International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

Current Challenges in Text Mining and Data Mining and its Applications in Bioinformatics Tools

Sarangam Kodati[1], Dr. R P. Singh[2]

[1.] *Research Scholar, Department of Computer Science and Engineering, Sri Satya Sai University of Technology and Medical Science, Sehore,Bhopal,Madhya Pradesh ,(India)*

[2.] *Professor, Department of Computer Science and Engineering,Sri Satya Sai University of Technology and Medical Science,Sehore,Bhopal, Madhya Pradesh , (India)*

ABSTRACT

in this article text mining, data mining, basic concepts are explained and research area of bioinformatics are reported. in the domain of bioinformatics text mining and data mining application are highlighted. in bioinformatics few of the current changes and opportunities of text mining and data mining are discussed.

Keywords: Text Mining, Data Mining, , Bioinformatics Tools.

I.INTRODUCTION

Bioinformatics and computational biology are an interdisciplinary fields of interpreting biological data with computer discipline and information technology . In current years researchers are focused on computational biology and bioinformatics due to have fast growth developments in genomics and proteomics. The genomics and proteomics have reported biological data and these requires difficult in computational analyses. hence investigation will develop and go on to produce and combine large amount of genomic data and proteomic data.

To explain biological problems in bioinformatics exploit the application and development of text mining and data mining techniques. Investigating large biological data sets requires making sense of the data by conclude structure from the data. Such as bioinformatics data sets of analysis include protein structure prediction, cancer classification based on microarray data, gene types, gene expression data clustering and protein-protein interaction statistical modeling. Hence, a great possible to increase the text mining and data mining interaction in bioinformatics. therefore in this article basics of text mining, data mining and current challenges of text mining and data mining in bioinformatics are reported.

International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

II.TEXT MINING

Text mining is used to find fascinating information from large database and it also represent with text data mining. It can be work with semi structured or unstructured data from database and the text mining is to investigate large amount of usual language text. It identify lexical patterns to extract useful information and it is useful for organization since the information is in text format.

The text mining helps to converts the unstructured text into structured data

In text mining the following steps are can be integrated

- From structured data identify the pattern
- using text mining methods analyze the pattern
- from the text extract the useful information

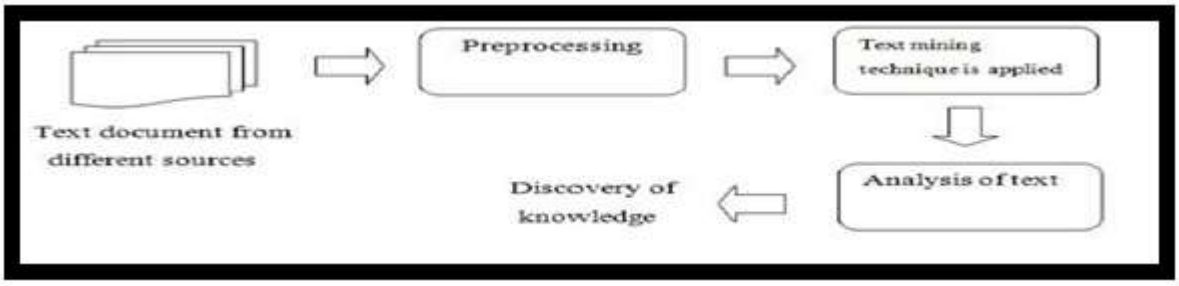

Figure 1: Text mining processing method

III. TEXT MINING TECHNIQUES

In text mining process used the tools such as summarization, clustering, visualization, classification and information extraction, etc.,

Summarization:

Summarization method is used to summarize as short data without change meaning of content instead of large data documents. therefore the complete document set replace with short summery.

Clustering:

Clustering technique is used to dividing similar text into same cluster and each cluster contain a number of similar documents. Clustering is an unsupervised method and it differs from categorization method.

Visualization:

The visualization method used to discover the information with represents different color, relationship distance from group of documents or a single document text. It give batter and understandable information.

