Application of Data mining in the Field of Bioinformatics

[1]B.Vinothini, [2]D.Shobana and [3]P.Nithyakumari
[1,3]Scholar, [2]Assignment Professor,
[1,2,3]Department of Information and Technology, Sri Krishna College of Arts and Science,
Coimbatore, TamilNadu, India

Abstract: This paper elucidates the application of data mining in bioinformatics. It also highlights the databases of bioinformatics. Biological data analysis and the link between data mining and bioinformatics is explained.

Keywords: Data mining, Bioinformatics, PDB, SWISS-PROT, MEDline, EMBL.

I. INTRODUCTION TO BIOINFORMATICS

Bioinformatics deals with biology and biological data. Bioinformatics, or computational biology, is the interdisciplinary science of interpreting biological data using information technology and computer science. The importance of this new field of inquiry will grow as we continue to generate and integrate large quantities of genomic, proteomic, and other data. A particular active area of research in bioinformatics is the application and development of data mining techniques tological problems. Analyzing large biological data sets requires making sense of the data by inferring structure or generalizations from the data. Examples of this type of analysis include protein structure prediction, gene classification, cancer classification based on microarray data, clustering of gene expression data, statistical modeling of protein-protein interaction, etc. Therefore, we see a great potential to increase the interaction between data mining and bioinformatics [1].

Bioinformatics involves the manipulation, searching and data mining of DNA sequence data. The development of techniques to store and search DNA sequences [2] have led to widely applied advances in computer science, especially string searching algorithms, machine learning and database theory.

II. DATABASE OF BIOINFORMATICS

There are many rapidly growing databases in the field of Bio informatics.

A. Protein Data Bank

The PDB archive is a repository of atomic coordinates and other information describing proteins and other important biological macromolecules. Structural biologists use methods such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy to determine the location of each atom relative to each other in the molecule [3]. They then deposit this information, which is then annotated and publicly released into the archive by the PDB.

B. SWISS-PROT

SWISS-PROT [4] is an annotated protein sequence database, which was created at the Department of Medical Biochemistry of the University of Geneva and has been a collaborative effort of the Department and the European Molecular Biology Laboratory (EMBL), since 1987.

The SWISS-PROT protein sequence database consists of sequence entries. Sequence entries are composed of different line types, each with their own format. For standardization purposes the format of SWISS-PROT follows as closely as possible that of the EMBL Nucleotide Sequence Database.

The SWISS-PROT database distinguishes itself from other protein sequence databases by three distinct criteria:

(i) Annotations,
(ii) Minimal redundancy and
(iii) Integration with other databases.

C. MEDLINE

Medical Literature Analysis and Retrieval System Online, or MEDLARS Online is a bibliographic database of life sciences and biomedical information. It includes bibliographic information for articles from academic Journals covering medicine, nursing, pharmacy, dentistry, veterinary medicine, and health care. MEDLINE also covers much of the literature in biology and biochemistry, as well as fields such as molecular evolution.

D. The EMBL Nucleotide Sequence Database

The EMBL Database collects, organizes and distributes a database of nucleotide sequence data and related biological information. Since 1982 this work has been done in collaboration with GenBank (NCBI, Bethesda, USA) and the DNA Database of Japan (Mishima)[5]. Each of the three international collaborating databases DDBJ/EMBL/GenBank, collect a portion of the total sequence data reported world-wide. All new and updated database entries are exchanged between the International Nucleotide Sequence Collaboration on a daily basis. EMBL Database releases are produced quarterly and are distributed on CD-ROM. The most up-to-date data collection is available via Internet and World Wide Web interface.

III. DATA MINING

Data mining ,also popularly referred to as knowledge discovery from data (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large data bases, data warehouses , the web, other massive information repositories , or data streams.[6]

Data mining refers to extracting or “mining” knowledge from large amounts of data. Alternatively, others view data mining as simply as essential step in the process of knowledge discovery.

