ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 2, Issue 2, March 2013

Comparative Study of Multi-class Protein Structure Prediction Using Advanced Soft computing Techniques

Patel Mayuri Dinubhai , Dr.Hitesh B Shah
Information Technology Department, Electronics &Communication Department
G H Patel College of Engineering &Technology, Gujarat Technological University Vallabh Vidyanagar – 388 120

Abstract*—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. Protein structure prediction is depends on its amino acid sequence. Current studies prefer soft computing techniques for classification and regression task. Recently many researchers used various data mining and soft computing tool for protein structure prediction. Our objective is to enhance the prediction of 1D, 2D and 3D protein structure problem using advance soft computing techniques like linear and non-linear support vector machine with different kernel functions. The data base used for this problem is Protein data bank (PDB) select sets which is based on structural classification of protein (SCOP). All proteins in the PDB-40D that had more than 35% identity with proteins of the training set were excluded from the testing set. Using multi-layer neural network we can achieve 53.05% accuracy.

***Index Terms*— Bioinformatics, feature selection (FS), Scoop (Structural classification of protein), Support Vector machines (SVMs), Neural Networks (NNs).**

I. INTRODUCTION

Bioinformatics is an emerging and rapidly growing field of science. As a consequence, a large number of biological data are being collected due to genome-sequencing projects over the world. Therefore, computational tools are needed to analyze the collected data in the most efficient manner. For example, working on the prediction of the biological functions of genes and proteins (or parts of them) based on structural data. Recently support vector machines (SVM) have been a new and promising technique for machine learning. On some applications it has obtained higher accuracy than neural networks [1]. SVM has also been applied to biological problems. In this paper, we exploit the possibility of using SVM for important issues of bioinformatics: the prediction of 1-D, 2-D and 3-D structure from amino acid sequence. The prediction of protein secondary structure and 3-D fold recognition is a challenging field strongly related with function determination which is of high interest for the biologists and the pharmaceutical industry.

Protein structure prediction is one of the most important goals pursued by bioinformatics and theoretical chemistry; it is highly important in medicine (for example, in drug design) and biotechnology (for example, in the design of novel enzymes). The primary structure refers to amino acid sequence is called primary structure. The primary structure is held together by covalent or peptide bonds, which are made during the process of protein biosynthesis or translation. A specific sequence of nucleotides in DNA is transcribed into mRNA, which is read by the ribosome in a process called translation. The secondary structure consists of local folding regularities maintained by hydrogen bonds and is traditionally subdivided into three classes: alpha-helices (H), beta-sheets (E), and coil(C). Tertiary structure refers to three-dimensional structure of a single protein molecule. The alpha-helices and beta-sheets are folded into a compact globule. Quaternary structure is the arrangement of multiple folded protein or coiling protein molecules in a multi-subunit complex.

Many pattern recognition and machine learning methods have been proposed to solve this issue. Surveys are, for example, some typical approaches are as follows: (i) statistical information (ii) physico-chemical properties [11] ;(iii) sequence patterns (iv) multi-layered neural networks [3,12]; (v) graph-theory (vi) multivariate statistics (vii) expert rules (viii) nearest-neighbor algorithms and (iv)support vector machine[1,2,10,13].

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 2, Issue 2, March 2013

Among these machine learning methods, neural networks and Support Vector Machine may be the most popular and effective one for the secondary structure prediction. Up to now the highest accuracy is achieved by approaches using it. In this survey paper, we apply SVM for protein secondary structure prediction. We worked on similar data and encoding schemes as those in Protein Data Bank[1] and Rost &Sander [12](referred here as RS126) which has sharing less than 25% identity. The performance accuracy is verified by a ten-fold cross-validation. Ding and Dubchak [2] indicate that SVM easily returns comparable results as neural networks. Therefore, SVM and its various kernel functions is a promising direction for classification and protein structure prediction.

A. Primary structure

The primary structure refers to amino acid sequence is called primary structure. Each α-amino acid consists of a backbone part that is present in all the amino acid types, and a side chain that is unique to each type of residue. An exception from this rule is proline. The primary structure is held together by covalent or peptide bonds, which are made during the process of protein biosynthesis or translation. The primary structure of a protein is determined by the gene corresponding to the protein. A specific sequence of nucleotides in DNA is transcribed into mRNA, which is read by the ribosome in a process called translation. The sequence of a protein is unique to that protein, and defines the structure and function of the protein. The sequence of a protein can be determined by methods such as Edman degradation or tandem mass spectrometry.