Categorization:

The categorization method is used to categorize the text document into pre define class and it prepare classification on the basis of known and unknown example automatically. Categorization is an similar to text classification and it is a supervised method.

Information Extraction :

The information extraction method is used for large text documents and also used to look for pre define sequence of text. It include identification, sentence segmentation and it is an primary step to analyze unstructured text and its relationship for computer.

IV. DATA MINING

Data mining is an process to discovering potential, novel, interesting and previously unknown pattern from large amount of data. it refers to use for extracting or mining knowledge from large amount of data. data mining is also called as " knowledge discovery from data"(KDD). There are a number of other terms similar to data mining like knowledge extraction, data dredging and data archaeology.

V. DATA MINING TASKS

Data mining tasks are very diverse and different since there are several patterns in a large data base. to find several kinds of patterns needed the different kinds of methods and techniques. depends on kinds of patterns data mining tasks are classified as summarization, clustering, classification, association and trend analysis

Summarization :

It is the abstraction or generalized of data. A set of task-relevant data is summarized and abstracted, resulting a smaller set which gives a general overview of the data and usually with aggregation information.

Clustering :

Segmenting a population into a number of subgroups or clusters.

Classification:

Classification is learning a function that maps (classifies) a data item into one of several predefined classes.

Association:

Determining which things go together, also called dependency modeling.

The development of new data mining and knowledge discovery tools is a subject of active research. One motivation behind the development of these tools is their potential application in modern biology.

VI. BIOINFORMATICS

Bioinformatics is an study of bioinformatics processes in biotic systems and the term bioinformatics was invented by Paulien Hogeweg in 1979. It was primary used since 1980 has been in genomics and genetics. Particularly in those areas of genomics involving large-scale DNA sequencing. Bioinformatics can be defined as the application of computer technology to the management of biological information. Bioinformatics is the science of storing, extracting, organizing, analyzing, interpreting and utilizing information from biological sequences and molecules. It has been mainly fueled by advances in DNA sequencing and mapping techniques. 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. The primary goal of bioinformatics is to increase the understanding of biological processes.

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 text mining of biological literature and the development of biological and gene ontology's to organize and query biological data. It also 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 Applications:

- Molecular interactions
- Molecular modeling
- DNA sequence Analysis
- Phylogenetic analysis
- c)Protein Sequence Analysis
- Drug designing
- Molecular Dynamics Simulations
- Bioinformatics Tools

Table : Bioinformatics tools

Problems	**Tools or Database**
Sequence Alignment	BLAST, FASTA
Multiple Sequence Alignment	Clustal W, Macaw
Pattern Finding	GRAIL, FGENEH, tRNAscan-SE, NNPP, eMOTIF, PROSITE, ChloroP
Structure Prediction	Bend.it, RNA Draw, NNPREDICT, SWISSMODEL
DNA Microarray	GeneX, GOE, MAT, GeNet
Hardware's for Proteomics	2D Gel, MALDI-TOF

