*International Institution for Technological Research and Development*
*Volume 1, Issue 4, 2016*

# Protein Structure Classification using Fuzzy Logic

**Stebin K Sebastian**

**Department of Computer Engineering, B.V.M Engineering College,**
**Anand, Gujarat, India**
***binsteb@gmail.com***

**Abstract**

**Protein structure classification on a large database is one of the most challenging biological problems. Proteins can be classified according to their similarity in structure or sequential occurrence. Protein structure prediction techniques were introduced to process large amount of biological data to solve the complex biological problems in the real world. These techniques matched the sequence of protein structure to classify them into a certain category or family which exhibits similar characteristics. This system is a proposal of new hybrid technique of adaptive method and fuzzy logic system which works on voting algorithm selecting best possible results from the traditional techniques. Hence the results will be optimized on the basis of threshold values for each helix, beta and coil structure, and the self learning technique used in the system. Accurate prediction of protein structure will be very useful for the medical field as it will acquire the characteristic details of the proteins and help in proper diagnosis.**

***Keywords: Protein family, Artificial intelligence, Neural network, Fuzzy logic, Adaptive network, Secondary structure.***

## 1. Introduction

We A complex molecular structure which includes components like hydrogen, oxygen, carbon, nitrogen etc combines together to form amino acids. Then this group of amino acids combines together by the virtue of covalent bonds between the molecules of acids forms a particular structure of proteins. On the basis of structure and the sequence alignment proteins can be classified into certain family. To reduce the complexity of the protein classification problem adaptive networks and genetic algorithms were introduced. In adaptive networks the system change during the operation cycle on the basis of the results obtained from the previous operation and from that network learns to work on next step. Protein structure prediction is an efficient way of discovering protein models whose structures are not experimentally determined, and this can be made possible with the help of sequence alignment.

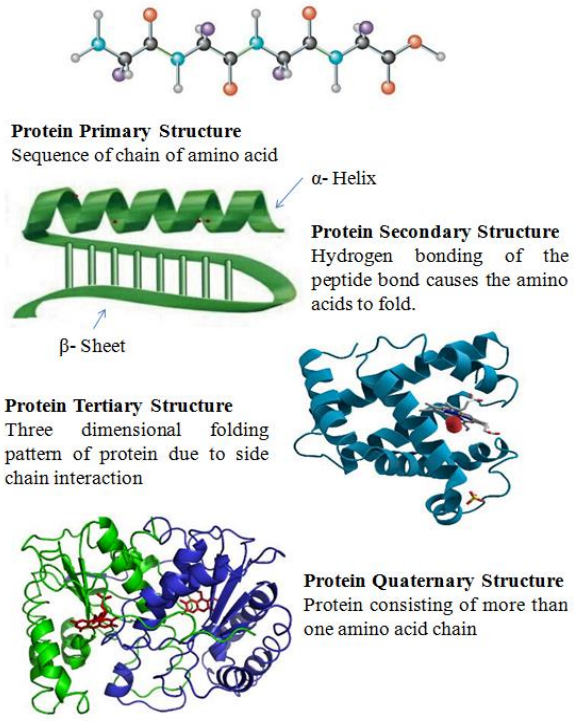


**Fig 1: Four levels of protein Structure**

Protein is a biomolecular structure consisting of four main basic categories: primary, secondary, tertiary and quaternary structure. These structures are formed due to the folding property of protein and formation of peptide bond between the molecules. There exists 20 different amino acids ("Magic 20") found in proteins which are: Alanine (A), Aspartic Acid (D), Phenylalanine (F), Histidine (H), Lycine (K), Methionine (M), Proline (P), ARginine (R), Threonine (T), Tryptophan(W), Cysteine (C), GlutamicAcid (E), Glicine (G), Isoleucine (I), Leucine (L),

AsparagiNe (N), Glutamine (Q), Serine (S), Valine (V), tYrosine (Y). The set of twenty amino acids can be represented by X = {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}. Here the letters A, R, N, etc. are the one-letter codes of the amino acid residues (total 20 possible amino acids). The length of the protein sequence will be N=(n1+n2+…+n20).

