Evolutionary Neuro-Fuzzy System for Protein Secondary Structure Prediction

Andey Krishnaji[1] and Allam Appa Rao[2]

[1] Assistant Professor, Dept. of Computer Applications, Swarnandhra College of Engineering & Technology, Narasapur, Andhra Pradesh, India 534 280
[2] Former Vice-Chancellor, JNTUK, Director, CRRao Advanced Institute for Mathematics, Statistics & Computer Science, University of Hyderabad Campus, Gachibowli, Hyderabad, India 500 046
{krishnaji.scet,allamapparao}@gmail.com

Abstract. Protein secondary structure prediction is an essential step for the understanding of both the mechanisms of folding and the biological function of proteins. Experimental evidences show that the native conformation of a protein is coded within its primary structure. This work investigates the benefits of combining genetic algorithms, fuzzy logic, and neural networks into a hybrid Evolutionary Neuro-Fuzzy System, especially for predicting a protein's secondary structure directly from its primary structure. The proposed system will include more biological information such as protein structural class, solvent accessibility, hydrophobicity and physicochemical properties of amino acid residues to improve accuracy of protein secondary structure prediction. The proposed system will experiment on three-class secondary structure prediction of proteins, that is, alpha helix, beta sheet or coil. The experimental results indicate that the proposed method has the advantages of high precision, good generalization, and comprehensibility. The method also exhibits the property of rapid convergence in fuzzy rule generation.

Keywords: Artificial Neural Networks, Fuzzy Logic, Genetic Algorithms, Protein, Secondary Structure.

1 Introduction

One of the greatest mysteries on the earth is life. Proteins are essential biochemical compounds for the life to exist on the earth. In order to understand both the mechanisms of folding and the biological activity of proteins, the knowledge of protein secondary structure is essential. Although X-ray diffraction has been accurate and successful in understanding the three dimensional structure of many crystallized proteins, it is quite time-consuming and expensive. Experimental evidences show that the native conformation of a protein is coded within its primary sequence, i.e, amino acid sequence. So, many methods have been developed to predict the secondary structure of proteins from the sequence data. Protein secondary structure prediction is an intermediate step in the prediction of 3D structure from amino acid sequence [2][3][7].

S.C. Satapathy et al. (Eds.): Proc. of Int. Conf. on Front. of Intell. Comput., AISC 199, pp. 19–26.
DOI: 10.1007/978-3-642-35314-7_3

1.1 Levels of Protein Structure

There are four levels of protein structure. They are:

Primary Structure: Primary structure refers to the sequence of the different amino acids of the peptide or protein. The primary structure is held together by covalent or peptide bonds, which are made during the process of protein biosynthesis or translation. The two ends of the polypeptide chain are referred to as the carboxyl terminus (C-terminus) and the amino terminus (N-terminus) based on the nature of the free group on each extremity. Counting of residues always starts at the N-terminal end (NH_2-group), which is the end where the amino group is involved in a peptide bond. The primary structure of a protein is determined by the gene corresponding to the protein.

Secondary Structure: Secondary structure refers to highly regular local sub-structures. These secondary structures are defined by patterns of hydrogen bonds between the main-chain peptide groups. There are three types of local segments:

- **Helices:** Where residues seem to be following the shape of a spring. The most common are the so-called alpha helices.
- **Extended or Beta-strands:** Where residues are in line and successive residues turn their back to each other.
- **Random coils:** When the amino-acid chain is neither helical nor extended.

Tertiary Structure: Tertiary structure refers to three-dimensional structure of a single protein molecule. The alpha-helices and beta-sheets are folded into a compact globule. The folding is driven by the *non-specific* hydrophobic interactions (the burial of hydrophobic residues from water), but the structure is stable only when the parts of a protein domain are locked into place by *specific* tertiary interactions, such as salt bridges, hydrogen bonds, and the tight packing of side chains and disulfide bonds. The disulfide bonds are extremely rare in cytosolic proteins, since the cytosol is generally a reducing environment.