B. Protein secondary structure

The secondary structure consists of local folding regularities maintained by hydrogen bonds and is traditionally subdivided into three classes: alpha-helices (H), beta-sheets (E), and coil(C).Secondary structure contained localized and recurring fold of a polypeptide chain, where two main regular structures are the α-helix and β-sheet. Hydrogen bond is responsible for secondary structure-helix may be considered the default state for secondary structure. These secondary structures are defined by patterns of hydrogen bonds between the main-chain peptide groups. They have a regular geometry, being constrained to specific values of the dihedral angles ψ and φ on the Ramachandran plot.

C. Tertiary structure

The multi-class protein fold recognition or tertiary structure problem is central in molecular biology and it can be formulated as follows: given the primary structure of a protein, how the 3-D fold can be deduced from it. Tertiary structure is an important approach where same structure without relying on sequence similarity. Different types of methods have been developed for fold recognition [3]. These methods [8] are divided into two methodological approaches:

(a) The informatics based methods that involve the sequence based methods and the structure based methods, and
(b) The biophysics based methods.

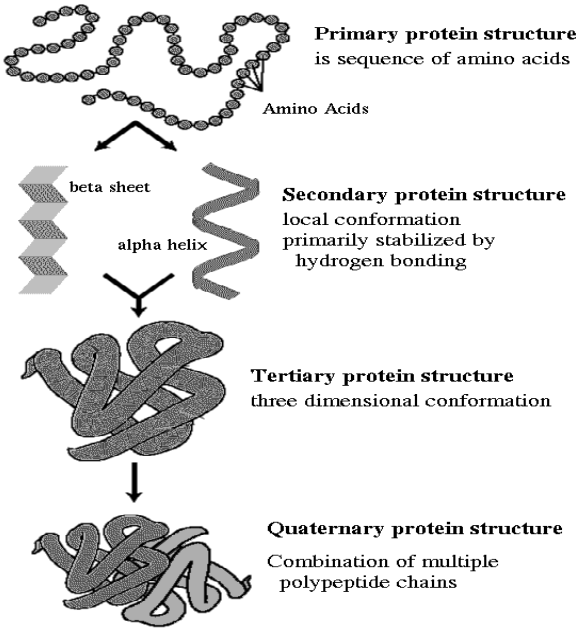

Fig 1 Four Levels of Protein Structure [13]

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 2, Issue 2, March 2013

In fold recognition by threading or structure based, we must take the amino acid sequence of a protein and evaluate how well it fits into one of the known three-dimensional (3D) protein structures [8].

In fold recognition or tertiary sequence based methods is very common. Machine learning techniques, such as genetic algorithms, support vector machines [1, 2, 13], Using Fuzzy Rule-Based Classifier [9] and Multi Layer Perceptron12] Ensemble of Probabilistic Neural Networks [7], have been adopted to exploit protein sequence or secondary structure information. The amino acid composition (protein sequence), in specific, has been employed in many areas of bioinformatics, like protein structural class prediction [3], discrimination of DNA binding proteins and discrimination of outer membrane proteins. However, although significant improvement has been made in the field of fold recognition, the accuracy of the existing methods remains limited and there is a need to develop new methods. Using this block diagram we are solve protein folding problem.

