International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 4
ISSN: 2454-4248
345 – 348

Determining the Secondary Structure of Elapid Toxins using Multi-Layer Perceptrons and Kohonen Networks

Akash Nag
Dept. of Computer Science
The University of Burdwan
Burdwan 713104, India
E-mail: nag.akash.cs@gmail.com

Sunil Karforma
Dept. of Computer Science
The University of Burdwan
Burdwan 713104, India
E-mail: sunilkarforma@yahoo.com

***Abstract*—** In this paper, a two-stage neural network consisting of a feed-forward neural network and a Kohonen self-organizing map, has been used to predict secondary structure. We have applied our methods to determine the structure of 245 proteins containing neurotoxins, cytotoxins, cardiotoxins and three-finger toxins, derived from venoms of Elapid snakes. In doing so, the system achieved a Q_3 score of 70%, which is quite remarkable.

Keywords*-protein structure prediction; secondary structure; sequence analysis; artificial neural network; multi-layer perceptron; Kohonen network; self organizing map; snake venom toxins; Elapid snakes; Elapidae*

*****

I. INTRODUCTION

Protein databases have been growing exponentially over the past two decades, owing to a deluge of new sequence data from various sequencing projects. As a result, protein structure prediction is now more important than ever. Numerous structure prediction methods are available today that employ various techniques, but the most reliable is the comparative modeling approach [1]. However, a disadvantage of this method is that it requires the structure of a protein, which is largely homologous to the query protein, to be known before it can predict the structure of the latter. In the absence of a suitable homologous protein being available, we need to look at other prediction techniques, such as fold recognition [2-4] and *ab initio* methods [5-13].

Secondary structure prediction is often the precursor step to predicting tertiary structure. The most common secondary structure prediction methods have been either based on stereochemical [5] or statistical principles [6-7]. More recently however, a number of other approaches have arrived at the forefront due to a family of related proteins being available simultaneously for analysis. This lends to applying multiple sequence alignment (MSA) on the set of sequences, and determining additional information about the family regarding insertions, deletions, and mutations. Employing multiple sequence alignment to structure prediction was successfully performed by Niermann et al. [14] on the alpha-subunit of tryptophan synthase, which was later generalized by Zvelebil et al. [8]. But MSA based prediction methods were popularized by Benner & Gerloff [15] by the successful prediction of cAMP-dependent kinases. The popularization of soft computing approaches, and neural networks in particular, gradually led to more automated methods of structure prediction. Most popular algorithms employing neural networks in this field are the PHD method [9], DeepCNF [16], JNet [17] and PSIPRED [18]. Most of these algorithms employ feed-forward neural networks – the latter employing two of them connected sequentially, to determine the Q_3 structure of a protein.

The proposed method is based on combining Kohonen self-organizing maps [19-20] with Artificial neural networks [21-24], feeding the output of the former to the latter. An overview of these two types of neural networks is presented briefly in section 2. The proposed method is presented in Section 3, and the results obtained by our method are presented in Section 4.

II. NEURAL NETWORKS

Neural networks are a popular soft computing approach that has been developed from a computational model of the human brain [21]. The most popular neural networks are those of the feed-forward variety, which is explained in Section 2.1. The other popular methods are Self-organizing maps (see Section 2.2), Hopfield networks [25], etc.

A. Feed-Forward Artificial Neural Networks

A feed-forward artificial neural network (ANN) is a neural network containing a series of nodes called neurons, which are organized in a sequential array of stages called layers. The term feed-forward arises due to the fact that neurons in a layer pass information to neurons only in the next subsequent layer and never in the backward direction (except during training). The first layer is called the input-layer, and the last being the output layer. These two layers are separated by zero or more hidden layers. Each neuron receives data from all neurons in its previous layer, and it activates depending on a threshold function. A weight value is associated with each inter-neuron connection, which is updated during the training phase in order to reduce the error at each iteration, through an algorithm known as back-propagation [26]. The input received by a neuron is the output produced by the connected neuron in its previous layer, adjusted by the weight of the edge connecting the two. A typical feed-forward ANN, containing two hidden layers and three output nodes, is shown in Fig. 1.

