www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 1, January 2014 Page No. 3593-3597

Protein Secondary Structure Prediction Using Improved Support Vector Machine And Neural Networks

Anureet Kaur Johal, Prof. Rajbir Singh
Student of M.Tech
Department of CSE LLRIET, Moga
reetsoldier@gmail.com
Associate Prof. & Head
Department of I.T LLRIET, Moga
cheema_patti@yahoo.com

ABSTRACT
To solve the Protein folding problem is one of the most important task in computational biology. Protein secondary structure prediction is key step in prediction of protein tertiary structure. There have emerged many methods such as meta predictor based, neighbor based and model based methods to predict protein structure. The model based approaches employ machine learning techniques like neural networks and support vector machines to learn a predictive model trained on sequence of known structure. Historically machine learning methods have shown amazing results Therefore objective of this paper is to compare the performance of Neural Networks (NN) and Support Vector Machines (SVM) in predicting the secondary structure of proteins from their primary sequence. For each NN and SVM, we created classifiers to distinguish between helices (H) strand (E), and coil (C). Finally the output obtained illustrates that out of these top most novel methods for classification purpose Neural Networks performs much better then support vector machine and produces better efficiency in much lesser time.

General Terms
Frequency profiling, Pssm, back propagation, binary classifiers, Multiple sequence alignments

Keywords
Amino Acid, Protein folding problem, support vector machine, neural networks.

1. INTRODUCTION
In present scenario protein folding problem is the most significant in molecular biology. Protein folding basically refers to the prediction the 3-D structure of protein from its amino acid sequence. In order to predict the structure the problem is subdivided into various levels. As this experiment will provide various advantages in the field of drug design and protein engineering. Also, as the number of sequences are growing at rapid rate than our ability to solve their structure experimentally such as X –ray cryptography is creating an ever widening sequence structure gap and inclining the pressure to predict secondary structure. More over these method claims much higher and expensive infrastructure and computational power. Various methods have been developed that can resolve protein folding problem based on different algorithms, like Statistical Analysis ,Information theory, Bayesian Statistics and Evolutionary Information, Nearest Neighbor Methods. These methods claims, the accuracy levels between 60–80%.
The Present work analyses the prediction of secondary structure of proteins from their sequences using

two new novel and popular methods for prediction are SVM and NN.

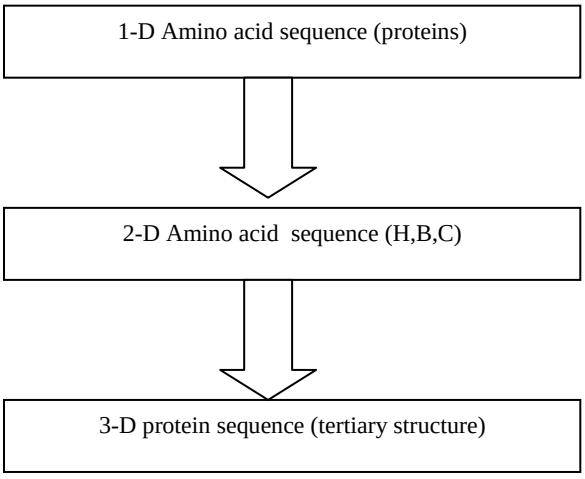

Fig. 1: Protein folding problem

Various applications of Neural Networks mostly implements Supervised Learning. For supervised learning, training data which contains both the input and the required output are given initially. After complete training, sequence is presented to the Neural Network and the machine will calculate the result value which will be nearer to the required output. Further more training is performed if output doesn't lie between the required output values, which illustrates that the parameters of the network are adjusted til the output is close to the target value [2]. For training of Neural Networks, Resilient Backpropagation [12] is used.

Support Vector Machines is another machine learning technique implemented for classification and regression [5]. For

classification, Support Vector Machines functions by finding a separating hyperplane in the space of possible inputs. This hyperplane attempts to divide the positive values from the negative values. Normally the division is chosen on the basis of the largest distance from the hyperplane to the nearest of the positive and negative values. Data points that are at the margin are called Support Vectors. These data points are very much needed in the theory of Support Vector Machines as they can be used to classify information contained in the dataset. [5]. The hyper plane with a maximum margin allows more accurate classification of new points [14]. In other cases where data is not easily separated by using hyper plane then, Kernels functions are used to perform the mapping. Moreover pre - processing of data is carried out through frequency profiling in this paper. Svm provides various advantages over other methods such as, effective avoidance of over fitting, ability to handle large feature spaces, information condensing of the given data, and pattern recognition problems like, hand written digit recognition, speaker identification and text categorization.

