Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 6, June 2013, pg.187 – 192

RESEARCH ARTICLE

Protein Structure Prediction using Genetic Algorithm

Subhendu Bhusan Rout[1], Satchidananda Dehury[2], Bhabani Sankar Prasad Mishra[3]
[1]Dept of CSE&A, IGIT Sarang, Odisha, India
[2]Department of Systems Engineering, AJOU University, South Korea
[3]School of Computer Engineering, KIIT University, Odisha, India

[1] subhendu.as@gmail.com; [2] Satchi.lopa@gmail.com; [3] bspmishra@gmail.com

Abstract— Protein Structure Prediction is the process of prediction of the three dimensional structure of a protein from its amino acid sequence. Proteins are large biological molecules which contain large amount of amino acid sequence. The Bioinformatics industry is in the fledgling condition and gaining more attention of researchers. For the development of a new drug there need a high level of research with large amount of data from different cluster and locations. In the field of biology protein structure prediction plays a vital role for the development a new drug. In recent years many techniques are being used for the protein structure prediction. There are many Soft Computing methods like Fuzzy Logic, Artificial neural network, Genetic algorithms, Swarm optimization, etc are used for this purpose to distinguish, compare or process the various structure of protein. It is always a big task for researchers to develop new tools and methods for the purpose of processing of data as well as development of drugs. Protein Structure Prediction is the process of prediction of the secondary, territory & quaternary structure of the protein from its amino acid sequence. Genetic Algorithm is a computational technique which creates the mimic of the process. In this paper we have proposed a technique, which based on Genetic Algorithm technique for the prediction of protein structure. This technique will be helpful to work with huge amount of data and for the prediction of protein structure in a large scale. This technique can be adopted by the medicine researchers to develop various drugs after study and analyzing the changes of protein structure.

Key Terms: - Protein Structure; Bioinformatics; Soft computing; Artificial neural network; Fuzzy logic; Genetic Algorithm; Swarm Optimization

I. INTRODUCTION

Bioinformatics is the application of computer technologies in the field of biological information. Due to the rapid growth in computer technology it is now easier to process various type of biological information through computer. Right now large number of researchers is working upon bioinformatics and many research works also implemented every day. Now there are much more research data and information near us which is very precious and helpful for our future research. Like the secondary data these can be reuse the previous data to modify the newer one to develop a new one. Some cases the improper result after application to various genomic body of a previously developed drug may need some better research or study, which can be fruitful in same gene structure or some other genomic body also. Proteins are large biological molecules consisting of one or more chains of amino acids. Proteins perform a vast array of functions within living organisms, i.e. catalyzing metabolic reactions, replicating DNA, responding to stimuli, and transporting molecules from one location to another. The

Proteins differ from one another primarily from their sequence of amino acids, which is dictated by the nucleotide sequence of their genes and which usually results in folding of the protein structure into a certain three-dimensional structure that determines its activity.

Proteins are the very tiny particle of a living body. Protein Structure prediction is nothing but the prediction of the three dimensional structure of a protein from its amino acid sequence. The protein structure shape changes from time to time after applying the drugs as shown in Fig 1. From the change in its protein structure in the secondary, territory & quaternary structure of the protein a researcher can study what are the effects of a particular drug to that amino acid sequence. So many times a drug may be designed in a proper way but it may not work properly for all type of living body. In that case we are having huge amount of research data, but inefficient study of the changes in the DNA, RNA & Protein structure may cause the harmfulness of the drug or may not work properly. In that case it needs some high level of study or research which may bring out the proper knowledge from the previous data. A highly talented researcher or a high level of technology may bring out the problems or the modifications to short out the problems according to that a new product may design to fit the actual need.

The paper is organised as follows. The Section 2 provides a clean idea about the role of soft computing in bioinformatics. In the Subsection 2.1, 2.2, 2.3 etc give some lesson about various soft computing methods. In section 3 we have discussed about the necessity of protein structure. The section 4 provides the proposed idea of application of genetic algorithm in protein structure prediction. Finally the paper concludes in section 5.

