Need of Machine Learning In Bioinformatics

K.Jayanthi, C. Mahesh

Abstract*: The complexity of biological data has emerged the need of powerful and modern data analysis with computer tools and techniques. In order to satisfy the requirements of various processes in the biological data machine learning can provide different kinds of learning algorithms and this helps the computer can automatically learn by experience and construct a model for future output cases. This paper have discussed about the various biological data need to be processed with bioinformatics and techniques which are used in machine learning for achieving the analysis of biological data for creating model for them, and also some of the applications of machine learning in bioinformatics explained briefly.

Keywords: About bioinformatics, machine learning, machine learning methods, applications of machine learning in bioinformatics

I. INTRODUCTION

In Recent years, Rapid increases in biological data require computational analysis. Bioinformatics is one of the main applications of computer technology for the management of information in biological data. Biological molecules and sequences can provide the information needed to process various tasks like extracting the data, arranging the data, the properly analysing the data, and interpretation of the data which can be used for achieving different kinds of process in bioinformatics. Main target of bioinformatics is sequencing DNA and some kind of mapping. By the rapid development over various molecular, genetic and genomic researches the field related to molecular biology can produce a mass amount of information. There are so many biological processes are in the decade of biological research area that leads to increase the understanding of the process by using the computer science and information technology. The bioinformatics goal is to improve the different perspective of understanding the biological information. This can be achieved by machine learning especially for accurate classification and prediction of tremendous amount of data from changing environment [2]. Analysing huge collection of biological data includes gene finding, gene expression, gene classification, microarray data, disease gene prediction, statistical framework of protein- protein interaction, clustering of similar gene data and etc., Feature extraction from different kinds of input and output relationships of biological data can be accurately analysed by machine learning. This paper discuss about how these biological data be processed and making sense of data by using machine learning techniques.

A. Bioinformatics Research Area

a) Sequence Analysis

In computational biology, the most primordial operation is sequence analysis. These operations identify which sequence is identical and which part differs in analysis of biological data and genome mapping processes. This analysis manipulating the DNA, RNA, peptide sequence to understands its features, structure, function and evolution. The DNA sequencing process determines the order of adenine, guanine, cytosine, and thymine.

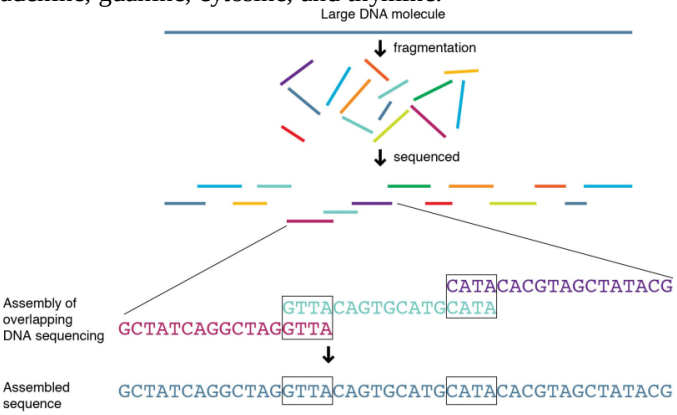


Figure 1: Genome Sequencing

b) Genome Annotation

Annotation is detailed description of individual genes and its RNA product. It also marks the gene and other features of biological in a DNA sequence. Annotation process can determine the coding regions and non-coding regions of a particular genome, genome functions and genome positions.