In [7], the following definition is given: Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules. Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table. [8]

IV. NEED FOR DATA MINING IN BIOINFORMATICS

The entire human genome, the complete set of genetic information within each human cell has now been determined. Understanding these genetic instructions promises to allow scientists to better understand the nature of diseases and their cures, to identify the mechanisms underlying biological processes such as growth and ageing and to clearly track our evolution and its relationship with other species. The key obstacle lying between investigators and the knowledge they seek is the sheer volume of data available. This is evident from the following figure which shows the rapid increase in the number of base pairs and DNA sequences in the repository of GenBank.

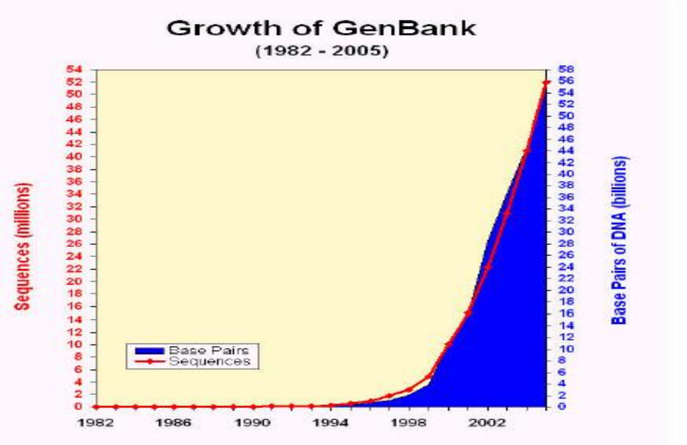

Figure 1: Growth of Genbank

Biologists, like most natural scientists, are trained primarily to gather new information. Until recently, biology lacked the tools to analyze massive repositories of information such as the human genome database. Luckily, the discipline of computer science has been developing methods and approaches well suited to help biologists manage and analyze the incredible amounts of data that promise to profoundly improve the human condition. Data mining is one such technology.[10]

V. BIOLOGICAL DATA ANALYSIS

Biological data mining is a very important part of Bioinformatics.

Following are the aspects in which data mining contributes for biological data analysis −

- Semantic integration of heterogeneous, distributed genomic and proteomic databases.
- Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.
- Discovery of structural patterns and analysis of genetic networks and protein pathways.
- Association and path analysis.
- Visualization tools in genetic data analysis.

VI. APPLICATIONS OF DATA MINING IN BIOINFORMATICS

Applications of data mining to bioinformatics include

1. Gene finding
2. protein function domain detection,
3. Function motif detection,
4. Protein function inference
5. disease diagnosis,
6. Disease prognosis,
7. Disease treatment optimization,
8. Protein and gene interaction network
9. Reconstruction,
10. Data cleansing, and
11. Protein sub-cellular location prediction.
12. Analysis of protein and dna sequences [9]

CONCLUSION

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, 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. Bioinformatics are fast growing research area today. It is important to examine what are the important research issues in bioinformatics and develop new data mining methods for effective analysis.

References

1. http://arxiv.org/ftp/arxiv/papers/1205/1205.1125.pdf
2. Bergeron, Bryan. Bioinformatics Computing. New Delhi: Pearson Education, 2003..
3. .http://www.rcsb.org/pdb/101/static101.do?p=education_discussion/Looking-at-Structures/intro.html
4. Bairoch A. and Apweiler, R. (1999) Nucleic Acids Res., 27, 49–54. [PMC free article] [PubMed]
5. .http://nar.oxfordjournals.org/content/26/1/8.full
6. Han, J. &Kamber, M. (2012). Data Mining: Concepts and Techniques. 3rd.ed. Boston: Morgan Kaufmann Publishers.
7. Hwang, H.G., Chang, I.C., Xingquan Zhu, Ian Davidson, "Knowledge Discovery and Data Mining: Challenges and Realities", ISBN 978- 1-59904-252, Hershey, New York, 2007.
8. Joseph, Zernik, "Data Mining as a Civic Duty – Online http://www.anderson.ucla.edu/faculty/ jason.frand/teacher/technologies/palace/datamining.htm
9. .http://arxiv.org/ftp/arxiv/papers/1205/1205.1125.pdf
10. http://ethesis.nitrkl.ac.in/4154/1/Application_of_Data_Mining_Techniques_in_Bioinformatics.pdf