SOFT COMPUTING PARADIGM FOR PROTEIN FOLDING PROBLEM

Dr. R.C Jain*

Geetika S. Pandey**

ABSTRACT

As the proteins are responsible for the various important functions in the body so its essential for a protein to be in a native state to perform them properly. The native state is decided by the protein structure. Initially all the proteins are arranged as a linear amino acid sequence but gradually due to bonds formation between the amino acids they get folded. Improper folding or the misfolding is thus referred as the Protein Folding Problem. This problem has a real world significance as it may cause many severe diseases called the prion diseases. Protein sequence prediction is the solution for this problem. Remarkable efforts are being made to find an optimum solution for this problem but still its awaited. Out of the various techniques, the soft computing based solutions have proved to be the most efficient ones. In this paper we discuss how soft computing is better than other traditional techniques and also provide a survey on the most promising soft computing techniques which could more efficiently solve the problem. Our emphasis is more on ab initio approaches.

Keywords *: amino acid , prion , protein sequence prediction, soft computing, ab initio*

---

*Director, SATI(D),Vidisha(M.P)

** A.P, CSE Dept., SATI(D), Vidisha (M.P)

1. INTRODUCTION

Advancement in the field of bioinformatics has led to the development of numerous databases and tools which efficiently recognize a protein sequence family and solve the related queries. But since the size of the protein sequence databases is increasing exponentially so its required to develop such efficient systems which could efficiently recognize as well as predict the homology of particular amino acid sequence. Tremendous efforts are been made in this direction but this purpose is most efficiently solved by numerous soft computing approaches. To begin with, let's get more familiar with the proteins, their architecture and functions.

1.1 Proteins : a brief introduction

Proteins are the most abundant and functionally diverse molecules in living systems. Virtually every life process depends on this class of molecules. They play crucial functional roles in all biological processes : be it enzymatic catalysis, signaling messengers or structural elements. Proteins are linear hetero-polymers of twenty different amino acids actually the building blocks. A given protein has always the same amino acid sequence which is determined by DNA sequence and a unique three dimensional structure which is determined by protein sequence.

1.2. Protein architecture:

The protein architecture includes primary structure, secondary structure, super secondary structure, tertiary structure and quaternary structure of proteins. The primary structure is the linear chain of amino acids joined by peptide bonds.

The secondary structure has no amino acid side chains, they have regular patterns of hydrogen bonds and backbone torsion angles, the types of secondary structure are α-helix and β-sheet. Basically it's a local folding maintained by short distance interactions.

The super secondary structure consists of the local arrangements of secondary structure elements like the beta-hairpin turn ,formed due to hydrogen bond formation between two parallel beta strands; beta-alpha-beta unit , beta-meander.

The tertiary structure is a result of additional folding maintained by more distant interactions.

The quaternary structure further is maintained by interchain interactions.

1.3. **Protein Function:**

Proteins perform several functions, some of them are mentioned here- forming antibodies which are the specialized proteins involved in defending the body from antigens. Contractile proteins are involved in muscle contraction and movement. Enzymes are the proteins that facilitate biochemical reactions. Then there are certain hormonal proteins which help to coordinate certain bodily activities like insulin, oxytocin and somatotropin .Structural proteins are fibrous and stringy and provide support like keratins strengthen protective coverings such as hair, quills, feathers , horns. Storage proteins stores amino acids. Transport proteins are carrier proteins which move molecules from one place to another around the body.

2. SOFT COMPUTING APPROACHES

Many experimental and computational methods are available for protein structure prediction. Experimental methods such as Nuclear Magnetic Resonance (NMR) and X-ray crystallography methods produce accurate structures but they are very time consuming and expensive.

The computational method for protein structure prediction is basically divided into : comparative / homology modeling, fold recognition ('threading'), Ab initio ('de novo'). In this paper we mainly discuss ab intio methods, where no template is available for use and so structure is predicted by folding simulation. The de novo methods are based on short segments independently sample distinct distributions of local conformations from known structure. Folding happens when orientations and conformations allow low free energy interactions, which could be optimized by methods like monte carlo search procedure.

A number of soft computing techniques are being continuously introduced as the base of conformational search algorithms and as Yang [5] mentioned that biological inspired algorithms are more efficient than conventional algorithms under appropriate conditions, soft computing techniques proposed for the same purpose till date are back propagation algorithm[12], radial basis function[4], self organizing map, adaptive resonance theory, Fuzzy ARTMAP, genetic algorithm , GA coupled with Multi layer preceptor, Adaptive GA coupled with BP[10], GA with Fuzzy ARTMAP[2], Ant Colony Optimization, Swarm Intelligence and Honey Bees Colony optimization [1].

