Secondary Structure Prediction of Protein Using Support Vector Machine and Neural Networks

1Sumnit Kaur, 2Er.Navneet Bawa
1Mtech (CSE) ACET, Mwanawala
2Associate Professor ACET, Manawala

ABSTRACT
Bioinformatics is the science of managing, mining, and interpreting information from biological sequences and structures. Bioinformatics methods are used to predict protein structure that mostly depends on the amino acid sequence. In computational biology the most important task is to solve the protein folding problem. To predict tertiary structure first we need to predict the secondary structure of protein. Secondary structure is intermediate part between primary and tertiary structure. In order to predict the secondary structure many methods are used i.e. Choufasman, Gor and machine learning techniques like Neural network and Support Vector Machine. In this paper, comparison of both Statistical technique SVM and NN is drawn. In this we compare the performance of neural network and support vector machine from predicting the secondary structure of protein from the protein structure is evaluated. For each NN and SVM, classifiers that will classify Alpha, Beta and Helix structure from the protein structure to distinguish between helices (H) strand (E), and coil (C).

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

1.INTRODUCTION
Nowadays protein folding problem is most significant. Many methods are used that can resolve protein folding problem. Protein folding basically refers to the prediction the 3-D structure of protein from its amino acid sequence. Protein – Proteins are the basic building block. Proteins are the biochemical compounds consisting of one or more polypeptides. In order to predict the secondary structure of proteins many methods and techniques are used.

(A). Choufasman Method – Analyzed the frequency of the 20 amino acid in alpha helix, beta sheet and turn. When 4 of 5 amino acid have a high probability of being in alpha helix, it predict an alpha helix. When 3 of 5 amino acid have a higher probability of being in a strand it predict strand.

(b) GOR – GOR method not only contain the propensities but it also take the conditional propensities. GOR method is 60%-70% accurate. We calculate the accuracy by using Q3 formula.

(c) PHD- Combine neural network with sequence profiles
- 6-8 Percentage points increases in prediction accuracy over standard neural network
- Use second layer “Structure to structure” network to filter predictions.
- Use alignments from iterative sequence
- Better prediction due to better sequence profiles

(d) Neural Network- In neural network data samples are collected first and then create a network. Various applications of Neural Networks mostly implements Supervised Learning. In neural network two steps are perform – Pre-processing and Division into subsets. In neural network, we train the data which contains both the input and the required output are given. After the training part is completed the calculation of result take place by the sequence which is presented to the neural network. For training of Neural Networks, Resilient Backpropagation is used.

(e) Support Vector Machine - Is implemented for classification and regression. It is a machine learning technique in classification. Support VM finds a separating hyperplane in the space of possible value. This hyperplane attemptsto divide the positive and negative values. The hyperplane with a maximum margin allows more accurate classification of new points. In some cases, Kernel function are used to perform in which data is not easily separated using hyperplane. Various advantage SVM provide such as , pattern recognition problem, Hand written digit recognition , text recognition. In SVM, for inputs we use multiple sequence alignments to the network.
2. METHODOLOGY

From the database almost 62 proteins data is used, which contain the information regarding protein name, it’s both primary and secondary sequence. In secondary structure prediction the most important part is to train the neuron network and support vector machine to respond to the sequence of protein when the prediction of the secondary structure are known. The required classification of primary sequence into the secondary sequence are performed using mat lab. Preparation for the data for processing is done. Pre processing is performed in first step through frequency profiling. Purpose of performing preprocessing is to converting the data set in letters into number. Assignment of a secondary structure is the second step of Implementation of Matlab codes. Using structure assignment called PSSM 8 categories are reduced to 3 categories. Implementation are for two class problems and the following binary classifiers are created i.e., the one-against-all classifier and the one versus one classifier. Finally at last the comparison between both machine learning algorithms take place.

![Flowchart depicting methodology followed.](image)

2.1 SUPPORT VECTOR MACHINE

- Support Vector Machines are used data has exactly two classes.
- An SVM classifies data by finding the best hyperplane that separates all data points of one class from other class.
- Best hyperplane means one with the largest margin between two classes.
- Sometimes small number of mislabeled example decrease the performance

(a) Binary classifiers

In Support Vector Machine there are 6 binary classifiers including three one-virus-rest classifiers („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. The programs for constructing the SVM binary classifier were written in the C++ language.
(c) Gaussian kernel
In Support Vector Machines, the Gaussian Kernel is used:
SVM has two parameters: the kernel and the cost parameters C. For this study, a kernel parameter of $\gamma = 0.1$ will be used and is fixed for all experiments... Their cost parameter was set to 1.5 to construct the classifiers.

2.2 NEURAL NETWORKS
(a) Neural Network Architecture.
- Create a neural network (a computer program).
- ’Train’ it uses protein with secondary structure
- Then give new proteins with unknown structure and determine the structure with neural network.

Neural networks are composed of simple elements operating in parallel... The connections between elements largely determine the network function. Each inputs in neural network are associated with weights and bias. 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.

(b) Weights
- Input values at each layer are multiplied by weights.
- Weights are initially random

Neural network have been trained to perform complex functions in various fields. Neural Network considered are usually of the feed forward type.

(b) PSSM INPUT PROFILE
A PSSM, or Position–Specific Scoring Matrix common method which the problem is break 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.
PSSM scores are generally shown as positive or negative integers...

