, feed-forward neural network, amino acid, DSSP

1. Introduction

Protein secondary structure forms from a sequence of amino acid which is primary structure of the protein. It is generally a three dimension structure. The initial methods for predicting the secondary structure was presented in 1974. In this year, Chou and Fasman (1974) proposed a statistical method for this purpose. Shortly thereafter, in 1978, the method GOR (Garnier Osguthorpeb Robson) was proposed (Garnier & Osguthorpe, 1978). A common feature of these initial methods known as AB Initio was that they predicted the secondary structures of the amino acids of a protein with respect to the same protein independently of its protein family. In other words, for each sequence of the amino acids of a protein, a new structure is predicted. The predictions made in AB Initio methods are conducted through maximizing or minimizing a function that is experimentally obtained. It should be mentioned that the inaccuracy of this empirical function can be counted as one of the limitations of this method. Regarding the fact that in this method, the statistical parameters obtained from determined structures are used, the limited number of proteins with known structures at the time of presentation of these methods results in the inaccurate values of parameters and is counted as another limitation to these methods. The accuracies obtained from the application of these methods are around 50 to 60 percent. The next generations of prediction methods are stable from the 80s to the early 90th. In classification methods of this generation, they have tried to use a larger set of determined structures and also to engage more environmental information to obtain higher accuracy. Among these methods are algorithms GOR2, GOR3 and GOR4. It should be noted that the predictions of the secondary generation showed higher accuracy were compared with the predictions made of classification methods in the first generation. The prediction methods of protein secondary structure proposed since the 1990s, known as similarity-based methods are the third generation of prediction methods of protein secondary structure (Adamczak et al., 2004; Dor & Zhou, 2007; Rost & Sander, 1999). In this method, using multiple correspondence of amino acid sequences of a given protein with amino acids amino acid sequences of other proteins with determined structures; we tried to predict the nearest structure as an optimal structure. The main idea of this method is that proteins with more than 25% similar amino acids and sharing three-dimensional structures are usually similar to each other. Mention should be made that, the reverse is not true, and most of similar structures are only about 10% identical in amino acids. Regarding the idea used in the third generation in the prediction methods of protein secondary structure, the prediction accuracy of these methods in turn will increase compared to that of the secondary generation. In a recent study of this generation, in predicting secondary structures of amino acids in a protein, the family data to which this protein belongs we reused. In fact in such kinds of methods, the common feature of a family of proteins is used in predicting secondary structures. It should be mentioned that the protein family information is used through multiple correspondence and profile tables. Although in all prominent methods, the focus is on taking advantage of other features and also on improving the training algorithm, in the proposed method, it has been tried to increase the precision of predicting the secondary structure of protein, combining several conversion methods simultaneously along with applying neural network.

Protein secondary structure is one of the issues raised in the field of bioinformatics and biology and the process of its prediction is based on amino acids through using techniques of artificial intelligence. In this study in the second part, we provide an explanation of protein structures and their prediction. In the third section, the neural networks and the feed-forward status are presented. The fourth section of this paper deals with the proposed method of this study and presents an approach to predict protein secondary structure based on feed-forward neural networks and combining their results in a fuzzy manner, which increases the prediction accuracy compared with prediction methods of the secondary structure based on neural network. In the fifth section, the results of the application of the proposed method are investigated. A glance at the topics of the future researches for enhancing the efficiency of this method will be observed in the sixth section.

2. Protein Structures

Proteins have different structures which are named under the titles of the primary, secondary, etc. As follows, a brief description of each is provided. The first structure of a protein refers to the linearity or the sequence of amino acids in the polypeptide chain. The secondary structure is related to the conformation of hydrogen bonds between certain Residues in polypeptide chains which lead to regular and repetitive structure patterns. These structures include the alpha helix, beta- folded sheets (regular secondary structures) and coils (irregular secondary structures). Coils are conformed, in turn, from other sub-structures including irregular coils and circles. The third structure is the final formulation of polypeptide chain which is the very three-dimensional alignment of amino acids in it. The main factor in the conformation of this structure is the hydrophobic interactions between non-polar side chains and, in certain proteins, the disulfide bonds. The third structure explains for all the features of the three-dimensional folding of a polypeptide. Proteins that are composed of a single polypeptide chain (monomeric proteins), the third structure is the highest level of structure. In proteins which consist of two or more polypeptide chains or sub-units, the subunits are aligned together by non-covalently bonds whose spatial alignment is called the fourth structure (Lodish et al., 2007; Lehninger & Nelson, 2008; Naghavi, 2009).

