Asian J. Adv. Basic Sci.: **2015,** *3(2),* 17-19
ISSN (Print): 2454 – 7492
ISSN (Online): 2347 – 4114
www.ajabs.org

New Applications of Soft Computing in Protein Folding Problem

P. P. Mishra[1*] & J. K. Verma [2]

[1&2] Department of Mathematics, Govt. P. G. College, Panna, Madhaya Pradesh, INDIA
[*] Correspondence: E-mail: mishrapp@yahoo.com & jk.verma87@yahoo.com

(Received 24 Feb, 2015; Accepted 10 Mar, 2015; Published 11 Mar, 2015)

ABSTRACT: Applications of soft computing in Protein Folding Problems is playing a very important role as it provides techniques that are well suited and intrinsic in nature to obtain results in an efficient way with a precise level of quality. The imprecision and uncertainty that the data and problems in protein folding have can also well be modeled and handled by the application of soft computing techniques. Proteins are large polymeric molecules which form the major part of all living matter. These proteins are responsible for various important functions in the body and it is strictly essential for a protein to be in a native state to perform these functions properly. To perform the central role the proteins have to undergo various bond formations and during this bond formation these get folded. The improper folding of the proteins is called protein folding problem. In this paper the techniques of soft computing have been discussed and it has been shown that how are these techniques superior to traditional techniques.

Keywords: Fuzzy sets; Soft Computing; Protein Folding.

INTRODUCTION: The application of computer technology to the solution and management of biological information is called Bioinformatics. Nowadays this technology is widely used to analyze the databases and recognize protein sequencing. Protein structure prediction and classification has widely been studies in the literature[1-6]. The above mentioned tools can better be employed if we are well versed in the understanding of proteins, their structure and functions.

Understanding Proteins: Any sequence of polymerized amino acids is called protein. Amino acids are Zwitterionic in nature. Thus, proteins are large polymeric molecules which form the most essential part of all living matter through peptide bonds.

Protein Structure: Amino acids are joined end-to-end by peptide bonds during protein synthesis. Various forces work during protein formation and the interaction of molecules gives rise to-bonds creating energy between covalently bonded atoms which is approximated by a harmonic force, angles which also account for the energy in non-equilibrium bond angles, torsions representing the energy due to twisting of a non-single bond and the Pauli repulsion, vanderwaals force and the electrostatic energy as a coulomb potential. Proteins contain tens of thousands of atoms, and their structure is dependent on their interactions with water molecules.

Protein Functions: Proteins can be thought of as little molecular machines, which process other molecules in very specific, repeatable and controllable ways[7]. 'Machine design' is very compactly coded in the sequence of amino acids. The sequence of amino acids is coded in the DNA sequence of the gene for that protein. These proteins form antibodies which are useful in defending the body from antigens. There are some other proteins viz. contractile proteins, Enzymes, Hormonal proteins including insulin, oxytocin and somatotropin etc., structural proteins like keratin which are fibrous and form hair, quills, feathers and horns etc. Transport proteins and storage proteins are other specific proteins for various biochemical transportation and storage. The central roles of proteins in cells make them vulnerable for particular job impropriety.

USE OF SOFT COMPUTING: For protein structure prediction we depend mainly on algorithms and computational methods. The most advanced adaptive genetic algorithm AGA is used for alignment and comparison of DNA, RNA, and protein sequences[8] and protein structure prediction and clustering[9]. But for these and some other techniques like NMR and x-ray crystallography which are of little importance, we come across the problems that computer simulations of protein folding and dynamics pose in the study of protein folding. We know that the duo-core processors can perform very little simulations in a long duration for large atom proteins; therefore we will have to search for the avenues like abinitio, fuzzy ARTMAP and swarm intelligence for our purpose.

In the remaining part of the paper we discuss some important techniques mentioned above. Our source of knowledge is based on the conclusions and discus-

sions of the respective papers describing these techniques.

