WORLD JOURNAL OF PHARMACY AND PHARMACEUTICAL SCIENCES
SJIF Impact Factor 7.421
Volume 8, Issue 1, 1443-1452 Review Article ISSN 2278 – 4357

STUDY OF MICROARRAY BIOLOGICAL DATA USING DATA MINING

Roma Chandra[1]*,[2]

[1]IILM College of Engineering & Technology, Greater Noida, Uttar Pradesh, India.
[2]Shri Venkateshwara University, Gajraula, Amroha, Uttar Pradesh.

Article Received on
23 Nov. 2018,
Revised on 13 Dec. 2018,
Accepted on 01 Jan. 2019,
DOI: 10.20959/wjpps20192-13012

***Corresponding Author**
Roma Chandra
IILM College of Engineering & Technology, Greater Noida,Uttar Pradesh, India.

ABSTRACT

Data mining is the process of information retrieval to extract meaningful information from very large data. It can also be defined as one of the knowledge discovery database processes. It is also used to extract biological data analyzing microarray data producing meaningful results. This review article briefs about the process of data mining as well as its relation with biological data. It includes mining of biological data and explains role of bioinformatics to study branches like genomics and proteomics.

INTRODUCTION

In 1980's, with the introduction of new technologies and an exponential growth in data, data storage and data management was the biggest challenge that led to the emergence of a new discipline, data mining. Traditional data analysis methods were inefficient to process large amounts of raw data with noisy data. Data mining on the other hand was based on knowledge extraction from large database converting raw data to normalized data. Data mining is thus said to be an interdisciplinary branch of research that works in coordination with areas that includes the study of machine learning, databases, statistical analysis, data retrieval, pattern recognition, microarray analysis, etc. Data mining can be defined as the process of discovering interesting knowledge from large amount of data stored in databases, data warehoused or any other repositories. The main aim of data mining is to study the databases applying mining tasks and discover meaningful, useful patterns and relationships in data. Data mining can also be defined as essential step of the KDD (knowledge discovery from data) process. It uses different mining tasks like for classification, regression, clustering, prediction or association rules.

The raw data that needs to be mined under goes the following steps of KDD process.

1. *Data Extraction & Collection*
2. *Data Cleaning*
3. *Data Integration*
4. *Data Selection*
5. *Data Transformation*
6. *Data Mining*
7. *Data Modeling*
8. *Pattern Discovery& Evaluation*
9. *Data Visualization & Knowledge Presentation*

Data Extraction & Collection

Data extraction and collection is done from various databases, data warehouses and data repositories. It is the first step for the process of data mining. Data mining is done based on specific information about the data so that the same can be searched in various databases and other repositories. Mining meaningful data from raw data is a complicated process. Data extraction using information or data retrieval systems from various databases is done and after collection of data further steps are followed. First sample is collected and based on the essential relationships revealed in the sample the complete data is mined from already prepared databases and data warehouses. Researchers usually depend on prior information based on knowledge of biology to identify essential relationships.

Data Cleaning

Data collected is raw data and not clean, containing errors, missing values, noise, outliers or inconsistent data. Different methods can be applied to get rid of such anomalies.Thus, we can say that once the data extracted & collected, it has to be preprocessed and cleaned which is done in the following steps.

Data characterization: It basically deals with documentation of data in a relevant and meaningful manner, so that it could be understandable and interpret easily. This task is basically done by program by creating a high-level description of the nature and the content of the data to be mined.

Consistency analysis: It analyzes the variability of data independent of domain based on statistical analysis of data. Outliers, missing values and other insignificant values are removed from the knowledge-discovery process based on predefined statistical conditions.

Domain analysis: It is process of analyzing as well as validating the data based on larger context of biology. It is something which goes beyond simply verifying that data value which is a text string or an integer, or that it's statistically consistent with other data on the same parameter, to ensure that it makes sense in the context of the biology. Domain analysis basically requires knowledge from experts from the field so that and create the heuristics in biology which can be applied.

Data enrichment: It involves improving the data obtained from various sources by minimizing the limitations of a single data source. Various data sources are improved by adding value to them which may include merging the best information from various different sources as well as moderating the information as per requirement.