II. Role of Soft Computing in Bioinformatics

Bioinformatics and computational biology present many complex optimization and data mining problems that can be addressed using soft computing methods. The amount of information from biological experiments and the applications involving large-scale high-throughput technologies is rapidly increasing nowadays. Soft computing and machine learning algorithms are well-suited for many bioinformatics problems including gene selection, clustering and classification, signal processing and image analysis. The scope of this special session covers the development of computational models and algorithms for bioinformatics data mining. This chapter will highlight the various applications of soft computing methods to a broad range of topics from bioinformatics.

2.1. Fuzzy Sets

Most biological systems behave in a fuzzy manner, with the interaction and activity of various genes exhibiting different expression-level variations over time. A single gene may be involved in different biological processes. Fuzzy clustering allows genes to simultaneously belong to multiple clusters and participate in multiple pathways. This is a more natural reflection of the biological reality of cellular metabolism. Information integration from multiple sources is utilized to generate biologically meaningful results. Fuzzy systems enable the incorporation of user friendly domain knowledge about some genes into the network. Fuzzy aggregation may be used to combine this information from databases of genes and their products, along with their interactions, within a natural framework in terms of easily understandable linguistic variables (like high, low, big, and small). Fuzzy Adaptive Resonance Theory (FART)-based “associated matrix” method has been developed to cluster gene expression profiles followed by the extraction of genetic networks. The gene expression profiles of Saccharomyces cerevisiae (yeast), responding fewer than two oxidative stresses, were initially pre-processed using FART [1].

The genetic interactions inferred by forming a 2D-associated matrix were quantitatively evaluated and validated in terms of the KEGG metabolic map, BRITE4 protein interaction map, etc. The number of clusters was controlled by the vigilance parameter of FART. Fuzzy rules of an activator-repressor model of gene interactions were used to transform expression values into qualitative descriptors. The algorithm searches for regulatory triplets, consisting of the activator, repressor, and target genes, from yeast data. The entire range of the gene expression data was first ascertained and then classified into low, medium, and high states of varying degree, based on a set of membership functions. Pairing of activator and repressor genes was achieved using Affymetrix gene expression data, where the gene pairs determined the various regulatory elements and their predicted target genes (based on a set of heuristic rules). However, the method was found to be computationally intensive, and was limited to determining possible interactions between one positive and one negative regulator per gene. Clustering has been incorporated, as an interface to a fuzzy logic-based method, in order to improve the computational efficiency, while incorporating information about coactivators or corepressors and the regulatory triplets. The authors claim that this resulted in a gain of almost 50 percent in overall computation time. Experiments were reported on gene expression data gathered from both yeast and rat CNS.5 Interactions between multiple genes were also investigated using a scalable linear variant of fuzzy logic[3]. Incorporation of domain knowledge is expected to improve the performance of clustering. Moreover, enhancing the resolution over regions of interest enables better focussing on useful information. Domain knowledge involving biological processes from several sources, along with a knowledge base of various genes, gene products, and their interactions, aided in the reconstruction of the genetic regulatory networks. A window of varying size was employed to control the level of resolution within the clusters, which were then merged and/or split.

2.2. Artificial Neural Network

The information about 3D structure of proteins is very important factor in medical research. The experimental methods like crystallography used to determine the territorial structure is very time consuming and sometimes it also become failure. It is first easy to determine the secondary structure from amino acid sequence. Using the secondary structure thus determined we can predict the territory structure through methods like threading. The territory structure can give insights into the function of the protein.