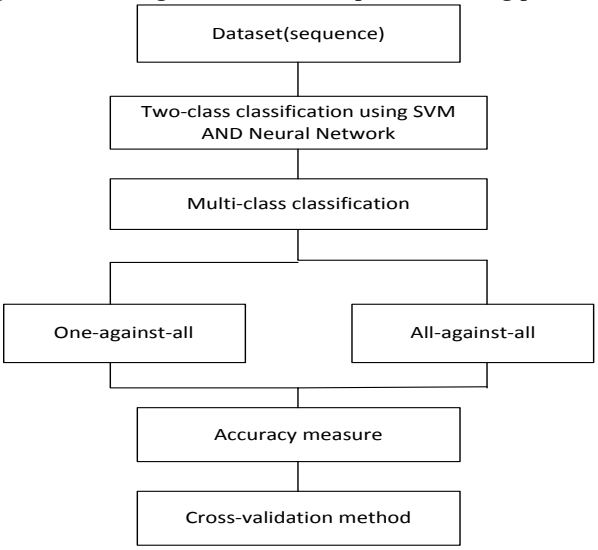

Fig.2 Block Diagram for solve fold recognition problem

D. Quaternary structure

Quaternary structure is the arrangement of multiple folded protein or coiling protein molecules in a multi-subunit complex. Many proteins are actually assemblies of more than one polypeptide chain, which in the context of the larger assemblage are known as protein subunits. In addition to the tertiary structure of the subunits, multiple-subunit proteins possess a quaternary structure, which is the arrangement into which the subunits assemble .Enzymes composed of subunits with diverse functions are sometimes called holoenzymes, in which some parts may be known as regulatory subunits and the functional core is known as the catalytic subunit. Examples of proteins with quaternary structure include hemoglobin, DNA polymerase, and ion channels. Other assemblies referred to instead as multi-protein complexes also possess quaternary structure.

II. PREDICTION METHOD FOR MULTI-CLASS CLASSIFICATION

Many discriminative method like SVM and its variation, NN and kernel methods are often most accurate and efficient when dealing with only two classes. While large number of classes, higher level multi-class methods are developed that utilize two class classification methods as the basic building blocks.

A. One-against-all Method

The earliest used implementation for SVM multi-class classification is probably the one-against-all method. It constructs *K* SVM models where *k* is the number of classes. The *i*th SVM is trained with all of the examples in the *i*th class with positive labels, and all other examples with negative labels. Thus given *l* training data (*x*1, *y*1),....,(*xl*, *yl*), where , *i* = 1,..., *l* and is the class of *xi*, the *i*th SVM is by solving the following problem

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 2, Issue 2, March 2013

$$\min_{(w^i,b^i,\xi^i)} \frac{1}{2}(w^i)^T w^i + c\sum_{j=1}^{l}(\xi^i)_j$$

$$(w^i)^T \phi(x_j) + b^i \geq 1 - \xi_j^i, if\ (y_j = i),$$
$$(w^i)^T \phi(x_j) + b^i \leq -1 - \xi_j^i, if\ (y_j \neq i),$$
$$\xi_j^i \geq 0,\ j = 1, 2, ..., l \qquad (1)$$

B. All-against-all Method

The second method is called the all-against-all method. This method constructs *K*(*K*-1)/2 classifiers where each one trains data from two classes. For training data from the *i*th and the *j*th classes, we solve the following binary classification problem[1]:In this case, the correct class will get the maximum possible votes, which is K-1 for all class-class pairs; and votes for other K-1 classes would be randomly distributed, leading to[K(K-1)/2-K-1]/(K-1)=(K-2)/2 per class on average. Thus we aspect on average *signal-to-noise ratio of*

$$r = 2(k-1)/(k-2) \cong 2 \qquad (2)$$

a fairly large margin. And furthermore, the output class is uniquely generated. In practise, the number of votes for each protein has large variations. The most popularly voted class do not necessarily get maximum possible number of votes; the number of votes for each class tends to decrease gradually from maximum to minimum. For example, the margin between correct class and incorrect classes is not as large as K-1 versus (K-1)/2[2].The class with the highest vote, regardless of whether this vote is a maximum possible vote or not. F set or solving this problem initially require two-class classification method like SVM and Neural Network.

III. TWO CLASS-CLASSIFICATION METHOD

The two multi-class classification methods requires two-class classifier as their building blocks which is describe below:

A. Support vector Machine (SVM)

The support vector machine (SVM) method is a new and promising classification and regression technique proposed by Vapnik and his co-workers (Cortes & Vapnik, 1995; Vapnik,1998). It is especially important for the field of computational biology because it is used for pattern recognition problems including protein remote homology detection, microarray gene expression analysis [12], protein fold-recognition[2,5] , protein structure prediction, promoter recognition[11], prediction of protein–protein interactions[13]. SVM represents novel learning techniques that have been introduced in the framework of structural risk minimization (SRM) inductive principle and in the theory of VC (Vapnik Chervonenkis) bounds. SVM has a number of interesting properties, including effective avoidance of over fitting, the ability to handle high-dimensional feature spaces(determined by w,b), and information condensing of the given data set, etc. Large feature indicate a boundary that maximize the margin between data sample into two classes, therefore give good generalization properties. The decision boundary is defined by the function:

$$f(x) = w \cdot \Phi(x) + b \qquad (3)$$

Protein sequence is depending upon this function, protein x is classified into either of the two classes.

