Significance of Data Mining in Bioinformatics

K . Ushasri
Lect. In Microbiology
SGGDC, PILER

Dr . A . Ravi Prasad
Lect. in Computers
SGGDC, PILER

J . Kishore Kumar Reddy
Lect in Computer Science
SGGDC, PILER

S . Saravana
Lect. in Computers
PVKN DC, CHITTOOR

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

Keywords: Gene mapping, Micro Array, cDNA, KDD

INTRODUCTION

Data Mining is knowledge discovery from data, Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data , Alternative names are Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, knowledge digging, information harvesting, business intelligence, etc.Data mining is essential step in the process of knowledge discovery.
The grand area of research in bioinformatics includes the better understanding of molecular level of organization is very first step to go for further research in the area of Bio – informatics, Genomics and proteomics, sequences analysis, micro array technology etc.,

Sequence analysis

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

Databases are the system which is used to store, search and retrieve any type of data. Biological databasesare libraries of life sciences information, collected from scientific experiments, published literature, high throughputexperiment technology, and computational analyses of various research areas like genomics,proteomics, metabolomics, microarray gene expression, phylogenetic information, clinical trials. Relationaldatabase concepts of computer science and information retrieval concepts of digital libraries are important forunderstanding biological databases. Biological database design, development, and long-term management arecore areas of the discipline of bioinformatics. The available biological nucleotide databases are DDBJ(DNA Data Bank of Japan), Gen Bank and EMBL(European Molecular Biology Laboratory). The available protein sequence databases are Swisport, PIR(Protein information resource), (PRF)Protein research foundation.

The applications of these biological databases are:-

- *Accuracy*
- *Data mining*
- *Consistency*
- *Bioinformatics issues*
- *Tools for analyzing, querying and visualization*
- *Make information available globally*
- *Systematic results from experiments and analysis*
- *Non-redundancy and redundancy deduction*
- *Database design and implementations*
- *Cross- references*

1. Genome annotation

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

2. Analysis of gene expression

The expression of many genes can be determined by measuring mRNA levels with various techniques such as microarrays, expressed DNA sequence tag (EST)

sequencing, serial analysis of gene expression (SAGE) tagsequencing, massively parallel signature sequencing (MPSS), or various applications ofmultiplexed in-sit hybridization etc. All of these techniques are extremely noise-prone and subject to bias in the biologicalmeasurement. Here the major research area involves developing statistical tools to separate signal from noise inhigh-throughput gene expression studies.

3. ***Analysis of protein expression***

Gene expression is measured in many ways including mRNA and protein expression; however protein expression isone of the best clues of actual gene activity since proteins are usually final catalysts of cell activity. Proteinmicroarrays and high throughput (HT) mass spectrometry (MS) can provide a snapshot of the proteins present in abiological 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 sequencingefforts are used to identify previously unknown point mutations in a variety of genes in cancer. Bioinformaticianscontinue to produce specialized automated systems to manage the sheer volume of sequence data produced, and theycreate new algorithms and software to compare the sequencing results to the growing collection of human genomesequences and germline polymorphisms. New physical detection technologies areemployed, such as oligonucleotidemicroarrays to identify chromosomal gains and losses and single-nucleotide polymorphism arrays to detect knownpoint mutations. Another type of data that requires novel informatics development is the analysis of lesions found tobe 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 onthe gene that codes for it. In most of the cases, this primary structure uniquely determines a structure in its nativeenvironment. Knowledge of this structure is vital in understanding the function of the protein. For lack of betterterms, structural information is usually classified as secondary, tertiary and quaternary structure. Protein structureprediction is one of the most important for drug design and the design of novel enzymes. A general solution to suchpredictions 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-codingfunctional 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 haverecently 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, and visualization and communication tools for the integration of large quantities of biological data with the goal of computer modeling. It involves the use ofcomputer simulations of biological systems, like cellular subsystems such as the networks of metabolites andenzymes, signal transduction pathways and gene regulatory networks to both analyze and visualize the complexconnections of these cellular processes. Artificial life is an attempt to understand evolutionary processes via thecomputer simulation of simple life forms.

High-throughput image analysis

Computational technologies are used to accelerate or fully automate the processing, quantification and analysis oflarge amounts of high-information-content biomedical images. Modern image analysis systems augment anobserver's ability to make measurements from a large or complex set of images. A fully developed analysis systemmay 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 cloneoverlaps in DNA mapping, Bioimage informatics, etc.

Protein-protein docking

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

Data Mining Techniques In Genome Mapping

Identification of protein coding regions of the genomes and assignment of functions to these genes on the basis of sequence similarity homologies against other genes of known functions. Gene expression data mining is the identification of intrinsic patterns and relation shipsin transcriptional expression data generated by largescale gene expression experiments. Improvements in genome,gene expression and proteome database mining algorithms will enable the prediction of protein function in

the context of higher order processes such as regulation of gene expression, metabolic pathways and signaling cascades.

Data mining techniques in micro array of gene expression

Micro arrays are one of the latest breakthrough in experimental molecular biology that allow monitoring expression of tens of thousands of genes simultaneously .the expression of many genes can be determined by measuring mRNA levels with multiple technique including micro arrays, expressed cDNA tag (EST)sequencing ,serial analysis of gene expression (SAGE) tag sequencing ,massively parallel signature sequencing (MPSS) etc.

Data mining in metabolomics

Disease diagnosis using molecularprofiles has gained more attention during the last decades. Among the molecular diagnosesstudy, metabolomics has been a recently emerging field as promising field for early detection of diseases.

CONCLUSION

In this paper we briefly discussabout the various data miningtechniques, tools, applications issues and trends. Data mining has wide application domain almost in every industry where the data is generated that's why data mining is consider done of the most important frontiers in database and information systems and one of the most promising interdisciplinary developments in Information Technology. Many challenges have been identified in the field of bio informatics. Hence thisreview would be helpful to researchers to focus on thevarious issues of data mining.

REFERENCES:

1. Jiong, Lei Liu; Yang, A. and Tung, K. H. Data Mining Techniques for Microarray Datasets, Proceedings of the 21st International Conference on Data Engineering (ICDE 2005).
2. Pevzner, P. A. (200). Computational Molecular Biology: An Algorithmic Approach The MIT Press.
3. Soinov, L. (2006). Bioinformatics and Pattern Recognition Come Together. Journal of Pattern Recognition Research(JPRR), Vol 1 (1) p.37-41Aluru, S., ed. (2006). Handbook of Computational Molecular Biology. Chapman & Hall/Crc,
4. Baxevanis, A.D.; Petsko, G.A.; Stein, L.D. and Stormo, G.D., eds. (2007). Current Protocols in Bioinformatics. Wiley.
5. Mount, D. W. (2002). Bioinformatics: Sequence and Genome Analysis Spring Harbor Press. Gilbert, D. (2004). Bioinformatics software resources. Briefings in Bioinformatics, Briefings in Bioinformatics.
6. Jiong, Lei Liu; Yang, A. and Tung, K. H. Data Mining Techniques for Microarray Datasets, Proceedings of the 21st International Conference on Data Engineering (ICDE 2005).
7. Pevzner, P. A. (200). Computational Molecular Biology: An Algorithmic Approach The MIT Press.
8. Soinov, L. (2006). Bioinformatics and Pattern Recognition Come Together. Journal of Pattern Recognition Research(JPRR), Vol 1 (1) p.37-41