Artificial neural networks are computational models which have the capability to adapt or learn, to generalize or to organise or cluster data. They attempt to model the functioning of the brain. In protein secondary structure prediction the neural network learn to predict the correct secondary structure prediction from the amino acid sequence. The basic unit of a artificial neural network is a neuron. The neurons interact with other neuron through weighted connection. Using ANN for prediction contain three phases like

- Training
- Testing
- Prediction

Training a neural network consisting of presenting to the network a set of known input output pairs the weights of the network are adjusted so as to minimize the error between the desired output and the output calculated by the network. An epoch consists of presenting all the training pairs in the set to network a single time. Training consists of many epochs with the order of training pairs randomly jumbled at each epoch. The network weights adjusted is carried out using the popular Back propagation rule for feed forward networks [2].

The representation of the input and output of the neural network is an important factor to be considered during the network design. The neural network is trained to recognize and adapt to patterns in the input. So the representation of the input plays a major role in determining the efficiency of the network. The output representation of the network classifies the current residue to be predicted into 3 distinct classes.

Alpha Helix- H
Beta Strand-E
Coil or loop-C

The Predictor consists of two different neural networks. Both networks consist of three layers, input layer, hidden layer and output layers. The first network is fed with fixed size windows of PSIBLAST profiles of the input amino acid sequence. The output layer of this network use orthogonal encoding to represent the three different classes H, E & C.

2.3. Neurocomputing

The application of ANNs for data-rich environments and their robustness to noise makes them appropriate for modelling genetic interactions from gene expression data. Some such type of models, employed for extracting genetic regulatory effects, include perceptrons [4], self-organizing maps [5], and recurrent neural networks (RNNs) [6]. The total Environment of feedback connections and dynamic memory units make RNNs very much preferable for modelling global gene regulatory networks. The output of each neuron is fed back to its input, after unit delay, and is connected to other neurons. In the gene regulatory network, each gene in the network is considered as a neuron. Therefore, the RNN is able to model interactions between genes, including self-regulation. Back propagation training is employed by unfolding the temporal operation of the network onto a layered feed forward one. Each time step is mapped onto an additional layer of the network. After that, the gradient descent is typically performed in order to minimize the mean-squared error between the expected output and the generated network output values across time. In this case the mean-squared error should be differentiable with respect to the individual weights of the network. The convergence of the method is sometimes also affected by the limited number of samples in gene expression data [1].

The RNN was used to model the dynamics of gene expression in the lambda phage9 regulatory system [53]. However there may several probable answer, involving positive and/or negative feedback, was considered. Here, every node in the networks provides the expression value of a gene at a certain instant of time. In the other hand the nodes in the subsequent layers referred to the expression values at the next time points. The connection between nodes in successive layers was indicative of the regulatory action between the corresponding genes.

Since the regulatory process of a gene depends both on transcription and translation, a pair of linked networks were employed to independently model them.

2.4. Swarm Optimization

Another variant of evolutionary computation known as Particle swarm optimization (PSO) [7] has been employed for evolving the RNN and determining its parameters while inferring the architecture and connection weights of genetic regulatory networks from gene data. A sparsely connected weight matrix provides a model of gene regulatory interactions. The particle swarm optimization works in parallel on a set of candidate solutions, providing information about the previous optimal solutions. However, some of the other desirable characteristics of PSO include the flexible balancing of global and local searches, computational efficiency in terms of time space, and easy implementation. PSO is particularly suitable for evolving neural networks as a continuous optimization problem without getting stuck in local minima. Here, the i^{th} particle of the swarm, at position xi, is considered to move in a multidimensional space with velocity v_i. Each particle randomly searches the network structure space by updating itself with its own memory, along with the information gathered from the other particles in the swarm. While the particles previous best solution is given as V_i, the corresponding best position of the whole swarm is stored as P_g. Again, P_g can be replaced by a locally best solution obtained within a certain topological neighbourhood. A fitness function is used to measure the deviation of the network output E(t) from the real measurement target of the expression value. However, a sufficient number of runs are required to assure the quality of the inferred regulatory networks. Thus PSO is applied for continuous optimization, Ant colony optimization (ACO) has been used in conjunction with PSO to train an RNN [9], and infer the gene regulatory network from time-series gene expression data. The ACO is a discrete optimization technique employed in combinatorial problems. Hence, ACO was employed for determining the structure of the RNN, and the PSO helped in generating the optimal parameters for a network bearing a given structure.

