Explore the world of Bioinformatics with Data Mining

[1]Swapnaja More, [2] Dr. Ajit More
[1]Research Scholar, [2]Research Guide

[1]Departement of Computer Science, Y.M. College, Bharati Vidyapeeth Deemed University, Pune, Maharashtra, India
[2]Departement of Computer Applications and system Studies, IMED,
Bharati Vidyapeeth Deemed University, Pune, Maharashtra, India.

Abstract: Since last decade the studies in genomics, proteomic, and various biological researches have generated huge amount of data from biological background. To interpret the data requires sophisticated computational analysis to conclude the results. To solve biological problems one of the most active areas to conclude the structure and principles of biological datasets is the use of data mining Some typical examples of biological analysis performed by data mining involve protein structure prediction, gene classification, analysis of mutations in cancer and gene expressions. It is important that the implementation of data mining progresses to continue the development of an active area of bioinformatics research since biological data and research have become vaster. This paper highlights information from varied sources in order to discuss an overview of the application of data mining in bioinformatics and a conclusive summary also point out some of the present-day demanding situations and possibilities of data mining in bioinformatics.

Index Terms - **Protein Sequences Analysis, Data Mining, Machine learning Bioinformatics Tools.**

I. INTRODUCTION

In latest years, rapid evolution in genomics and proteomics have generated a big amount of biological data. Drawing conclusions to handle these data requires sophisticated computational analyses techniques. Computational biology or bioinformatics, is the multidisciplinary science of interpreting biological data using information technology and computer science. The significance of this new area of inquiry will develop as we continue to generate and combine huge quantities of genomic, proteomic, and other data. Application and development of data mining techniques to solve biological problems is the specific active area of research in bioinformatics(Yang, Andrian Troup et al. 2017). Analysing large biological data sets requires experience of the data by means of inferring structure or generalizations from the data. Protein structure prediction, gene classification, cancer classification based on microarray data, clustering of gene expression data, statistical modeling of protein-protein interaction, etc are the examples of this type of analysis(Gao, Xu et al. 2018). Therefore, we see a great capability to increase the interaction between data mining and bioinformatics.

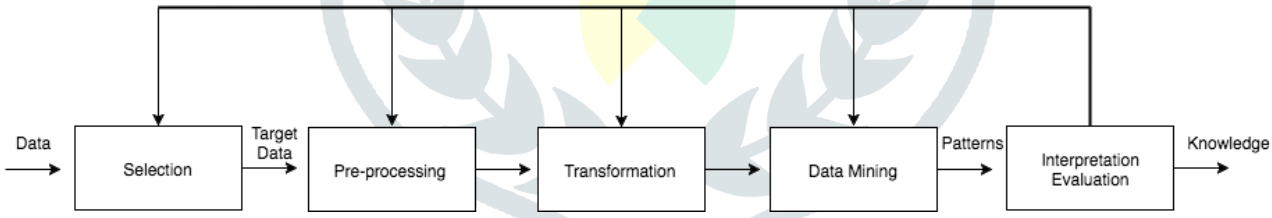

Figure 1: Data Mining Protocol

II. BIOINFORMATICS

For the research of informatic processes in biotic systems Paulien Hogeweg in 1979 coined term named Bioinformatics(Sheng-Yong Niu, Jinyu Yang et al. 2017). It has been particularly used in those areas of genomics involving large-scale DNA sequencing in genomics and genetics since late 1980s. Bioinformatics can be described as the implementation of computer technology to the management of biological information. Bioinformatics is the science of organizing, storing, extracting, analyzing, interpreting and utilizing information from biological molecules and sequences(Faiza 2016) . It has been particularly powered by advances in mapping techniques and DNA sequencing. Developments in genomic and other molecular research technologies give rise to huge amount of data and application information technologies have blended to provide a tremendous amount of information related to molecular biology. The primary objective of bioinformatics is to increase the interpretation of biological processes.

III. SOME OF THE AREAS OF RESEARCH IN BIOINFORMATICS INCORPORATE

A. Sequence analysis

Most primitive operation in computational biology is the sequence analysis technique. This operation includes to find which part of the biological sequences are similar and which part vary during scientific analysis and genome mapping techniques. DNA sequence to sequence alignment, sequence databases, repeated sequence searches and other bioinformatics methods on a computer can carried out by sequence analysis in biological terms(Zeng, Wang et al. 2017).