The structure of proteins is generally described in four separate levels; these different levels are shown in Figure 1 (http://www.proteomesoftware.com).

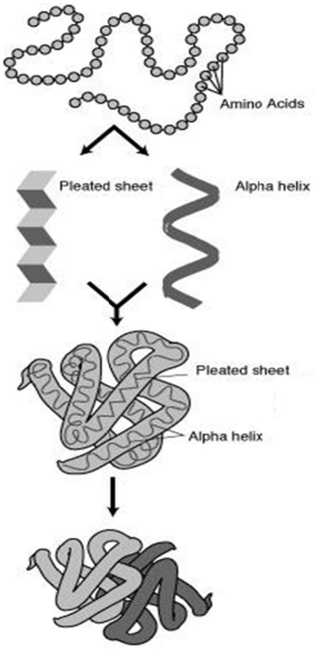

Figure 1. The protein structures from first to fourth

3. Neural Network

Since the nineteenth century, neurophysiologists simultaneously but separately tried to discover the training and analyzing system of the brain. In a parallel manner, mathematicians tried to build a mathematical model that has the ability to acquire and generally analyze the issues. The initial attempts at simulating were made through using a logical model by McCullouch and Walter Pitts which are today the basic building blocks of most of artificial neural networks. This model provides hypotheses about the function of neurons and the function of this model is based on the total of inputs and output. If the total of the inputs is greater than the threshold value, the neurons are, so to speak, stimulated. In addition to neurophysiologists, psychologists and engineers were influential in the development of simulation of neural networks, and in 1958, the perceptron network was introduced by Rosenblatt. In this network which had been presented in three-layer form, training ability had been provided.

One type of neural networks is the feed-forward network which has the same number of inputs and outputs and can be used as an associative memory. Information storage and retrieval is done based on content, not on the basis of the address; training methods are based on the Hebb and Delta rule (Gurney, 1997).

4. The Proposed Method

In the proposed method of this study, we tried to improve the accuracy in predicting protein secondary structure through using the conformation of different neural networks based on inference techniques from DSSP and converting them into an alpha helix structure, beta-sheets and coils. In the combination of these neural networks, the fuzzy theory has been applied, and the results obtained for each network are produced as membership degree for the determined class. As follows, the explanation associated with inference techniques and DSSP conversion into the three-state structure as well as details of the method is presented.

4.1 Inference Methods from DSSP

Regarding the fact that the structure of the results registered in the conducted experiments for accessing DSSP is different from the three-state structure of the protein secondary structure, therefore a method is needed to pre-process the experimental data for converting DSSP to the standard structure. For implementing the process of conversion, we can use the contents of Table 1.

Table 1. The methods for converting DSSP to the secondary structure

	Alpha substitute	Beta substitute	coil substitute
Method n^o 1	H,G,I	E	The rest
Method n^o 2	H,G	E,B	The rest
Method n^o 3	H,G	E	The rest
Method n^o 4	H	E,B	The rest
Method n^o 5	H	E	The rest

Regarding the fact that the nature of the conducted experiments for accessing DSSP is different from the secondary structure, we can select none of the above-mentioned methods as a superior one.

4.2 Data Set

The data sets used in this study are two sets CB513 (Cuff & Barton, 2000) and RS126 (Rost & Sander, 1999) which include 513 and 126 proteins with similar amino acids, respectively. The files placed in these two sets contain different information of proteins. Focusing on the DSSP information and amino acids chains, the proposed method in this study is applied for the training process of the network and extracting the results.

4.3 The Algorithm of the Proposed Method

As mentioned previously, we cannot select a method as a superior one from among the methods converting DSSP to the secondary structure. Regarding this fact, we have tried, in this method, to train the feed-forward neural networks based on the above-mentioned five conversion methods. The application of this method was done in the software environment of MATLAB. In this process, five neural networks were produced in form of feed-forward, and the data set were pre-processed based on the methods of converting DSSP to the secondary structure. Out of the converted data sets, 20% of them were separated for testing and the rest were used proportionately 80% for training and 20% for network validating. Furthermore, mentioned should be made that in the pre-processing, the binarization operation of the data sets was conducted too, and the input and output of the networks are in a binary manner. After carrying out the training process, the output of the networks which is indicative of the degree of membership for each of the three secondary structures are totaled in a fuzzy manner and through maximizing the obtained sum, the prediction of the secondary structures will be done.