1. PSS matrix is a category of blossom matrix. Blossom matrix also contains ds matrix. So there're resemblances.

2. The main thing about spasm matrix is the assignment of positive and negative

(c) Resilient Back propagation
The algorithm has two passes through the network; the forward and backward pass. For forward pass, during training a sample is presented to the network as input. In case of backward pass, weights are adapted to ensure that the minimum error between the targets and the actual output is achieved. 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. Based on the value of the error, the connection weights are adjusted.

3. LITERATURE REVIEW
Ibrahim Darwish1, Amr Radi2, Salah El-Bakry3 and El-Sayed M. El-Sayed(2015) [1] discussed that Precise prediction of protein secondary structures from the associated amino acids sequence is of great importance in bioinformatics and yet a challenging task for machine learning algorithms.

Hanna Hendy, ET. Al (2015) - [2] considered that Protein secondary structure prediction has been and will continue to be a rich research field. This paper presents a technical study on recent methods used for secondary structure prediction using amino acid sequence. The paper shows different approaches for predicting the protein structures that showed different accuracies that ranged from 50% to over than 90%. The most commonly used technique is Neural Networks

Pradeep Singh, Prof Rajbir Singh, et.al. (2015) – [3] has proposed the Machine learning technique is introduced as a method for the classification of proteins into functionally distinguished classes. Protein function classification is one of the most important problems in modern computational biology. Studies are conducted on a number of protein classes including RNA-binding proteins; protein homodimers, proteins responsible for drug absorption, proteins involved in drug distribution and excretion, and drug metabolizing enzymes.

Shavian Agarwal, et.al.(2014)- [4] considered that the tertiary structure of protein is difficult to predict accurately directly from a protein sequence. The intermediate step is required to predict the structure which project the one dimensional structure into the three dimensional structure.

Annulet Kaur Johal, Prof. Rajbir Singh (2014)-[5] explained that how 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.

Patel Mauri Dinuba, Dr.Hitesh B Shah (2013) — [5] suggested that the Bioinformatics or computational biology is field of science in which biology, computer science and information technology merges into a single discipline. In modern computation biology, research of protein secondary structure plays a major role in protein tertiary structure prediction. Vector machine with different kernel functions.

Shusha Shankar Ray and Shankar K (2013) - [6] discussed the Prediction of RNA structure is invaluable in creating new drugs and understanding genetic diseases. Several deterministic algorithms and soft computing-based techniques have been developed for more than a decade to determine the structure from a known RNA sequence.

Anil Kumar Mandle1, Pranitha Jain, et.al. (2012)- [7] consider how to Support Vector Machine (SVM) is used for predict the protein structural. Bioinformatics method use to protein structure prediction mostly depends on the amino acid sequence. In this paper, work predicted of 1D, 2-D, and 3-D protein structure prediction

Fee Xia, Yong Dou, Gouging Lei, Young Tan (2011)-[8] implementation on FPGA to accelerate the GOR-IV package for 2D protein structure prediction. To improve computing efficiency, they partition the parameter table into small segments and access them in parallel aggressively exploit data reuse schemes to minimize the need for loading data from external memory
Rajbir Singh, Suman deep Kaur Diol, Provender S. Sandhu (2010)-[9] discussed that improved protein structure determination and prediction as a focal research subject in the field of bioinformatics due to the importance of protein structure in understanding the biological and chemical activities of organisms.

P.V. Nageswara Rao, ET. al. (2010)-[10] illustrated that the proteins are key biological molecules with diverse functions. Successful secondary structure prediction provides a starting point for direct tertiary structure modeling, in addition it improves sequence analysis and sequence-structure binding for structure and function determination.

Rete Kamal Kaur, Man jot Kaur, Amino Kaur (2010)-[11] proposed an overview about biomedical research and healthcare which is continue to progress in the genomic/post genomic era, a number of important challenges and opportunities exist in the broad area of bioinformatics.

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 is:

> Avian polypeptide

GPSQPTYPGDAPVEDLIRFYDNLQQYLNVTRHRY
CCCCCCCCCTTCHHHHHHHHHHHHHHHHHTTCCC

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

5. RESULTS AND DISCUSSION

Fig 5.1. Neural Network

Fig 5.2. the calculation of alpha, beta and coil
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,

$$Q_3 = \frac{\sum (I=H, E, C) \# \text{of residues correctly predicted}}{\# \text{of residues in class}} \times 100$$

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

<table>
<thead>
<tr>
<th>Binary classifiers</th>
<th>NN% ACCURACY</th>
<th>SVM% ACCURACY</th>
</tr>
</thead>
<tbody>
<tr>
<td>H/~H</td>
<td>76.63</td>
<td>75.32</td>
</tr>
<tr>
<td>E/~E</td>
<td>74.69</td>
<td>72.15</td>
</tr>
<tr>
<td>C/~C</td>
<td>73.67</td>
<td>71.28</td>
</tr>
<tr>
<td>H/E</td>
<td>75.52</td>
<td>74.65</td>
</tr>
<tr>
<td>E/C</td>
<td>76.18</td>
<td>72.04</td>
</tr>
<tr>
<td>C/H</td>
<td>76.00</td>
<td>74.35</td>
</tr>
</tbody>
</table>

FIG 5.4: COMPARISON GRAPGH OF ACCURACY

The results in bar graph depicts 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%. 

---

Volume 5, Issue 6, June 2016
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. By complaining both techniques Neural Networks and Support Vector Machine, Neural Network provides much better accuracy as compared to.
2. Comparing the time between both techniques NN take much lesser training and computation time...
3. Impairing the memory and processors SVM requires much larger memory and powerful processor.
4. Finally NN provides much better results in all the classifiers

REFERENCES