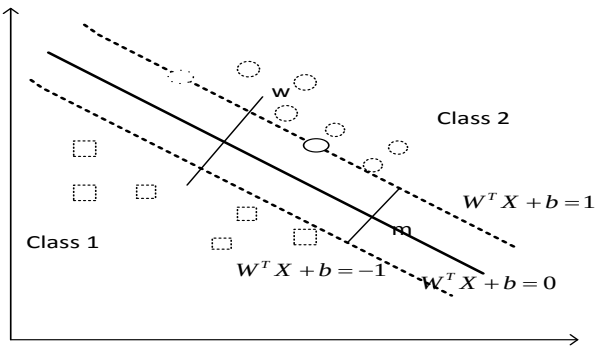

Fig 3 Optimal Hyper plane

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 2, Issue 2, March 2013

In fig 3 for non-linear problems where classes cannot be separated in the original feature space, one can easily converted into higher dimensional space [indicated by $\Phi(x)$], making it easier to optimal hyper plane, i.e. better decision boundary.

B. Neural Network (NNs)

In this Research, we use multi-layer perceptron in this research as a three-layer feed forward network with weight adjusted by conjugate gradient minimization factor. In NNs training there is always problem of generalization; the number of NNs parameters was adaptively adjusted to variable training set sizes by changing the number of hidden units. The perceptron classifies the input vector X into two categories. If the weights and threshold T are not known in advance, the perceptron must be trained. Ideally, the perceptron must be trained to return the correct answer on all training examples, and perform well on examples it has never seen. The training set must contain both type of data (i.e. with "1" and "0" output) which is shown in fig 3.

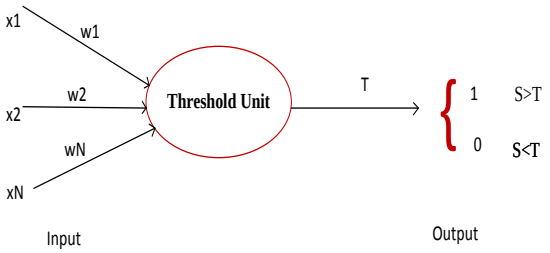

Fig.4 The perceptron

In this work, we use multi-layer perceptron as a three-layer feed forward network in figure 5 with weight adjusted by conjugate gradient minimization factor. Various NNs architecture were tested; but using three layer (1-hidden and 2-output layer) architecture achieves a good performance while having a minimum number of nodes.

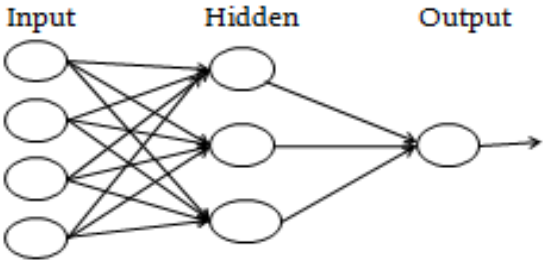

Fig 5 Multi-layer Perceptron

The perceptron computes the dot product S = xw. The output F is a function of S: it is often set discrete (i.e. 1 or 0), in which case the function is the step function.

For continuous output, often use a sigmoid:

$$F(x) = \frac{1}{1+e^{-x}} \tag{4}$$

Neural networks are trained just like perceptron, by minimizing an error function:

$$E = \sum_{i=1}^{Ndata} (NN(x^i) - t(x^i))^2 \tag{5}$$

IV. DATASET

The dataset which we used was selected from the (Ding and Dubchak ,2001).In the database 128 folds, which have seven or more proteins and represent all major structural classes: α ,β, α +β and α/β. since the accuracy of any machine learning tool depends on the number of representative for training, we used 27 most populated fold in this research. This dataset is available on (http://ranger.uta.edu/~chqding/protein/)

A. Feature vector extraction

Feature vector extraction method performed using machine learning tool and its variation. It is a pre-Processing step for protein secondary prediction and protein fold (tertiary structure) and approach focuses on modifying data set to improve the accuracy of the classification..It is extracted from original (primary) sequence based on three

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 2, Issue 2, March 2013

descriptors: 'Composition', composition of three constituents (e.g. polar, neutral and hydrophobic residues in Hydrophobicity);'Transition', the transition of frequencies (polar to neutral and neutral to hydrophobic,etc.); and 'Distribution', the distribution pattern of constituents. We are extracting three classes α,β and coil using Soft Computing technique from original amino acid sequence. Which is given in fig 2.