B. Gene expression analysis

Measuring mRNA levels with diverse strategies used to the expression of many genes such as microarrays, expressed cDNA sequence tag (EST) sequencing, massively parallel signature sequencing (MPSS), serial analysis of gene expression (SAGE) tag sequencing, or various applications of multiplexed in-situ hybridization etc. These all techniques are particularly noise susceptible and subject to bias in the biological measurement. Here the fundamental studies involve developing statistical tools to separate signal from noise in high-throughput gene expression studies(Chuang Kee, OngQi Bin et al. 2019).

C. Analysis of mutations in cancer

In most cancers, the genomes of affected cells are rearranged in complex or maybe unpredictable ways(Gaonkar, Patankar et al. 2018). Massive sequencing efforts are used to become aware of earlier unknown point mutations in a variety of genes in cancer. Control the huge volume of sequence data produced, and create new algorithms and software to examine the sequencing results to the developing collection of human genome sequences are all maintained by bioinformaticians. New physical detection technologies are employed, such as microarrays to discover chromosomal gains and losses and single-nucleotide polymorphism arrays to detect recognised point mutations(Nicolussi, Belardinilli et al. 2019).

D. Genome annotation

In the genomics context, annotation is the procedure of marking the genes and other biological features in a DNA sequence. The first genome annotation software program was designed in 1995 by Dr. Owen White(Escobar, Godoy-Lozano et al. 2018).

E. Protein structure prediction

The amino acid sequence of a protein called primary structure can be easily determined from the sequence on the gene that codes for it. This primary structure uniquely determines the structure in its native environment in most of the cases. Knowledge of this structure is essential to understand the function of the protein. Structural information commonly categorised as secondary, tertiary and quaternary structure. Protein structure prediction is one of the most essential for drug design and the design of novel enzymes. A general prediction for such an open problem is challenge for the researchers(Pearson 2016).

F. Comparative genomics

Study of the relationship of genome structure and function across different biological species is called comparative genomics. Discovery of new, non-coding functional elements of the genome hence gene finding is an vital utility of comparative genomics. Comparative genomics(Tringe, von Mering et al. 2005) make use of both similarities and differences in the proteins, RNA, and regulatory regions of different organisms. In computer science, computational approaches to genome comparison have recently become a common research topic(Reuter, Spacek et al. 2015).

G. Analysis of protein expression

Gene expression is measured in lots of approaches together with mRNA and protein expression, but protein expression is one of the best clues of actual gene activity considering the fact that proteins are usually final catalysts of cell activity. Image of the proteins present in a biological sample can be provided by Protein microarrays and high throughput (HT) mass spectrometry (MS). Bioinformatics is greatly involved in making perception of protein microarray and HT MS data(Koonin 2015).

H. Modeling biological systems

Modeling biological systems is a remarkable problem of systems biology and mathematical biology. Computational systems biology targets to evolve and use systematic algorithms, data structures, and visualization and communication tools for the combination of large quantities of biological information with the goal of computer modeling. It includes the use of computer simulations of biological structure, like cellular subsystems along with the networks of metabolites and enzymes, signal transduction pathways and gene regulatory networks. Computer simulation of simple life forms to understand evolutionary processes via the artificial existence(Escalona, Rocha et al. 2016).

I. High-throughput image analysis

Computational technologies are used large amounts of high information content biomedical images to boost up the completely automate the processing, quantification and evaluation. Modern image analysis systems increase an observer's ability to make measurements from a big or complex set of images. An entirely developed analysis system may completely update the observer. Biomedical imaging is becoming more essential for both diagnostics and research. Clinical image analysis and visualization, inferring clone overlaps in DNA mapping, Bioimage informatics, etc. are some of the examples of research in this area(Reuter, Spacek et al. 2015).

J. Protein-protein docking

Since last two decades, tens of thousands of protein three-dimensional structures have been resolved by Protein nuclear magnetic resonance spectroscopy (protein NMR) and X-ray crystallography. whether or not it's far realistic to expect possible protein-protein interactions only based on these 3D shapes, without doing protein-protein interaction experiments is the current critical question raised for the biological scientist. A variety of methods have been developed to tackle the Protein-protein docking problem; however, it seems that there is still much work to be done in this field.

IV. BIOINFORMATICS TOOLS

Following are the some of the important tools for bioinformatics (Table 1)

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

Table 1: Bioinformatics Tools.(Sheng-Yong Niu, Jinyu Yang et al. 2017),(Raza, K. 2012)

V. DATA MINING

Extracting knowledge from huge amounts of data is termed as "mining" and the process refers to data Mining (DM). It is the technology of finding new exciting patterns and relationship in massive amount of data. It is defined as the method of discovering meaningful new correlations, patterns, and trends by digging into large amounts of data stored in warehouses. Data mining is also termed as Knowledge Discovery in Databases (KDD). It calls for intelligent technologies and the willingness to explore the possibility of hidden knowledge that settled in the data. Data mining proceed towards perfectly suited for Bioinformatics, since it is data-intensive, but lacks a complete theory of life's organization at the molecular level. The large databases of biological information create each challenges and opportunities for development of novel KDD strategies. To extract useful knowledge from large datasets gathered in biology, and in other related life sciences areas carried out by data mining facilitates (Hu, X.2011).

