International Journal of Electronics, Communication & Soft Computing Science and Engineering
ISSN: 2277-9477, Volume 2, Issue 9

# Data Mining in Bioinformatics: Study & Survey of Data Mining and its Operations in Mining Biological Data

**Prof. Sapna V M** **Prof. Khushboo Satpute**

*Abstract*: - **Like any other data, biological data is a very vast one. Due to emergence of system biology it is necessary to develop various platforms and techniques to analyze and organize the biological data in meaning full manner for which it to be mined and processed carefully. As the complexity associated with biological data is high ,it has to be studied considering various criteria's and also it is mandatory to study all available databases and then has to undergo several processing mining techniques to finally put in a format which is easy to assess and produce the information of interest. There are various techniques and method for mining biological data. Here we will put forth all possible techniques and operations involved in data mining and will compare them in order to find the advantages and disadvantages of different methods**

## I. INTRODUCTION

Biological data as said is a vast one it includes the information right from DNA to RNA to protein. This includes data from Human genome project-which is genomic sequences,
data from microarray experiments which is gene expression, data from proteomics experiments which is protein identification and quantification, data from high - throughput SNP arrays which is SNP data .of the biological processes underlying these data lags far behind. Thus this shows that in past few years enormous amount of biological data has been accumulated as a result of which mining i,e data mining in bioinformatics takes a greater challenge . There is a strong interest in employing methods of knowledge discovery and data mining to generate models of biological systems. Mining biological databases imposes challenges which knowledge discovery and data mining have to address.
Due to fast growth in biotechnology and biodata analysis.Methods have led to the emergence of a promising new field: Bioinformatics. [1] On the other hand, in past few years progress in data mining researches led to the development of numerous efficient and scalable methods For mining biological data i,e interesting patterns and knowledge in large databases, ranging.From efficient classification methods to clustering, outlier analysis, frequent, Sequential, and structured pattern analysis methods, and visualization and spatial/temporal data analysis tools.

## II. DATA MINING

Data mining is defined as the process of automatically extracting meaningful patterns from usually very large quantities of seemingly unrelated data. It is an alternative to manual searching which is time-consuming and a very cumbersome. Data mining has had considerable success in various fields and environment. Data mining isn't an endpoint, but is one stage in an overall knowledge-discovery process. [1] It is an iterative process in which preceding processes are modified to support new hypotheses suggested by the data. The main aim of data mining is to explore the databases through automated means and discover meaningful, useful patterns and relationships in data. Data mining can be defined as one particular step of the KDD process: the identification of interesting structures in data. It uses different algorithms for classification, regression, clustering or association rules.

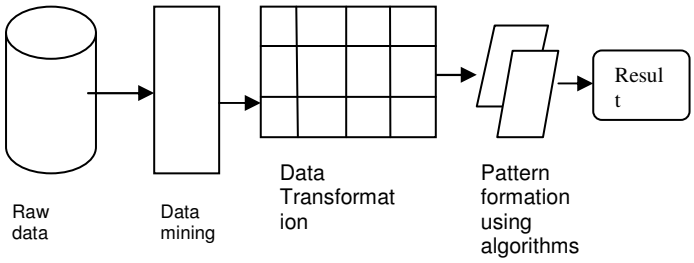


*Fig 1: Operations in data mining*

## III. OPERATIONS IN DATA MINING

*Selection/sampling*

Data mining is conducted against data accumulated in OLTP repositories, data warehouses, data marts and archived data. The steps for data mining follow the following pattern:

- **Data Extraction**
- **Data Cleansing**
- **Data Transformation /Reduction**

- **Data Mining Methods**
- **Applying Data Mining Algorithm**
- **Modeling Data**
- **Pattern Discovery**
- **Data Visualization**

### *A. DATA EXTRACTION*

Data selection and sampling from extracted data by data warehouses, databases data marts oltp repositories is a first challenging step in data mining. [2]Data mining requires a controlled vocabulary, usually implemented as part of a data dictionary, so that a single word can be used to express a given concept.
As millions and thousands of records and variables are gathered in data warehouses and data bases initial mining of meaningful data is quite a complicated process. Typically restricted to computationally tenable samples of the holding in an entire data warehouse. The evaluation of the relationships that are revealed in these samples can be used to determine which relationships in the data should be mined further using the complete data warehouse. With large, complex databases, even with sampling, the computational resource requirements associated with non-directed data mining may be excessive. In this situation, researchers generally rely on their knowledge of biology to identify potentially valuable relationships and they limit sampling based on these heuristics.