International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

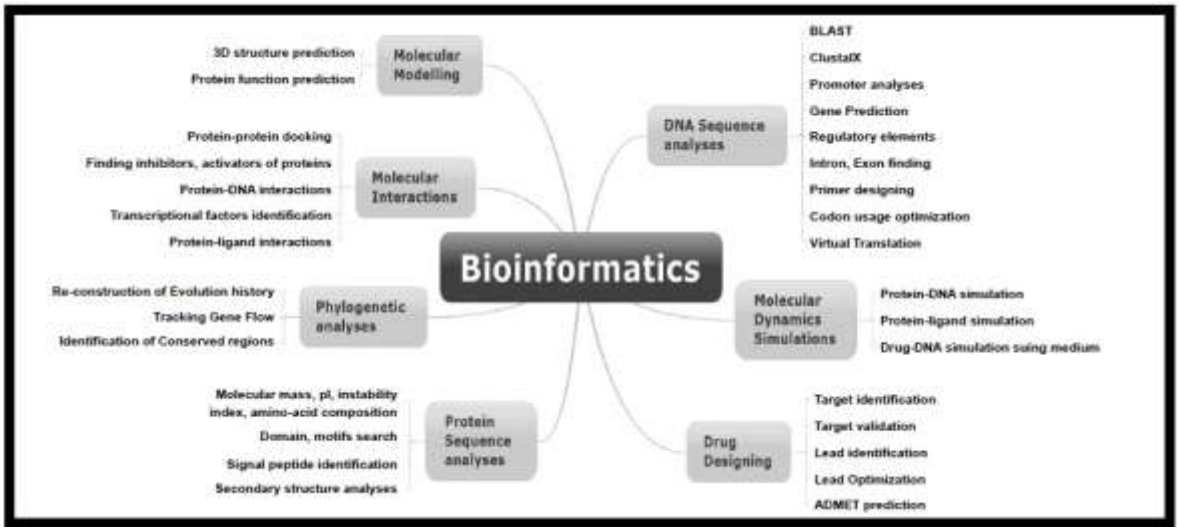

Figure 2: Applications of Bioinformatics tools in various areas of biological science

VII. TEXT MINING AND DATA MINING APPLICATION IN BIOINFORMATICS

Applications of text mining and data mining to bioinformatics include gene finding, protein function domain detection, function motif detection, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction.

For example, microarray technologies are used to predict a patient's outcome. On the basis of patients' genotypic microarray data, their survival time and risk of tumor metastasis or recurrence can be estimated. Machine learning can be used for peptide identification through mass spectroscopy. Correlation among fragment ions in a tandem mass spectrum is crucial in reducing stochastic mismatches for peptide identification by database searching. An efficient scoring algorithm that considers the correlative information in a tunable and comprehensive manner is highly desirable.

VIII. CHALLENGES IN BIOINFORMATICS

Sequence

analysis :

Sequence analysis is the most primitive operation in computational biology. This operation consists of finding which part of the biological sequences are alike and which part differs during medical analysis and genome mapping processes. The sequence analysis implies subjecting a DNA or peptide sequence to sequence alignment, sequence databases, repeated sequence searches, or other bioinformatics methods on a computer.

Genome

annotation :

In the context of genomics, annotation is the process of marking the genes and other biological features in a

DNA
sequence. The first genome annotation software system was designed in 1995 by Dr. Owen White.

Analysis of gene expression :

The expression of many genes can be determined by measuring mRNA levels with various techniques such as microarrays, expressed cDNA sequence tag (EST) sequencing, serial analysis of gene expression (SAGE) tag sequencing, massively parallel signature sequencing (MPSS), or various applications of multiplexed in-situ hybridization etc. All of these techniques are extremely noise-prone and subject to bias in the biological measurement. Here the major research area involves developing statistical tools to separate signal from noise in high-throughput gene expression studies.

Analysis of protein expression:

Gene expression is measured in many ways including mRNA and protein expression, however protein expression is one of the best clues of actual gene activity since proteins are usually final catalysts of cell activity. Protein microarrays and high throughput (HT) mass spectrometry (MS) can provide a snapshot of the proteins present in a biological sample. Bioinformatics is very much involved in making sense of protein microarray and HT MS data.

Analysis of mutations in cancer:

In cancer, the genomes of affected cells are rearranged in complex or even unpredictable ways. Massive sequencing efforts are used to identify previously unknown point mutations in a variety of genes in cancer. Bioinformatics continue to produce specialized automated systems to manage the sheer volume of sequence data produced, and they create new algorithms and software to compare the sequencing results to the growing collection of human genome sequences and germ line polymorphisms. New physical detection technologies are employed, such as oligo nucleotide microarrays to identify chromosomal gains and losses and single-nucleotide polymorphism arrays to detect known point mutations. Another type of data that requires novel informatics development is the analysis of lesions found to be recurrent among many tumors.

Protein structure prediction:

The amino acid sequence of a protein (so-called, primary structure) can be easily determined from the sequence on the gene that codes for it. In most of the cases, this primary structure uniquely determines a structure in its native environment. Knowledge of this structure is vital in understanding the function of the protein. For lack of better terms, structural information is usually classified as secondary, tertiary and quaternary structure. Protein structure prediction is one of the most important for drug design and the design of novel enzymes. A general solution to such predictions remains an open problem for the researchers.