A. MINING FEATURES

Prediction and description are two high-level primary goals of data mining in practice.
Mining new patterns from the data are:

- Classification: Classification is learning a function that classifies a data item into one of several predefined classes.
- Estimation: Given some input data and coming up with a value for some unknown continuous variable.
- Prediction: This is same as classification & estimation only difference is that the records are classified according to some future behaviour or estimated future value.
- Association rules: It is determination of which things go together, also called dependency modeling.
- Clustering: This termed as segmenting a population into a number of subgroups or clusters.
- Description & visualization: It indicates representation of the data using visualization techniques.

B. TEXT MINING

The increase in available biological publications force to the issue of the increase in difficulty in searching through and compiling all the relevant available information on a given topic across all sources. This task is known as knowledge extraction. This is important for biological data collection which can then, in turn, be fed into machine learning algorithms to generate new biological knowledge. Machine learning can be used for this knowledge extraction task using techniques such as Natural Language Processing (NLP) to extract useful information from human-generated reports in a database (Chauhan, N. S.2019).

Knowledge is gained through the use of differing machine learning methods used which falls into two categories: supervised and unsupervised learning. The first three tasks that is classification, estimation and prediction are the examples of supervised learning. The unsupervised learning includes association rules, clustering and description & visualization. In unsupervised learning, the goal is to establish some relationship among all the variables where no variable is singled out as the target. Unsupervised learning aims to find patterns without the use of a particular target field. The progress of new data mining and knowledge discovery tools is a subject of active research areas. One inspiration behind the development of these tools is their potential application in modern biology.

VI. APPLICATION OF DATA MINING IN BIOINFORMATICS

In the past, there are limitations to obtain and process microbial big data, scientists were not able to obtain a full understanding of the microbiota. High dimensional complicacy of the microbiota cannot meet by the sequencing technologies or the analysis tools. The integration of the current sequencing methods would be necessary to conduct a comprehensive study on microbiota in the future. First, the taxonomic information at various levels can be obtained by amplicon sequencing and metagenomic sequencing(Fernando Meyer 2019). Second, the functional annotation can be estimated by metagenomics and established by the multi-omics including metaproteome, metagenome, metatranscriptome, and metabolome(Ward and The Institute for Genomic Research 2006). Third, the connection between functions and phylogeny of a single microbe cell can be established by single-cell sequencing. Finally, the interactions between all chromosomes can be detected by sequencing techniques. The integration of these methods can answer the questions "who is there," "what are they doing," and "how are they doing" from a macroscopic level.

The comprehensive analysis of big data, followed by strict in vivo and in vitro experiments, is required to determine the causality of clinical diseases by microbes for specific medicine. Some areas from bioinformatics perspective where we can apply data mining techniques such as protein function domain detection, function motif detection, gene finding, protein function inference, disease diagnosis, disease treatment optimization, protein and gene interaction network reconstruction, and data cleansing(Yang, Andrian Troup et al. 2017). For example, microarray technologies are used to predict a patient's outcome like survival time and risk of tumour metastasis or recurrence can be estimated on the basis of patients' genotypic microarray data. Machine learning presents effective solution as various classification methods can be used to perform this identification. The most commonly used methods are radial basis function networks, deep learning, Bayesian classification, decision trees, and forest. Next-generation sequencing (NGS) has enabled methods for precise definition of breakpoints of genome structural variation of different sizes and types(Buermans and den Dunnen 2014; Kulski 2016). Machine learning can be used for peptide identification through mass spectroscopy. An efficient scoring algorithm that examines the correlative information in comprehensive ways is highly desirable(Afgan, Krampis et al. 2015).