There is some annotation software system available which is used finding and mapping various genes, introns and exons, repeats, regulatory elements and mutations. Genome software systems needs genome database to retrieve information from the genome data like name of the gene, protein product and DNA sequence behaviours [3]. Some of the software tools for annotation

1. Blast2Go
2. GREAT
3. DAVID
4. Genes2Go
5. KOBAS
6. FUNC
7. Annotare

Revised Manuscript Received on September 07, 2019
***** Correspondence Author
K.Jayanthi*,Department of Computer Science and Engineering, Veltech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Chennai, India.
Dr.C.Mahesh, Department of Information Technology , Veltech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Chennai,

8. GOTA

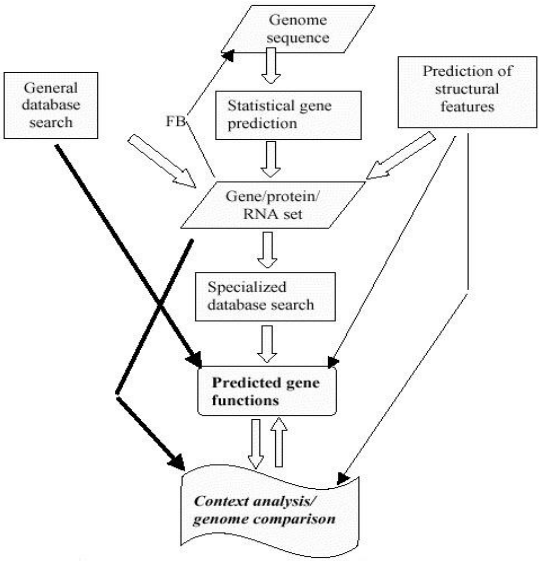


Figure 2: Genome Annotation

c) Gene Expression Analysis

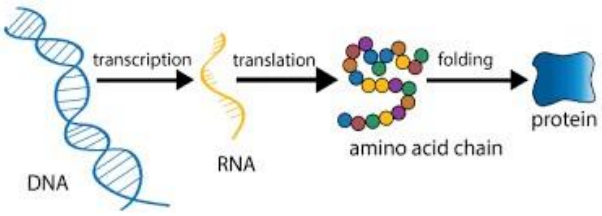
Gene product contained more information about RNA or protein. All the time these complete data not needed. By using gene expression analysis we can find the relevant expression from the gene. It has steps to convert the DNA sequence to express as RNA and protein products [4]. Using promoters and activators gene expression can be processed for analysing. Here measurement of mRNA levels with different techniques for example

1. Expressed cDNA sequence tag(EST) sequencing
2. Microarrays
3. Serial analysis of gene expression(SAGE) tag sequencing
4. Massively parallel signature sequencing (MPSS)

Developing statistical tools for separating signal from noise in high throughput gene expression is the major research area from this gene expression analysis.

d) Protein Expression Analysis

Protein expression is describing which proteins are arranged, modified and regulated in organisms. These analyses are required to study the structure of the protein, interaction of protein and functions, comparisons of the structures and functional predictions. It also predicts the actual gene activity [5]. Some of the measures can provide a snapshot of microarrays, mass spectrometry (MS), and high throughput (HT) in a biological sample. These measurements are making very sensible for protein expression analysis in bioinformatics.



e) Mutation Analysis

Genomes are rearranged in complex or it may be in unpredictable way if the gene is having affected cells. Many techniques are used to identify the mutations in the gene to determine the disease impacts from the genome product [6]. The main process of genetic approach begins with mutants only.The following steps are necessary for identifying the mutants from the gene:

1. Designing an Effective mutation – detection system for analysing.
2. A mutagen used to generate a huge collection of mutants that will display variants in the process.
3. By using complementation test group the mutations into genes.
4. The genes can be characterized by structure and function of each.
5. Isolating the gene based on DNA technology.

Depends upon the disease the affected cells of the gene mutation are different. Based on the requirement detection system be designed to identify the mutants.

