RESEARCH PAPER | Biotechnology | Volume : 4 | Issue : 5 | May 2014 | ISSN - 2249-555X

Role of Data Mining in Bioinformatics: A Review

KEYWORDS: Bioinformatics, Data mining, Microarrays, Soft computing techniques.

Ashish Shrivastava
Faculty of Biotechnology, Department of Biotechnology, C.S.A. Govt. P.G. Nodal College, Sehore, 466001(M.P), India.

Mickey Sahu
Faculty of Computer Science, Department of Computer Application, C.S.A. Govt. P.G. Nodal College, Sehore, 466001(M.P), India.

ABSTRACT *The field of computer applications in bioinformatics is a complex fusion of two distinct scientific disciplines - computer technology and bioscience. Bioinformatics is the science of managing, mining, and interpret information from biological sequences and structures. Advances such as genome-sequencing initiatives, microarrays, proteomics, and functional and structural genomics have pushed the frontiers of human acquaintance. This article highlights some of the basic concepts of bioinformatics and data mining. It also highlights some of the current challenges and opportunities of data mining in bioinformatics.*

INTRODUCTION

In recent years, rapid developments in genomics and proteomics have generated a large amount of biological data. Drawing conclusions from these data requires difficult computational analyses. Bioinformatics, or computational biology, is the interdisciplinary science of interpreting biological data using information technology and computer science. The meaning of this new field of inquiry will grow as we continue to generate and integrate large quantities of genomic, proteomic, and other data [8]. A scrupulous active area of research in bioinformatics is the application and development of data mining techniques to solve biological problems. Analyzing large biological data sets requires making sense of the data by inferring structure or generalizations from the data. Therefore, we see a great potential to increase the interaction between data mining and bioinformatics.

BIOINFORMATICS

The term bioinformatics was coined by Paulien Hogeweg in 1979 for the study of informatics processes in biotic systems. It was most important used since late 1980s 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 involves the manipulation, searching and data mining of DNA sequence data. The development of techniques to store and search DNA sequences [5] have led to widely useful advances in computer science, especially string searching algorithms, machine learning and database theory. In other applications such as text editors, even simple algorithms for this problem usually suffice, but DNA sequences because these algorithms to exhibit near-worst case behavior due to their small number of distinct characters. Data sets representing entire genomes' worth of DNA sequences, such as those produced by the Human Genome Project [4], are difficult to use without observations, which label the locations of genes and regulatory elements on each chromosome. Regions of DNA sequence that have the characteristic patterns associated with protein or RNA coding genes can be identified by gene finding algorithms [11], which allow researchers to predict the presence of particular gene products in an organism even before they have been inaccessible experimentally. The primary goal of bioinformatics is to increase the understanding of biological processes.

Some of the impressive area of research in bioinformatics includes:

- Analysis of gene expression
- Analysis of mutations in cancer
- Analysis of protein expression
- Comparative genomics
- Genome annotation
- High-throughput image analysis
- Modeling biological systems
- Protein-protein docking
- Protein structure prediction
- Sequence analysis

DATA MINING

Researchers identify two fundamental goals of data mining: prediction and description. Prediction makes use of presented variables in the database in order to predict future values of interest and description focuses on finding patterns describing the data and the subsequent presentation for user explanation [1]. The relative importance of both prediction and description differ with respect to primary application and the technique. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and explanation of data and for the extraction of interesting knowledge that could help in decision-making [2]. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD) [5], refers to as "the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable pattern in data".

The following points shows data mining as a step in an iterative knowledge discovery process.

1. **Data cleaning:** Also known as data cleansing, it is a phase in which raucous data and irrelevant data are removed from the collection.
2. **Data integration:** At this stage, multiple data sources, often heterogeneous, may be combined in a general source.
3. **Data selection:** At this step, the data relevant to the analysis is decided on and retrieved from the data collection.
4. **Data mining:** It is the crucial step in which intelligent techniques are applied to extract data patterns potentially useful.
5. **Pattern evaluation:** In this step, strictly interesting patterns representing knowledge are acknowledged based on given measures.
6. **Knowledge representation:** It is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses mental picture techniques to help users understand and interpret the data mining results.

Volume : 4 | Issue : 5 | May 2014 | ISSN - 2249-555X

The KDD is an iterative process. Once the discovered knowledge is presented to the user, the evaluation measures can be enhanced, the mining can be further polished new data can be selected or further transformed, or new data sources can be integrated, in order to get different, more appropriate results.

