www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 2 Issue 5 May, 2013 Page No. 1506-1511

SOFT COMPUTING METHODOLOGIES IN BIOINFORMATICS AND ITS ADVANCE TOWARDS BIOLOGICAL DNA

[1]Dr Tryambak A. Hiwarkar, [2]Sridhar Iyer
[1]Asso Prof, Computer Science and Engineering, MBITM, Dongargarh (C.G.)
[2]Research Scholar, Computer Science, CMJ University, Shillong

Abstract:

Bioinformatics is a promising and innovative research field in 21st century. Despite of a high number of techniques specifically dedicated to bioinformatics problems as well as many successful applications, we are in the beginning of a process to massively integrate the aspects and experiences in the different core subjects such as biology, medicine, computer science, engineering, chemistry, physics, and mathematics. Bioinformatics is a fast growing field in the scientific community. It involves a wide range of problems, for example, DNA sequence analyses, RNA secondary structure predictions, phylogenetic analyses and microarray analyses.

Recently the use of soft computing tools for solving bioinformatics problems have been gaining the attention of researchers because of their ability to handle imprecision, uncertainty in large and complex search spaces. The paper will focus on soft computing paradigm in bioinformatics.

Keywords: Nucleic Acid Sequence, Nucleotides, DNA Sequencing

I. INTRODUCTION:

An, **bioinformatics** is an interdisciplinary field that develops and improves upon methods for storing, retrieving, organizing and analyzing biological data. A major activity in bioinformatics is to develop software tools to generate useful biological knowledge.

Bioinformatics has become an important part of many areas of biology. In

experimental molecular biology, bioinformatics techniques such as image

and signal processing allow extraction of useful results from large amounts of raw data. In the field of genetics and genomics, it aids in sequencing and annotating genomes and their observed mutations. It plays a role in the textual mining of biological literature and the development of biological and gene ontologism to organize and query biological data. It plays a role in the analysis of gene and protein expression and regulation. Bioinformatics tools aid in the comparison of genetic and genomic data and more generally in the understanding of evolutionary aspects of molecular biology. At a more integrative level, it helps analyze and catalogue the biological pathways and networks that are an important part of systems biology. In structural biology, it aids in the simulation and modeling of DNA, RNA, and protein structures as well as molecular interactions. Bioinformatics uses many areas of computer science, mathematics and engineering to process biological data. Complex machines are used to read in biological data at a much faster rate than before.

Databases and information systems are used to store and organize biological data. Analyzing biological data may involve algorithms in artificial intelligence, soft computing, data mining, image processing, and simulation. Advancement in soft computing techniques demonstrates the high standards of technology, algorithms, and tools in bioinformatics for dedicated purposes such as reliable and parallel genome sequencing, fast sequence comparison, search in databases, automated gene identification, efficient modeling and storage of heterogeneous data, etc. The basic

problems in bioinformatics like protein structure prediction, multiple alignment, phyla genetic inference etc. are mostly NP-hard in nature. For all these problems, soft computing offers on promising approach to achieve efficient and reliable heuristic solution. On the other side the continuous development of high quality biotechnology, e.g. micro-array techniques and mass spectrometry, which provide complex patterns for the direct characterization of cell processes, offers further promising opportunities for advanced research in bioinformatics. So Soft Computing Methodologies in Bioinformatics 190 bioinformatics must cross the border towards a massive integration of the aspects and experience in the different core subjects like computer science and statistics etc. for an integrated understanding of relevant processes in systems biology. This puts new challenges not only on appropriate data storage, visualization, and retrieval of heterogeneous information, but also on soft computing methods and tools used in this context, which must adequately process and integrate heterogeneous information into a global picture.

1.1 Goals:

In order to study how normal cellular activities are altered in different disease states, the biological data must be combined to form a comprehensive picture of these activities. Therefore, the field of bioinformatics has evolved such that the most pressing task now involves the analysis and interpretation of various types of data. This includes nucleotide and amino acid sequences, protein domains, and protein structures.[9] The actual process of analyzing and interpreting data is referred to as computational biology. Important sub-disciplines within bioinformatics and computational biology include:

- The development and implementation of tools that enable efficient access to, use and management of, various types of information.
- The development of new algorithms (mathematical formulas) and statistics with which to assess relationships among members of large data sets. For example, methods to locate a gene within a sequence, predict protein structure and/or function, and cluster protein sequences into families of related sequences.

The primary goal of bioinformatics is to increase the understanding of biological processes. What sets it apart from other approaches, however, is its focus on developing and applying computationally intensive techniques to achieve this goal. Examplesinclude: pattern recognition, data mining, machine learning algorithms, and visualization. Major research efforts in the field include sequence alignment, gene finding, genome assembly, drug design, drug discovery, protein structure alignment, protein structure prediction, prediction of gene expression and protein–protein interactions, genome-wide association studies, and the modeling of evolution.

