ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 8, August 2016

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

[1] Sunnit Kaur [2] Er. Navneet Bawa
[1] M.Tech (CSE) [2] Associate Professor
Amritsar College of Engineering & Technology, Manawala Asr
[1]cheemasunnit@gmail.com [2]Bawa.navneet@gmail.com

***Abstract*— Precise prediction of protein secondary structure from the associated amino acids sequence is of great importance and also challenging task .Protein is an important molecule that performs a wide range of functions in biological system. The secondary structure of protein plays a key role in designing of drugs. Various tools are used for the secondary structure prediction of proteins such as Support vector machine and neural network, fuzzy logic. NN is machine learning methodology in which the network is trained using the recognized data sets.SVM is a supervised machine learning method is based on principle of the structural risk minimization. In this paper, both SVM and NN techniques are compared. For each NN and SVM, classifiers classifies the sequence in the 3-Level subclasses: Helix (H),Sheet(E)and coil(c).The objective is to acquire the maximum predict IVE accuracy with the minimal zed error. From the comparative study of SVM and NN it is concluded that technique takes lesser time than SVM**

***Key wards:* -- Amino Acid, Protein folding problem, support vector machine, neural networks, data set**

I. INTRODUCTION

Prediction of Secondary structure of protein is a big problem .But, once the prediction of done one can easily discover the drug of a particular dieases.Proteins are the biochemical compounds consisting of one or more polypeptides. Inred to predict the secondary structure of proteins many methods and techniques are used

(A)Choufasman Method –Analyzed the frequency of the20 amino acid in alpha helix, beta sheet and turn.
When 4 of 5amino acid have a high probability of being in alpha helix, it predicts an alpha helix.
When 3 of 5amino 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 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 Back propagation is used.

(e)Support Vector Machine - Is implemented for classification and regression .It is a machine learning technique in classification Support VM finds a separatinghyperplane in the space of possible value. Thishyperplaneattemptsto divide the positive and negative values The hyper plane 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 hyper plane .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

II. STRUCTURE OF PROTEINS

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 8, August 2016

Proteins are the fundamental components of all living cells. Proteins are the main building block of body. They are used to make muscle Proteins do most of the work in cells. Proteins are made up of linear sequence of twenty amino acid

Amino Acid- Amino Acids are made up combination of carboxyl atom and hydrogen atom.

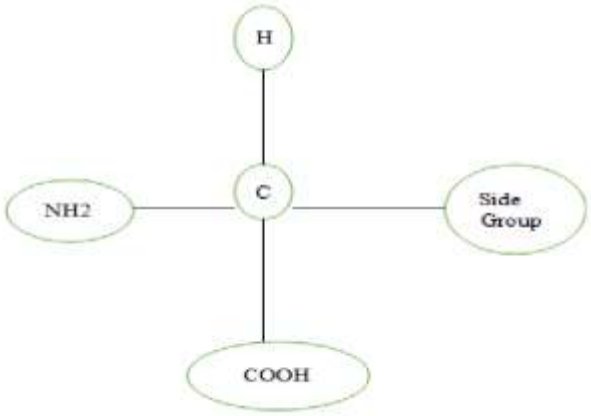

Fig 1: Structure of Amino Acid

Amino acids serve as the building blocks of proteins, which are linear chains of amino acids. There are 20 amino acids that are naturally incorporated into polypeptides. Nine standard amino acid are called "essential" for humans because they cannot be created from other compound by the human body. An alpha amino acid has the generic formula HHNCHRCOOH, where R is an organic substitute; the amino group is attached to the carbon immediately.

III. PROTEIN CLASSIFICATION

Classification of protein done into three classes

Simple Proteins: The protein which are made of amino acid and joined by peptide bond,

Conjugated Proteins : Which are composed of simple protein combined with non –proteinous substance.

DerivedProteins: The proteins which are not occurring naturally but are obtained from proteins.

a. Protein Structure

***Primary Sequence*:-**The primary structure refers to amino acid sequence The structural unit of proteins, joined together by polypeptide bonds .This is known as Primary Sequence.

