Bioinformatics in Neurocomputing Framework

Shubhra Sankar Ray, Sanghamitra Bandyopadhyay, Pabitra Mitra and Sankar K. Pal
Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108
Email: {shubhra_r,sanghami,pabitra_r,sankar}@isical.ac.in

CODEC-04, pp. 94, Januray 1-3, 2004, kolkata

Abstract - **Bioinformatics is an interdisciplinary research area of biology and computer science. This article provides an overview of neural network applications in different bioinformatics tasks. The relevance of intelligent systems and neural networks to these problems is first mentioned. Different tasks like gene sequence analysis, gene finding, protein structure prediction and analysis, microarray analysis and gene regulatory network analysis are described along with some classical approaches. Different neural network based algorithms to address the aforesaid tasks are then presented. Finally some directions for future research are provided.**

Keywords: **biological data mining, gene sequence analysis, protein structure, microarray, gene regulatory network, multiplayer perceptron, self organizing map**

I. INTRODUCTION

Over the past few decades, major advances in the field of molecular biology, coupled with advances in genomic technologies, have led to an explosive growth in the biological information generated by the scientific community. This deluge of genomic information has, in turn, led to an absolute requirement for computerized databases to store, organize and index the data, and for specialized tools to view and analyze the data.

Bioinformatics can be viewed as the *use of computational methods to make biological discoveries* [1]. It is an interdisciplinary field involving biology, computer science, mathematics and statistics to analyze biological sequence data, genome content & arrangement, and to predict the function and structure of macromolecules. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be derived. There are three important sub-disciplines within bioinformatics:

a) The development of new algorithms and models to assess different relationships among the members of a large biological data set;
b) The analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures; and
c) The development and implementation of tools that enable efficient access and management of different types of information

This article provides a survey of the various neural network based techniques that have been developed over the past few years for different bioinformatics tasks. First we describe the primary bioinformatics tasks along with their biological basis. Next different neural network based algorithms available to address them are explained. Finally, some conclusions and future research directions are presented.

II. BIOINFORMATICS TASKS

The different biological problems studied within the scope of bioinformatics can be broadly classified into two categories: genomics and proteomics which include genes, proteins, and amino acids. We describe below different tasks involved in their analysis along with their utility.

A. Gene Sequence Analysis

The evolutionary basis of sequence alignment is based on the principles of *similarity* and *homology* [3]. Similarity is a quantitative measure of the fraction of two genes which are identical in terms of observable quantities. Homology is the conclusion drawn from data that two genes share a common evolutionary history; no metric is associated with this. The tasks of sequence analysis are-

Sequence Alignment: An alignment is a mutual arrangement of two or more sequences, which exhibits where the sequences are similar, and where they differ. An optimal alignment is one that exhibits the most correspondences and the least differences. It is the alignment with the highest score but may or may not be biologically meaningful. Basically there are two types of alignment methods, global alignment and local alignment. Global alignment [4] maximizes the number of matches between the sequences along the entire length of the sequence. Local alignment [5] gives a highest scoring to local match between two sequences.

Pattern Searching: Pattern searching is search for a nucleic pattern in a nucleic acid sequence, in a set of sequences or in a databank. (e.g., INFOBIOGEN) [6]. It is the potential for uncovering evolutionary relationships and patterns between different forms of life. With the aid of nucleotide and protein sequences, it should be possible to find the ancestral ties between different organisms. So far, experience indicates that closely related organisms have similar sequences and that more distantly related organisms have more dissimilar sequences. Proteins that show a significant sequence conservation indicating a clear evolutionary relationship are said to be from the same *protein family*. By studying *protein folds* (distinct protein building blocks) and families, scientists are able to reconstruct the evolutionary relationship between two species and to estimate the time of divergence between two organisms since they last shared a common ancestor.

Gene Finding: In general DNA strand consists of a large sequence of nucleotides, or bases. For example there are more than 3 billions bases in human DNA sequences. Not all portions of the DNA sequences are *coding* and coding zones indicate that they are a template for a protein. In the human genome only 3%-5% of the sequence are coding, i.e., they constitute the gene. Due to the size of the database, manual searching of genes, which code for proteins, is not practical. Therefore automatic identification of the genes from the large DNA sequences is an important problem in bioinformatics [7].