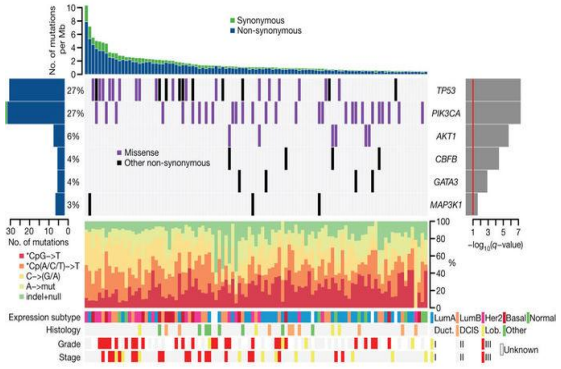


Figure 3: Sample analysis of breast cancer mutation analysis

f) Protein structure prediction

Protein structure prediction is determining the amino acid sequence of a protein on the gene that codes for this prediction. This protein structure is generally a primary structure and also it is unique one in the gene environment. Understanding the knowledge of this structure is important for the function of the protein. This type of prediction maximum involved in drug design for a various types of diseases respective to the protein structure and also designing of novel enzymes.

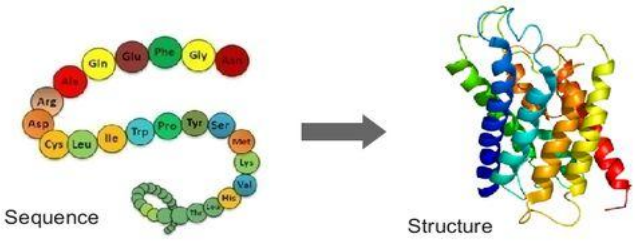


Figure 4: Protein Structure Prediction

g) Modeling biological systems

An important task is needed for mathematical biology to developing an efficient system to solve various processes by modeling biological systems. These kind of computational systems are used to develop and make use of efficient algorithms, visualization and data structures and communication tools for the combination of huge quantities of biological data with the aim of modeling in the computer.

h) High Throughput Image Analysis

Biomedical image analysis provides accurate diagnostics and it the systems make measurements from large or complex set of images, it also supports fully automatic the various processes and quantification analysis of high information content biological images. By developing a complete set of analysis system may replace the observer. Some examples of image analysis like clinical image analysis, DNA mapping, Bio image informatics, visualization.

i) Comparative genomics

It is the study of relationship of different kinds of genome structure and process across various biological environments. It provides more application in genomics especially in gene finding to discover a non-coding functional elements of the genome can be recognized. Comparative genomics accomplishes both identical and differences in the proteins, RNA, and regions of various organisms. One of the recent research papers of computer science is comparison of different kinds of genomic data. It provides the detailed view of how organisms are related to each other in the genetic level; by analysis these researchers can easily understand the structure of molecular levels. It also identifies the unique characteristics of each organism in the gene.

B. Bioinformatics Tools

To solving various problems in biological data using computer technologies we have some tools for achieving the goals of bioinformatics. Such as

Bioinformatics Research Area	Tool (Application)	References
Sequence alignment	BLAST	http://blast.ncbi.nlm.nih.gov/Blast.cgi
	CS-BLAST	ftp://toolkit.lmb.uni-muenchen.de/csblast/
	HMMER	http://hmmer.janelia.org/
	FASTA	www.ebi.ac.uk/fasta33
Multiple sequence alignment	MSAProbs	http://msaprobs.sourceforge.net/
	DNA Alignment	http://www.fluxus-engineering.com/align.htm
	MultAlin	http://multalin.toulouse.inra.fr/multalin/multalin.html
	DiAlign	http://bibiserv.techfak.uni-bielefeld.de/dialign/
Gene Finding	GenScan	genes.mit.edu/GENSCAN.html
	GenomeScan	http://genes.mit.edu/genomescan.html
	GeneMark	http://exon.biology.gatech.edu/
Protein Domain Analysis	Pfam	http://pfam.sanger.ac.uk/
	BLOCKS	http://blocks.fhcrc.org/
	ProDom	http://prodom.prabi.fr/prodom/current/html/home.php
Pattern Identification	Gibbs Sampler	http://bayesweb.wadsworth.org/gibbs/gibbs.html
	AlignACE	http://atlas.med.harvard.edu/
	MEME	http://meme.sdsc.edu/
Genomic Analysis	SLAM	http://bio.math.berkeley.edu/slam/
	Multiz	http://www.bx.psu.edu/miller_lab/
Motif finding	MEME/MAST	http://meme.sdsc.edu
	eMOTIF	http://motif.stanford.edu

