Artificial Neural Networks Solving Bioinformatics Problems

Bharat Bhushan
Asstt. Prof., Govt. College for Women, Bhodia Khera, Fatehabad, India
E-mail: mehta.bhushan@gmail.com

Abstract- Bioinformatics is a promising and innovative research field. Bioinformatics is a new science that is glowing out in the recent years. It is a multidisciplinary science that is made out of different kinds of other scientific fields like Biology, Computer Science, Mathematics and others. Since most of the problems in bioinformatics are inherently hard, researches have used artificial intelligence techniques to solve such problems. Artificial neural networks are one such method used in many situations and have proved to be very effective. This paper will focus on issues related to construction of a neural network to solve bioinformatics problems and describes some of its current applications.

Keywords- Bioinformatics, Artificial neural network, Multilayer, perception

1. Introduction

Bioinformatics is a promising and novel research area in the 21st century. Bioinformatics is the combination of biology and information technology. It is the branch of science that deals with computer based analysis of large biological data sets. Bioinformatics is a multifaceted discipline combining many scientific fields including computational biology, statistics, mathematics, molecular biology, and genetics [1]. So Bioinformatics is conceptualizing biology in terms of macromolecules and then applying “informatics” techniques to understand and organize the information associated with these molecules, on a large scale.

This field is data driven and aims at understanding of relationships and gaining knowledge in biology. In order to extract this knowledge encoded in biological data, advanced computational technologies, algorithms and tools need to be used. Basic problems in bioinformatics like protein structure prediction, multiple alignment of sequences, phylogenic inferences, etc are inherently non-deterministic polynomial-time hard in nature. To solve these kinds of problems artificial intelligence (AI) methods offer a powerful and efficient approach. Researchers have used AI techniques like Artificial Neural Networks (ANN), Fuzzy Logic, Genetic Algorithms, and Support Vector Machines to solve problems in bioinformatics [2]. Artificial Neural Networks is one of the AI techniques commonly in use because of its ability to capture and represent complex input and output relationships among data. The concept of ANN is basically introduced from the subject of biology where neural network plays an important and key role in human body. The purpose of this paper is to provide an overall understanding of ANN and its place in bioinformatics to a newcomer to the field.

2. Bioinformatics

Bioinformatics is the combination of biology and information technology. It is the branch of science that deals with computer based analysis of large biological data sets. So Bioinformatics is conceptualizing biology in terms of macromolecules and then applying “informatics” techniques to understand and organize the information associated with these molecules, on a large scale [3].

It is a science used to manage, analyze, organize, and classify the huge amount of biological data by using well developed algorithms, computational and statistical techniques, designing and construction of software tools and theories to solve different problems arising from biological data and help in generating, storing, accessing and analyzing data and information that are related to molecular biology. Noting that the suffix “informatics” is from European origin; “informatique” means and indicates computer science in French and Bio means Biology [3]. Fig. 1 below illustrates all the sciences that make up the Bioinformatics field.

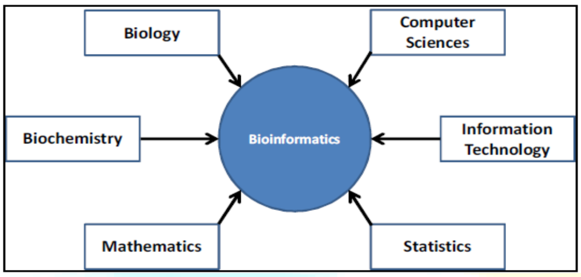

Figure 1: Bioinformatics multidisciplinary sciences

There are five main aims of Bioinformatics [3]:

1. To organize the biological data in an easy manner that helps biologists and researchers to store and access exiting information.
2. To develop and design software tools that help in the analysis and management of data.
3. To use these biological data in the analysis and interpretation of the results in a biological meaningful manner.
4. To assist researchers in the pharmaceutical industry to understand the protein structures that lead and help in the drugs industry development.
5. To help and assist physicians in the medical fields to understand gene structures that will help in detecting and diagnosing disease like cancer.

3. Neural Networks

A Neural Network is an interconnected assembly of simple processing elements, units or nodes, whose functionality is loosely based on the animal neuron. The processing ability of the network is stored in the inter-unit connection strengths, or weights, obtained by a process of adaptation to, or learning from, a set of training patterns [4].