 ISSN: 2454-4248
345 – 348

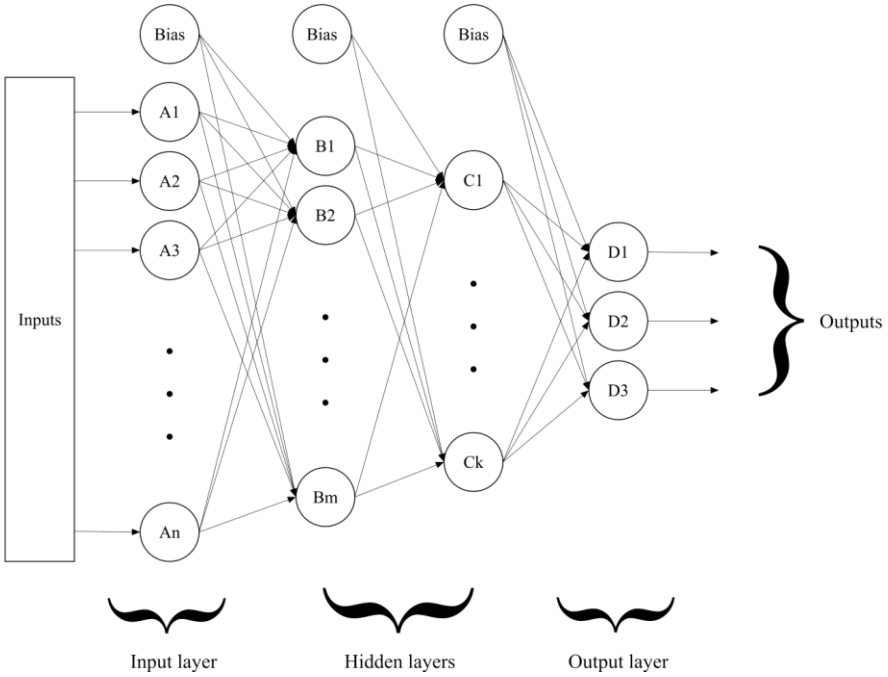

Figure 1. A typical feed-forward artificial neural network

B. Kohonen Self-Organizing Maps

A self-organizing map [20], developed by T. Kohonen, is a tool to map high dimensional data onto a low dimensional grid, often for better visualization purposes. A SOM contains an input layer and an output layer – the latter often being arranged in the form of a two dimensional grid, although more than two dimensions are also possible. Each element in the input layer is connected to every node in the output grid, each of which has an associated weight vector. Whenever a data is presented at the input, the connected output elements are activated. The winning neuron is determined to be the one which has a weight vector closest to the data presented. The weights of the winning neuron and those of its immediate vicinity are adjusted to reduce the error further. This vicinity radius keeps on decreasing with each iteration, and reaches zero, upon which the training phase is deemed to have ended. During the classification phase, the winning node is the output when a data is presented at the input. The structure of a typical SOM is shown in Fig. 2.

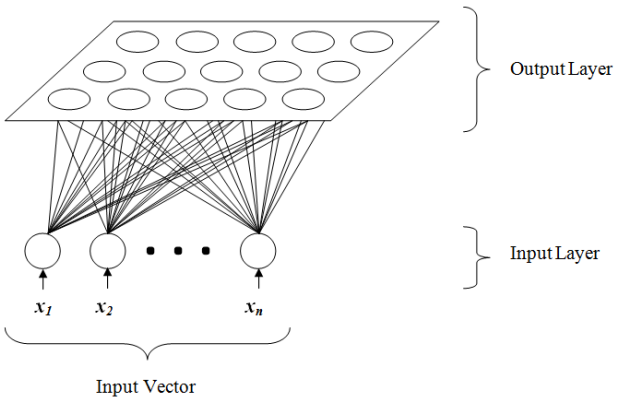

Figure 2. A typical self-organizing map

III. DATA AND METHODS

A. Data

The objective of our study is to determine the secondary structure of Elapid venom toxins. The data for testing our method consisted of 245 snake venom protein sequences containing long and short neurotoxins, cytotoxins, three-finger toxins, cardiotoxins, etc. All these proteins are derived from venoms of snakes in the Elapidae family. For each protein sequence, the secondary structure was isolated in terms of helices, beta strands, and turns. This structural information was consolidated from various computational as well as experimental sources, and was retrieved from UniProtKB/Swiss-Prot [27].

B. The Network Architecture

The network used by our approach consists of a SOM which is sequentially connected to a feed-forward ANN. The SOM used in our network has an input layer size of 13 nodes, and the output grid is a two-dimensional grid with five rows and five columns. The ANN used in our algorithm contains 4 layers: one input layer, two hidden layers, and one output layer. The output layer contains 3 nodes corresponding to the three types of structures to be predicted. The input layer contains 90 nodes, while the two hidden layers contain 61 and 30 nodes respectively. 25 of the input 90 nodes in the ANN are directly fed into by the 25 output nodes of the SOM.

