A Comparative Study of the Protein Secondary Structure Prediction methods

Shivani Agarwal
AssistantProfessor
IMS Engineering College, Ghaziabad

Arushi Baboota
Student
IMS Engineering College, Ghaziabad

Atul Kumar
AssociateProfessor
IMS Engineering College, Ghaziabad

ABSTRACT

Computationally biology is the innovative research for better drug designing. A number of classifiers and techniques are used for prediction of secondary structure prediction of proteins. The basic aim of this paper shows the comparative study by using these three models: - Artificial Neural Network, Fuzzy Logic, and Hidden Markov Model and to acquire the optimum end result.

Keywords

Artificial Neural Network, Fuzzy Logic, Hidden Markov Model, Soft Computing, DSSP.

1. INTRODUCTION

One of the basic problems in bioinformatics is the prediction of protein secondary structure. Accurate protein secondary structure prediction not only helps in understanding the function and the three dimensional structure of a protein, but is also valuable in determining sub cellular locations and improving the sensitivity of fold recognition methods. Various algorithms have been developed for protein secondary structure prediction

Proteins are energetic structures where conformational variation is very much important for their function [8]. Proteins are naturally made from one set of 20 amino acids. In most of the amino acids, a carbon (the alpha carbon) is bonded to four diverse groups, a hydrogen atom (H), a carboxyl group (-COOH), an amino group (-NH2) and a variable group (R). This protein structure is not stable because of the charge present on the atoms due to which the folding of proteins takes place and these folding results into the various structures such as Primary, Secondary, Tertiary and Quaternary. [9]

The indispensable tribulations in bioinformatics like prediction of the protein structure, phylogenetic deductions etc. are mostly NP-hard in character. Hard computing requires an accurately stated numerical analysis and hence the precise model but a lot of computation time whereas soft computing is lenient of vagueness, uncertainty, and it tries to guess the final output. The soft computing techniques involve the use of fuzzy logic, neural networks and probabilistic analysis.

Artificial neural network and Fuzzy logic provides a better way for solving the complex and robust problems. Fuzzy sets allow for quick processing of information by association of vaguely similar patterns while providing the means to deal scientifically with subjectivity - a territory that traditional science has essentially ignored. Artificial neural networks are fast computational tools with learning and adaptive capabilities whereas fuzzy logic has emerged as a mathematical tool to deal with the ambiguities or uncertainties in human perception and reasoning while the hidden Markov model (HMM), is a type of probabilistic model.

1.1 Biological Neural Network

Biological neural network is a group of interconnected neurons that task to pass signaling and detect or recognize the targeted output, this also known as electrochemical process. Neurons also called brain cells which have the ability of prediction and remembrance. Each neuron is connected to other neuron by using dendrites (that a tree likes structure).they passes the information using synapses, passing information by using Axon. If the sum of all the neurons are comes to the threshold function that it fires otherwise not that known as EXCITATORY State otherwise INHIBITRATORY State.

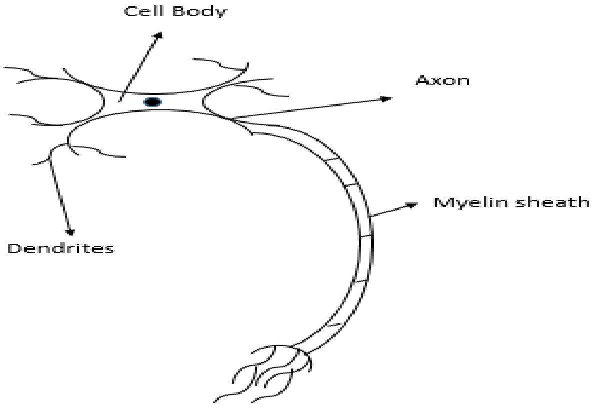

Figure 1 A Biological Neuron

1.2 Artificial Neural Network

Neural Networks take their encouragement from the human mind. As a human mind is a complex structure of the nerve cells, similarly a replica of the human mind is made such that the functioning of the mind can be implemented using the processing units in place of nerve cells. The artificial neural network cannot be made as complex as the biological neural network as it has to be used for training the machine. With such features, an artificial neural system has great latent underlying applications in speech and image recognition where extreme computation can be done in parallel and the computational elements are connected by weighted associations. Well-taught ANNs can predict complex biological patterns, structures, or functions of recently revealed sequences [7].