2. METHODOLOGY

In The present methodology major goal is to train the Neural Network and a Support Vector Machine to respond to the sequence of proteins when the predictions of the secondary structures are known. To solve this protein folding problem, programs are set up in matlab environment. The dataset used is obtained from ncbi.com which consists of various proteins from the database. As Preprocessing of data is done first which is carried out through frequency profiling which means converting the data sent in letters into numbers. Later, secondary structure assignment is performed. In Secondary structures are classified into 8 categories H,G,E,B,I,T,S and the last category is for unclassified structures. These are reduced to 3 categories of H, E and C by using a secondary structure assignment called PSSM.4. Further more, 6 binary classifiers are created. Finally comparison of both the two machine learning algorithms is done.

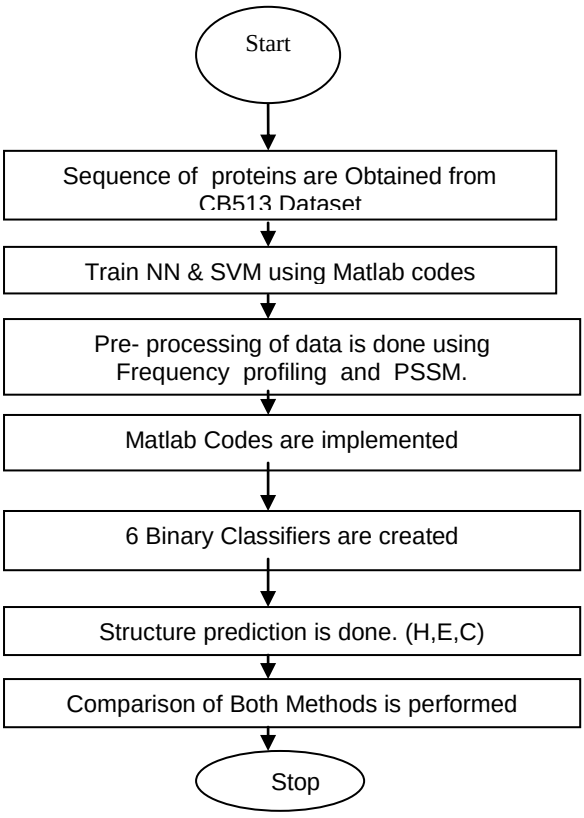

Fig. 2: Flowchart depicting methodology followed.

2.1 SUPPORT VECTOR MACHINE

(a)Frequency Profile

The approach is as follows:

• Multiple sequence alignments are used as inputs to the network.

• During training, the data base of protein families aligned to proteins of known structure is used.

• For prediction, the data base is scanned for all homologues of the protein to be predicted and the family profile of amino acid frequencies at each alignment position is fed into the network.

• Each type of secondary structure has an equal proportion of 33% of the states during training. This ensures balanced training as the proportions are not based on the data base.

The first network has three layers: input layer, hidden and output layer which has three output neurons where (1,0,0)^t codes for alpha . The outputs of the first network are used as inputs into the second network, which are propagated through the network to obtain new output comprising of the three states. Thus a real number is given to every alphabet in the sequence in order for it to be processed. The data consisting of 126 proteins, in which no two sequences have more than 25% of identical residues, were used for the prediction. The data are from the database in [7]. Three quality indices were used: OA the percent correct for the three states (2.2), the percentage of correct predictions for each state (2.1) and correlation coefficients (2.)

(b) Multiple windows

These are sliding windows of the same size (15) over the entire dataset. The steps are as follows:

- All the windows are extracted using linear indexing. Thereafter, load them into a bigger array.
- This is then processed using MATLAB vectorised operations.
- The program that we have executed in MATLAB gives the indices of all the sliding windows of the matrix.
- Indices in every case is the index of the centre most cell of that window.
- This, of course mandates that the length of the windows be an odd number.

(c) Binary classifiers

Six SVM binary classifier including three one-versus-rest classifier ('one': positive class, 'rest': negative class) names H/~H, E/~E and C/~C and three one-versus-one classifier named H/E, E/C, C/H were constructed. For example, the classifier H/E is constructed on the training samples having helices and sheets and it classifies the testing sample as helix or sheet. The programs for constructing the SVM binary classifier were written in the C++ language.