Bioinformatics now entails the creation and advancement of databases, algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of biological data.

Over the past few decades rapid developments in genomic and other molecular research technologies and developments in information technologies have combined to produce a tremendous amount of information related to molecular biology. Bioinformatics is the name given to these mathematical and computing approaches used to glean understanding of biological processes.

1.2 Approaches:

Common activities in bioinformatics include mapping and analyzing DNA and protein sequences, aligning different DNA and protein sequences to compare them, and creating and viewing 3-D models of protein structures. There are two fundamental ways of modeling a Biological system (e.g., living cell) both coming under Bioinformatics approaches.

- Static Sequences – Proteins, Nucleic acids and Peptides
- Interaction data among the above entities including microarray data and Networks of proteins, metabolites
- Dynamic
- Structures – Proteins, Nucleic acids, Ligands (including metabolites and drugs) and Peptides (structures studied with bioinformatics tools are not considered static anymore and their dynamics is often the core of the structural studies)
- Systems Biology comes under this category including reaction fluxes and variable concentrations of metabolites

- Multi-Agent Based modeling approaches capturing cellular events such as signaling, transcription and reaction dynamics a broad sub-category under bioinformatics is structural bioinformatics.

II. LITERATURE SURVEY:

2.1 Nucleic Acid Sequence:

A **nucleic acid sequence** is a succession of letters that indicate the order of nucleotides within a DNA (using GACT) or RNA (GACU) molecule. By convention, sequences are usually presented from the 5' end to the 3' end. Because nucleic acids are normally linear (unreached) polymers, specifying the sequence is equivalent to defining the covalent structure of the entire molecule. For this reason, the nucleic acid sequence is also termed the primary structure.

The sequence has capacity to represent information. Biological DNA represents the information which directs the functions of a living thing. In that context, the term **genetic sequence** is often used. Sequences can be read from the biological raw material through DNA sequencing methods.

Nucleic acids also have a secondary structure and tertiary structure. Primary structure is sometimes mistakenly referred to as *primary sequence*. Conversely, there is no parallel concept of secondary or tertiary sequence.

2.2 Nucleotides:

Nucleic acids consist of a chain of linked units called nucleotides. Each nucleotide consists of three subunits: a phosphate group and a sugar (ribose in the case of RNA, deoxyribose in DNA) make up the backbone of the nucleic acid strand, and attached to the sugar is one of a set of nucleobases. The nucleobases are important in base pairing of strands to form higher-level secondary and tertiary structure such as the famed double helix. The possible letters are A, C, G, and T, representing the four nucleotide bases of a DNA strand adenine, cytosine, guanine, thymine covalently linked to a phosphodiester backbone. In the typical case, the sequences are printed abutting one another without gaps, as in the sequence AAAGTCTGAC, read left to right in the 5' to 3' direction. With regards to transcription, a sequence is on the coding strand if it has the same order as the transcribed RNA. One sequence can be complementary to another sequence, meaning that they have the base on each position is the complementary (i.e. A to T, C to G) and in the reverse order. For example, the complementary sequence to TTAC is GTAA. If one strand of the double-stranded DNA is considered the sense strand, then the other strand, considered the antisense strand, will have the complementary sequence to the sense strand.

2. 3 Genomics:

Genomics is a discipline in genetics that applies recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyze the function and structure of genomes (the *complete* set of DNA within a single cell of an organism).[1][2] The field includes efforts to determine the entire DNA sequence of organisms and fine-scale genetic mapping. The field also includes studies of intragenomic phenomena such as heterocyst, epitasis, pleiotropyand other interactions between loci and alleles within the genome. In contrast, the investigation of the roles and functions of single genes is a primary focus of molecular biology or genetics and is a common topic of modern medical and biological research. Research of single genes does not fall into the definition of genomics unless the aim of this genetic, pathway, and functional information analysis is to elucidate its effect on, place in, and response to the entire genome's networks.

2.4 DNA Sequencing Technology Developed:

In addition to his seminal work on the amino acid sequence of insulin, Frederick Sanger and his colleagues played a key role in the development of DNA sequencing techniques that enabled the establishment of comprehensive genome sequencing projects. In 1975, he and Alan Cousin published a sequencing procedure using DNA polymerase with radio labeled nucleotides that he called the *Plus and minus technique*. This involved two closely related methods that generated short oligo nucleotides with defined 3' termini. These could be fractionated by electrophoresis on a polyacrylamide gel and visualized using autoradiography. The procedure could sequence up to 80 nucleotides in one go and

was a big improvement on what gone before but was still very laborious. Nevertheless, in 1977

his group was able to sequence most of the 5,386 nucleotides of the single-stranded bacteriophage φX174, completing the first fully sequenced DNA-based genome. The refinement of the *Plus and Minus* method resulted in the chain-termination, or Sanger method (see below), which formed the basis of the techniques of DNA sequencing, genome mapping, data storage, and bioinformatics analysis most widely used in the following quarter-century of research.