Quaternary Structure: Quaternary structure is a larger assembly of several protein molecules or polypeptide chains, usually called subunits in this context. The quaternary structure is stabilized by the same non-covalent interactions and disulfide bonds as the tertiary structure. Many proteins do not have the quaternary structure and function as monomers.

1.2 Neural Networks, Fuzzy Logic, and Genetic Algorithms (GA)

Neural Networks are information processing systems. They can be thought of as black box devices that accept inputs and produce outputs. Neural Networks map input vectors onto output vectors. Fuzzy Logic provides a general concept for description and measurement. Fuzzy logic systems can be used to encode human reasoning into a program to make decisions or control machinery. More information on neural networks and fuzzy logic can be found in [1][5][7][8][9].

Genetic Algorithms are search algorithms that are based the mechanics of natural selection and natural genetics. Genetic algorithms consist of three fundamental operations: reproduction, crossover and mutation. More information on genetic algorithms can be found in [8][10].

2 Problem Statement

Protein structure prediction is the prediction of the three-dimensional structure of a protein from its amino acid sequence. The primary sequence of the protein contains the full information to determine the three dimensional structure. The primary sequence of a protein can be represented as

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

where the letters are the one letter codes of the amino acid residue (total 20 possible amino acids), and n1, n2, n3, …,n20 represent the number of times the corresponding amino acid code repeats in the protein sequence and n=(n1+n2+…+n20) is the length of the protein to be predicted. The secondary structure of the sequence having length n is $\{L^{m1},H^{m2},E^{m3}\}$, where H, L, E are different secondary structure classes and m1, m2, m3 represent the number of times the corresponding secondary structural class repeats in the secondary structure of the protein. So, the problem of secondary structure prediction can be represented as a mapping problem as follows:

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

3 Hypothesis

In order to develop and implement a better protein secondary structure prediction system, the hypothesis of this work can be stated as follows:

"Constructing and designing advanced well organized artificial neural networks architecture combined with fuzzy logic and genetic algorithms to extract more information from neighboring amino acids can increase accuracy of secondary structure prediction of proteins".

4 Methodology

This section briefly describes the methodological framework used in developing and implementing a method to achieve a better prediction method for the protein secondary structure from its primary sequence (amino acid sequences). Due to the complex and dynamic nature of biological data, the application of conventional methods of machine learning approaches including neural networks without augmentation does not achieve good performance. This work proposes to use a hybrid

computational intelligence technique, which combines artificial neural network approach with Fuzzy Logic and Genetic Algorithms into Evolutionary Neuro-Fuzzy System to include more biological information to achieve a better and more accurate prediction method for protein secondary structure. The architecture of the proposed Evolutionary Neuro-Fuzzy System shown in Fig. (1):