Artificial neural network accepts a set of inputs to generate the weighted sum which is a product of inputs and connected weights, and then promotes the resulted weighted sum to the other layers. To predict the secondary structure of proteins accurately and hastily, a machine-based learning approach is adopted by the various scientists and researchers which is known as artificial neural network which uses the known patterns to train the network and classifies the unknown patterns using the knowledge gained. The output classification is ended into 3 classes of the DSSP.

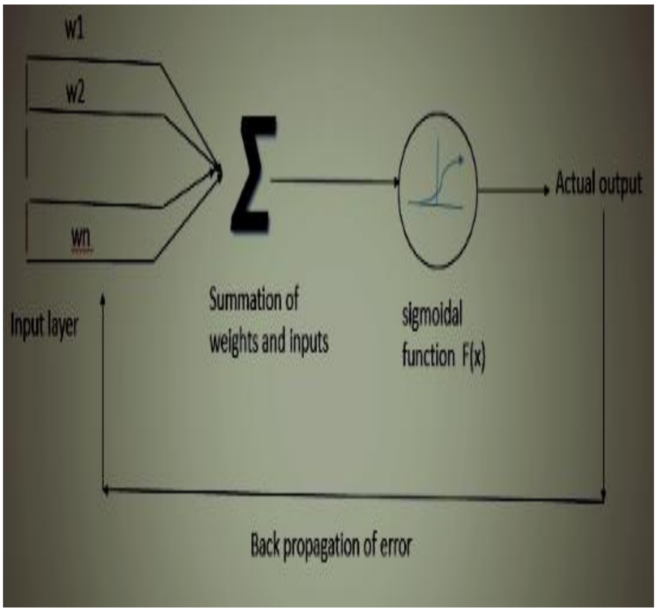

Figure 2 Artificial Neural Network Model with Back Propagation.

1.3 Hidden Markov Model

In Markov models the state is directly visible to the observer, and the state transition probabilities are the only parameters. In a Hidden Markov model, the state is not directly visible [6], but output which is dependent on the state, is visible. Each state has a probability distribution over the possible output tokens.

Hidden Markov models are especially known for their application in speech, handwriting, gesture recognition, bioinformatics etc.

The elements of HMM (λ = (N, M, A, B, π)) include [5]-

a) N: Number of states in the model.