Mention should be made that the results registered in the results section have been obtained based on averaging the training states of the network with different parameters for window size and also the number of the mid-layers beside using the method 10-folds.

Regarding the structure and architecture of this method, as it was mentioned, the results obtained in the method was started using a set of neural networks that based on the specific structure, each was gone under training process and the final result was provided by the end summary of the results of each network. Figures 2 and 3 are the displayers of the training structure and the test of the method, respectively that explanations related to each have been presented in the following.

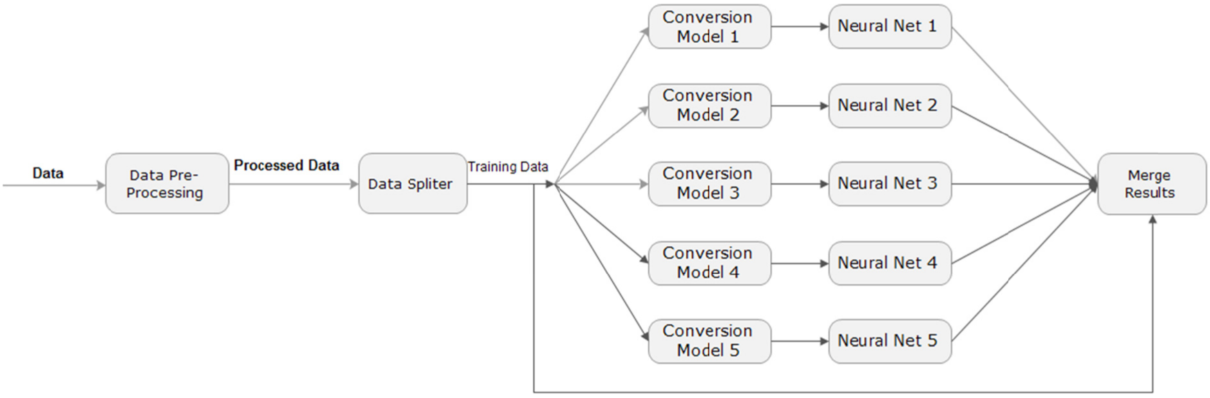

Figure 2. The training structure

As it is observable in Figure 2 as the training structure, the studied data including samples of the data set of 126RS and cb513 was placed under Pre - processing operations as input and titled Data and based on the determined window, data binary set is created. After finishing Pre-processing, based on the desired percentages for Train, Validation and Test, data is separated in separation phase In the next step and applying 5 methods of protein structure inference from amino acids, 5 data sets are provided in Conversion phase.The output of this phase that is in the form of binary data sets will be used as the neural network input and with regard to their training. After finishing neural networks training, in section Merge Result, the results of each of the five networks are combined with each other and offer the output of this method. It should be noted that although this method has used Voting method as integrator, it will be possible to improve the accuracy of Integration using another mechanism as neural network that in this way, a set of data is needed for a new integrator.