Figure 5: Bioinformatics Tools

II. MACHINE LEARNING IN BIOINFORMATICS

In traditional computer system which is used to solve the biological data not able to adopt with the complex data and also not to deal with rapid increment of data in the real world. Machine Learning plays a vital role to solve the problems in existing techniques of bio informatics. Biological data needs to be diagnosing by various techniques to predict different kinds of process in genome data products. That can be automatically learned by the computers and produce tremendous amount of support in the field of bioinformatics. Machine learning provides adaptive environment for learning in various aspects such as learning by experience, learning by analogy, learning by example. Through these kinds learning capabilities of computer systems can automatically improve the performance over time based on the past results [12].

III. APPLICATIONS OF MACHINE LEARNING IN BIOINFORMATICS

Machine learning in bio informatics includes neural networks, Markov model, support vector machines, and graphical models are useful in handling uncertainty and randomness of data and mainly they are used for genome analysis such as prediction of coding or non-coding region of genome also for prediction of RNA etc.,

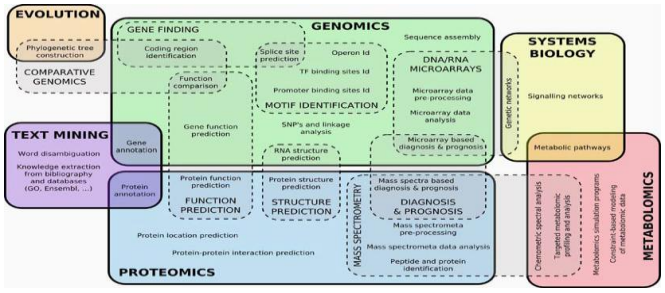


Figure 6: Bioinformatics Applications

Application of genomic in machine learning are Gene Expression Data Analysis using Clustering Algorithms Gene Finding Algorithms using Simple Markov models, Hidden Markov Model (HMM), Parameter Estimation, Viterbi Algorithm. Motif Finding Algorithms using different discrete distribution, Genome Alignment using HMM. Integration of different biological data for feature extraction using Classification Algorithms, Random Forest, Decision Trees, Regression, Naive bayes. Differentiating the various genes/tfs based on different kinds of group tables like ANOVA, t-test, F-test.

IV.MACHINE LEARNING ALGORITHMS

Machine learning algorithms can address various problems in bioinformatics. The algorithms can be as follows [15]

1. Supervised Learning

2. Unsupervised Learning
3. Semi supervised Learning
4. Reinforcement Learning
5. Optimization

A. Supervised Learning

A database contains training data that consists of pairs as input cases along with desired outputs. By using these training samples of data the algorithm construct a function used to predict accurate model that makes future predictions of output values is unknown. Here two types of tasks are there,

1.Classification
2.Regression

Classification – When the label is determined as a finite set of discrete values that task is known as Classification.

Regression – When the label is determined as a finite set of continuous values that task is known as Regression.

B. Unsupervised Learning

A database contains training data of input cases, its main function is to partition the training samples into clusters so that the data in each cluster displays a high level of proximity. Here the labels are not given to train the database.

C. Semi - supervised Learning

A database contains training data that has both labeled and unlabeled examples, the main goal is to design a model that able to exactly predict the target of future output cases for which its output value is not known. Typically, this database holds a small amount of labeled data together with a huge amount of unlabeled data.

D. Reinforcement Learning

It allows software agents to automatically manipulate the optimal behavior within a specific situation, to maximize it level of performance. A sample beneficial feedback is needed for the agent to learn its behavior.

E. Optimization

The process of searching for an optimal solution in a space of multiple solutions of possibility as searching for the model that best fits the data, optimization methods can be in modelling.