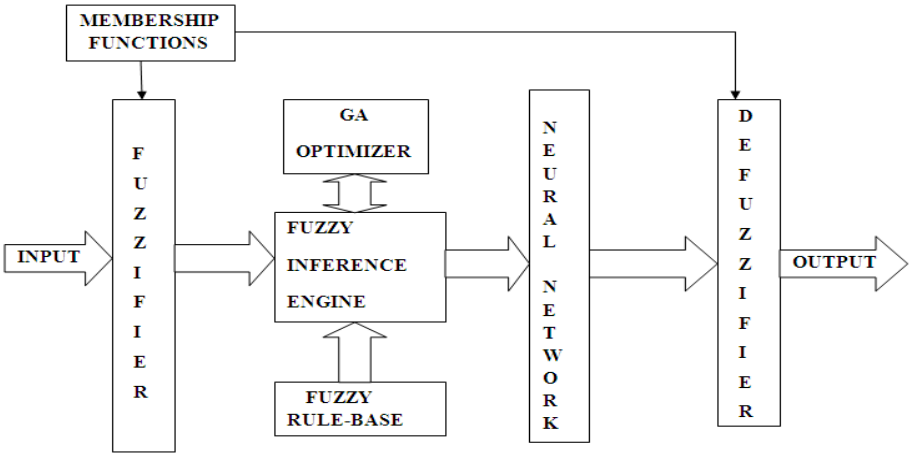

Fig. 1. Architecture of Evolutionary Neuro-Fuzzy System

4.1 Description of Evolutionary Neuro-Fuzzy System

It uses fuzzy data, fuzzy rules, and fuzzy inference. The fuzzy rules and the membership functions make up the system knowledge base [4]. It can handle different types of production rules depending up on the type of the antecedent and the consequent part in the rule: crisp to crisp, crisp to fuzzy, fuzzy to crisp, or fuzzy to fuzzy. The membership functions transform the approximate measurements of protein features into membership values. Fuzzy inference engine activates all the satisfied rules at every cycle and maps the primary fuzzified features to other secondary fuzzy features by using the fuzzy production rules. The membership values are supplied to the input of a pre-trained neural net for classification [1][6].GA Optimizer is a genetic algorithm based optimizer which optimizes the parameters of the rule-base.

4.2 Major Operational Steps of GA Optimizer

(i) Initialize the population
(ii) Calculate the fitness for each individual in the population
(iii) Reproduce selected individuals to form a new population
(iv) Perform evolutionary operations, such as crossover and mutation on the population
(v) Loop to step (ii) until the required condition is met.

4.3 Feature Selection and Normalization

There are 20 amino acids which are named as: alanine(A), aspartic acid(D), phenylalanine(F), histidine(H), lycine(K), methionine(M), Proline(P), ARginine(R), threonine(T), tryptophan(W), cysteine(C), glutamic acid(E), glicine(G), Isoleucine(I), leucine(L), asparagiNe(N), glutamine(Q), serine(S), valine(V), tYrosine(Y). These are called magic 20. 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}. The Amino Acid Index Datsbase (AAindex) contains a number of physicochemical properties of amino acids[5]. The most appropriate features such as hydrophobicity and other physicochemical properties for our work have been selected carefully from AAindex databse. Now an amino acid normalization function is defined as follows:

$$\mathbf{y = f(x)}$$

where x is a member of X and the value of y will be in the closed interval [0,1].

4.4 Data Collection

Protein Sequence data are very much essential for this work for training and testing the proposed Evolutionary Neuro-Fuzzy System. Protein sequence files are openly available in the protein data bank (PDB). The PDB website, *http://www.rcsb.org/pdb,* is used as the main source of data for this work. The data stored in the .pdb files are basically the protein primary structure sequences and the three dimensional coordinates of all the atoms of the amino acid molecules i.e. the residues in the sequence. The format in which the secondary structure is given in these files is not suitable for protein secondary structure prediction. For this reason, the PDB data are transformed into secondary structural data, geometrical features and solvent exposure of proteins. In order to transform PDB sequence data into the suitable and compatible input to proposed Evolutionary Neuro-Fuzzy System, Amino Acid Encoding Normalization methods are used. The data set entries that match the following criteria are include: (a) protein sequences with a length of greater than 80 amino acids (b) protein sequences that have no breaks (c) protein sequences of those proteins whose structures are determined by X-ray diffraction method.

5 Results and Discussion

The data set in our experiment is extracted as described in Data Collection section (sec. 4). The experimental details of Evolutionary Neuro-Fuzzy System are shown in the following screen shots. They show the results when the program runs on the chosen protein data. Figure (2) shows the start up window which contains two main buttons: one for moving to prediction interface and the other to move to prediction analysis. Figure (3) shows a window which allows user to browse the sequences for which secondary structures can be assigned. Figure (4) shows a window which displays tetra peptide averages and corresponding plots. The experimental results indicate that the proposed system has the advantages of high precision, good generalization, and comprehensibility. The method also exhibits the property of rapid convergence in fuzzy rule generation.