In the rest of the paper we discuss only the most efficient and latest techniques among the above mentioned. The results mentioned in their respective papers are the source of our decision regarding their efficiency over the other methods.

The paper is further organized as – section 3 mentions the databases, datasets used in most of the research work, section 4 &5 contains the description of the techniques mentioned in this section, section 6 contains survey result and conclusion.

3. DATASET

There are many databases available in open source websites such as PDB, Swiss Prot, Protein Information Resources (PIR), UniProt , SCOP, ASTRAL, CATH, FSSP. As per the research convenience the scientist choose any of these databases , Mansour et al[6] and Osman et al[12] preferred PDB for their research. Whereas others like Rath et al used UniProt dataset. According to Shakir Mohamed et al[3]. Protein Data Bank (PDB) and Universal Protein Resource (UniProt) are serving a particular segment of the protein analysis community. Structural Classification of Proteins (SCOP) version 1.69 database organizes protein sequences according to a hierarchy of protein classes, folds, super families and families with each level of the hierarchy showing a different type of structural relationship between individual protein sequences. SCOP only consists of classification of proteins according to a protein sequence ID but does not supply any of the sequences. They needed the database with no more than 95% sequence identity and so they preferred sequences from ASTRAL Database 1.69 and SCOP. Mohamed et al[2] also used G-Protein Coupled Receptor Database (GPCRDB) for some another work. Ding and Dubchak dataset is also quite popular among the scientists , Hashemi et al[4] used the same for his work.

4. FEATURE EXTRACTION TECHNIQUES

For the feature extraction technique Rath et al[10] used the n- gram hashing function which extracts and counts the occurrences of the patterns of n consecutive residues (i.e a sliding window of size n) from a sequence string. To reduce the size of the input vector which is the major drawback of the n-gram method, they used singular value decomposition technique. The features which they used to construct the real valued input matrix are isoelectric point, molecular weight, atomic composition and the length of amino acid.

According to Hashemi et al [4] , the original training and testing dataset respectively contain 313 and 385 proteins. They mentioned six extracted features namely amino acid composition, predicted secondary structure, hydrophobicity, normalized vander waal's volume, polarity and polarizability, from protein sequences with their dimensions.

For feature subset selection Mohamed et al[3] used GA because of its stochastic nature , GA makes a good tool for feature selection since it can quickly explore the possible search space and determine the global maximum of a fitness solution.

the synaptic weights in order to minimize the error ,applying the delta rule.

5. METHODOLOGIES AND COMPARISON

Basically the neural network are built from simple units called nodes or neurons or cells which are analogy to the real thing. It consists of layers the first layer is the input layer, each node on this layer represents some value of the feature extracted by some technique. The intermediate layer or the hidden layer (if any) simply proceeds or propagates information further to the last layer that is the output layer which gives the final result.

In case of back propagation the information propagates from first layer to the last layer and the error flows backward from the output layer through the hidden layer thus changing

Along with the back propagation algorithm Rath et al[10] used adaptive genetic algorithm , they proposed AGA as an improvement over the simple GA by adaptively varying the probabilities of crossover (Pc) and mutation (Pm) depending upon the average fitness values of the solution in a generation. The AGA stuck at local minima fewer times compared to SGA and adaptively decrease the pc and pm values to protect the high fitness solutions and completely disrupt the average fitness solution. Thus combination of BP and AGA is more efficient than the former coupled with SGA.

Very recently Mansour et al [6] proposed genetic algorithm for protein structure prediction problem based on the cubic 3D hydrophobic polar (HP) Model. They compared their work with Unger and Moult[7] , Patton et al[8] and Johnson et al[9]. Unger et al used GA and Monte Carlo method to fold proteins on 2D and later on 3D lattice. Patton et al , whose work was better, used standard GA and reached higher number of hydrophobic contacts with less number of energy evaluations. Through the experiments Mansour et al proved their work to be more efficient than that of Patton et al in 70% of the 10 cases, whereas the remaining 30% of the cases are identical.

A hybrid of MLP , RBF and Bayesian ensemble method is applied by Hashemi et al [4]. According to them the main idea of classifier ensemble methods is to acquire better classification results by fusing the outcomes of some base classifiers. So they compared their work with the different base classifiers like and produced the best results using RBF and Bayesian classifier i.e 58.96% which was better than the second most efficient method of SVM proposed by Chung et al [11] which produced 56.0%.