***Secondary Structure*:-**The next level of protein structure which consists of folding and twisting of the primary structure sequence. Three classes are there: alpha-helices (H), beta-sheet (E) and coil(C).

***Tertiary Structure*:-**The tertiary structure is the three dimensional structure of the polypeptide chain.

***Quaternary Structure*:-**This structure can be determined using a variety of experimental techniques that require a sample of protein in a variety of experimental conditions.

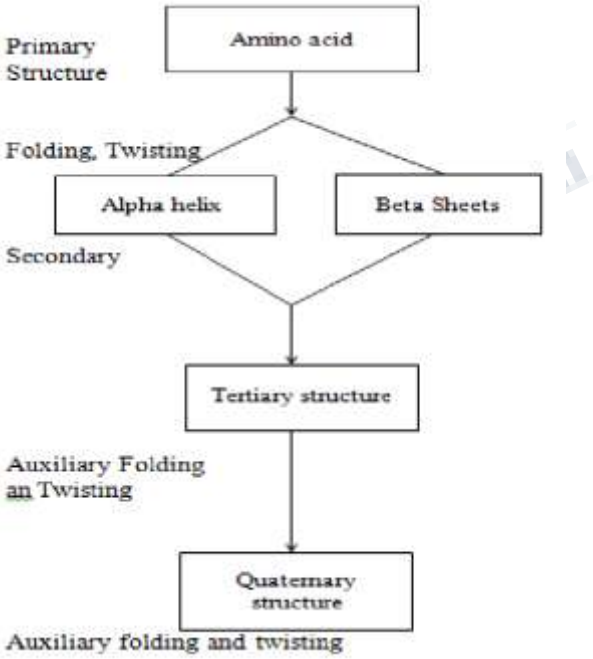

Fig.2: Structure of Protein

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

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 8, August 2016

performed using mat lab...Preparation for the data for processing is done .Pre processing is performed n 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 mat lab 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 –against –one classifier. Finaly at last the comparison between both machine learning algorithms take place.

Steps depicting methodology followed.

- Startmatlab tool.
- Then 62 protein sequence are obtained from Protein Data Bank of CB513 Dataset.
- Both NN and SVM are trained using Matlab.
- Using Frequency profiling and PSSM.Pre-processing of data is done
- Six Binary Classifiers are created
- Prediction of secondary structure is done.
- Both Methods comparison is performed.
- Stop the tool.

Support Vector Machine

- Support Vector Machine is supervised learning techniques.
- Support Vector Machines data has exactly two classes.
- SVM advantage of avoidance of data overfittig.
- An SVM classifies data by finding the best hyper lane that separates all data points of one
- Class from other class.
- Best hyper lane means one with the largest margin between two classes.
- Sometimes small number of mislabeled example decrease the performance
- SVM uses a high dimensional feature space.
- SVM has ability to choose large feature space
- SVM has been used in many applications like image, text classification, and cancer classification.

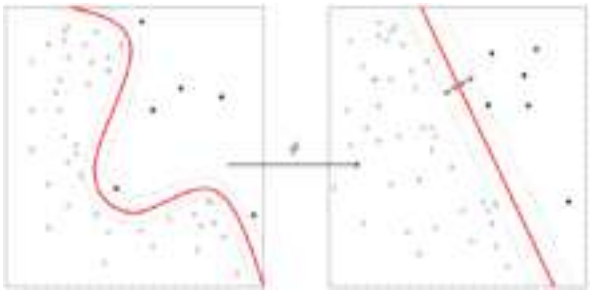

Fig: 4Hyperlane

(a) ***Binary classifiers***

In Support Vector Machine there are 6 binary cashiers 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 = 0.1 will be used and is fixed for all experiments... Their cost parameter was set to 1.5 to construct the classifiers.