(d) Gaussian kernel

In Support Vector Machines, the following Gaussian Kernel is used:

K (xi, xj) = exp(− k xi − xj k^2)

SVM has two parameters: the kernel and the cost parameters C. For this study, a kernel parameter of = 0.1 will be used and is fixed for all experiments. Hua and Sun [24] used this parameter in their study. Their cost parameter was set to 1.5 to construct the classifiers. Since we are interested in the error estimates for accuracy in SVM, the cost parameter will be varied to ensure better parameters for the model. The cost parameters used are 0.1, 0.3, 0.5, 0.7, 0.9, 1 and 5.

2.2 NEURAL NETWORKS

(a) Neural Network Architecture.

Neural networks are composed of simple elements operating in parallel. These elements are inspired by biological nervous systems. As in nature, the connections between elements largely determine the network function. You can train a neural network to perform a particular function by adjusting the values of the connections (weights) between elements.

Typically, neural networks are adjusted, or trained, so that a particular input leads to a specific target output. The next figure illustrates such a situation. There, the network is adjusted, based on a comparison of the output and the target, until the network output matches the target. Typically, many such input/target pairs are needed to train a network.

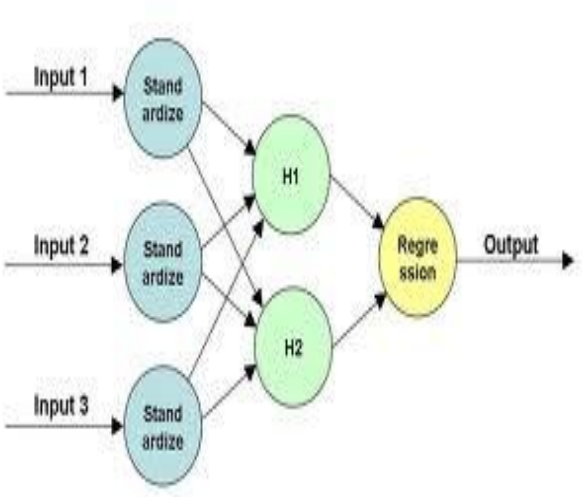

Fig. 3: Neural Networks

Neural networks have been trained to perform complex functions in various fields, including pattern recognition, identification, classification, speech, vision, and control systems. NN considered are usually of the feed forward type. The approach uses a network that receives an input vector representing a segment of primary amino acid sequence. The input layer encodes a moving window in the amino acid sequence and secondary structure prediction is made for the central residue in the window. The length of the window can be varied.

(b) PSSM INPUT PROFILE

A common method is PSSM algorithm, which breaks the problem down into 2-dimensional sub-problems that may be solved analytically, eliminating the need for a numerical optimization algorithm.

PSSM algorithm:

A PSSM, or Position-Specific Scoring Matrix, is a type of scoring matrix used in protein BLAST searches in which amino acid substitution scores are given separately for each position in a protein multiple sequence alignment. Thus, a Tyr-Trp substitution at position A of an alignment may receive a very different score than the same substitution at position B. This is in contrast to position-independent matrices such as the PAM and BLOSUM matrices, in which the Tyr-Trp substitution receives the same score no matter at what position it occurs.

PSSM scores are generally shown as positive or negative integers. Positive scores indicate that the given amino acid substitution occurs more frequently in the alignment than expected by chance, while negative scores indicate that the substitution occurs less frequently than expected. Large positive scores often indicate critical functional residues, which may be active site residues or residues required for other intermolecular interactions.

1. Pssm matrix is a category of blossum matrix. blossum matrix also contains dssp matrix. so there're resemblances.
2. The main thing about pssm matrix is the assignment of positive and negative integers (or any 2 integers to suggest the occurrence of that alphabet) that suggest the occurrence of that alphabet i.e. its frequency.

3. So we have to combine :

(a.) The secondary structure prediction as suggested in our paper i.e H,G to H.
(b). And the +1 and -1 assignment of svm... (or in the case of neural networks 1 and 2)
so steps 3a and 3b together implement pssm which concludes +1 more often indicates h (in h/~h) i.e more +1 implies more h.

(c) Resilient Back propagation

The training algorithm used in this thesis for Neural Networks is Resilient Back propagation [40]. The algorithm has two passes through the network; the forward and 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 [6]. The algorithm depends on the initial update values and the step size that has a default setting. The update values determine the size of the weight steps while the maximum step size has to be provided to avoid weights that are too large for the network. The approach provides faster convergence to the minimum because of the few parameters required to obtain optimal convergence times [12]. For the Resilient Back propagation algorithm, the size of the weight change is determined by the update-value.