Neural Network is just a web of inter connected neurons which are millions and millions in number. With the help of this interconnected neurons all the parallel processing is done in human body and the human body is the best example of Parallel Processing.

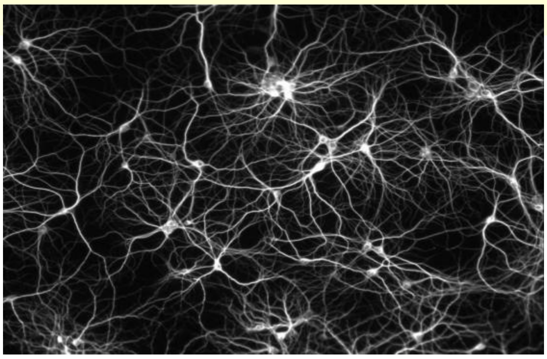

Figure 2: Neural Network in Human Body

A neuron is a special biological cell that process information from one neuron to another neuron with the help of some electrical and chemical change. It is composed of a cell body or soma and two types of out reaching tree like branches: the axon and the dendrites. The cell body has a nucleus that contains information about hereditary traits and plasma that holds the molecular equipments or producing material needed by the neurons. The whole process of receiving and sending signals is done in particular manner like a neuron receive signals from other neuron through dendrites. The Neuron send signals at spikes of electrical activity through a long thin stand known as an axon and an axon splits this signals through synapse and send it to the other neurons.

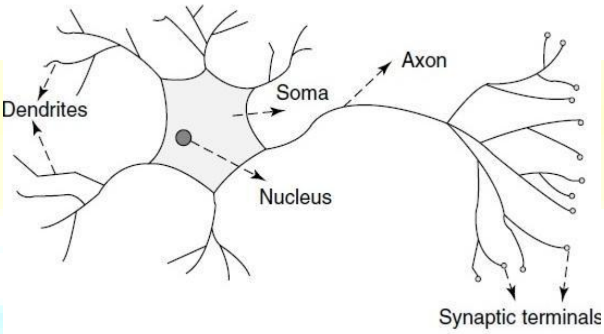

Figure 3: Human neurons

An Artificial Neuron is basically an engineering approach of biological neuron. It has device with many inputs and one output. ANN consist of large number of simple processing elements that are interconnected with each other and layered also.

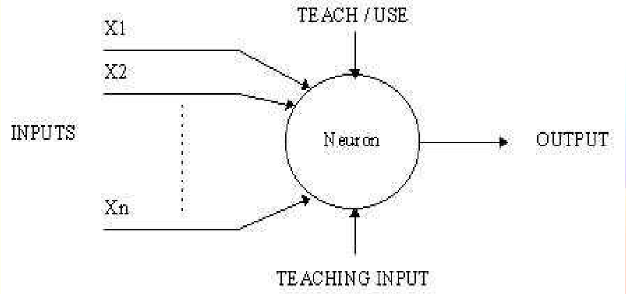

Figure 4: Artificial Neuron

3.1. Architecture Structures of Neural Networks

Neural networks are not only different in their learning processes but also different in their structures or topology [4]. They are broadly classified into recurrent (involving feedback) and nonrecurrent (without feedback) ones [5].

In a little more details, Neural network architectures are divided into the following three classes [6]:

3.1.1. Single-layer perceptrons (feed forward networks)

The single-layer perceptrons was among the first and simplest learning machines that are trainable. In [6], perceptron denotes the class of two-layer feed forward networks, 1) whose first-layer units have fixed function with fixed connection weights from the inputs, and 2) whose connection weights linking this first layer to the second layer of outputs are learnable.

The model of training in perceptrons is supervisory, because the steps in the algorithm involve the comparison of actual outputs with desired outputs associated with the set of training patterns. An input layer of source nodes can project onto an output layer of neurons, but not vice versa. The LMS algorithm can be used in the supervisory training.

3.1.2. Multi-layer perceptrons (feed forward networks)

Multi-layer feed forward structures are characterized by directed layered graphs and are the generalization of those earlier single layer structures [5]. A typical multi-layer feed forward network consists of a set of sensory units that constitute the input layer, one or more hidden layers of computation nodes, and an output layer of computation nodes. The input signal propagates through the network in a forward direction on a layer-by-layer basis. These neural networks are also commonly referred to as multi-layer perceptrons (MLPs) [6].