Neural Networks

(a)The most accepted method of prediction that is used is Neural Networks.. They are useful for classification and function approximation. There are different types of networks used for different applications. The most commonly used is the Multilayer Feed Forward Networks. For these networks, there are 3 types of layers - input layer, hidden layers and output layer.

-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. Each input in neural network is associated with weights and bias.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 8, August 2016

`Typically, neural networks are trained, so that a particular input leads to a specific target output.. There, the network is adjusted, based on a comparison of the output and the target, until the network output matches the target.

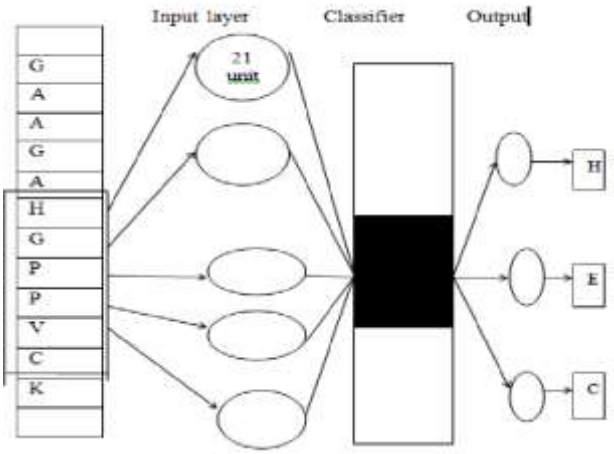

Fig Output for secondry structure prediction of protein

Neural networks have been trained to perform complex functions in various fields, including pattern recognition, identification, classification, speech, vision.

IV. DATASETS

4.1 The form of the data

The dataset consist of 62 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
GPSQPTYPGDDAPVEDLIRFYDNLQQYLNVVTRHRY
CCCCCCCCCTTSCHHHHHHHHHHHHHHHHHTTCC
C

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.