B. *Protein Analysis*

Proteins are polypeptides, formed within cells as a linear chain of amino acids [8]. Within and outside of cells, proteins serve a myriad of functions, including structural roles (cytoskeleton), as catalysts (enzymes), transporter to ferry ions and molecules across membranes, and hormones to name just a few. There are twenty different amino acids that make up essentially all proteins on earth. Different tasks involved in protein analysis are as follows:

Multiple Sequence Alignment: Multiple amino acid sequence alignment techniques [1] are usually performed to fit one of the following scopes:
(a) determination of the consensus sequence of several aligned sequences; (b) help in the prediction of the secondary and tertiary structures of new sequences; and (c) preliminary step in molecular evolution analysis using phylogenetic methods for constructing phylogenetic trees.

In order to characterize protein families, one needs to identify shared regions of homology in a multiple sequence alignment; (this happens generally when a sequence search revealed homologies in several sequences) .The clustering method can do alignments automatically but are subjected to some restrictions. Manual and eye validations are necessary in some difficult cases. The most practical and widely used method in multiple sequence alignment is the hierarchical extensions of pairwise alignment methods, where the principal is that multiple alignments is achieved by successive application of pairwise methods.

Protein Motif Search: Protein motif search [7,8] allows search for a personal protein pattern in a sequence (personal sequence or entry of Gene Bank). Proteins are derived from a limited number of basic building blocks (domains). Evolution has shuffled these modules giving rise to a diverse repertoire of protein sequences, as a result of it proteins can share a global or local relationship. Protein motif search is a technique for searching sequence databases to uncover common domains/motifs of biological significance that categorize a protein into a family.

Structural Genomics: Structural genomics is the prediction of 3-dimensional structure of a protein from the primary amino acid sequence [9]. This is one of the most challenging tasks in bioinformatics..

The four levels of protein structure (Figure 1) are (a) *Primary structure* is the sequence of amino acids that compose the protein, (b) different regions of the sequence form local *secondary structures*, such as alpha helices and beta strands, (c) *Tertiary structure* is formed by packing secondary structural elements into one or several compact globular units called domains, and (d) Final protein may contain several polypeptide chains arranged in a *quaternary structure*.

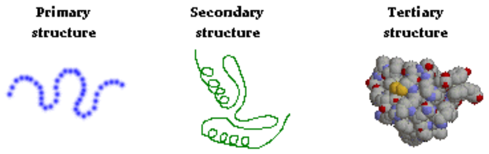

Figure 1: Different levels of protein structures

Sequence similarity methods predict secondary and tertiary structure based on homology to know proteins. Secondary structure predictions methods include Chou-Fasman [9], GOR, neural network, and nearest neighbor methods. Tertiary structure prediction methods include energy minimization, molecular dynamics, and stochastic searches of conformational space.

C. *Microarrays*

Microarray technology [10] makes use of the sequence resources created by the genome projects and other sequencing efforts to answer the question, what genes are expressed in a particular cell type of an organism, at a particular time, under particular conditions. Gene expression is the process by which a gene's coded information is converted into the structures present and operating in the cell. Expressed genes include those that are transcribed into mRNA and then translated into protein and those that are transcribed into RNA but not translated into protein (e.g., transfer and ribosomal RNAs). For instance, they allow comparison of gene expression between normal and diseased (e.g., cancerous) cells. There are several names for this technology - DNA microarrays, DNA arrays, DNA chips, gene chips, others.

Microarrays exploit the preferential binding of complementary single-stranded nucleic acid sequences. A microarray is typically a glass (or some other material) slide, on to which DNA molecules are attached at fixed locations (spots). There may be tens of thousands of spots on an array, each containing a huge number of identical DNA molecules (or fragments of identical molecules), of lengths from twenty to hundreds of nucleotides. (According to quick napkin calculations by Wilhelm Ansorge and Quackenbush in Heidelberg on 4 October, 2001, the number of DNA molecules in a microarray spot is 10^7-10^8). For gene expression studies, each of these molecules ideally should identify one gene or one exon in the genome, however, in practice this is not always so simple and may not even be generally possible due to families of similar genes in a genome. Microarrays that contain all of the about 6000 genes of the yeast genome have been available since 1997. The spots are either printed on the microarrays by a robot, or synthesized by photo-lithography (similarly as in computer chip productions) or by ink-jet printing.

There are different ways how microarrays can be used to measure the gene expression levels. One of the most popular micorarray applications allows to compare gene expression levels in two different samples, e.g., the same cell type in a healthy and diseased state. Conceptually, a gene expression database can be regarded as consisting of three parts – the gene expression data matrix, gene annotation and sample annotation.

Microarrays are already producing massive amounts of data. These data, like genome sequence data, can help us to gain insights into underlying biological processes only if they are carefully recorded and stored in databases, where they can be queried, compared and analyzed by different computer software programs. In many respects gene expression databases are inherently more complex than sequence databases (this does not mean that developing, maintaining and curating the sequence databases are any less challenging).