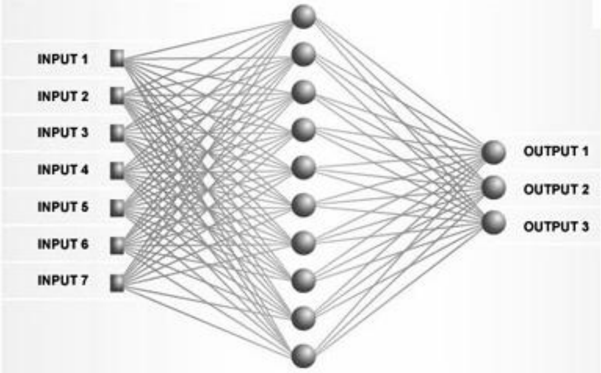

Figure 5: Multilayer Perceptron

MLP neural network structure with 7 input neurons, hidden layer and output layer of 3 output neurons. Each neuron in a layer is connected with each neuron in the previous layer, signals can travel only forward hence the name feedforward network.

3.1.3. Recurrent networks

In the neural network literature, neural networks with one or more feedback loops are referred to as recurrent networks. A recurrent network distinguishes itself from a feed forward neural network in that it has at least one feedback loop. Such a system has very rich temporal and spatial behaviors, such as stable and unstable fixed points and limit cycles, and chaotic behaviors. These behaviors can be utilized to model certain cognitive functions, such as associative memory, unsupervised learning, self-organizing maps, and temporal reasoning [4].

4. Designing Neural Networks for Bioinformatics

When designing NN for bioinformatics applications, there are common designing issues that needs to be addressed. Fig. 6 summarizes these issues [7].

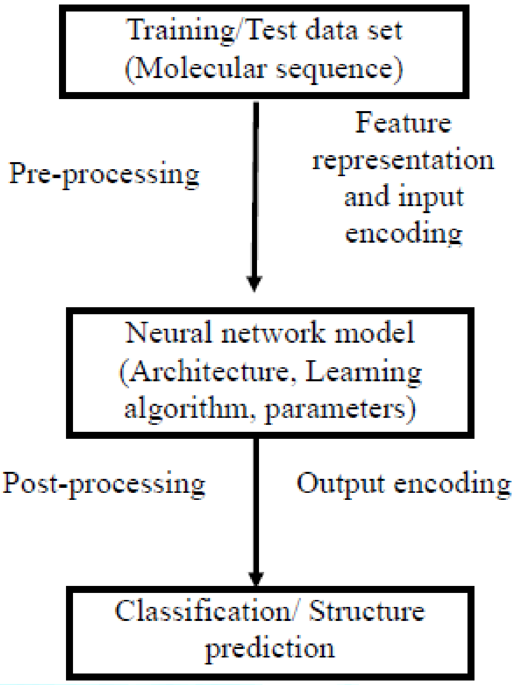
Training/Test data set (Molecular sequence)
Pre-processing
Feature representation and input encoding
Neural network model (Architecture, Learning algorithm, parameters)
Post-processing
Output encoding
Classification/ Structure prediction

Figure 6: Designing Neural Networks for Bioinformatics

Preprocessing of data involves feature presentation and input encoding. This is an important element which determines the performance and the information entered to the NN. In order to get the full benefit of the NN, the designer has to represent prior knowledge about sequence structure and functions [7]. This allows the extraction of salient features in the given sequence. This information can be encoded and used in the NN.

After identifying the features needed to be represented in the NN application, there are several ways to represent these data to maximize information extraction. They are as follows:

- Real number measurements in a continuous scale e.g. Mass
- Vectors of distance or frequencies e.g. PAM matrix
- Categorised into classes
- Using alternative alphabet to represent AA with similar features
- Hierarchical classes

After identifying the features that need to be represented in the NN application, data need to be encoded. Encoding can be local (involving single or neighboring residues in short sequence segments) or global (involving long range relationship in entire sequence) [7]. The sequence encoding method can be direct or indirect. Direct encoding converts each residue in to a vector and it preserves the positional information. Indirect encoding on the other hand provides overall information measures of the entire sequence [7].