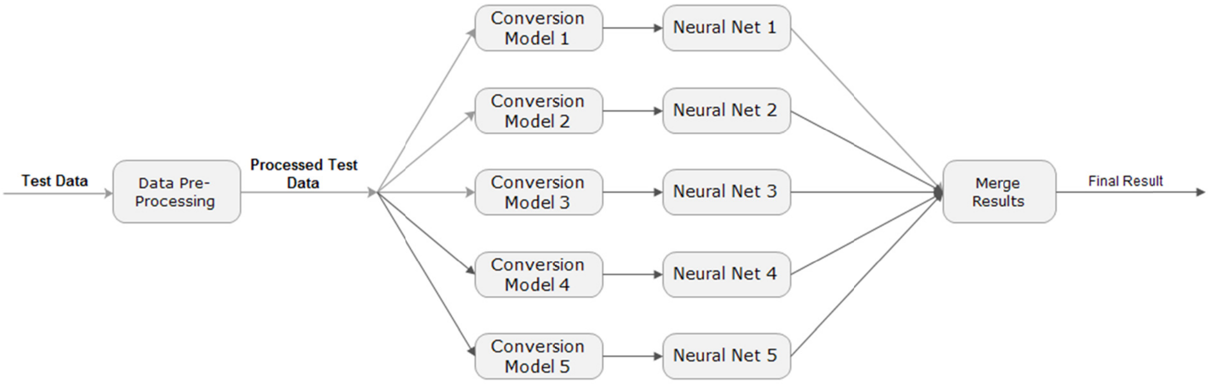

Figure 3. The test structure

As it can be seen in the test phase, after preprocessing process on input data and converting it into a binary structure and separating based on the determined time window by the user, the data are separated based on 5 mentioned methods and are given to neural networks. The result of the five networks is transferred into merging elements and the end result of this method is presented.

5. Results

The results obtained from the application of the proposed method in the present study on the two data sets CB513 and RS126. After pre-processing operation, the process of separating the data into two sections of training and testing with proportions of 80% and 20%, respectively and randomly with 10 times repetition was carried out so that the functionality of the proposed method can be proved. The results obtained from using the neural network are singly observable based on the methods of converting DSSP into the secondary structure for the data sets CB513 and RS126 and with the window time of 17, respectively in Tables 2 and 3.

Table 2. The results of the neural network on the data set RS126

	Prediction accuracy (%)
Conversion method n^{o} 1	64
Conversion method n^{o} 2	62.4
Conversion method n^{o} 3	63.9
Conversion method n^{o} 4	64.2
Conversion method n^{o}5	65.4

Table 3. The results of the neural network on the data set CB513

	Prediction accuracy (%)
Conversion method n^{o} 1	63.7
Conversion method n^{o} 2	63.5
Conversion method n^{o} 3	64
Conversion method n^{o} 4	64.6
Conversion method n^{o}5	65.2

The results obtained from the application of the proposed method of this study on data sets CB513 and RS126 are listed in Tables 4 and 5.

Table 4. The results of the proposed method on the data set RS126

	Alpha accuracy	Beta accuracy	Coil accuracy	Efficiency improvement %
Conversion method n^{o}1	0.7689	0.7608	0.7963	13.4
Conversion method n^{o}2	0.7608	0.7616	0.7850	14.4
Conversion method n^{o}3	0.7689	0.7608	0.7963	13.5
Conversion method n^{o}4	0.7763	0.7594	0.7792	12.9
Conversion method n^{o}5	0.7841	0.7560	0.7924	12.3

Table 5. The results of the proposed method on the data set CB513

	Alpha accuracy	Beta accuracy	Coil accuracy	Efficiency improvement %
Conversion method n°1	0.7501	0.7942	0.7765	13.6
Conversion method n°2	0.7460	0.7510	0.7676	11.9
Conversion method n°3	0.7502	0.7542	0.7763	12
Conversion method n°4	0.7552	0.7566	0.7736	11.5
Conversion method n°5	0.7602	0.7604	0.7811	11.5

5.1 Comparison with other Methods

The comparison between the results of the proposed method of the present study and the methods common in the field for the data sets CB513 and RS126 is shown in Tables 6, and 7.

Table 6. The comparison of the results on the data set RS126

	Alpha accuracy	Beta accuracy	Coil accuracy
The method presented by W.Qu	84.35	72.62	83.01
The method SSNet2 by S. Babaei	71.93	74.47	78.82
The method SSNet by S. Babaei	72.23	74.19	78.39
The proposed method of this study	74.41	75.60	79.24

Since in the study carried out by S. Babaei (Babaei et al., 2008) the proposed method is implemented only on the data set RS126, the results corresponding to the implementation of the method on CB513 are not listed in Table 7.

Table 7. The comparison of the results on the data set CB513

	Alpha accuracy	Beta accuracy	Coil accuracy
The method presented by W.Qu	83.52	70.31	82.16
The proposed method of this study	75.01	79.42	77.65

According to the table 6, the results of the implementation of the proposed method of the present study on the data set RS126 are better than those obtained from S. Babaei's method in each case of Alpha, Beta, and Coil.

As to the comparison between the proposed method of the present study and that presented by Qu et al. (2011), it should be noted that in the latter, is not even and only in the alpha and beta does it increase and the recognition of beta is done with a low accuracy, whereas the improvement of the accuracy in the method proposed in the present study is even between three classes.