$$\{A^{n1}R^{n2}N^{n3}D^{n4}C^{n5}Q^{n6}E^{n7}G^{n8}H^{n9}I^{n10}L^{n11}K^{n12}M^{n13}F^{n14}P^{n15}S^{n16}T^{n17}W^{n18}Y^{n19}V^{n20}\} \rightarrow \{L^{m1}H^{m2}E^{m3}\}$$

1. Primary structure of protein refers to linear sequence of amino acids in a polypeptide chain, and is held together by covalent bonds, which are made during the process of protein bio synthesis or translation.

2. Secondary structures are the most regular sub structure with a basic property of having a certain regular shape which can again be categorized into three different levels referred as three class: Alpha helix (H), Beta sheets (E) and Random coils or loops (L). There exists an eight class categorization technique which can be classified as these classes: α-helix (H), $3_{10}$helix (G), π-helix (I), β-strand (E), bridge (B), β-turn (T), bend (S) and coil (C). Secondary structure defined by the hydrogen bonds of biopolymer and can be observed as an atomic-resolution structure.

3. Tertiary structure will have a single polypeptide chain backbone with one or more protein secondary structure, the protein domain. Amino acid side chains may interact and bond in a number of ways, the interaction of side chain within a particular protein determines the tertiary structure.

4. The protein which contains two or more polypeptide chains are known as subunits. Quaternary structure is the arrangement of more than one protein molecule in a multi-subunit complex. The nomenclature here can be a little confusing because we call a single polypeptide chain a protein if it can function of its own.

## 2. Basic Methods

Various methods of prediction follow adaptive technologies in which the results are generated from the previously generated output. Here the machine learning techniques optimizes the performance creation using example data or past experience, hence some of the methods used for protein structure classification are:

1. **Artificial neural network (ANN)** consists of large number of inter connected processing elements known as neurons which adapts certain properties according to the data sample. They are an information processing system which can be taught as a black box device that accepts input and produces output accordingly and maps input vector onto output vector. There are mainly five application areas which are: pattern completion, noise removal, classification, optimization, stimulation and control.
2. **Fuzzy logic** system presents a general concept for description and measurement which can be used to encode human reasoning into a program to control machineries and decision making. It's a control dynamic system which is used to adjust the regularly changing conditions. Fuzzy logic methods are frequently outperforming the classic mathematical and statistical modeling techniques with real world data and applications.
3. **Support vector machines (SVM)** are a supervised and associated learning algorithm that analyzes the data used for classification and regression. SVMs map input samples into a higher-dimensional space where a maximal separating hyperplane among the instances of different classes is constructed. The method works by constructing another two parallel hyperplanes on each side of this hyperplane. The SVM method tries to find the separating hyperplane that maximizes the area of separation between the two parallel hyperplanes. It is assumed that a larger separation between these parallel hyperplanes will imply a better predictive accuracy of the classifier.
4. **Knowledge based system (KBS**) is a system that uses artificial intelligence techniques in problem-solving processes to support human decision-making, learning, and action. Two types of sub-systems exist in knowledge based systems which are: inference engine and knowledge base. The knowledge base system represents world facts, often in some form of subsumption ontology. Whereas an inference engine represents conditions about the world and logical assertions, usually represented in the form of IF-ELSE rules. Knowledge-based artificial intelligence (KBAI) is the way of using a knowledge bases to inform feature selection or large statistical for machine based adaptive algorithms used in Artificial Intelligence. The main use of knowledge base systems is to train the features of AI algorithms improves the accuracy, recall and precision of these methods. Better results to information queries, including pattern recognition can be attained due to this improvement.