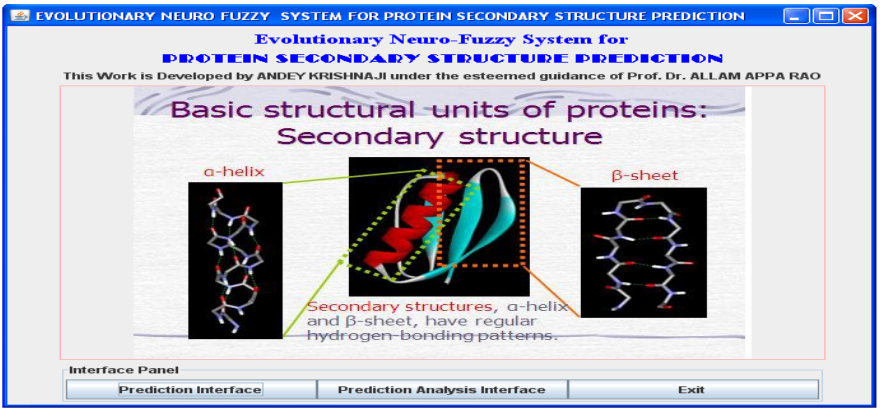

Fig. 2. Start-up window of Evolutionary Neuro-Fuzzy System

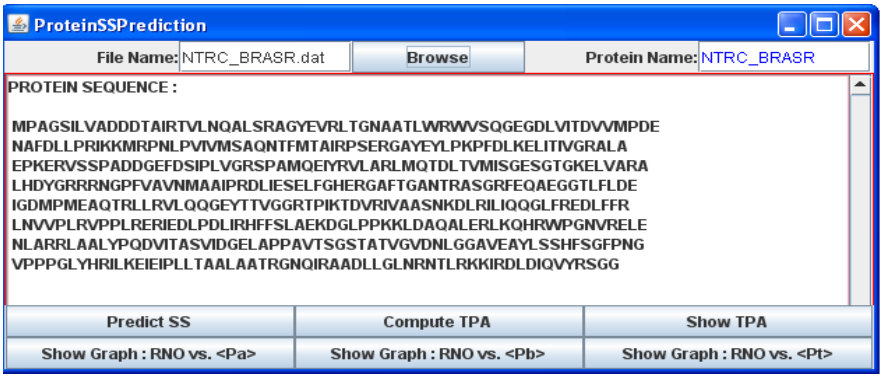

Fig. 3. Select a protein whose secondary structure is predicted

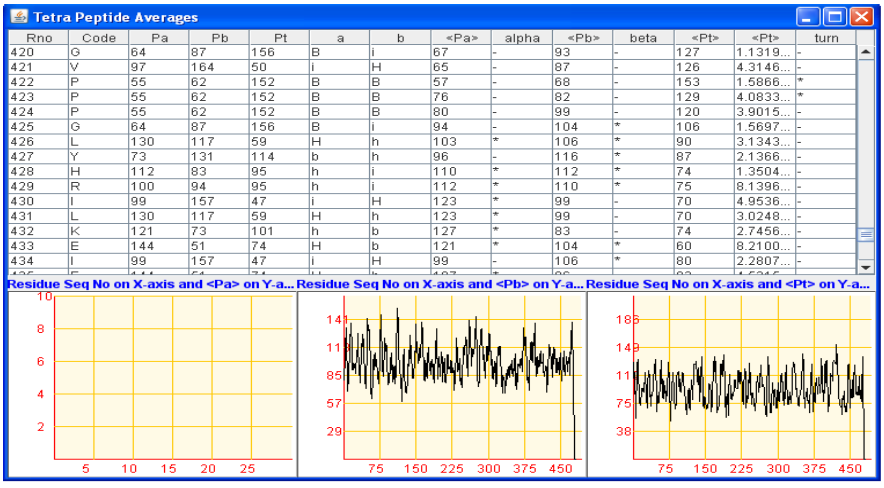

Fig. 4. Compute tetra peptide averages(TPAs)

Secondary Structure Assignment: The following screen shot (fig.5) shows how the system assigns secondary structure classes to amino acid residues in the given protein sequence. The experimental results indicate that the proposed Evolutionary Neuro-Fuzzy System has the advantages of high precision, good generalization, and comprehensibility. The method also exhibits the property of rapid convergence in fuzzy rule generation.