D. Gene Regulatory Network Analysis

Another important and interesting question in biology is how gene expression is switched on and off, i.e., how genes are regulated [1]. Since almost all cells in a particular organism have an identical genome, differences in gene expression and not the genome content are responsible for cell differentiation (how different cell types develop from a fertilized egg) during the life of the organism.

Gene regulation in eukaryotes, is not well understood, but there is evidence that an important role is played by a type of proteins called *transcription factors*. The transcription factors can attach (bind) to specific parts of the DNA, called transcription factor *binding sites* (i.e., specific, relatively short combinations of A, T, C or G), which are located in so-called *promoter* regions. Specific promoters are associated with particular genes and are generally not too far from the respective genes, though some regulatory effects can be located as far as 30,000 bases away, which makes the definition of the promoter difficult.

Transcription factors control gene expression by binding the gene's promoter and either activating (switching on) the gene's transcription, or repressing it (switching it off). Transcription factors are gene products themselves, and therefore in turn can be controlled by other transcription factors. Transcription factors can control many genes, and some (probably most) genes are controlled by combinations of transcription factors. Feedback loops are possible. Therefore we can talk about *gene regulation networks*. Understanding, describing and modelling such gene regulation networks are one of the most challenging problems in functional genomics. Microarrays and computational methods are playing a major role in attempts to reverse engineer gene networks from various observations. Note that in reality the gene regulation is likely to be a stochastic and not a deterministic process. Traditionally molecular biology has followed so-called reductionist approach mostly concentrating on a study of a single or very few genes in any particular research project. With genomes being sequenced, this is now changing into so-called systems approach.

III. ARTIFICIAL NEURAL NETWORK (ANN) ALGORITHMS IN BIOINFORMATICS

Neural network models try to emulate the biological neural network with electronic circuitry. NN models have been studied for many years with the hope of achieving human like performance, particularly in the field of pattern recognition. Recently, ANN have found a widespread use for classification tasks and function approximation in many fields of medicinal chemistry and bioinformatics. For these kinds of data analysis mainly two different types of networks are employed, "supervised" neural networks (SNN) and "unsupervised" neural networks (UNN). The main applications of SNN are function approximation, classification, pattern recognition and feature extraction, and prediction tasks. These networks require a set of molecular compounds with known activities to model structure-activity relationships. In an optimization procedure, these known "target activities" serve as a reference for SAR modeling. This principle coined the term "supervised" networks. Correspondingly, "unsupervised" networks can be applied to classification and feature extraction tasks even without prior knowledge of molecular activities or properties. The following sections describe different neural network applications to the bioinformatics tasks described previously.

A. Sequence Analysis

GenTHREADER is a neural network architecture that predicts similarity between gene sequences [11]. The effects of sequence alignment score and pairwise potential are the network outputs. Using GenTHREADER was successfully used in the following cases: ORF MG276 from Mycoplasma genitalium was predicted to share structure similarity with 1HGX;. MG276 shares a low sequence similarity (10% sequence identity) with 1HGX.

A back-propagation neural network can grossly approximate the score function of the popular BLAST family of genomic sequence alignment and scoring tools. The resultant neural network may provide a processing speed advantage over the BLAST tool, but may suffer somewhat in comparison to the accuracy of BLAST. Further study is necessary to determine whether a neural network with additional hidden units or structural complexity could be used to more closely approximate BLAST. However, closer approximation may also limit the speed performance advantages enjoyed by the neural network approach.

B Protein Analysis

The most successful techniques for prediction of the protein three-dimensional structure rely on aligning the sequence of a protein of unknown structure to a homologue of known structure. Such methods fail if there is no homologue in the structural database, or if the technique for searching the structural database is unable to identify homologues that are present.

Qian et al [12] used X-ray crystal structures of globular proteins available at that time to train a NN to predict the secondary structure of non-homologous proteins. Since every residue in a PDB entry can be associated to one of three secondary structures (HELIX, SHEET or neither: COIL) the authors were able to design a NN that had 21 input nodes (one for each residue including a null residue) and three output nodes coding for each of the three possible secondary structure assignments (HELIX, SHEET and COIL). It was easiest to restrict the input and output nodes to binary values (1 or 0) when loading the data onto the network during training. This explains why three output nodes are needed: HELIX was coded as 0,0,1 on the three output nodes; SHEET is coded as 0,1,0 and COIL is coded as 1,0,0. A similar binary coding scheme was used for the 20 input nodes for the 20 amino acids. Since the authors wished to consider a moving window of seven residues at a time, their input layer actually contained 20 x 7 nodes plus one node at each position for a null residue. Over 100 protein structures were used to train this network. After training, when the NN was queried with new data, a prediction accuracy of 64% was obtained.