## 3. Accuracy computation

The unit of accuracy of protein structure prediction can be determined using two methods:

1) Average $Q_3$
2) Segment Overlap (sov)

The residue in the terms of (H, E and L) $Q_3$ percentage is measured which can be observed from the following equation:

$$Q_3 = \sum_{(i=H,E,C)} \frac{\text{Predicted}_i}{\text{Observed}_i} * 100$$

Here H represents the number of α-helix, E represents β-sheets and C represents random coil or loops sequence. Now Predicted_i is the total number of correctly predicted residue in state I (H, E, L) and Observed_i is the total number of residue in the sequence.

Segment overlap is done for each data set which attempts occupy segment prediction and their ignorance level varies from 35% (random protein pairs) to average of 91% for homologous protein pairs. Segment overlapping can be calculated by:

$$\text{sov} = 100 \times \frac{1}{N} \sum_{i \in H,E,L} \sum_{s} \frac{\text{minov}(S_{obs}, S_{pred}) + \delta(S_{obs}, S_{pred})}{\text{maxov}(S_{obs}, S_{pred}) \times \text{len}(X)}$$

Where, *N* denotes the total number of residues, minov and maxov are the minimum and maximum overlap in the extent of the segment. δ is the minimum variation with a ratio of 1.0 where there are only minor deviations at the ends of segments. Per-class accuracy criterion $Q_i^{\%obs}$ as given below for $class_i$ is defined as the percentage of correctly classified residues in the $class_i$, to all residues observed in $class_i$. Where, $M_{ii}$ is the number of residues is observed in $class_i$ and classified as i, and $obs_i$ is the total number of residues observed in $class_i$.

## 4. Proposed System

The proposed system is an adaptive fuzzy voting system which is a hybrid method of previously studied systems where the adaptive system will learns from the voting algorithm and predict the protein secondary structure based on the accuracy of matching sequence.

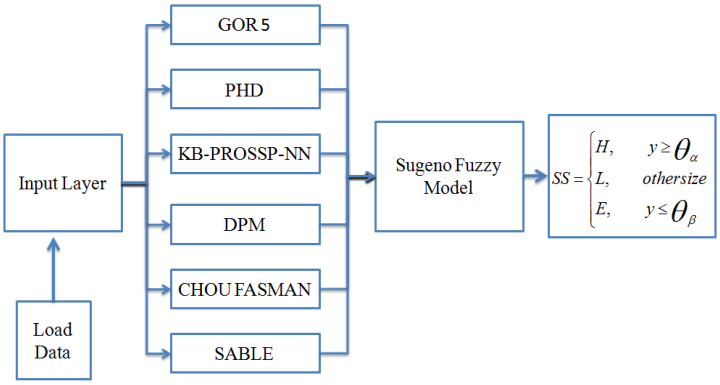


**Fig.2: System Flow**

**1) Adaptive Network:** Adaptation (or learning) is a process of frequently changing a system while the operation is going on in a dynamically changing environment. Without adaptation there is no intelligence. A multilayered feed forward neural network acts as the adaptive network. As shown in figure, the adaptive network consists of six layers.

**2) Voting algorithm:** The main motivation for designing this voting system comes from the unique characteristics of different SS predictors. SS prediction of different methods can in fact improve the quality of the SS prediction. Study shows Fuzz-SSVS improves the SS prediction of RS126, CB396 and CB513 by 1.2%, 1.0% and 0.9%, respectively. Hence this small improvement is obtained only by combining two/six SS prediction methods and using this method combined with adaptive system will increase the accuracy of prediction further.

| Previous System Accuracy | Current System Accuracy |
|---|---|
| DSC (70.1%) | GOR V (78.6%) |
| MUL (67.3%) | PHD (71.6%) |
| NNSSP (73.5%) | DPM (69.7%) |
| PHD (71.6%) | Chou-Fasman (71.4%) |
| PRED (68.4%) | SABLE (78.6%) |
| ZPRED (68.6%) | KB-PROSSP-NN (82.28%) |

Dataset used here is retrieved SS.txt a collection of protein structure sequences containing 290228 sequences with their variant length. From this 100 protein sequences are retrieved with the length of 164 amino acid sequence.

These sequences are then passed through severs of prediction for accruing the secondary sequence and passing it through the fuzzy logic system.
Hence by combining the results of current method the prediction accuracy will have a significant increase from 69.9% to 75.36%.