SOFT COMPUTING TECHNIQUES

Soft Computing refers to a collection of computational techniques in computer science, artificial intelligence, machine learning and some engineering disciplines, which attempt to study, model and analyze very complex phenomena [10]. In advance computational approaches could model and precisely analyze only relatively simple systems. More complex systems arising in biology, medicine, the humanities, management sciences, and similar fields often remained intractable to conventional mathematical and analytical methods. This is where soft computing provides the solution. Key areas of soft computing [6] include the following:

Fuzzy logic

Fuzzy logic is derived from fuzzy set theory dealing with reasoning that is approximate rather than exactly deduced from classical predicate logic. It can be thought of as the application side of fuzzy set theory advertising with well thought out real world authority values for a complex problem. Fuzzy logic can be used to control household appliances such as washing machines (which sense load size and detergent concentration and adjust their wash cycles accordingly) and refrigerators.

Neural Networks

Artificial neural networks are computer models built to emulate the human pattern recognition function through a similar parallel processing structure of multiple inputs. A neural network consists of a set of essential processing elements (also called neurons) that are distributed in a few hierarchical layers. Most neural networks contain three types of layers: input, hidden, and output. After each neuron in a hidden layer receives the inputs from all of the neurons in a layer ahead of it (typically an input layer), the values are added through applied weights and converted to an output value by an activation function (e.g., the Sigmoid function). Then, the output is passed to all of the neurons in the next layer, providing a feed forward path to the output layer.

Statistical Inferences

Statistics provides a solid theoretical foundation for the problem of data analysis. Through hypothesis validation and/or investigative data analysis, statistical techniques give asymptotic results that can be used to describe the likelihood in large samples. The basic statistical exploratory methods include such techniques as examining distribution of variables, reviewing large correlation matrices for coefficients that meet certain thresholds, and examining multidimensional frequency tables.

Rule induction

Induction models belong to the logical, pattern decontamination based approaches of data mining. Based on data sets, these techniques produce a set of if-then rules to represent significant patterns and create prediction models. Such models are fully transparent and provide complete explanations of their predictions.

Genetic Algorithm

They are search algorithms based on the mechanics of ordinary genetics i.e. operations existing in nature. They combine a Darwinian 'survival of the fittest' approach with a structured, yet randomized, information exchange. The advantage [9] is that they can search complex and large amount of spaces efficiently and locate near optimal solutions rapidly.

CONCLUSION

In this report an in-depth study of the different bioinformatics and data mining was made. Bioinformatics and data mining are developing as interdisciplinary science. Data mining approaches seem ideally suited for bioinformatics, since bioinformatics is data-rich but lacks a broad theory of life's organization at the molecular level [6]. However, data mining in bioinformatics is held back 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. 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 data mining methods for scalable and successful investigation. Data mining is therefore a helpful technique to solve the problem of enormous data faced by researchers in their quest to solve the puzzles of our life.

REFERENCE

[1] Aluru, S., ed. (2006). Handbook of Computational Molecular Biology. Chapman & Hall/Crc. | [2] Ashish Shrivastava and Mickey Sahu (2013), Computer abetted drug design: Application of drug target like protein prediction using learning algorithms of support vector machine, International organization of scientific research (IOSR), Journal of pharmacy and biological science. | [3] Bergeron, Bryan (2013). Bioinformatics Computing. New Delhi: Pearson Education. | [4] Gilbert, D. (2004). Bioinformatics software resources. Briefings in Bioinformatics, Briefings in Bioinformatics. | [5] Han, Jiawei and Kamber, Micheline (2010). Data mining: concepts and techniques, San | Francisco: Morgan Kaufmann Publishers. | [6] Han and Kamber (2007). Data Mining concepts and techniques, Morgan Kaufmann Publishers. | [7] Jiong, Lei Liu, Yang, A. and Tung, K. H. (2012). Data Mining Techniques for Microarray Datasets, Proceedings of the 21st International Conference on Data Engineering. | [8] Lee, Kyoungrim (2008). Computational Study for Protein-Protein Docking Using Global Optimization and Empirical Potentials, Int. J. Mol. Sci. | [9] Li, J., Wong, L. and Yang, Q. (2009). Data Mining in Bioinformatics, IEEE Intelligent System, IEEE Computer Society. | [10] Mickey Sahu and Ashish Shrivastava (2013), Mobile Learning-Telecommunication Infrastructure and Usage in Rural and Remote Areas Students: A Review, Indian journal of applied research. | [11] Pujari, Arun (2010) Data Mining Techniques. Nancy: Universities Press. | [12] Richard, R.J.A. and Sriraam (2005). A Feasibility Study of Challenges and Opportunities in Computational Biology: A Malaysian Perspective, American Journal of Applied Sciences. |