DATABASES AND DATASETS: There are several databases among which some are primary sequence databases while others are more specialized. A few to name are Gen Bank–an annotated collection of all available DNA sequences, PIR-Protein Information Resource, SWIDD–PROT & TREMBL, TIGR, ALIGN, BLOCKS, DOMO, SBASE, CATH, FSSP, RCSB Protein Data Bank, WORMBASE, PDB, UniProt, SCOP, ASTRAL, GPCRDB, TG, EDD and many others. Most widely used SCOP is now 1.73 and 1.75, Abdollah Dehzangi et. al.[10] have used structural classification of proteins (SCOP). In the latest version of the SCOP, the number of structural classes has increased to 11 groups. Abdollah Dehzangi et. al.[11] has used TG and EDD to investigate the performance of Evolutionary and structural features selection technique.

FEATURE EXTRACTION APPROACH: In their study Abdollah et. al.[11] concatenated features driven from the three main sources (sequential, physiochemical and evolutionary based features) to form a feature vector which is used for the protein structural class prediction problem. In the first step, PSSM was calculated by applying the PSIBLAST on NCBI'S non redundant database for their explored benchmarks. The PSSM consisted of two LX20 matrices where L is the length of a protein and the columns of the matrices represent 20 amino acids. The extracted features were then combined and the potential of all categories of attributes was considered and explored.
Some use GA for its stochastic nature and can be useful for several physical features of importance.

METHODOLOGIES AND COMPARISON: In pattern recognition, neural networks play an important role. AGA, SGA, HP model and monte-carlo are also widely used. Support vector machine is also considered as the state of art classification technique. It was introduced by Vapnik, V. N.[12] aiming at finding the Maximum Margin Hyper plane (MMH) based on the concept of support vector theory to minimize classification error. In the literature[11] it was used to transform the input data to higher dimensionality using the kernel function to find support vectors. The classification of some known points in input space X_i is y_i which is defined to be either -1 or +1. If x′ is a point in input space with unknown classification then:

$$y' = \text{sign} \sum_{i=1}^{n} \left(a_i y_i K(x_i, x') + b \right)$$

Where y′ is the predicted class of point x_i. The function K is the kernel function, n is the number of support vectors and a_i are adjustable weights and b is the bias.
Further, instead of using a single classifier, some researchers have used an ensemble of different classifiers for protein structural class prediction task. A well defined ensemble of these classifiers is capable of addressing statistical, computational and representational issues better than an individual classifier.
For an ensemble classifier, diversity and individual accuracy of its component classifiers are two main criteria that define its classification performance. Fuzzy ARTMAP classifiers coupled with GA have also been used.

CONCLUSION: In our study, we have surveyed various techniques and investigated the use of physicochemical attributes of the amino acids along with the evolutionary based information. The features have been surveyed for various abinitio methods, computational techniques and different classifiers. It has been found that with a skillful input feature vector the use of support vector machine with a little modification is the most efficient classifier.

REFERENCES:

1. Wang and G. B. Huang (2005) Protein sequence classification using extreme learning machine, *Proc. IJCNN'05*, Montreal, QC, Canada, 1406-1411.
2. M. Xiong, J. Li and X. Fang (2004) Identification of genetic networks, *Genetics*, 1bb, 1037-1052.
3. H. Saigo, J. P. Vert, N. Veda and T. Akutsu (2004) Protein homology detection using string alignment kernels, *Bioinformatics*, .20, 1682-1689.
4. I. D. Szustakowski and Z. weng (2000) Protein structure alignment using a genetic algorithm, *Proteins*, 38 (4), 428-440.
5. C. I. Branden and J.Tooze (1999) Introduction to protein structure, *Garland Publications*, New York, 2nd Ed.
6. N. Qian and T. J. Sejnowski (1988) Predicting the secondary structure of globular proteins using neural network models, *J. Mol. Biol.*, 202 (4), 865-884.
7. Daniel B. Dix, Mathematical models of protein folding.
8. Thong Wei et.al. (2007) Parallel protein secondary structure prediction schemes using thread and open MP over hyperthreading technology", the journal of super computing, 41(1), 1-16.
9. A. P. Engelbrecht (2005) Fundamentals of Computational Swarm Intelligence, Wiley.
10. Abdeollah Delizangi et. al., A combination of feature extraction methods with an ensemble of dif-

ferent classifiers for protein structural class prediction problem.

11. Abdollah Dehzangi et. at., Enhancing Protein fold Prediction accuracy using evolutionary and structural features.

12. Vapnik, V. N. (1995) The nature of statistical learning theory", Springer – Verlag, New York.