## 5. Results

The proposed system develops an improved prediction system by collecting results from online severs of secondary structure prediction which takes 164 amino acid sequence as an input

QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAE
DTREMPFHVTKQESKPVQMMCMNNSFNVATLPAE
KMKILELPFASGDLSMLVLLPDEVSDLERIEKTINFE
KLTEWTNPNTMEKRRVKVYLPQMKIEEKYNLTSLM
ALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELS
EDGIEMAGSTGVIEDIKHSPESEQFRADHPFLFLIKH
NPTNTIVYFGRYWSP

Result is generated from six different predictions servers and those results are normalized from crisp data and passes though fuzzy logic system.

i. **GOR V (Garnier-Osguthorpe-Robson):** The GOR method inspects alpha helix, beta sheet, random coil or turn sequences to predict secondary structure at each position established on 17 amino acid sequence windows. The initial specification of the method provides 4 scoring matrices of 17×20 size, where the columns defines the log-odds score, which resembles the probability of finding a provided amino acid at every single place in the 17 residue sequence.
ii. **The Chou–Fasman** technique is only concerned with the probability of each and every individual amino acid which will define helix, strand, or turn. Unlike the complex GOR technique, it does not resemble the conditional probabilities of any amino acid to create any particular secondary structure depending on what structure its neighbor posses.
iii. **PHD:** Predict Protein (PP) is a self supervised service that searches up to date public sequence databases, creates alignments, and predicts aspects of protein structure and function. Users send a protein sequence and receive a single file with results from database comparisons and prediction methods.
iv. **DPM(double prediction method):** was built to elevate the success rate in the secondary structure prediction of proteins by taking into consideration the pre predicted class of proteins. This technique is also known as the 'double prediction method' because it contains first prediction of the secondary structure from a new technique which uses constraints of the method described by Chou and Fasman, and then predicts the classes of proteins from their amino acid composition.

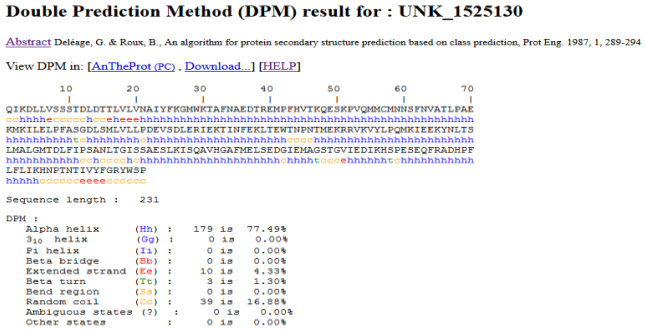


**Fig 3 DPM prediction**

v. **PREDATOR:** a prediction technique of secondary structure which takes a single protein sequence as an input to predict and optimally use a set of sequences which are not aligned as additional information to predict the query sequence. The mean prediction accuracy of PREDATOR is 68% for a single sequence and 75% for a set of related sequences.
vi. **SABLE:** The SABLE server prediction is based on amino acids real value relative solvent accessibility and with the use of evolutionary profiles and predicted accessibility of the solvent residue is identified as a fingerprint in the whole set of protein structure
vii. **KB-PROSSP-NN**: A novel approach of predicting protein secondary structure with the help of 2-tier architecture of neural network and knowledge base is used here. Combination of 5-residue word with analogues structure organizes the knowledgebase. The knowledge base is used as an adequate lookup table followed by hierarchical and lateral validation of 5-residue words and structure. The choice of five-residue words depends on optimization of memory usage and word occurrence frequency. Increasing the value of n reduces the occurrence of the n-residue words, leading to insufficient statistical information for prediction.