Comparative

genomics:

Comparative genomics is the study of the relationship of genome structure and function across different biological species. Gene finding is an important application of comparative genomics, as is discovery of new, non-coding functional elements of the genome. Comparative genomics exploits both similarities and differences in the proteins, RNA, and regulatory regions of different organisms. Computational approaches to genome comparison have recently become a common research topic in computer science.

Modeling biological

systems:

Modeling biological systems is a significant task of systems biology and mathematical biology. Computational systems biology aims to develop and use efficient algorithms, data structures, visualization and communication tools for the integration of large quantities of biological data with the goal of computer modeling. It involves the use of computer simulations of biological systems, like cellular subsystems such as the networks of metabolites and enzymes, signal transduction pathways and gene regulatory networks to both analyze and visualize the complex connections of these cellular processes. Artificial life is an attempt to understand evolutionary processes via the computer simulation of simple life forms

High-throughput image

analysis:

Computational technologies are used to accelerate or fully automate the processing, quantification and analysis of large amounts of high-information-content biomedical images. Modern image analysis systems augment an observer's ability to make measurements from a large or complex set of images. A fully developed analysis system may completely replace the observer. Biomedical imaging is becoming more important for both diagnostics and research. Some of the examples of research in this area are: clinical image analysis and visualization, inferring clone overlaps in DNA mapping, Bioinformatics, etc.

Protein-protein

docking:

In the last two decades, tens of thousands of protein three-dimensional structures have been determined by X-ray crystallography and Protein nuclear magnetic resonance spectroscopy (protein NMR). One central question for the biological scientist is whether it is practical to predict possible protein-protein interactions only based on these 3D shapes, without doing protein-protein interaction experiments. A variety of methods have been developed to tackle the Protein-protein docking problem, though it seems that there is still much work to be done in this field.

IX. CONCLUSION

Bioinformatics text mining and data mining are developing as interdisciplinary science. Text mining and Data mining approaches seem ideally suited for bioinformatics, since bioinformatics is data-rich but lacks a comprehensive theory of life’s organization at the molecular level.

However, text mining and data mining in bioinformatics is hampered by many facets of biological databases, including their size, number, diversity and the lack of a standard ontology to aid the querying of them as well as the heterogeneous data of the quality and provenance information they contain. Another problem is the range of levels the domains of expertise present amongst potential users, so it can be difficult for the database curators to provide access mechanism appropriate to all. The integration of biological databases is also a problem. Challenges in text mining data mining and bioinformatics are fast growing research area today. It is important to examine what are the important research issues in bioinformatics and develop new text mining and data mining methods for scalable and effective analysis.

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] Soinov, L. (2006). Bioinformatics and Pattern Recognition Come Together. *Journal of Pattern Recognition Research (JPRR), Vol 1 (1) p.37-41*

[5] Varsha C. Pande , Dr. A.S. Khandelwal ,A Survey Of Different Text Mining Techniques, IBMRD's *Journal of Management and Research, Online ISSN: 2348-5922, Volume-3, Issue-1, March 2014*

[6] Vishal Gupta, Gurpreet S. Lehal,A Survey of Text Mining Techniques and Applications, *Journal of Emerging Technologies in Web Intelligence, Vol-1,No-1, August 2009*

[7] Mr. Rahul Patel, Mr. Gaurav Sharma, A survey on text mining techniques, *International Journal Of Engineering And Computer Science, ISSN:2319-7242, Volume 3 ,Issue 5 ,May 2014*

[8] PatilMonali S, KankalSandip S, A Concise Survey on Text Data Mining, *International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 9, September 2014, ISSN (Online) : 2278-1021,ISSN (Print) : 2319-5940*

[9] Mahesh T R, Suresh M B, M Vinayababu, Text Mining: Advancements, Challenges and Future directions, *International Journal of Reviews in Computing, ISSN: 2076-3328 E-ISSN: 2076-3336.*