C. Methodology

The proposed prediction method, shown in Fig. 3, consists of two stages: training phase, and prediction phase. Each phase in turn, is split into two sub-phases: the first for the self-

 ISSN: 2454-4248
345 – 348

organizing map (SOM), and the second for the feed-forward artificial neural network (ANN). The dataset was divided into half, with one half being used for training the neural networks, and the other to be used for validation. During the SOM training, for each proteins sequence a consecutive subsequence of 13 amino-acid residues were fed into it. Each amino-acid was converted to a real number normalized in the range 0-1. The window size was tested for 5 through 19 residues, but the size of 13 was found to be optimal with respect to both accuracy obtained and feasibility of execution time. With each input, the window was shifted one amino-acid to the right and the process was continued till the entire length of the sequence, and in turn, for the entire test set of proteins. When the SOM outputs had become stable, its 25 node output along with the same window of 13 amino-acids was fed into the ANN. However, this time, the amino-acids were not fed as a real number, but were encoded into binary, thereby requiring 5 bits for each amino-acid. Therefore, the total input size for the ANN was 90 (25 + 5×13). The ANN produced a 3 bit output, which was then decoded to the three possible structures, namely: alpha helices, beta strands or turns.

IV. Results and Discussion

The proposed algorithm was implemented in Java. The testing was performed on an Intel Celeron M single-core 1.6GHz processor with 2GB memory. For the 122 proteins in the training set, the bulk of the execution time was taken for training the ANN. The SOM was trained in 28 seconds but the ANN required 56 minutes for the same dataset. The Q_3 accuracy for the structure prediction results against the 123 test proteins are shown in Fig. 4. As we can see, our proposed method has a very high accuracy, in par with well known algorithms in this field. The plot of the reducing MSE (mean squared error) during the ANN training phase is shown in Fig. 5.

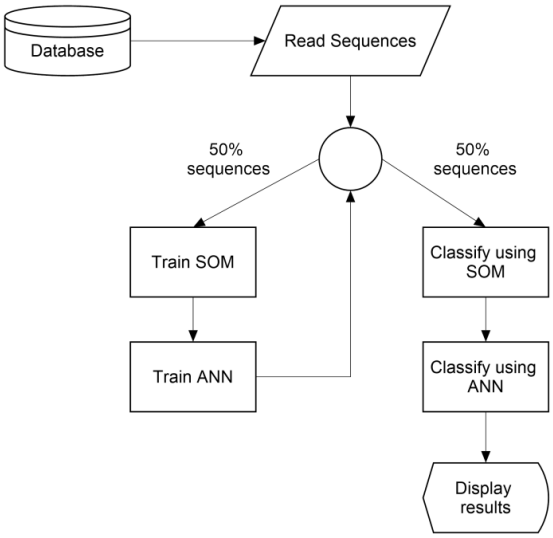

Fig. 3. The proposed method

The mean Q3 accuracy was found to be 70.53% and the median accuracy was 73.3%. Only 3 sequences out of 123 were found to have an accuracy lower than 20%, while 94 sequences achieved an accuracy of at least 60%, with 8 of those reaching above 90%. The standard deviation was found to be 16%. In contrast, methods such as PSIPRED consistently reach a median accuracy of 76% with a standard deviation of 7-8% over many datasets including CASP3. Other methods such as DSC [13] fare worse than our method with a median accuracy of 67.3% over 16 CASP3 targets. PSIPRED may be marginally better than the proposed method but the advantage of our method is that, unlike PSIPRED or the PHD method, it does not require the generation of sequence profiles or multiple sequence alignments prior to prediction. This makes porting from client to server-side implementations much easier, and also results in much faster execution times.

V. Conclusions

Neural networks have been successfully used in the past for predicting protein secondary structures, such as in PSIPRED, JNet and DeepCNF. This is the first time neural networks have been mixed with SOM. The impressive results presented above suggest that the advantage gained by generating alignments and sequence profiles prior to prediction can be overcome to a large extent by incorporating SOM with ANNs. More studies are underway to finding ways of improving the neural network architecture presented above in order to increase the accuracy up from 70% to 80%. More studies are also required to see how our algorithm scales with the increase in the number of sequences, as well as their variability in terms of both length and structure. The primary objective of this study was to predict the structure of Elapid venom toxins.