b) M: Number of states in the output alphabet

c) A: transition matrix,

$A = a_{ij}, i, j > N$

d) B: Emission matrix,

$B = b_j, j < N, k < M$

e) Π: State probabilities,

$\pi = < \pi_1, \pi_2, \dots . >$

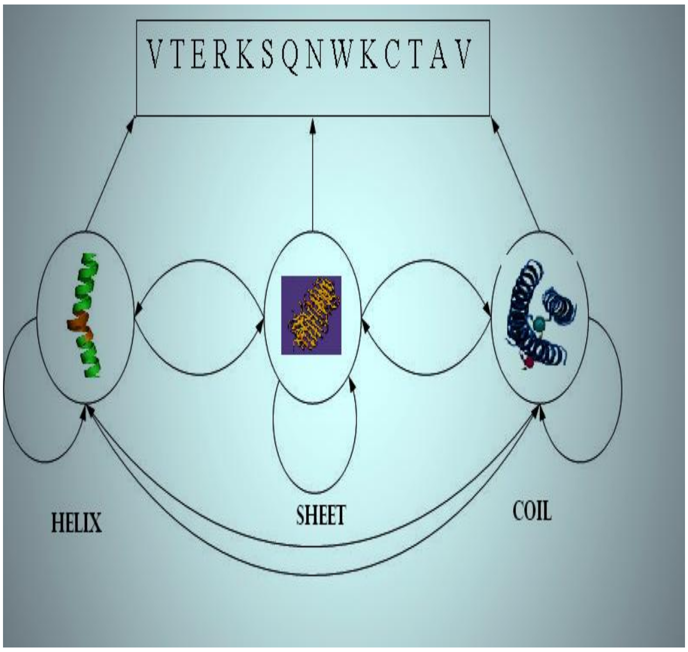

Figure 3: Hidden Markov Model

1.4 Fuzzy Logic

The concept of fuzzy logic was proposed by Lotfi A Zadeh in 1972. The concept of fuzzy logic is not the binary concept it is multivalve concept that uses in control systems. As intricacy increases, accurate statements lose sense and significant statements lose exactness and it becomes much trickier and ultimately impossible to make a precise statement about the behavior of the system. The fuzzy systems convert these rules to their mathematical equivalents. Fuzzy Logic provides a trouble-free way to bring about a definite termination based upon ambiguous, imprecise, noisy, or missing input information. Fuzzy logic concept deals with the problem of pattern classification Simpson proposed a Classifier for reducing this problem that is Fuzzy min-max neural network classifier (FMNN).FMNN concepts uses a concept that is hyperbox fuzzy set concept that is uses for classification of linearly non seperable of multiclass of data,but this is a very complex task to classify data. FMNN having a problem if classify the new data so it can be changed by the learning algforithm it is very difficult task.so for minimizing that problem we use the classifier named as FMCN(Fuzzy min-max classifier with a compensatory neurons) that concept use for minimizing the overlapping of hyperboxes.It learns the data by using the known databases and when new data comes than it checks all the boxes boundary values if the new data value touches the boundary value or threshold value so that it goes for that class that's why it removes ambiguities between the data of hyperboxes.

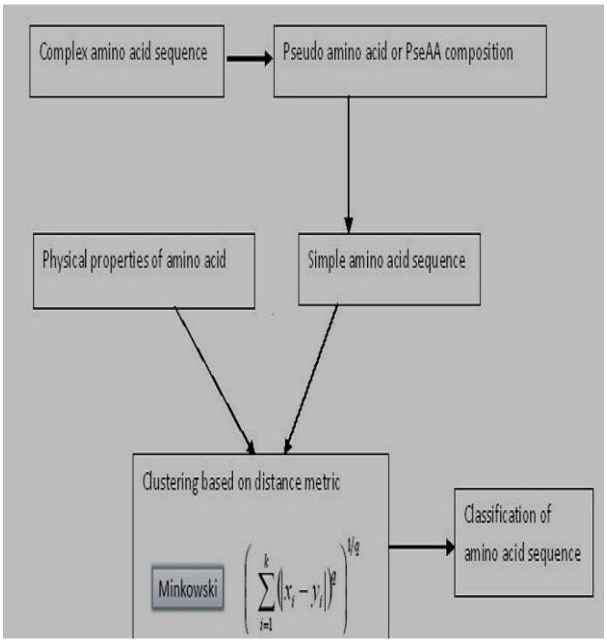
Complex amino acid sequence
Pseudo amino acid or PseAA composition
Physical properties of amino acid
Simple amino acid sequence
Clustering based on distance metric
Classification of amino acid sequence
Minkowski
$\left(\sum_{i=1}^{k}(x_i - y_i)^q\right)^{1/q}$