III. NEED OF PROTEIN STRUCTURE PREDICTION

Protein Structure Prediction is the process of prediction of the three dimensional structure of a protein from its amino acid sequence. It is always a big challenge for researchers to develop new tools and methods for the processing of data as well as development of drugs. In order to develop a new drug it needs to process huge amount of data to study the behavior of various types of genes. In recent years many techniques are being used for the protein structure prediction. Recently the Bioinformatics industry is in the fledgling condition and gaining more attention of researchers. Various Soft Computing methods like Fuzzy Logic, Artificial neural network, genetic algorithms, swarm optimization, etc are used for this purpose to distinguish, compare or process various type of data. After application of various drugs or from any reaction of external agents the structure of the amino acid sequence changes. The study of these changes provides some information to the medicine researchers about the reaction and changes of the amino acid sequence. From these changes they can study the reaction and behavior of genes. As the number of these type of gene is very high so there should be a platform which can process huge amount of this type of genomic data. Thus application of computer makes some clumsy information distinguish easily, but there should be some technique to connect the data with the computer system. So it is a major chapter for the recent researchers and scientists for the prediction of protein structure for the designing of drugs.

The increased accessibility of genomic data and especially, that of large-scale expression data has opened new possibilities for the search for target proteins. The development of such technique to prompted large-scale investments with the new technology will be adopted by many pharmaceutical companies. The respective screening experiments rely critically on appropriate bioinformatics support for interpreting the generated data. Specifically, methods are required to identify interesting different expressed genes and to predict the function and structure of target proteins from differential expression data generated in an appropriate screening experiment. Proteins spontaneously fold into intricate three-dimensional (3D) structures. The advances of techniques such as X-ray crystallography and nuclear magnetic resonance have brought about the determination of more than 10,000 protein structures. Ever since there were thousands of 3D structures, comparing or aligning them has been an important technique for elucidating fundamental principles of protein structure, function, and evolution. Structure alignment involves establishing equivalencies between residues in two proteins based on their 3D coordinates. A structure alignment is essentially a list of residue pairs from two proteins that should superpose closely after one protein is translated and/or rotated rigidly.

IV. PROTEIN STRUCTURE PREDICTION USING GA

Genetic Algorithms (GAs) are adaptive heuristic search algorithm based on the evolutionary ideas of natural selection and genetics. Genetic Algorithm is a model of machine learning which derives its behaviour in the form of a metaphor of the processes. GAs represents an intelligent exploitation of a random search used to solve optimization problems. This is better than conventional AI techniques as the GAs do not break easily even if the inputs changed slightly or in the presence of reasonable noise. GAs simulate natural evolution , mimicking processes the nature uses: Selection, crosses over, mutation and accepting. Genes code for proteins that may, in turn, regulate the expressions of other genes. If we go for a comparison we can observe that although caterpillars and butterflies have identical genomes, the difference in expression of the genes in the DNAs lead to their obvious physical difference. Such events motivate the modelling of gene expression networks. Rebuilding of genetic networks from gene expression profiles allows the discovery of various functions ranging over diverse domains like molecular biology, biochemistry, bioengineering, and pharmaceutics.