2.5 Genome Project:

Genome projects are scientific endeavors that ultimately aim to determine the complete genome sequence of an organism (be it an animal, a plant, a fungus, bacterium, an archaean, a protest or a virus) and to annotate protein-coding genes and other important genome-encoded features. The genome sequence of an organism includes the collective DNA sequences of each chromosome in the organism. For a bacterium containing a single chromosome, a genome project will aim to map the sequence of that chromosome. For the human species, whose genome includes 22 pairs of autonomies and 2 sex chromosomes, a complete genome sequence will involve 46 separate chromosome sequences. The Human Genome Project was a landmark genome project that is already having a major impact on research across the life sciences, with potential for spurring numerous medical and commercial developments.

2.6 Genome Annotation:

Genome annotation is the process of attaching biological information to sequences. It consists of three main steps:

1. identifying portions of the genome that do not code for proteins
2. identifying elements on the genome, a process called gene prediction, and
3. Attaching biological information to these elements.

Automatic annotation tools try to perform all this by computer analysis, as opposed to manual annotation (a.k.a. duration) which involves human expertise. Ideally, these approaches co-exist and complement each other in the same annotation pipeline.

The basic level of annotation is using BLAST for finding similarities, and then annotating genomes based on that. However, nowadays more and more additional information is added to the annotation platform. The additional information allows manual annotators to deco volute discrepancies between genes that are given the same annotation. Some databases use genome context information, similarity scores, experimental data, and integrations of other resources to provide genome annotations through their Subsystems approach. Other databases (e.g. Ensembl) rely on both curate data sources as well as a range of different software tools in their automated genome annotation pipeline.

Structural annotation consists of the identification of genomic elements.

- ORFs and their localization
- gene structure
- coding regions
- location of regulatory motifs

Functional annotation consists of attaching biological information to genomic elements.

- biochemical function
- biological function
- involved regulation and interactions
- expression

These steps may involve both biological experiments and in silico analysis. Proteo genomics based approaches utilize information from expressed proteins, often derived from mass spectrometry, to improve genomics annotations. A variety of software tools have been developed to permit scientists to view and share genome annotations. Genome annotation remains a major challenge for scientists investigating the human genome, now that the genome sequences of more than a thousand human individuals and several model organisms are largely complete.[9][10] Identifying the locations of genes and other genetic control elements is often described as defining the biological "parts list" for the assembly and normal operation of an organism.[1] Scientists are still at an early stage in the process of delineating this parts list and in understanding how all the parts "fit together".

III. SOFT COMPUTING TECHNIQUES IN BIOINFORMATICS:

There are a number of reasons why soft computing approaches are widely used in practice, especially in bioinformatics

1. Traditionally, a human being builds such an expert system by collecting knowledge from specific experts. The experts can always explain what factors they use to assess a situation, however, it is often difficult for the experts to say what rules they use (for example, for disease analysis and control). This problem can be resolved by soft computing mechanisms. Soft computing mechanism can extract the description of the hidden situation in terms of those factors and then fire rules that match the expert's behavior.
2. Systems often produce results different from the desired ones. This may be caused by unknown properties or functions of inputs during the design of systems. This situation always occurs in the biological world because of the complexities and mysteries of life sciences. However, with its capability of dynamic improvement, soft computing can cope with this problem.
3. In molecular biology research, new data and concepts are generated every day, and those new data and concepts update or replace the old ones. Soft computing can be easily adapted to a changing environment. This benefits system designers, as they do not need to redesign systems whenever the environment changes.
4. Missing and noisy data is one characteristic of biological data. The conventional computer techniques fail to handle this. Soft computing based techniques are able to deal with missing and noisy data.
5. With advances in biotechnology, huge volumes of biological data are generated. In addition, it is possible that important hidden relationships and correlations exist in the data. Soft computing methods are designed to handle very large data sets, and can be used to extract such relationships.