(d) Accuracy Measure

Q_3 is one of the most commonly used performance measures in the protein secondary structure prediction and it refers to the three-state overall percentage of correctly predicted residues. This measure is defined as,

$$Q_3 = \sum_{(i=H,E,C)} \frac{\#\text{ofresiduescorrectlypredicted}}{\#\text{ofresiduesinclass(i)}} * 100(\%)$$

3. LITERATURE REVIEW

Previous research discusses the use of a novel method for the prediction of the protein secondary structure from the amino acid sequence. Present method is based on the different encoding schemes of SVM and NN. A significant improvement is obtained by combining multiple windows with frequency profiling and PSSM. Additional improvement in the predictions is obtained by using back propagation method with no hidden units. Presented results when compared with earlier papers[1],[14] shows greater improvement in the accuracy of all the 6 classifiers computational time and accuracy.

4. DATASETS

4.1 The form of the data

The dataset consist of 60 proteins from CB513 dataset.obtained from PDB. The data is structured in rows by protein name, primary and secondary structure. An example of a protein Acprotease is:
>Acprotease

IVGTVPMTDYGNDVEYYGQVTIGTPGKSFNLNFDTGSSNL
WVGSVQCQASGCKGGRDKFNPSDGSTFK
SQPTYPGDDAPVEDLIRFYDNLQQYLNVVTRHRY*
CCCCCCCCCTTSCHHHHHHHHHHHHHHHHHHTTCC*

The primary structure is a sequence of amino acids, which are represented by a 1 letter code as explained before in frequency profiling technique and pssm. The secondary structures are made of 8 classes : H,B,E, G, I, T, S and rest marked a dash (−).

Table 1: Secondary Structure Assignment

DSSP 8-classes	**3-class**
α-helix (H) ,3/10 helix (G)	Helix(H)
β-sheet (E), β- Bridge(B)	Strand(E)
π-helix (I),Turn(T), Bend(S), Coil(C)	Coil(C)

4.2 The data comprises 10766 samples and the secondary structure composition is given in Table 1.

Table 2: Training time vs. total number of samples

Total number of secondary structure states for the dataset	Total number	Percentage
H	3047	28.3
E	2288	21.3
C	5431	50.4

For One-Against-All classifiers, all the 10766 samples are used in formulating Neural Networks and Support Vector Machines, while for the One-against-One classifiers, samples differ based on the classifier under consideration. For

example, in H/E, only two classes are considered: alphas and betas. This means that coils are excluded from the data set.

Table 3: Gives the total number of samples for all the classifiers.

Binary Classifier	Total Samples
H/~H	10777
E/~E	10777
C/~C	10777
H/E	5535
E/C	7719
C/H	8478

5. RESULTS AND DISCUSSION

5.1 Time taken by classifiers of both methods

Table 4: Time taken by both the methods

BINARY CLASSIFIERS	SVM(time taken)	NN(time taken)
H/~H	2 min 11 sec	1 min 6 sec
E/~E	1 min 23 sec	2 min 2 sec
C/~C	3 min 45 sec	2 min 3 sec
H/E	2 min 30 sec	1 min 2 sec
E/C	3 min 20 sec	3 min 5 sec
C/H	2 min 12 sec	1 min 0 sec

5.2 Comparisons of NN and SVM

Table 5: Comparison of Neural Networks and Support Vector Machine classifiers Accuracy

Binary classifiers	NN% ACCURACY	SVM% ACCURACY
H/~H	75.63	74.32
E/~E	74.65	73.15
C/~C	74.67	71.27
H/E	75.59	74.65
E/C	72.15	71.04
C/H	76.00	74.33

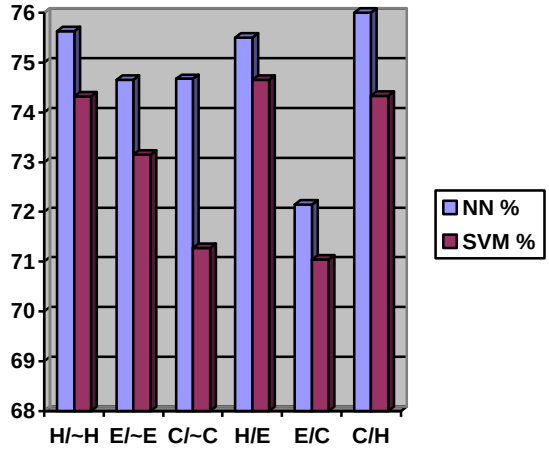

FIG.4: COMPARISON GRAPGH OF ACCURACY

The results in bar graph depictes that performance of NN is much better than SVM. For the One-against-All classifiers, NN achieved the highest prediction accuracy of about 75.63% where as SVM achieved only 74.32% only. Also, in One-Against-One classifiers NN again achieved the highest accuracy of about 76%.where as SVM achieved about only 74.65%.