Rost et al. [13] took advantage of the fact that a multiple sequence alignment contains more information about a protein than the primary sequence alone. Instead of using a single sequence as input into the network, they used a sequence profile that resulted from the multiple alignments. This resulted in a significant improvement in prediction accuracy to 71.4%. Recently, more radical changes to the design of NNs including bi-directional training and the use of the entire protein sequence as simultaneous input instead of a shifting window of fixed length has led to prediction accuracy above 71%.

The task of applying ANNs to the problem of protein structure prediction requires a certain number of input “nodes" and connect each one to every node in a hidden layer. Each node in the hidden layer is then connected to every node in the final output layer. The connection strength between each and every pair of nodes is initially assigned a random value and is then modified by the program itself during the training process. Each node will "decide" to send a signal to the nodes it is connected to based on evaluating its transfer function after all of its inputs and connection weights have been summed. Training proceeds by holding particular data (say from an entry in the Protein Data Bank) constant onto both the input and output nodes and iterating the network in a process that modifies the connection weights until the changes made to them approach zero. When such convergence is reached, the network is ready to receive new experimental data. Now the connection weights are not changed and the values of the hidden and output nodes are calculated. Selection of unbiased and normalized training data, however, is probably just as important as the network architecture in the design of a successful NN.

The prediction of protein secondary structure by use of carefully structured neural networks and multiple sequence alignments have been investigated by Riis and Krogh[14]. Separate networks are used for predicting the three secondary structures ff-helix, fi-strand and coil. The networks are designed using a priori knowledge of amino acid properties withrespect to the secondary structure and of the characteristic periodicity in ff-helices. Since these single-structure networks all have less than 600 adjustable weights over-fitting is avoided. To obtain a three-state prediction of ff-helix, fi-strand or coil, ensembles of single-structure networks are combined with another neural network. This method gives an overall prediction accuracy of 66.3% when using seven-fold cross-validation on a database of 126 non-homologous globular proteins. Applying the method to multiple sequence alignments of homologous proteins increases the prediction accuracy significantly to 71.3% [14].

C. *Microarray*

Clustering is commonly used in microarray experiments to identify groups of genes that share similar expression. Genes that are similarly expressed are often co-regulated and involved in the same cellular processes. Therefore, clustering suggests functional relationships between groups of genes. It may also help in identifying promoter sequence elements that are shared among genes. In addition, clustering can be used to analyze the effects of specific changes in experimental conditions and may reveal the full cellular responses triggered by those conditions.

Bayesian neural network is used with structural learning with forgetting for searching for optimal network size and structure for *microarray* data in order to capture the structural information of gene expressions [15,16]. The process of Bayesian learning starts with a Feed forward Neural Network (FFNN) and prior distribution for the network parameters. The prior distribution gives initial beliefs about the parameters before any data are observed. After new data are observed, the prior distribution is updated to the posterior distribution using Bayes rules. Multi-Layer Perceptron (MLP) is mainly considered as the network structure for Bayesian learning. Since the correlated data may include high levels of noise, efficient regularization techniques are required to improve the generalization performance. This involves network complexity adjustment and performance function modification. To do the latter, instead of the sum of squared error (SSE) on the training set, a cost function is automatically adjusted. PLANN (Plausible neural network) is another universal data analysis tool based upon artificial neural networks and is capable of plausible inference and incremental learning [17]. This tool has been applied to research data from molecular biological systems through the simultaneous analysis of gene expression data and other types of biological information.

D. *Gene Regulatory Network*

A novel clustering technique used for identifying gene regulatory networks is the adaptive double self-organizing map (ADSOM) [18]. It has a flexible topology and performs clustering and cluster visualization simultaneously, thereby requiring no a-priori knowledge about the number of clusters. ADSOM is developed based on a recently introduced technique known as double self-organizing map (DSOM). DSOM combines features of the popular self-organizing map (SOM) with two-dimensional position vectors, which serve as a visualization tool to decide how many clusters are needed. Although DSOM addresses the problem of identifying unknown number of clusters, its free parameters are difficult to control to guarantee correct results and convergence. ADSOM updates its free parameters during training and it allows convergence of its position vectors to a fairly consistent number of clusters provided that its initial number of nodes is greater than the expected number of clusters. The number of clusters can be identified by visually counting the clusters formed by the position vectors after training. The reliance of ADSOM in identifying the number of clusters is proven by applying it to publicly available gene expression data from multiple biological systems such as yeast, human, and mouse. ADSOM's performance in detecting number of clusters is compared with a model-based clustering method. It may be noted that gene regulatory network analysis is a very recent research area, and neural network applications to it are scarce.