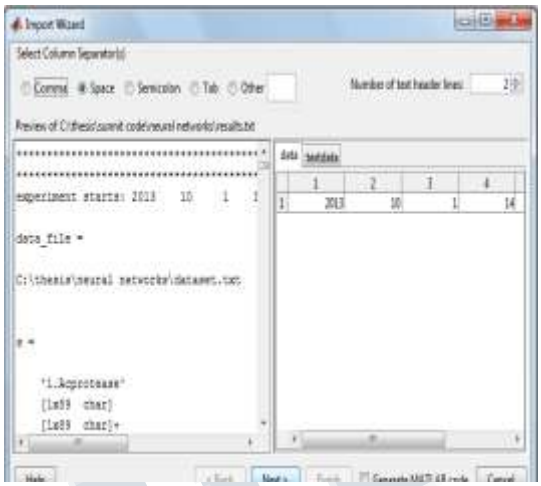

Fig: Dataset of protein

V. RESULTS AND DISCUSSION

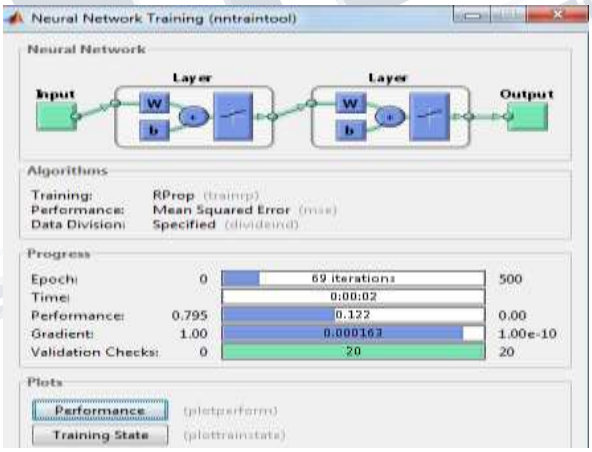

Fig 5.1. Neural Network

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 8, August 2016

Fig 5.2 the calculation of alpha, beta and coil

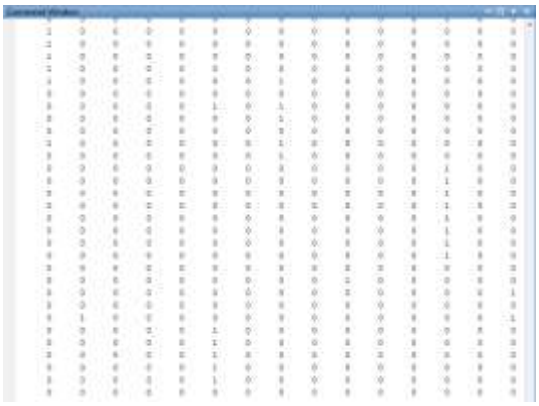

Fig 5.3 the calculation of alpha,beta,coil by using Support vector machine

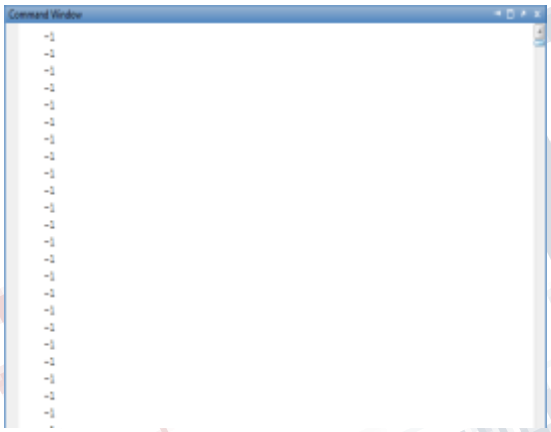

Fig 5 shows the result of calculated alpha, beta and coil using support vector machine

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 percentages of correctly predicted residues. This
Measure is defined as,

$$Q3= \sum (I=H, E, C) \frac{\#\ ofresiduescorrectlypredicted}{\#ofresiduesinclass} * 100$$

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

Binary classifiers	NN% ACCURACY	SVM% ACCURACY
H/~H	76.63	75.32
E/~E	74.69	72.15
C/~C	73.67	71.28
H/E	75.52	74.65
E/C	76.18	72.04
C/H	76.00	74.35

Table5.1: Time taken by both machine learning 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

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 8, August 2016

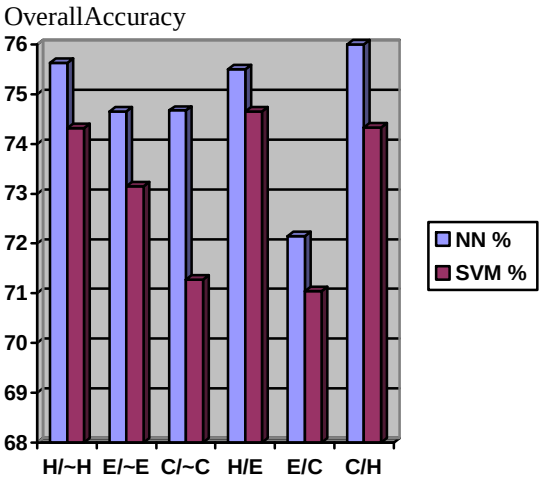

Fig 5.4: Comparison Grapgh Of Accuracy

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%.