VII. CONCLUSION

Bioinformatics and data mining are evolving as interdisciplinary science. Data mining perspective looks ideally suited for bioinformatics, due to the fact bioinformatics is data-rich but lacks a comprehensive theory of life's organization at the molecular level. We believe that there are several motives why these genomic data mining approaches have been successful and represent a promising direction for future work. However, data mining in bioinformatics is constrained by many angles of biological databases, such as their size, number, diversity and the lack of a standard ontology. Another issue is the range of levels the domains of expertise present among potential users, so it can be difficult for the database curators to provide access mechanism appropriate to all. The integration of biological databases is likewise a problem. Data mining and bioinformatics are rapidly expanding research area today. It is essential to have a look at what are the vital research troubles in bioinformatics and develop new data mining methods for scalable and effective analysis. Ultimately, we think for machine learning to really grow well, it's going to come down to better bioinformatics data. Heath and bioinformatics data right now have pretty poor statistical power.

VIII. ACKNOWLEDGMENT

This work has been supported by Department of Computer Science, Y.M. College, Pune and Department of Computer Applications and system Studies, IMED, Bharati Vidyapeeth, Pune for strong intramural support for experimental, computational work and for the critical reading of this manuscript.

REFERENCES

[1] Afgan, E., K. Krampis, et al. (2015). Building and provisioning bioinformatics environments on public and private Clouds. 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO).

[2] Buermans, H. P. J. and J. T. den Dunnen (2014). "Next generation sequencing technology: Advances and applications." Science direct.

[3] Chauhan, N. S. (2019 ·). "Explore the world of Bioinformatics with Machine Learning."

[4] Chuang Kee, OngQi Bin, et al. (2019). Pipeline of High Throughput Sequencing. Encyclopedia of Bioinformatics and Computational Biology. Oxford, Academic Press: 144-151.

[5] Escalona, M., S. Rocha, et al. (2016). "A comparison of tools for the simulation of genomic next-generation sequencing data." Nature Reviews Genetics17: 459.

[6] Escobar, A., E. Godoy-Lozano, et al. (2018). "Analysis of sequencing strategies and tools for taxonomic annotation: Defining standards for progressive metagenomics." Scientific Reports8.

[7] Faiza, M. (2016). "Big Data in Bioinformatics " Cloud Computing.

[8] Fernando Meyer, A. B., Peter Belmann, Stefan Janssen, Alice C. McHardy and David Koslicki (2019). "Microbiomes and Metagenomics" BMC Part of Springer Nature.

[9] Gao, L., T. Xu, et al. (2018). "Oral microbiomes: more and more importance in oral cavity and whole body." Protein & cell9(5): 488-500.

[10] Gaonkar, P., S. Patankar, et al. (2018). "Oral bacterial flora and oral cancer: The possible link?" Journal of Oral and Maxillofacial Pathology22(2): 234-238.

[11] Hu, X. (2011). Data mining and its applications in bioinformatics: Techniques and methods. 2011 IEEE International Conference on Granular Computing

[12] Raza, K. (2012). "Application of Data Mining In Bioinformatics." Indian Journal of Computer Science and Engineering Vol 1 (No 2): 114-118.
[13] Koonin, E. V. (2015). "Why the Central Dogma: on the nature of the great biological exclusion principle." Biology Direct10: 52-52.
[14] Kulski, J. K. (2016). "Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications." Books : Next Generation Sequencing - Advances, Applications and Challenges.
[15] Nicolussi, A., F. Belardinilli, et al. (2019). "Next-generation sequencing of BRCA1 and BRCA2 genes for rapid detection of germline mutations in hereditary breast/ovarian cancer." PeerJ. 2019 Apr 22;7:e6661. doi: 10.7717/peerj.6661. eCollection 2019.
[16] Pearson, W. R. (2016). "Finding Protein and Nucleotide Similarities with FASTA." Current protocols in bioinformatics53: 3.9.1-3.9.25.
[17] Reuter, J. A., D. V. Spacek, et al. (2015). "High-throughput sequencing technologies." Molecular cell58(4): 586-597.
[18] Sheng-Yong Niu, Jinyu Yang, et al. (2017). "Bioinformatics tools for quantitative and functional metagenome and metatranscriptome data analysis in microbes ".
[19] Tringe, S. G., C. von Mering, et al. (2005). "Comparative metagenomics of microbial communities." Science308(5721): 554-7.
[20] Ward, N. and R. The Institute for Genomic Research, MD, USA; and Center of Marine Biotechnology, Baltimore, MD, USA (2006). "New directions and interactions in metagenomics research."
[21] Yang, Andrian Troup, et al. (2017). "Scalability and Validation of Big Data Bioinformatics Software." Computational and Structural Biotechnology Journal15: 379-386.
[22] Zeng, F., Z. Wang, et al. (2017). "Large-scale 16S gene assembly using metagenomics shotgun sequences." Bioinformatics. 2017 May 15;33(10):1447-1456. doi: 10.1093/bioinformatics/btx018.