3.1 Relevance of Artificial Neural Network in Bioinformatics:

An Artificial Neural Network (ANN) is an information processing model that is able to capture and represent complex input-output relationships. The motivation the development of the ANN technique came from a desire for an intelligent artificial system that could process information in the same way the human brain. Its novel structure is represented as multiple layers of simple processing elements, operating in parallel to solve specific problems. ANNs resemble human brain in two respects: learning process and storing experiential knowledge. An artificial neural network learns and classifies a problem through repeated adjustments of the connecting weights between the elements. A typical neural network (shown in Figure 1) is composed of input units X1, X2,... corresponding to independent variables, a hidden layer known as the first layer, and an output layer (second layer) whose output units Y1,... correspond to dependent variables (expected number of accidents per time period).

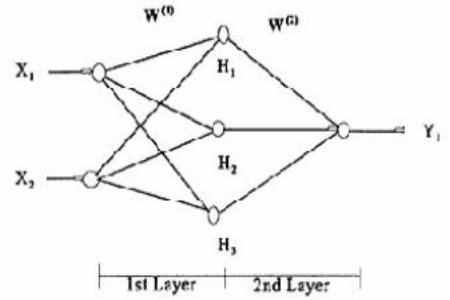

Figure 1: A simplified Artificial Neural Network

In between are hidden units H1, H2, corresponding to intermediate variables These interact by means of weight matrices W(1) and W(2) with adjustable weights. The values of the hidden units are obtained from the formulas:

$$H_j = f\left(\sum_k W_{jk}^{(1)} X_k\right)$$
$$Y_i = f\left(\sum_j W_{ij}^{(2)} H_j\right).$$

In One multiplies the first weight matrix by the input vector X = (X1, X2,...) and then applies an activation function f to each component of the result. Likewise the values of the output units are obtained by applying the second weight matrix to the vector H = (H1, H2,...) of hidden unit values, and then applying the activation function f to each component of the result. In this way one obtains an output vector Y= (Y1, Y2,...).The activation function f is typically of sigmoid form and may be a logistic function, hyperbolic tangent, etc.:

$$f(u) = \frac{1}{1 + e^{-u}}, \qquad f(u) = \frac{e}{e}$$

Usually the activation function is taken to be the same for all components but it need not be. Values of W(1) and W(2) are assumed at the initial iteration. The accuracy of the estimated output is improved by an iterative learning process in which the outputs for various input vectors are compared with targets (observed frequency of accidents) and an average error term E is computed:

$$E = \frac{\sum_{n=1}^{N} (Y^{(n)} - T^{(n)})^2}{N}.$$

Here
N = Number of highway sites or observations
Y(n) = Estimated number of accidents at site n for n = 1, 2,..., N
T(n) = Observed number of accidents at site n for n = 1, 2,..., N.
After one pass through all observations (the training set), a gradient descent method may be used to calculate improved values of the weights W(1) and W(2), values that make E smaller. After reevaluation of the weights with the gradient descent method, successive passes can be made and the weights further adjusted until the error is reduced to a satisfactory level

IV CONCLUSION:

We conclude that the Bioinformatics is a promising and innovative research. In this paper we present the various techniques that used in soft computing.

References:

[1] Waterman, Michael S. (1995). Introduction to Computational Biology: Sequences, Maps and Genomes. CRC Press. ISBN 0-412-99391-0.

[2] Mount, David W. (May 2002). Bioinformatics: Sequence and Genome Analysis. Spring Harbor Press. ISBN 0-879-69608-7.

[3] Claverie, J.M.; Notredame, C. (2003). Bioinformatics for Dummies. Wiley. ISBN 0-7645-1696-5.

[4]Hogeweg, P. (2011). "The Roots of Bioinformatics in Theoretical Biology". In Searls, David B. PLoS Computational Biology **7** (3): e1002021

[5] A. Kel, A. Ptitsyn, V. Babenko, S. Meier-Ewert and H. Lehrach (1998), "A genetic algorithm for designing gene family -specific oligonucleotide sets used for hybridization: The G protein couple dreceptor protein superfamily", Bioinformatics, Vol. 14, No. 3, pp. 259–270.

[6] A. Narayanan, E. Keedwell, and B. Olsson (2003), "Artificial Intelligence Techniques forBioinformatics", Applied Bioinformatics, Vol.1, No. 4, pp. 191-222.

[7] A. P. Gultyaev, V. Batenburg and C. W. A. Pleij (1995), "The computer simulationof RNAfolding pathways using a genetic algorithm", J. Mol. Biol., Vol. 250, pp. 37–51.

[8] A. R. Lemmon and M. C. Milinkovitch (2002), "The metapopulation genetic algorithm: Anefficient solution for the problem of large phylogeny estimation", Proc. Nat. Acad. Sci., Vol.99, No. 16, pp. 10516–10521.

[9] A. S. Wu and I. Garibay (2002), "The proportional genetic algorithm: Gene expression in agenetic algorithm", Genetic Programm. Evol. Hardware, Vol.3, No. 2, pp. 157–192