Frequency and Distribution Analysis: It includes searching for various data mining processes adding up values on the basis of frequency of data occurrence. This is done to maximize the contribution of common findings while minimizing the effect of rare occurrences on the conclusions made from the data-mining output.

Data Integration

Data integration is an important step that involves integration of data from various sources.

Data Selection

Data selection refers to selection of data as a sample on which task is applied.

Data Transformation

The data even after cleaning are not ready for mining as we need to transform them into forms appropriate for mining. The technique used to transform data includes data normalization, aggregation and scaling up & down the data according to requirement. It involves data transformation from one representation to another based on various range of values. Various scales are used in normalization process like absolute scales, nominal scales, ordinal scales, rank scales. For example, qualitative values, such as "high" and "low," and qualitative values from multiple sources regarding a particular parameter might be normalized to a numerical score from 1 to 10.

Data Mining

Now we are ready to apply data mining techniques on the data to discover the interesting patterns. The process of data mining is concerned with extracting relevant patterns from the data. Techniques like clustering, classification, regression and association analysis are among the different techniques used for data mining.

This is not a single method or approach but it includes various technology and techniques which are used for mining of wide range of biological data. Machine learning methods includes various data mining algorithms covering branches like statistics ,artificial intelligence, biological modeling, adaptive control theory and psychology as well few famous techniques like genetic algorithm and neural networks which are used for mining biological data.

Data Modeling

Data modeling is basically a process that structures and organizes the data according to the recent biological world scenario. Managing data and then modeling is done to represent the data in a meaningful manner exploiting various forms of data in a meaningful manner. The data present is taken from various formats which are stored and retrieved from various resources based on the various branches.

Pattern Discovery & Evaluation

The branch of biology had transformed all these years producing massive data and huge databases for DNA, RNA and proteins. These databases include information as important biological insights represented as conserved patterns, motifs and other useful information for various species. Pattern matching helps in finding conserved regions and pattern discovery helps in categorizing useful biological information into classes. For particular biological applications, even the definition of a relevant pattern may be difficult to state clearly, or may be unresolved. In bioinformatics, pattern recognition classifies character sequences which are representatives of nucleotide bases, molecular structures and three dimensional structures of proteins.

Data Visualization & Knowledge Presentation

Visualizing biological data is one of the most challenging parts of data mining process. In this modern digital society, how the data is visualized becomes the prime facto, when it comes to communicating or understanding complex concepts. Better the data visualized, better the

concepts will be clear. Visualization technologies can provide an intuitive representation of the relationships among large groups of objects or data points that could otherwise be incomprehensible, while providing context and indications of relative importance. The "sequence visualization" and "structure visualization" are types of data visualization techniques.

Mining biological data

Houle et al. (2000) had classification & given three successive levels for analysis of biological data,[11] It was identified on the basis of the central dogma of molecular biology that.

1. Genomics is the study of genome of an organism and use of its genome information to produce new biological knowledge.

2. Gene expression analysis deal with measurements of expression of gene in term of its mRNA expression analysis that characterizes biological processes and help in mechanisms of gene transcription.

3. Proteomics is the study of proteins, their structures and functions. Like with the Human Genome Project study focused more on sequence & map genomes which bought a major shift from structural genomics to dynamic functional genomics.[11] The term structural genomics refers to the DNA sequencing and mapping activities, while functional genomics refers to study of functional information to known sequences. There is study of specific DNA sequences with specific biological role is a problem that concerns bioinformatics scientists. Where transcription (the process of mRNA production from DNA) starts, another biologically meaningful sequence is the translation initiation site, which is the site where translation (protein production from mRNA) initiates. Although every cell in an organism with only few exceptions- has the same set of chromosomes, two cells may have very different properties and functions. This is due to the differences in abundance of proteins. The protein information is determined by amount of mRNA which is determined by the expression or non-expression of the corresponding gene. Microarray technology is used for gene expression analysis. A microarray experiment measures the relative mRNA levels of around thousands of genes, which has ability to compare the expression levels of different biological samples. These samples may correlate during a biological process or with different tissue types such as normal cells and cancer cells.