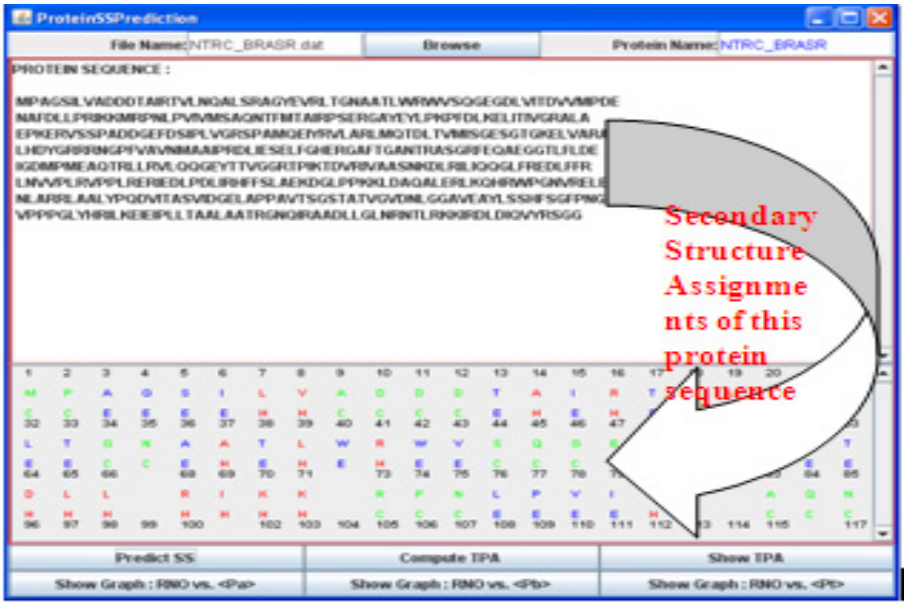

Fig. 4. Assigning Secondary Structure to the target Protein Sequence

6 Conclusion

This paper presents a technique called Evolutionary Neuro-Fuzzy System for protein secondary structure prediction. The main characteristic of this technique is to combine the best properties of both neural networks and fuzzy logic into Evolutionary Neuro-Fuzzy System. This paper also reports an experiment on 3-class secondary structure prediction of proteins using this technique. The experimental results indicate that the proposed system has the advantages of high precision, good generalization, and comprehensibility. The method also exhibits the property of rapid convergence in fuzzy rule generation.

References

1. Jang, J.-S.R.: ANFIS: Adaptive-Network-Based Fuzzy inference system. IEEE Transactions on Systems, Man and Cybernetics 23(0018-9472), 665–685 (1993)
2. Baker, D., Sali, A.: Protein Structure Prediction and Structural genomics. Science 294(5540), 93–96 (2001)
3. Mount, D.W.: Bioinformatics: Sequence and Genome Analysis. Gold Spring Harbor Laboratory Press
4. Sugeno, T., Yasukawa, M.: A fuzzy logic based approach to qualitative modeling. IEEE Transactions on Fuzzy Systems 1(1), 7–31 (1993)
5. Kawashima, S., Kenehisa, M.: AAIndex: Amino acid index database. Nucleic Acids Research 28, 374 (2000)
6. Kasabov, N.K.: Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering. MIT Press (1998)
7. Baldi, P., Brunak, S.: Bioinformatics: The machine learning approach. The MIT Press (2001)
8. Eberhart, R.C., Shi, Y.: Computational Intelligence: Concepts & Implementation. Morgan Kalfman Publishers (2007)
9. Takagi, H.: Introduction to Fuzzy Systems. Neural Networks, and Genetic Algorithms
10. Fausette, L.: Fundamentals of Neural Networks: Architectures, Algorithms, and Applications
11. Weise, T.: Global Optimization Algorithms: Theory and Applications, 2nd edn (2009)