Output encoding is not as complex as input encoding. It depends on the number of classification required in the application. However networks like self organizing maps automatically configure the number of output units. The value of the output units can be used qualitatively or quantitative measure of confidence level or activity level.

5. Applications of Artificial Neural networks in Bioinformatics

Two important aspects of Artificial Neural Networks: ability of "learn" and then make predictions after being "trained" are responsible for answering various questions in molecular biology. The network is able to process information and modify parameters of the weight functions between variables during the training stage. Once it is trained, it is able to make automatic predictions about the unknown. Interest in such networks has been stimulated by the recent development of a learning rule for the automatic assignment of connection strengths and thresholds. A neural network may simply be viewed as a highly parallel computational device and have been shown to be useful in a variety of tasks including modeling content addressable memory, solving certain optimization problems, automating pattern recognition and many more like speech recognition, medical diagnosis, image compression and financial prediction. The neural networks generally used for in various applications in bioinformatics are of feed-forward type. Some of the important applications are:

Gene prediction

In gene prediction, a neural network is constructed with multiple layers; the input, output, and hidden layers.

The input is the gene sequence with intron and axon signals and the output is the probability of an axon structure. Between input and output, there may be one or several hidden layers where the machine learning takes place. The machine learning process starts by feeding the model with a sequence of known gene structure.

The gene structure information is separated into several classes of features such as hexamer frequencies, splice sites, and GC composition during training (Fig. 7).

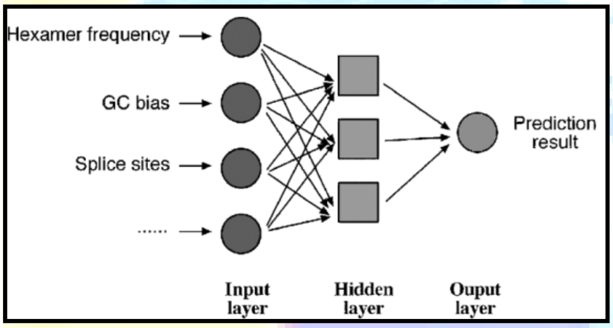

Figure 7: Gene prediction using neural network approach

When the algorithm predicts an unknown sequence after training, it applies the same rules learned in training to look for patterns associated with the gene structures.

Protein structure prediction

The third generation protein structure prediction algorithms extensively apply sophisticated neural networks to analyze substitution patterns in multiple sequence alignments. The neural network has to be first trained by sequences with known structures so it can recognize the amino acid patterns and their relationships with known structures. During this process, the weight functions in hidden layers are optimized so they can relate input to output correctly. When the sufficiently trained network processes an unknown sequence, it applies the rules learned in training to recognize particular structural patterns.

Secondary structure prediction for globular proteins

In secondary structure prediction, the input is an amino acid sequence and the output is the probability of a residue to adopt a particular structure (Fig. 8).

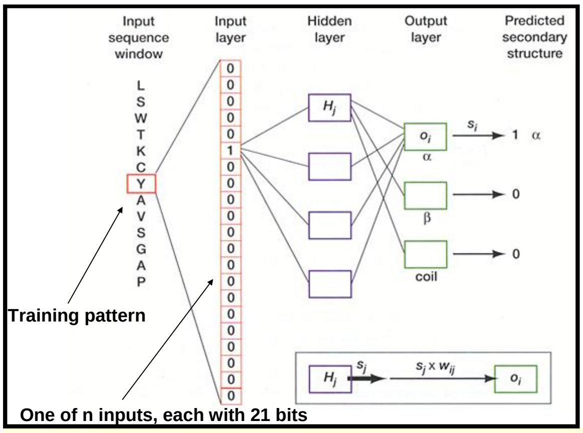

Figure 8: Protein structure prediction through artificial neural network based approach

Between input and output are many connected hidden layers where the machine learning takes place to adjust the mathematical weights of internal connections. A neural network is trained not by a single sequence but by a sequence profile derived from the multiple sequence alignment for protein structure prediction. This approach has been shown to improve the accuracy to above 75%, which is a breakthrough in secondary structure prediction

Coiled coil prediction

Coiled coils are super helical structures involving two to more interacting α-helices from the same or different proteins. The coiled coil conformation is important in facilitating inter or intraprotein interactions. Proteins possessing these structural domains are often involved in transcription regulation or in the maintenance of cytoskeletal integrity. Coiled coils have an integral repeat of seven residues (heptads) which assume a side chain packing geometry at facing residues.