Serial Analysis of Gene Expression (SAGE) is a method that allows the quantitative profiling of a large number of transcripts. A transcript is a sequence of mRNA produced by transcription. It is a very expensive method in contrast to microarrays, thus there is a limited amount of publicly available SAGE data. Proteomics is the prediction and study of various properties related ton proteins that include factors like active sites, modification sites, stability, shape, localization, protein domains, its secondary structure and possible interactions. It also include study of interaction of one protein to another.

Mining biological data

Data mining is the discovery of useful knowledge from databases. It is the main step in the process known as Knowledge Discovery in Databases (KDD), although the two terms are often used interchangeably. Other steps of the KDD process are the collection, selection, and transformation of the data and the visualization and evaluation of the extracted knowledge.[1] Data mining employs algorithms and techniques from statistics, machine learning, artificial intelligence, databases and data warehousing etc. Some of the most popular tasks are classification, clustering, association and sequence analysis, and regression. All these mining tasks utilizes algorithms and produces knowledge as per need constructive models which are either predictive or descriptive. A predictive model makes a prediction about data using known examples, while a descriptive model identifies patterns or relationships in data.[12]

Many general data mining systems such as SAS Enterprise Miner, SPSS, S-Plus, IBM Intelligent Miner, Microsoft SQL Server 2000, SGI MineSet, and Inxight VizServer can be used for biological data mining. There are few specific data mining tools for analyzing biological data like Gene Spring, Spot Fire, Vector NTI, COMPASS, Statistics for Microarray Analysis and Affymetrix Data Mining Tool. National Center for Biotechnology Information and European Bioinformatics Institute are few sources with freely available mining tools for analyzing biological data.

Data Mining in Genomics

The most common include neural networks, Bayesian classifiers, decision trees, and Support Vector Machines (SVMs).[3] Sequence recognition algorithms exhibit performance tradeoffs between increasing sensitivity (ability to detect true positives) and decreasing selectivity (ability to exclude false positives).[11] To solve this type of recognition problems the traditional methods of data mining are not useful and thus to overcome such problems methods like feature generation and feature selection are adapted. Besides these techniques,

applications like clustering algorithms are also at genomic level to group structurally related DNA sequences.

Gene Expression Data Mining

Microarray data analysis can be done by methods like selection, clustering and classification. The main types of microarray data analysis[3] include: gene selection, clustering, and classification. Microarray datasets in contrast with other application domains contain a small number of records (less than a hundred), while the number of fields (genes)is typically in thousands. The same case is in SAGE data. This increases the likelihood of finding "false positives". Selection of features is an issue under data analysis. In case of gene expression analysis the features selected are the genes which are selected by gene selection process finding genes which are strongly related to a particular class. The benefit of this process is dimensionality reduction of datasets. Classification causes large reduction of gene datasets but clustering is the most used method for analyzing genes. Clustering method applied is either one- way clustering or a two-way clustering. One way clustering is used to categorize genes with similar function where as two way clustering categorizes gene clusters and samples. The famous of clustering methods includes hierarchical clustering methods which are best used in gene expression analysis. In microarray analysis classification is applied to discriminate diseases or to predict outcomes based on gene expression patterns and perhaps even identify the best treatment for given genetic signature.

Data Mining in Proteomics

Many modification sites can be detected by simply scanning a database that contains known modification sites. However, in some cases, a simple database scan is not effective. The use of neural networks provides better results in these cases. Similar approaches are used for the prediction of active sites. Neural network approaches and nearest neighbor classifiers have been used to deal with protein localization prediction.[3] Neural networks have also been used to predict protein properties such as stability, globularity and shape. Whishart refers to the use of hierarchical clustering algorithms for predicting protein domains. There are many techniques used for the prediction of secondary structure of proteins including the task of data mining. Earlier statistical methods were used to deal with such prediction problems but later more accurate methods such as information theory, bayes theory, nearest neighbors, hidden markov models, self optimization methods and neural networks were developed and used. Combined methods such as integrated multiple sequence alignments with neural network or

nearest neighbor approaches improve prediction accuracy. A density based clustering algorithm (GDBSCAN) is presented by Sander et al. (1998) that can be used to deal with protein interactions. This algorithm is able to cluster point and spatial objects according to both, their spatial and non-spatial attributes.