Figure 4: Fuzzy Logic

2. COMPARISON

Table 1: Comparison of methods

S. No	Techniques	Method
1	Artificial Neural Network	The preprocessing of the data is the foremost step done using frequency profiling i.e. number conversion. 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 Position Specific Scoring Matrix. The training technique incorporated is Resilient Back propagation which leads to a faster convergence.[1]
2	Hidden Markov Model	The amino acid sequence is seen for to get better the hidden, secondary protein structure. Amongst the 20 different amino acids, the models with a first-order Markov chain for the hidden process and a zero-order Markov chain for the known process. The forward- backward algorithm of HMM is used to forecast the structure with the maximum probability at each position of the transition matrix. Training and selection of the HMMs is done using a fourfold rigorous cross-validation course of action. 3 quarters out of four, of the cross-validation learning set is used to estimate the model parameters and the left over quarter, to test the performance. The procedure is replicated about four epochs with non-overlapping cross-validation test sets and the prediction accuracy is the mean values obtained on the four models.[2]
3	Fuzzy Logic	The pseudo amino acid concept can be used to represent a protein sequence with a distinct model without completely losing its sequence-order knowledge and hence is predominantly useful for analyzing a hefty amount of complicated protein. The fuzzy equivalence relation-based hierarchical clustering method is used which avoids any a priori assumption on the number of classes. This relation includes three properties to be satisfied- reflexive symmetric and max-min transitive. The distance between the elements is used to conclude the protein-protein interaction. Clustering analysis of the twenty amino acids is carried on, based on several physical properties for instance the number of codons, molecular weight, hydrophobicity, folding, twisting, curling etc. The impact of the amino acid properties on the classification procedure as well as the effect of the distance metric employed in the clustering process is studied to end with the amino acid classes.[3]

3. RESULT

Table 2: Accuracy achieved till date

S. No.	Technique	Accuracy
1	Artificial Neural Network	NN has achieved the maximum prediction accuracy of about 81%.
2	Hidden Markov Model	Approximately 74%, on single sequences have been achieved.
3	Fuzzy Logic	80% accuracy has been achieved.

4. CONCLUSION

In Bioinformatics, the relationship between amino acid sequence and three dimensional structures is quite essential to understand. This problem comes in the category of NP hard problem and thus its time and space complexity is too high. The whole procedure is performed as sequence of events. This sequence of the procedure starts with the encoding scheme of the amino acids and then it move to predict the secondary structure prediction by using these classifiers. Secondary Structure Prediction employs wide-ranging methods, among which the three main methods have been discussed here and the accuracy of each method is depicted in table 2.

5. REFERENCES

Anureet Kaur Johal, Prof. Rajbir Singh, “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.

[2] Juliette Martin, Jean-Francois, and Francois Rodolphe, “Choosing the Optimal Hidden Markov Model for Secondary Structure Prediction”, French National Institute of Agriculture Research, NOVEMBER/DECEMBER 2005.

[3] D.N. Georgiou, T.E. Karakasidis, J.J. Nieto, A. Torres, “Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition”, Author manuscript, published in "Journal of Theoretical Biology 257, 1 (2009) 17" DOI:10.1016/j.jtbi.2008.11.003.

[4] Ms. Shivani Agarwal, Ms. Arushi Baboota and Ms. Deepali Mendiratta, “Design and Implementation of an algorithm to predict Secondary Structure of Protein using Artificial Neural Network”, International Journal of Emerging Research in Management and Technology, Volume 2.

[5] Rabiner and Juang “Hidden Markov models” lecture by, 1993.

[6] Dr. Stefan Wegenkittl ,,Fachhochschule Salzburg “Pattern Recognition with Hidden Markov Models Dynamic Programming at its Best” by Univ.Doz.,Studien-gang Information stechnik.

[7] Thimmappa S. Anekonda, "Artificial Neural Networks and Hidden Markov Models for Predicting the Protein Structures: The Secondary Structure Prediction in Caspases", Computational Molecular Biology (Biochemistry 218-BioMedical Informatics 231).

[8] Wilkinson,"Computational prediction of protein-protein interactions", Alex BIOC218, Spring 2012.

[9] “Hidden Markov Models: Protein Secondary Structure Analysis”, Algorithms in Bioinformatics - Spring 2004, University of Illinois at Urbana-Champaign.