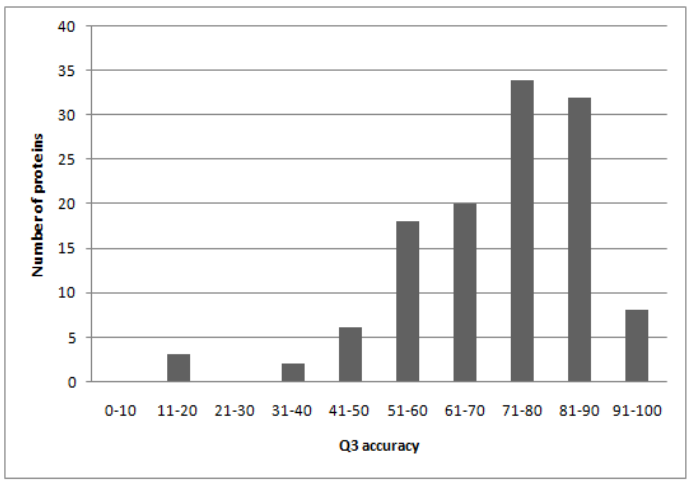

Fig. 4. Accuracy of the proposed method

 ISSN: 2454-4248
345 – 348

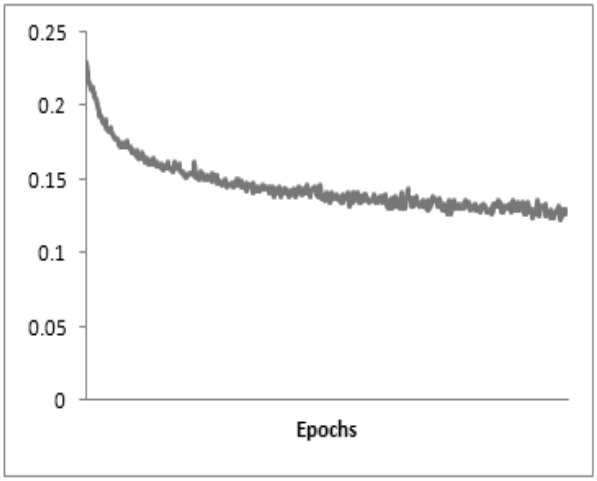

Fig. 5. Plot of MSE during training

REFERENCES

[1] Sali, A. (1995). Modelling mutations and homologous proteins. Curr. Opin. Biotechnol. 6, 437-451.

[2] Bowie, J. U., Luethy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. Science, 253, 164-170.

[3] Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). A new approach to protein fold recognition. Nature, 358, 86-89.

[4] Lemer, C., Rooman, M. J. & Wodak, S. J. (1995). Protein structure prediction by threading methods: evaluation of current techniques. Proteins: Struct. Funct. Genet. 23, 337-355.

[5] Lim, V. I. (1974). Algorithms for prediction of alpha helices and structural regions in globular proteins. J. Mol. Biol. 88, 873-894.

[6] Chou, P. Y. & Fasman, G. D. (1974). Conformational parameters for amino acids in helical, -sheet, and random coil regions calculated from proteins. Biochemistry, 13, 211-222.

[7] 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, 97-120.

[8] Zvelebil, M. J. J. M., Barton, G. J., Taylor, W. R. & Sternberg, M. J. E. (1987). Prediction of protein secondary structure and active sites using the alignment of homologous sequences. J. Mol. Biol. 195, 957-961.

[9] Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. 232, 584-599.

[10] Geourjon, C. & Deleage, G. (1995). SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Comp. Appl. Biosci. 11, 681-684.

[11] Salamov, A. A. & Solovyev, V. V. (1995). Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. J. Mol. Biol. 247, 11-15.

[12] Frishman, D. & Argos, P. (1996). Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. Protein Eng. 9, 133-142.

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

[14] Niermann, T., Kirschner, K. & Crawford, I. P. (1987). Prediction of tertiary structure of the alpha-subunit of tryptophan synthase. Biol. Chem. Hoppe-Seyler, 368, 1087-1088.

[15] Benner, S. A. & Gerloff, D. (1990). Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases. Advan. Enzyme Reg. 31, 121-181.

[16] Wang, S., Peng, J., Ma, J. & Xu, J. (2016). Protein secondary structure prediction using deep convolutional neural fields. Scientific reports 6.

[17] Cuff, James A., and Barton, G. J. (2000). Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins: Structure, Function, and Bioinformatics 40.3, 502-511.

[18] Jones, David T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. Journal of molecular biology 292.2, 195-202.

[19] Kohonen, T., and Honkela, T. (2007) Kohonen network. Scholarpedia 2.1, 1568.

[20] Kohonen, Teuvo (1982). Self-Organized Formation of Topologically Correct Feature Maps. Biological Cybernetics. 43(1), 59–69.

[21] McCulloch, W., and Pitts, W. (1943). A Logical Calculus of Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics. 5(4), 115–133.

[22] Hebb, Donald (1949). The Organization of Behavior. New York: Wiley. ISBN 978-1-135-63190-1.

[23] Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model For Information Storage And Organization In The Brain. Psychological Review. 65(6), 386–408.

[24] Ivakhnenko, A. G., Lapa, V. G. (1967). Cybernetics and forecasting techniques. American Elsevier Pub. Co.

[25] Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities, Proceedings of the National Academy of Sciences of the USA, vol. 79 no. 8, 2554–2558.

[26] Rumelhart, David E.; Hinton, Geoffrey E., Williams, Ronald J. (1986). Learning representations by back-propagating errors. Nature. 323(6088), 533–536.

[27] Boutet, Emmanuel, et al. (2007). UniProtKB/Swiss-Prot: the manually annotated section of the UniProt KnowledgeBase. Plant bioinformatics: methods and protocols, 89-112.