**Sugeno Model**: Fuzzy Logic based approach is fashioned to operate as a voting system, particularly the Fuzzy Secondary Structure Voting System (Fuzz-SSVS), to merge results of several Secondary Structure prediction algorithms against developing high-caliber results. As

depicted earlier, SSPP has been gathering attention for certain years and based on various conjecture, patterns of model and methodologies of several techniques which are already fashioned and carried out toward solving it. The prime encouragement behind designing this voting system appears from the particular tendency of distinct SS predictors. However, the above approaches are usually not according to their final results but primarily because they try to occupy distinct protein conditions in their Secondary Structure prediction process.

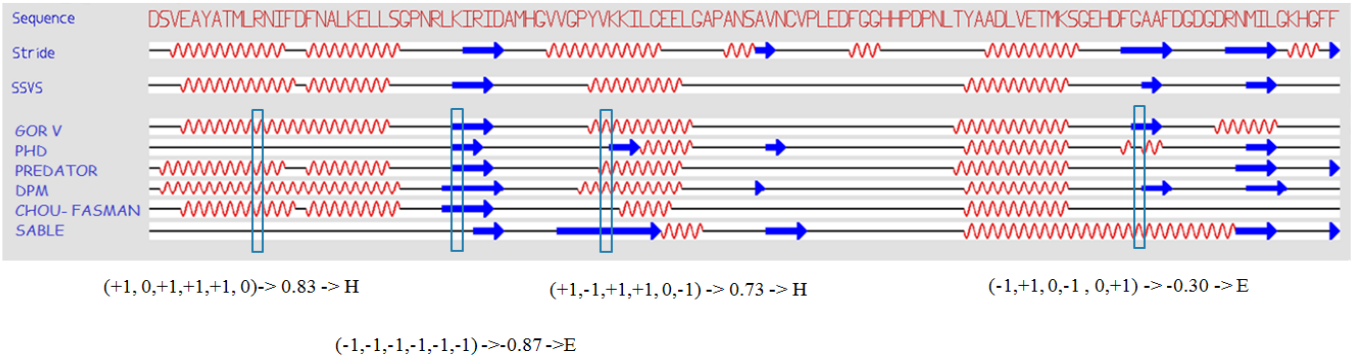


**Fig 4: Sugeno Fuzzy model**

## 6. Conclusion and Future Work

Proposed system increases the prediction accuracy and reduces the complexity of the prior methods used, the fuzzy logic system combined output provides an effective solution for the secondary structure classification. Traditional methods of protein structure classification are used in the system of voting algorithm hence the improved versions of traditional algorithm will provide more accurate result of protein prediction. Different prediction algorithms are available with more accuracy

### Acknowledgments

It is a matter of great pleasure for me to get this opportunity of expressing my deep and sincere gratitude to those who has always helped me throughout my work. I heartily thank my guides Dr. Maulika S Patel and Dr. Mayur M Vegad for providing me their valuable guidance and constant support in my work. They guide me with their excellent knowledge in the area of "Bioinformatics and Artificial intelligence." I would also like to thank all my colleagues who have helped me in some way or the other. I would like to thank my parents & family member for their love and support to build my moral during the work.

## References

[1] Javid Taheri, Albert Y. Zomaya, Flávia C.Delicato and Paulo F. Pires "A Fuzzy Logic Based Voting Scheme to Improve Protein Secondary Structure Prediction" *Computer Systems and Applications (AICCSA )* 2011 IEEE.

[2] Andey Krishnaji, Allam Appa Rao "An Improved Hybrid Neuro Fuzzy Genetic System for Protein Secondary Structure Prediction from Amino Acid Sequence" *Advances in Computing, Communications and Informatics* 2013 IEEE.

[3] Thanh Nguyen, Abbas Khosravi, Douglas Creighton &Saeid Nahavandi "Structural Classification of Proteins through Amino Acid Sequence using Interval Type-2 Fuzzy Logic System" *Fuzzy Systems (FUZZ-IEEE), International Conference* 2014 IEEE.

[4] Bassam M. El-Zaghmouri &Marwan AL-abed Abu-zanona "Protein Family Recognition based on Fuzzy Logic" *International Journal of Emerging Research in Management & Technology* ISSN: 2278-9359 (Volume-2, Issue-4) 2013.