IV. CONCLUSION AND SCOPE OF FUTURE RESEARCH

The rationale for applying computational approaches to facilitate the understanding of various biological processes includes:

- A more global perspective in experimental design; and
- The ability to capitalize on the emerging technology of database-mining - the process by which testable hypotheses are generated regarding the function or structure of a gene or protein of interest by identifying similar sequences in better characterized organisms.

Neural networks appear to be a very powerful artificial intelligence (AI) paradigm to handle these issues [19]. Other soft computing tools, like fuzzy set theory and genetic algorithms, integrated with ANN may also be based on the principles of Case Based Reasoning [20].

Even though the current approaches in biocomputing are very helpful in identifying patterns and functions of proteins and genes, they are still far from being perfect. They are not only time-consuming, requiring Unix workstations to run on, but might also lead to and assumptions due to necessary simplifications. It is therefore still mandatory to use biological reasoning and common sense in evaluating the results delivered by a biocomputing program. Also, for evaluation of the trustworthiness of the output of a program it is necessary to understand the mathematical / theoretical background of it to finally come up with a use- and senseful analysis.

REFERENCES

1. P. Baldi and S. Brunak, *Bioinformatics: The Machine Learning Approach*, MIT Press, Cambridge, MA, 1998.
2. R. B. Altman, A. Valencia, S. Miyano and S. Ranganathan, Challenges for intelligent systems in biology, *IEEE Intelligent Systems*, 16(6), pp. 14-20, 2001
3. H. Nash, D. Blair, J. Grefenstette, Proc. 2nd IEEE International Symposium on Bioinformatics and Bioengineering (BIBE'01), pp. 89, 2001, Bethesda, Maryland
4. S. B. Needleman and C. D. Wunsch,, *Journal of Molecular Biology*, 48, pp. 443-453, 1970.
5. T. F. Smith and M. S. Waterman, *Journal of Molecular Biology*, 147, pp. 195-197, 1981.
6. D. Gautheret, F. Major and R. Cedergren, Pattern searching/alignment with RNA primary and secondary structures: an effective descriptor for tRNA, *Comp. Appl. Biosc.* 6, pp. 325-331, 1990
7. J. W. Fickett, Finding genes by computer: The state of the art, Trends in Genetics, 12(8), pp. 316-320, 1996
8. Salzberg, S.L. , Searls, D.B., and Kasif, S. (Eds.) Computational Methods In Molecular Biology, North Holland: Elsevier Sciences, 1988
9. P. Chou, and G. Fasmann, "Prediction of the secondary structure of proteins from their amino acid sequence", Advances in Enzymology, 47, pp 45-148, 1978
10. J. Quackenbush, Computational analysis of microarray data, *National Review of Genetics*, 2, pp. 418-427, 2001
11. D. T. Jones, "GenTHREADER: An Efficient and Reliable Protein Fold Recognition.", Journal of Mol. Bio., 287, pp.797-815, 1999
12. N. Qian and T. J. Sejnowski, *Journal Molecular Biology,* 202**,** pp. 865-84, 1988
13. B. Rost and C. Sander, *Proc Natl Acad Sci U S A* 90, 7558-62, 1993
14. S.K. Riis and A. Krogh,"Improving Prediction of Protein Secondary Structure using Structured Neural Networks and Multiple Sequence Alignments", Journal of Computational Biology, 3, pp. 163-183, 1996
15. S. Liang, S. Fuhrman, and R. Somogyi. REVEAL: A general reverse engineering algorithm for inference of genetic network architectures, *In Pacific Symposium on Biocomputing,* 3, pp. 18-29, 1998.
16. Bayesian Neural Network for Microarray Data, Yulan Liang, E Olusegun Georgre, Arpad Kelemen Technical Report, Department of Mathematical Sciences , University of Memphis
17. PLANN Software, *PNN Technologies*, Pasadena, CA
18. H. Ressom, D. Wang, and P. Natarajan, Clustering gene expression data using adaptive double self-organizing map**,** Physiol. Genomics, 14, pp. 35–46, 2003
19. S. K. Pal, L. Polkowski and A. Skowron, Rough-neuro Computing: A way of computing with words, Springer, Berlin, 2003
20. S. K. Pal and S. C. K. Shiu, Foundations of Soft Case Based Reasoning, John Wiley, NY, 2004