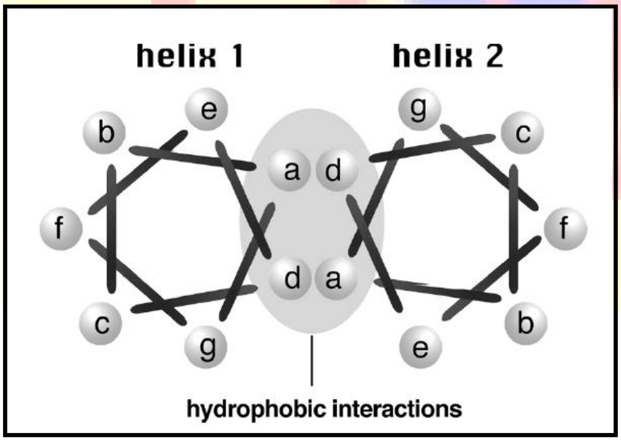

Figure 9: Coiled coils

For every seven residues, the first and fourth are hydrophobic, facing the helical interface; the others are hydrophilic and exposed to the solvent (Fig. 9). The sequence periodicity forms the basis for designing algorithms to predict this important structural domain. As a result the location of coiled coils can be predicted precisely.

Pattern recognition

Clustering by SOMs algorithm (Self-Organizing Maps) employs neural networks for pattern recognition. Process starts by defining a number of nodes. The data points are initially assigned to the nodes at random. The distance between the input data points and the centroids are calculated. The data points are successively adjusted among the nodes, and their distances to the centroids are recalculated. After many iterations, a stabilized clustering pattern are reached with the minimum distances of the data points to the centroids.

Prediction of disulfide bridges

A disulfide bridge is a unique type of posttranslational modification in which covalent bonds are formed between cysteine residues. Disulfide bonds are important for maintaining the stability of certain types of proteins. The disulfide prediction is the prediction of paring potential or bonding states of cysteines in a protein. Accurate prediction of disulfide bonds may also help to predict the three dimensional structure of the protein of interest. Advanced neural networks are often used to discern long distance pair wise interactions among cysteine residues.

6. Conclusion

It is evident that artificial intelligence techniques like Neural Networks are heavily used in the field of bioinformatics to solve hard problems. These methods have proved and established its value in the field of bioinformatics. Knowledge and ability to use neural networks method add definite advantage to bioinformaticians to solve many types of problems in the field of bioinformatics.

REFERENCES

[1] Fenstermacher, D. (2005). Introduction to bioinformatics: Research Articles. *Journal of the American Society for Information Science and Technology, Vol. 56, No.5,* 440-446.

[2] Seiffert, U., Hammer, B., Kaski, S., Villmann, T. (2006). Neural Networks and Machine Learning in Bioinformatics - Theory and Applications. *ESANN'2006 proceedings – European Symposium on Artificial Neural Networks Bruges (Belgium), d-side publi., ISBN 2-930307-06-4.*

[3] Al-Rajab, M. & Lu, J. Bioinformatics: an overview for cancer research.

[4] He, Q. (1999). Neural Network and Its Application in IR. *UIUCLIS—1999.*

[5] Bose, N. K.; & Liang P. (1996). Neural network fundamentals with graphs, algorithms, and applications. *McGraw-Hill, Inc.*

[6] Haykin, S. (1999). Neural networks: a comprehensive foundation. *(2nd ed.) Upper Saddle Rever*, New Jersey: Prentice Hall.

[7] Wu & McLarty (2000). Neural Networks and Genome Informatics. Methods in computational Biology and Biochemistry. *Vol. 1, Elsevier*.

[8] Shanthi, D., Sahoo, G., Saravanan, N. Designing an Artificial Neural Network Model for the Prediction of Thrombo-embolic Stroke. *International Journals of Biometric and Bioinformatics (IJBB), Volume 3, Issue 1,* 10-18.

[9] Li, E., Y. (1994). Artificial neural networks and their business applications. *Information & Management 27,* 303-313.