V.CONCLUSION

Bioinformatics is the study of various process involved in biological data with computer tools and techniques. Generally biological data is complex to analyse and in the rapid development of data need to be processes automatically and learning through the model which created by past experience this can be achieved only through machine learning algorithms. In this study paper describes different process of biological data with machine learning, what are the analysis needed to be processed in bioinformatics, machine learning methods and applications of machine leaning in bioinformatics discussed.

REFERENCES

1. Zaki , J.; Wang , T.L. and Toivonen, T.T. (2001). BIOKDD01: Workshop on Data Mining in Bioinformatics".
2. Li, J.; Wong, L. and Yang, Q. (2005). Data Mining in Bioinformatics, IEEE Intelligent System, IEEE Computer Society.
3. Liu, H.; Li, J. and Wong, L. (2005). Use of Extreme Patient Samples for Outcome Prediction from Gene Expression Data, Bioinformatics, vol. 21, no. 16, pp. 3377–3384
4. Yang, Qiang. Data Mining and Bioinformatics: Some Challenges, http://www.cse.ust.hk/~qyang
5. Berson, Alex, Smith, Stephen and Threaling, Kurt, "Building Data Mining Application for CRM", Tata McGraw Hill.
6. Zhang, Yanqing; C., Jagath, Rajapakse, Machine Learning in Bioinformatics, Wiley, ISBN: 978-0-470-11662-3Kuonen, Diego. Challenges in Bioinformatics for Statistical Data Miner, Bulletin of the Swiss Statistical Society, 46; 10-17.
7. Richard, R.J. A. and Sriraam, N. (2005). A Feasibility Study of Challenges and Opportunities in Computational Biology: AMalaysian Perspective, American Journal of Applied Sciences 2 (9): 1296-1300.
8. Tang, Haixu and Kim, Sun. Bioinformatics: mining the massive data from high throughput genomics experiments, analysis of biological data: a soft computing approach, edited by Sanghamitra Bandyopadhyay, Indian Statistical Institute, India.
9. Nayeem, Akbar; Sitkoff, Doree, and Krystek, Jr., Stanley. (2006) A comparative study of available software for highaccuracyhomology modeling: From sequence alignments to structural models, Protein Sci. April; 15(4): 808–824 N., Cristianini and M., Hahn. (2006) Introduction to Computational Genomics, Cambridge University Press. ISBN.
10. Prompramote S, Chen Y, Chen Y-PP.(2005) Machine learning in bioinformatics. In Bioinformatics Technologies (Chen Y-PP., ed.), Springer, Heidelberg, Germany, pp.117–153.
11. Somorjai RL, Dolenko B, Baumgartner R. (2003) Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. Bioinformatics 19:1484–1491.
12. Larra˜naga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Arma˜nanzas R, Santaf´e G, P´erez A, Robles V. (2006) Machine learning in bioinformatics.Briefings in Bioinformatics 7:86–112.
13. Alpaydin E. (2004) Introduction to Machine Learning, MIT Press, Cambridge, MA.5. Mitchell T. (1997) Machine Learning,McGraw Hill, New York.
14. Causton HC, Quackenbush J, Brazma A. (2003) A Beginner's Guide. Microarray Gene.

AUTHORS PROFILE

Mrs. K Jayanthi was awarded B. Tech. in Information Technology from Anna University, Chennai -2009. She was awarded M. Tech. in Information Technology from Anna University, Chennai - 2013.She is currently working as an Assistant professor, Department of Information Technology in Veltech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Chennai. Her research interests include machine learning and deep learning.

Dr.C. Mahesh has more than 20 years of experience in the field of teaching. He was awarded B.E in Electrical and Electronics Engineering from Madras University, Chennai. He was awarded M.E in Computer Science and Engineering, Anna University, Chennai. He was awarded Doctorate in Computer Science and Engineering in the year 2016. Currently he is working as an Associate professor in Veltech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Chennai. His area of interests includes Neural Networks and Natural language Processing.