Databases of bioinformatics

There are many rapidly growing databases in the field of Bioinformatics.[11]

S.no.	Databases and Retrieval Systems	Brief Summary of Content
01.	DDBJ	Primary nucleotide sequence database in Japan
02.	EMBL	Primary nucleotide sequence database in Europe
03.	AceDB	Genome database for *Caenorhabditiselegans*
04.	Entrez	NCBI portal for a variety of biological databases
05.	ExPASY	Portal system
06.	FlyBase	A database of the *Drosophila* genome
07.	FSSP	Protein secondary structures
08.	GenBank	Primary nucleotide sequence database in NCBI
09.	HIV databases	HIV sequence data and related immunological information
10.	Microarray Gene expression database	DNA microarray data and analysis tools
11.	OMIM	Genetic information of human diseases
12.	PIR	Annotated protein sequences
13.	PubMed	Biomedical literatureInformation
14.	Ribosomal database project	Ribosomal RNA sequences and phylogenetic trees derived from the sequences
15.	SRS	General sequence retrieval system
16.	SWISS-Prot	Curated protein sequence database
17.	TAIR	Arabidopsis information database

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 of 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.

Applications of data mining in bioinformatics

- *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*
- *Protein sub-cellular location prediction*
- *Analysis of protein and nucleotides sequences.*

REFERENCES

1. M. Andrade and P. Bork. Automated extraction of Information in molecular biology, *FEBS Letters*, 2000; 476.
2. V.Saranya. Bio-informatic Using Data Mining Technique, *International Journal of Advance Research in Compuer Science And Management Studies*, 2015; 3(10).
3. P. Baldi and S. Brunak. Bioinformatics: The Machine Learning Approach, Second Edition. A Bradford Book, *MIT Press*, 2001; 7-16.
4. Dr.S.Vijayaraniand Ms. S.Deepa. Protien Sequences Classification In Data mining - A Study. *IJITMC*, 2014; 2(2).
5. Mao JH and Weier HUG. Bioinformatics Applications in Biological and Clinical Studies, *J Data Mining Genomics Proteomics*, 2014; 5: 1.
6. B.Vinothini, D.Shobana and P.Nithyakumari. Application of Data mining in the Field of Bioinformatics, *International Journal of Trend in Research and Development*, 2016; 3(1).
7. Prof. Khushboo Satpute, Prof. Sapna V M .Data Mining in Bioinformatics: Study & Surveyof Data Mining and its Operations in Mining Biological Data. *International Journal of Electronics, Communication & Soft Computing Science and Engineering*, 2014; 2(9).

8. Stefano Lonardi and Jake Chen. Data Mining in Bioinformatics: Selected Papers from BIOKDD. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2010; 7(2).
9. JIN XIONG Texas A&M University .*Essential Bioinformatics*. Cambridge University press, 2006; 15-16.
10. Prof. Sapna V M, Prof. KhushbooSatpute Data Mining in Bioinformatics: Study & Survey of Data Mining and its Operations in Mining Biological data.
11. Houle, J.L., Cadigan, W., Henry, S., Pinnamaneni, A. and Lundahl, S. Database Mining in the Human Genome Initiative. Whitepaper, Biodatabases.com, Amita Corporation, 2004; 1(9).
12. Dunham, M.H.Data mining: Introductory and advanced topics. Prentice Hall, Upper Saddle River, New Jersey, USA, 2002.
13. Sander, J., Ester, M., Kriegel, P.-H. andXu, X.. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications. Data Mining and Knowledge Discovery, 1998; 2(2): 169-194.
14. Introduction to Data Mining and Knowledge Discovery (3rd ed). Two Crows Corporation, 1999.
15. Luscombe, N.M., Greenbaum, D. and Gerstein, M.What is Bioinformatics? A Proposed Definition and Overview of the Field. Methods of Information in Medicine, 2001; 40(4): 346-358.