[5] Shahriar Arab, Farzad Didehvar, Changiz Eslahchi & Mehdi Sadeghi "Helix segment assignment in proteins using fuzzy logic" *Iranian journal of Biotechnology* Vol. 5, No. 2, April 2007.

[6] Bo XUI, Hongfei Lin, Zhihao Yang, Kavishwar B. Wagholikar & Hongfang Liu "Classifying Protein Complexes from Candidate Subgraphs using Fuzzy Machine Learning Model" *International Conference on Bioinformatics and Biomedicine Workshops* IEEE 2012.

[7] R. D. King and M. J. E. Sternberg, "Identification and application of the concepts important for accurate and reliable protein secondary structure prediction," Protein Science, vol. 5, pp. 2298-2310, 1996.

[8] G. Barton and W. Taylor, "Prediction of protein secondary structure and active sites using the alignment

of homologous sequences," *Journal of Molecular Biology*, vol. 195, pp. 957- 961, 1988.

[9] A. A. Salamov and V. V. Solovyev, "Prediction of protein secondary sturcture by combining nearest-neighbor algorithms and multiply sequence alignments," *Journal of Molecular Biology*, vol. 247, pp. 11–15, 1995.

[10] B. Rost and C. Sander, "Prediction of protein secondary structure at better than 70 % Accuracy," *Journal of Molecular Biology*, vol. 232, pp. 584-599, 1993.

[11] P. Mehta, et al., "A simple and fast approach to prediction of protein secondary structure from multiple aligned sequences with accuracy above 70 %," Protein Science, vol. 4, pp. 2517-2525, 1995.

[12] M. J. J. M. Zvelebil, et al., "Prediction of Protein Secondary Structure and Active Sites Using the Alignment of Homologous Sequences," *Journal of Molecular Biology*, vol. 195, pp. 957-961, 1987.

[13] Inaki Inza, Borja Calvo, Ruben Arman anzas, Endika Bengoetxea,Pedro Larranaga, and Jose A. Lozano Chapter 2 "Machine Learning: An Indispensable Tool in Bioinformatics".

[14] Maulika S. Patel and Himanshu S. Mazumdar *"*Knowledge base and neural network approach for protein secondary structure prediction" *Journal of Theoretical Biology* 2014.

[15] Smith, Reid "Knowledge-Based Systems Concepts, Techniques, Examples" *Schlumberger-Doll Research* November 9, 2013.

[16] Muhammad Javed Iqbal, Ibrahima Faye, Abas Md Said and Brahim Belhaouari Samir "A Distance-Based Feature-Encoding Technique for Protein Sequence Classification in Bioinformatics" CYBERNETICSCOM 2013 IEEE.

[17] Jiang Xie, Minchao Wang, Dongbo Dai, Huiran Zhang and Wu Zhang "A Network structuring algorithm for Detection of protein family*" 34th Annual International Conference of the IEEE EMBS San Diego, California USA*, 28 Aug - 1 Sept, 2012.

[18] Russel C. Eberhart and Yuhui Shi. "Computational Intelligence: Concepts & Implementation" *Morgan Kalfman Publishers,* 2007.

[19] En-Shiun Annie Lee and Andrew K. C. Wong "Identifying Protein Binding Functionality of Protein Family Sequences by Aligned Pattern Clusters" *International Conference on Bioinformatics and Biomedicine Workshops* 2012 IEEE.

[20] Kailash Shaw and Debahuti Mishra "A Meta-heuristic Framework for Secondary Protein Structure Prediction using BAT-FLANN Optimization Algorithm" *Indian Journal of Science and Technology* July 2015.

[21] Sheng Wang, Jian Peng, Jianzhu MaJianzhu and Jinbo Xu "Protein Secondary Structure prediction using Deep Convolution Neural Feilds" *Scintefic Reports* Nov 2015.

[22] Seethalakshmi Sakthivel and Habeeb S.K.M "NNvPDB: Neural Network based Protein Secondary Structure Prediction with PDB Validation" Biomedical Informatics August 31, 2015.