6 CONCLUSIONS AND DISCUSSION

6.1 CONCLUSION

The main aim of this paper was to compare performance of Support Vector Machines and Neural Networks in predicting the secondary structure of proteins from their amino acid sequences. The following conclusions were derived:

1. Neural Networks provides much better accuracy as compared to SVM even when employed with simple network parameters and architecture.
2. Also, NN take much lesser training and computation time when compared to SVM.
3. Presented results also reveal that SVM requires much larger memory and powerful processor as compared to NN.
3. Finally NN provides much better results in all the classifiers. For the One-against-All classifiers, NN achieved the highest prediction accuracy of about 75.63% where as SVM achieved only 74.32% only. Also, in One-Against-One classifiers NN again achieved the highest accuracy of about 76%.where as SVM achieved about only 74.65%. This illustrates NN is far better machine learning technique then SVM from all the aspects.

6.2 FUTURE SCOPE

The future work of both the categories primarily deals with using different encoding schemes, which may increase the results of binary classifier's accuracy levels. More concrete case can be developed if other datasets are used to prove the supremacy of these new methods over other contemporary techniques. Multiple windows with different window sizes will be considered for future studies. After forming the best binary classifiers, the new tertiary classifiers can be designed and tested to prove that their performance is best among all the current research methods.

6. ACKNOWLEDGEMENT

I wish to express my sincere gratitude and indebtedness to my Supervisor, Prof. Rajbir Singh (Associate Prof. & Head, Department of Information Technology) for his valuable guidance, attention-grabbing views and obliging nature which led to the successful completion of this study. This research work would not have been possible without the support of Dr. H.S. Gulati (Principal, LLRIET, Moga) & Dr. Dinesh Grover (Director, CSE & IT) who was abundantly helpful and offered invaluable assistance, support and guidance. I am speechless to express my gratitude towards my family for their moral and financial support and encouragement without which I would not have been able to bring out this thesis.

7. REFERENCES

[1] Anjum B. Reyaz-Ahmed (2007) "Protein Secondary Structure Prediction Using Support Vector Machines, Nueral Networks and Genetic Algorithms" Digital Archive , pp 43.

[2] Bishop, C.M. (1995). Neural Networks for Pattern Recognition. Clarendon Press, Oxford.

[3] Campbell, C. and Cristianini, N. (1999). Simple Learning Algorithms for Training Support Vector
Machines. Technical Report CIG-TR-KA, University of Bristol, Engineering Mathematics,
Computational Intelligence Group.

[4] Chen, P.H., C.J. Lin and Schölkopf, B. (2005). A Tutorial on v-Support Vector Machines. Applied
Stochastic Models in Business and Industry 21(2): pp. 111-136.

[5] Gunn, S. (1998). Support Vector Machines for Classification and Regression. Technical Report, ISIS, Department of Electronics ad Computer Sciencne, University of Southampton.

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

[7] Hobohm, U., Scharf M., Schneider R. & Sander C. (1992). Selection of Representative Protein
Data Sets. Protein Sci. 1: pp. 409-417.

[8] Holley, H.L. & Karplus, M. (1989). Protein Secondary Structure Prediction with a Neural Network.
Proc. Nat. Acad. Sci., U.S.A. 86: pp. 152-156.

[9] Hornik, K., Stinchcombe, M. & White, H. (1989). Multilayer Feedforward Networks Are Universal
Approximators. Neural Networks, 2: pp. 359-366.

[10] Hua, S. & Sun, Z. (2001). A Novel Method of Protein Secondary Structure Prediction with High
Segment Overlap Measure: Support Vector Machine Approach. J. Mol. Biol. 308: pp. 397-407.

[11] Qian, N. & Sejnowski, T.J. (1988). Predicting the Secondary Structure of Globular Proteins Using
Neural Network Models. J. Mol. Biol. 202: pp. 865-884.

[12] Riedmiller, M. (1994). Rprop-Description and Implementation Details. Technical Report, University
of Karlsruhe, Germany.

[13] Wackerly, D.D., Mendenhall, W., & Schaeffer, R. (2002). Mathematical Statistics with Applications.
Duxbury Press, USA.

[14]Yang Jaewon (2008) "Protein Secondary Structure Prediction based on Neural Network Models and Support Vector Machines " CS229 Final Project, Dec 2008.

[15] Whitford, D. (2005). Proteins Structure and Function. John Wiley & Sons Ltd, England.