The difference between two genomic bodies may be the difference between their evolution, DNA RNA dissimilarity, protein structure difference. The proteins are basically differing from their amino acid sequence. The sequences may be in the form of α Helix, β sheet etc. The nature and shape of the amino acid sequences changes from time to time. This is because of some external agents or the application of some drug. Generally these types of gene data comes in a huge amount which need a dedicative platform to process these and interpret these. Though the basic shape changes in the secondary, and again its shape changes in the territory and quaternary positions. The process of studying the changes of behaviour and shapes in the primary, secondary, territory and quaternary is very much useful in the design of drugs. The drug scientists use these data for designing of various drugs. But these processes are a very time taking, expensive and clumsy process. There should be some good technique to make these processes in an easier way. Genetic Algorithm is such a good technique for these basic needs. The GA provides a metaphor of the processes. So by the application of Genetic Algorithm for protein structure prediction problem we can predict the territory and quaternary changes in the amino acid sequence from the metaphor. So the genetic algorithm is very much fruitful for these protein structure prediction problems.

V. Conclusion and Future Work

There are various types of macromolecules like polysaccharides, nucleic acids; proteins etc are in a living body. Bioinformatics is a better way to study the behavior of DNA, RNA and to prediction of the protein structure. In the field of computer science and Artificial Intelligence a genetic algorithm is a model of machine learning which derives its behavior from a metaphor of the processes of evolution of nature. This is done by the creation within a machine of a population of individuals represented by chromosomes. Thus application of bioinformatics provides a gateway to process huge amount of data but Genetic Algorithm is very helpful to predict the structure of Protein or study the behavior of DNA, RNA, etc in a small amount time. So in a way of providing a metaphor of the processes the genetic algorithm is very useful for designing of various drugs, after processing of huge amount data with less amount of time. Biochemistry is always a good and upcoming topic for every researcher as every day a lot of data are being processed & many changes to drugs are being developed. This Paper gives a proposed idea and our next work will focus upon the various real time effects and development of bioinformatics & genetic algorithm upon various macromolecules. It also includes how these processes can be carried out in a beneficiary mode with less amount of time.

References

[1] S. Mitra, R. Das, and Y. Hayashi, “Genetic Networks and Soft Computing”, IEEE/ACM Transction on computational biology and bioinformatics, Vol. 8(1), pp. 94-107, 2011.

[2] D.A. Joseph, “Protein Secondary Structure Prediction using Artificial Neural Networks” IIT Madras, India, 2002.

[3] N. Friedman, M. Linial, I. Nachman, and D. Pe’er, “Using Bayesian Networks to Analyze Expression Data,” J. Computational Biology, vol. 7, pp. 601-620, 2000.

[4] S. Kim, E.R. Dougherty, Y. Chen, K. Sivakumar, P. Meltzer, J.M. Trent, and M. Bittner, “Multivariate Measurement of Gene Expression Relationships,” Genomics, vol. 67, pp. 201-209, 2000.

[5] P. Toronen, M. Kolehmainen, G. Wong, and E. Castren, “Analysis of Gene Expression Data Using Self-Organizing Maps,” FEBS Letters, vol. 451, pp. 142-146, 1999.

[6] J. Vohradsky, “Neural Network Model of Gene Expression,” FASEB J., vol. 15, pp. 846-854, 2001.

[7] J. Kennedy, R. Eberhart, and Y. Shi, Swarm Intelligence. Morgan Kaufmann, 2001.

[8] R. Xu, D.C. Wunsch, II, and R.L. Frank, “Inference of Genetic Regulatory Networks with Recurrent Neural Network Models Using Particle Swarm Optimization,” IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 4, no. 4, pp. 681-692, Oct.-Dec. 2007.

[9] H.W. Ressom, Y. Zhang, J. Xuan, Y. Wang, and R. Clarke, “Inference of Gene Regulatory Networks from Time Course Gene Expression Data Using Neural Networks and Swarm Intelligence,” Proc. IEEE

Symp. Computational Intelligence in Bioinformatics and Computational Biology (CIBCB ’06), pp. 435-442, 2006.

[10] Sarkis M., Diepold K., Westad F., “A new algorithm for gene mapping: Application of partial least squares regression with cross model validation”, IEEE International Workshop on Genomic Signal Processing and Statistics, 2006.