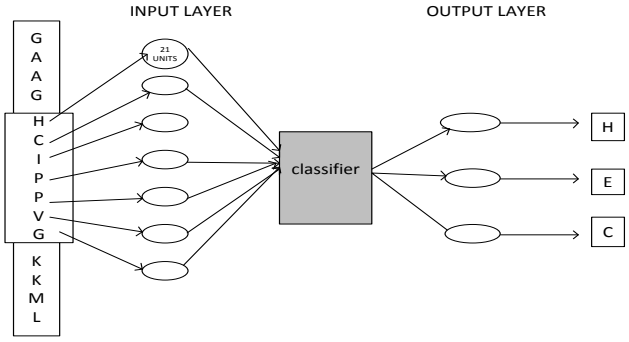

Fig 6 Input and Output Coding for Protein fold problem [13]

B. ASSESSMENT OF PREDICTION ACCURACY

The most common measure for the secondary structure prediction is the overall three-state accuracy (*Q*3). It is defined as the ratio of correctly predicted residues to the total number of residues in the database under consideration [2].*Q*3 is calculated by:

$$Q_3 = \frac{q_\alpha + q_\beta + q_{coil}}{N} \times 100 \qquad (6)$$

(6)

Where q_α =accuracy of alpha, q_β =accuracy of beta and q_{coil} =accuracy of coil .N is the total number of residues in the test data sets.

V. PROTEIN STRUCTURE PREDICTION

A. Primary structure prediction

In this work primary protein structure (DNA, RNA) is converted in to amino acid sequence. For this conversion different algorithm is available like Chou-Fasman, GOR and Neural Network(Qian and Sejnowski).

Original Sequence of protein

MKATCDISECPEGMMCSRLTNKCDCKIDINCRKTCPNGLKRDKLGCEYCECP

↓

CCCCCCCCCCCCCCEHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC

Amino acid Sequence

Protein Sequence: Input 1D

MKATCDISECPEGMMCSRLTNKCDCKIDINCRKTCPNGLKRDKLGCEYCECP

↓

CCCCCCCCCCCCCCHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC

Protein Structure: Output 1D (feed forward network)

B. Secondary structure prediction

In secondary structure prediction, using output of 1-D structure is translated into three classes. Helix (H) and beta (E) and coil and then predicate the accuracy of classification using different method which is see in fig1.7(a) and Fig(b).

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 2, Issue 2, March 2013

Table 1 Secondary structure prediction accuracy of the Neural Network versus that obtained with 4 different protein prediction servers using an identical set of PDB (take only 50 sequence for this result).

Server	q_{α} (%)	q_{β} (%)	q_{coil} (%)
SOMPA	20.36	16.66	37.18
GOR_4	18.16	16.63	40.00
Neural Network (multi net)	17.66	7.832	49.27
Neural Network (single net)	14.32	10.17	53.05
GOR	31.53	14.25	11.24
Chou-fashman	23.06	22.40	29.53

FIG.7BAR GRAPH FOR COMPARING NEURAL NETWORK METHOD WITH OTHER SERVER (INDIVIDUAL ACCURACY)

VI. CONCLUSION

Different algorithms are available prediction: GOR (Garnier Osguthorpe-Robson), Neural lNetwork, Chou-Fasman., jpred. We used neural network for 1-D and 2-D Structure prediction. In Neural Network tool they are used Qian and Sejnowski algorithm which is available online.In this research paper single network give highest total accuracy using PDB dataset (use only 50 sequences here).our future work is prediction secondary structure and tertiary structure using support vector machine and its variation. We are also implement Neural Network in Matlab-2011 using Ding and Dubchak,2001[1] dataset.We are comparing different database as well as comparing techniques like single Network, Feed Forward Neural Network ,Support Vector Machine and its variation.