Shakir et al [3] used the fuzzy ARTMAP classifiers coupled with GA. This work is then compared with MLP and RBF based classifiers and proved best as the error recorded was lowest.

The advantages of fuzzy artmap over the other methods mentioned in this section are that it creates a mapping of the input feature space , by the division of the input space into hyperboxes, which are related to each other by mapping field. Through the experiments its found that fuzzy artmap takes lesser training time as compared to MLP and RBF which is most desired for a system with larger set of data. Another advantage of this model over others is its ability for incremental learning i.e addition of knowledge to a previously trained system. This thus allows more accurate protein classification.

6. CONCLUSION

As mentioned above this survey has emphasized the most efficient ab initio methods based on soft computing approaches. The survey thus covers the best of soft computing approaches till date. On the basis of experimental results shown by respective researchers we thus conclude fuzzy artmap is the most efficient classifier. And for the feature selection adaptive genetic algorithm could be more promising. Future work is to implement and compare a hybrid of AGA and Fuzzy ARTMAP with some modifications.

REFERENCES:

1. Hesham Awadh. A. Bahamish, Rosni Abdullah and Rosalina Abdul Salam(2006), Protein conformational search using honey bee colony optimization, Proceedings of the 2nd IMT-GT Regional Conference, Malaysia.

2. Shakir Mohamed, David Rubin, Tshilidzi Marwala , 'Incremental Learning for Classification of Protein Sequences', Proceedings of International Joint Conference on Neural Networks, Orlando, Florida, USA, August 12-17, 2007
3. Shakir Mohamed, David Rubin and Tshilidzi Marwala,' Multi-class Protein Sequence Classification Using Fuzzy ARTMAP', 2006 IEEE International Conference on Systems, Man, and Cybernetics October 8-11, 2006, Taipei, Taiwan.
4. Homa Baradaran Hashemi, Azadeh Shakery, Mahdi Pakdaman Naeini,' Protein Fold Pattern Recognition Using Bayesian Ensemble of RBF Neural Networks', 2009 International Conference of Soft Computing and Pattern Recognition.
5. Xin-She Yang "Engineering Optimization via Nature-Inspired Virtual Bee Algorithms" IWINAC 2005, LNCS 3562, pp. 317-323, 2005.
6. Nashat Mansour, Fatima Kanj, Hassan Khachfe,' Evolutionary Algorithm for Protein Structure Prediction', 2010 Sixth International Conference on Natural Computation (ICNC 2010).
7. R. Unger and J. Moult, "Genetic algorithms for protein folding simulations," *Journal of Molecular Biology*, vol. 231, pp. 75-81, 1993.
8. A.L. Patton, W.F. Punch, and E.D. Goodman, "A standard GA approach to native protein conformation prediction," In Proceedings of the 6th International Conference on Genetic Algorithms, 1995.
9. C. Johnson and A. Katikireddy, "Genetic algorithm with backtracking for protein structure prediction," Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation, Washington, USA, 2006.
10. Swati Vipsita, Santanu Rath,' An Evolutionary Approach for Protein Classification Using Feature Extraction by Artificial Neural Network', Int'l Conf. on Computer & Communication Technology 2010.
11. I. F. Chung and C. D. Huang, "Recognition of structure classification of protein folding by NN and SVM hierarchical learning architecture," In Lecture Notes in Computer Sciences (Kaynak, O., Alpaydin, E., Oja, E. & Xu, L., eds.), vol. 2714, pp. 1159-1167. Springer, Istanbul, Turkey. a10, 2003.
12. Mohd Haniff Osman, Choong-Yeun Liong, and Ishak Hashim,' Hybrid Learning Algorithm in Neural Network System for Enzyme Classification', Int. J. Advance. Soft

Comput. Appl., Vol. 2, No. 2, July 2010 ISSN 2074-8523; Copyright © ICSRS Publication, 2010 www.i-csrs.org.

13. Rabindra Ku. Jena, Musbah M. Aqel, Pankaj Srivastava, Prabhat K. Mahanti,' Soft Computing Methodologies in Bioinformatics', European Journal of Scientific Research ISSN 1450-216X Vol.26 No.2 (2009), pp.189-203 © EuroJournals Publishing, Inc. 2009 http://www.eurojournals.com/ejsr.htm
14. G. A. Carpenter, S. Grossberg, N. Markuzon, J. H. Reynolds, and D. B. Rosen, "Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps.," IEEE Transactions on Neural Networks, vol. 3, pp. 698-713, 1992.