Mention should be made that this difference is observable in the listed figures in tables 6 and 7 proportionately with the data sets CB513 and RS126.

Regarding the even improvement of the results obtained from the method proposed in the present study, more improvement through combining it with the method presented by Qu and accessing more accuracy in predicting the protein secondary structure will be possible.

6. Application

The main idea of the method proposed in the present study is based on using a fuzzy combination of training neural networks according to data with different pre-processing and little attention is paid to optimize the parameters involving each of the networks. Regarding this fact, the development and improvement of the results through focus on optimization of the parameters of network training will be possible which should be taken into account in the future researches. In the method proposed in the present study, the results are obtained in a stable manner according to the size of time windows. Using changing time windows and specifying the appropriate parameters for determining the appropriate value of time window in each span can be considered as another idea beside the application of the method of back-up vector machines in a fuzzy manner for developing the task.

7. Conclusion

In this study we presented a modern and combinatorial method for predicting the protein secondary structure based on the sequence of amino acids which improved the efficiency of the prediction. The focus of this study is a combination of methods and introducing the concept of fuzzy while paying little attention to the optimization of the network parameters. Thus, through the development in optimizing the neural networks, we will be able to better improve the efficiency of prediction that will be the canon of focus in the future studies.

References

Adamczak, R., Meller, J., & Porollo, R. (2004). Combining Prediction of Secondary Structure and Solvent Accessibility in Proteins. *Proteins, 59*, 467-475. http://dx.doi.org/10.1002/prot.20441

Babaei, S., Geranmayeh, A., & Seyyedsalehi, S. A. (2008). Pruning Neural Networks for Protein Secondary Structure Prediction. In *BioInformatics and BioEngineering, 2008* (pp. 1-6). http://dx.doi.org/10.1109/BIBE.2008.4696702

Bahimore, D., Berk, A., Darnell, V., Lodish, H., Matsudaira, P., & Zipursky, S. (2007). Molecular Cell Biology (6th ed.). New York: Freeman, W. H.

Barton, G. J., & Cuff, J. A. (2000). Application of multiple sequence alignment profiles to improve Srotein Secondary Structure Prediction. *Proteins Struct. Funct. Genet., 40*, 502-511. http://dx.doi.org/10.1002/1097-0134(20000815)40:3<502::AID-PROT170>3.0.CO;2-Q

Chou, P. Y., & Fasman, D. (1974). Prediction of protein conformation. *Biochemistry, 13*(2), 222-245. http://dx.doi.org/10.1021/bi00699a002

Cox, M., Lehninger, A. L., & Nelson, D. L. (2008). Lehninger Principles of Biochemistry (4th ed.). New York: Freeman, W. H.

Dor, O., & Zhou, Y. (2007). Achieving 80% Ten-fold Cross-Validation accuracy for Secondary Structure Prediction by Large-Scale Traning. *Proteins, 66*, 838-845. http://dx.doi.org/ 10.1002/prot.21298

Garnier, J., Osguthorpe, D. J., & Robson, B. (1978). Analysis and Implications of Simple Methods for Predicting the Secondary Structure of Globular Proteins. *J. Mol. Biol., 120*(1), 97-120. http://dx.doi.org/10.1016/0022-2836(78)90297-8

Gurney, K. (1997). *An Introduction to Nerual Networks*. London and New York. http://dx.doi.org/10.4324/9780203451519

Kabsch, W., & Sander, C. (1983). Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-bonded and Geometrical Features. *Biopolymers, 22*(12), 2577-2637. http://dx.doi.org/10.1002/bip.360221211

Malboobi, M. A., Naghavi, M. R., & Rashidi M. S. (2009). *Bioinformatics*. Tehran University Press.

Qin, W., Qu, W., Sui, H., & Yang, B. (2011). Improving Protein Secondary Structure Prediction using a Multi-modal BP method. *Computers in biology and Medicine, 41*, 946-959. http://dx.doi.org/10.1016/j.compbiomed.2011.08.005

Rost, B., & Sander, C. (1999). Prediction of Protein Secondary Structure at better than 70% accuracy. *J. Mol. Biol., 292*, 195-202. http://dx.doi.org/10.1006/jmbi.1993.1413