ACKNOWLEDGMENT

First of all, I am grateful to GOD ALMIGHTY, the most merciful, the most beneficent, who gave me strength, guidance and abilities to complete this project work in successful manner. I am thankful to my project guide **Dr. Hitesh Shah**, Professor, EC Dept for his expert guidance, encouragement and suggestions throughout the preparation of this project work. I would like to thank **Dr. Apurva Shah** Professor and Head of IT Dept , who had always been prepared to offer help at any time in spite having their desk full of nagging priorities.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 2, Issue 2, March 2013

I will always remain grateful to my parents who had been a consistence source of encouragement and whose constant care about me provided me a new direction to work. Last but not the least, I also thankful to all my friends and all other staff members of IT Department for their motivation, support and provision of all necessary facilities during my project work.

REFERENCES

[1] Chries H.Q.Ding and Inna Dubchak, "Multi-class protein fold recognition using Support Vector Machines and Neural Network," Bioinformatics, vol. 17, no.4, pp. 349-358, 2001

[2] Jung-Yang Wang, "Application of Support Vector Machine in Bioinformatics,"2002.

[3] Alessio Ceronia, Paolo Frasconia and Gianluca Pollastrib, "Learning protein secondary structure from sequential and relational data," Neural Networks 18, pp. 1029–1039 2005.

[4] Jieyue He, Hae-Jin Hu, Robert Harrison , Phang C. Tai and Yi Pan b, "Transmembrane segments prediction and understanding using support vector machine and decision tree," Expert Systems with Applications 30, pp. 64-72, 2006.

[5] Hany Alashwal, Safaai Deris and Razib M. Othman, "Comparison of Domain and Hydrophobicity Features for the Prediction of Protein-Protein Interactions using Support Vector Machines, "World Academy of Science, Engineering and Technology 7, pp.431-437, 2007 .

[6] Themis P. Exarchos, Costas Papaloukas Christos Lampros and Dimitrios I. Fotiadis, "Mining sequential patterns for protein fold recognition, "Journal of Biomedical Informatics 41 , pp. 165-179,2008 .

[7] Yuehui Chen, Xueqin Zhang, Mary Qu Yang and Jack Y. Yang, "Ensemble of Probabilistic Neural Networks for Protein Fold Recognition,"IEEE Tran., pp.66-70, 2007.

[8] Robertas Damasevicius, "Analysis of Binary Feature Mapping Rules for Promoter Recognition in Imbalanced DNA Sequence Datasets using Support Vector Machine,"4th International IEEE Conference intelligent Systems , pp. 2008.

[9] Eghbal G. Mansoori, Mansoor J. Zolghadri and Seraj D. Katebi, "Protein Super family Classification Using Fuzzy Rule-Based Classifier,"IEEE Trans. Nanobioscience, vol. 8,no. 1,pp- 92-99, MARCH 2009.

[10] Ioannis K. Valavanis, George M., Spyrou and Konstantina S. Nikita, "A comparative study of multi-classification methods for protein fold recognition," Int. j. Comput. Intelligence in Bioinformatics and Systems Biology, vol. 1, no. 3, 2010.

[11] Abdollah Dehzangi and Bahador Ganjeh Khosravi, "Introducing Novel Physicochemical Based Features to Enhance Protein Fold Prediction Accuracy," International Conference on Computer Design and Applications (ICCDA 2010), no.1, pp. 592-596, 2010.

[12] Wu Qu, Haifeng Sui,Bingru Yang and Wenbin Qian, " Improving protein secondary structure prediction using a multi-modal BP method," Computers in Biology and Medicine 41,pp. 946-959,2011.

[13] Anil Kumar Mandle, Pranita Jain and Shailendra Kumar Shrivastava, "Protein Structure Prediction Using Support Vector Machine," International Journal on Soft Computing (IJSC), vol.3,no. 1,pp. 67-78, February 2012.

AUTHOR BIOGRAPHY

Mayuri patel,. B.E. (Information Technology, Gujarat University), Presently ,persuing Master of engineering in Information technology from G H Patel College of Engineering & Technology, Vallabh Vidyanagar ,2011.My interested area of research is protein structure prediction using soft computing technique and data mining.

Dr.hitesh shah, B.E. (Electronics), Sardar Patel University,M.Tech (Control & Guidance), IIT – Roorkee, PhD (Thesis Submitted), IIT – Delhi. Presently working as **Associate Professor** in Electronics & Communication Engineering Department, G H Patel College of Engineering & Technology, Vallabh Vidyanagar since 1st April, 2011.