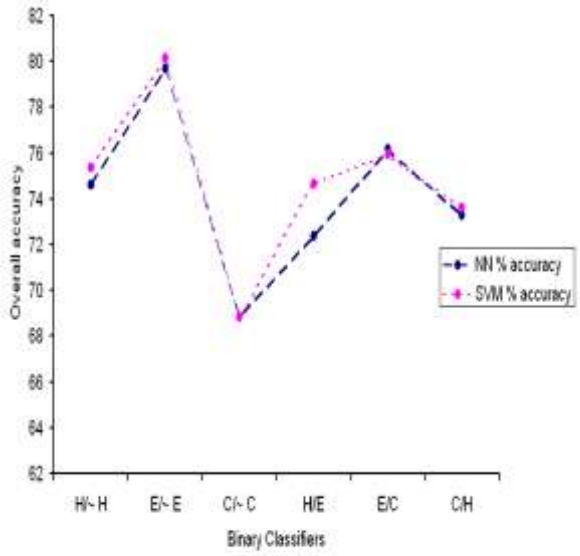

Fig 5.5 Comparison of Accuracy of NN and SVM

VI. CONCLUSIONS AND DISCUSSION

The main aim of this research is 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. When employed with simple network parameters Neural network provides much better accuracy as compared to svm..
2. Comparing with SVM, NN takes much lesser training and computational time.
3. Impairing the memory and processors SVM requires much larger memory and powerful processor. as compare to SVM.
4. Finally NN provides much better results in all the classifiers.

VII. FUTURE SCOPE

The future scope of this research is to improve the furthet accuracy in time . By implementing the Backward feedback accuracy is improved.After forming the best binary classifiers, the new tertiary classifiers will be designed and tested to prove that their performance is best among all the current research methods.

REFERENCES

1. Ibrahim Darwish1, Amr Radi2, Salah El-Bakry3 and El-Sayed M. El-Sayed4 (2015) **"***Protein Secondary Structure Prediction Using Artificial Neural Network Implemented on FPGA"*, International Journal of Bio-Medical Informatics and e-Health Volume 3, No.1, January - February 2015

2. Hanna Hendy, Weal Khaliah, Mohamed Rushdie, (2015) "*A Study Of Intelligent Techniques for Protein Secondary Structure Prediction* "International Journal "Information Models and Analyses" Volume 4, Number 1, 2015

3.Pradeep Singh ,Prof Rajbir Singh, et.al.(2015) "*Improved Protein Function Classification Using Support Vector Machine* "International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015, 964-968.

4. Shavian Agarwal, et.al. (2014) "Prediction *of Secondary Structure of Protein using Support Vector Machine* "International Journal of Computer Applications® (IJCA) (0975 – 8887)

5.Annulet Kaur Johal, Prof. Rajbir Singh (2014)" *Protein Secondary Structure Prediction Using Improved Support Vector Machine and Neural Networks"*International Journal of Engineering and Computer Science ISSN: 2319-7242 Volume 3 Issue 1, January 2014 Page No. 3593-3597

6. Patel Mauri Dinuba, Dr.Hitesh B Shah (2013)" *Comparative Study of Multi-class Protein Structure Prediction Using Advanced Soft computing Techniques"*International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 2, Issue 2, March 2013

7. Shusha Shankar Ray and Shankar K. Pal (2013*)" RNA Secondary Structure Prediction Using Soft Computing"* IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 10, NO. 1,

8.Anil Kumar Mandle1, Pranitha Jain2, and Shailendra Kumar Srivastava (2012)"*Protein Structure Prediction Using Support Vector Machine"*International Journal on Soft Computing (IJSC) Vol.3, No.1,

9FeeXia, Dou Y. ET. Al (2011) "*FPGA accelerator for protein secondary structure prediction based on the GOR algorithm*", BMC Bioinformatics 12(Supple 1):S5.

10..Kaur RK, Kaur M, Kaur A. (2010) "*Using Cluster Analysis for Protein Secondary Structure Prediction*", International Journal of Computer Applications (0975 – 8887) Vol. 4(12).

11.Kumar B, Jani NN (2010) "*Prediction of Protein Secondary Structure based on GOR Algorithm Integrating with Multiple Sequences Alignment"* , International Journal of Advanced Engineering & Applications.

12.Rao P.V.N, et.al (2010)"*Protein Secondary Structure Prediction using Pattern Recognition Neural Network*", International Journal of Engineering Science and Technology Vol. 2(6), pp. 1752-1757

13. Singh R, Diol SK, Sandhu PS (2010) "*Chou-Farman Method for Protein Structure Prediction using Cluster Analysis*", World Academy of Science, Engineering and Technology 72.