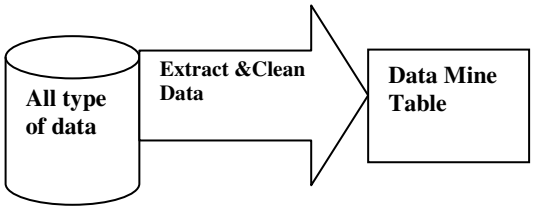


***Fig 2: Data extraction***

### *B. DATA CLEANSING*

The data collected are not clean and may contain errors, missing values, noisy or inconsistent data. So we need to apply different techniques to get rid of such anomalies. Once the is extracted it has t be preprocessed and cleaned. This is done in following steps:

***Data Characterization:*** It basically deals with documentation of data in an appropriate and meaningful manner ,so that any person could understand and interpret the data comfortabaly.This task s basically done by programmers and other staff involved in data mining project it involves creating a high-level description of the nature and the content of the data to be mined.

***Consistency Analysis:*** It is analyzing the variability of data independent of domain. Based on data values, it is primarily statistical analysis of data. Outliers and values determined to be significantly different from other data may be automatically excluded from the knowledge-discovery process, based on predefined statistical constraints.
For example, data associated with a given parameter that is more than three standard
Deviations from the mean might be excluded from the mining operation.

***Domain Analysis:*** It is validating the data values in a larger context of biology. It is something which goes beyond simply verifying that data value 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 requires that someone familiar with the biology create the heuristics that can be applied to the data. Data enrichment: involves strengthening of data from multiple data sources to minimize the limitations of a single data source. It basically involves studying various sources of data For example; two databases on inherited diseases might each be sparsely populated in terms of proteins that are associated with particular diseases. This deficit could be addressed by incorporating data from both databases, assuming only a moderate degree of overlap in the content of the two databases. [3][4]Frequency and Distribution Analysis: It finds the frequency of occurrence of data during the data mining process by placing the weights on values as a function of their frequency of 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.

### *C. DATA TRANSFORMATION*

The data even after cleaning are not ready for mining as we need to transform them into forms appropriate for mining. The techniques used to accomplish this are smoothing, aggregation, normalization etc.
Normalization: It represents the data in various forms depending on analysis and based on further processes to be implemented. It involves transforming data values from one representation to another, using a predefined range of final 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.

***Missing Value Analysis:*** The final preprocessing and cleaning activity, missing-value analysis, involves detecting,

characterizing, and dealing with missing data values. One way of dealing with missing data values is to substitute the mean, mode, or median value of the relevant data that are available.

### *D. 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 patterns from the data,

Techniques like clustering and association analysis are among the many different techniques used for data mining.

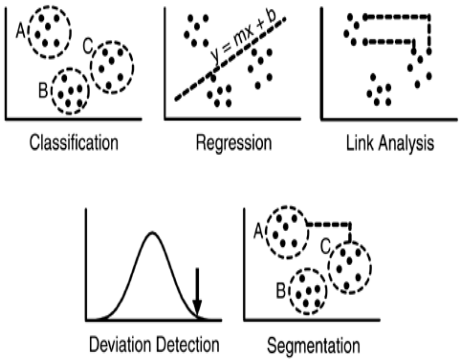


*Fig 3. Data mining techniques*

### *E. APPLYING DATA MINING ALGORITHM*

is not a single method or approach, but ii converges various technology and techniques to achieve proper mining of wide range of and also the data of interest biological data. Machine learning methods have wide applicability in data mining algorithms. It It includes statistics, biological modeling, adaptive control theory, psychology, and artificial intelligence (AI). [3][4] Basically genetic algorithm and neural networks take a major part as a technique to in biological data. Similarly, adaptive control theory, where parameters of System change dynamically to meet the current conditions, and psychological theories, especially those regarding positive and negative reinforcement learning, heavily influence machine learning methods. Artificial Intelligence techniques, such as pattern matching through inductive logic programming, are designed to derive general rules from specific examples.

**Table 1: Machine Learning Technologies and Their Applicability to Data-Mining Methods.**

| Machine Learning Technologies | Data Mining Methods | | | | |
|---|---|---|---|---|---|
| | Classification | Regression | Segmentation | Link Analysis | Derivation& Deviation |
| Inductive Logic Programming | X | X | | | |
| Genetic Algorithms | X | X | X | | |
| Neural Networks | X | X | X | | |
| Statistical Method | X | X | X | X | X |
| Decision Trees | X | | X | | |
| Hidden Markov Model | X | | | | |

### *F. DATA MODELING*

Data modeling basically is a process of structuring and organizing the data, and then these structured data are implemented in database management system. Today's biological world demands for heavy exploitation of data These data as are in various forms which has to be capsulated in a meaning full manner .The data are in disparate formats, remotely dispersed, and based on the different vocabularies of Various disciplines. Furthermore, data are often stored or distributed using formats that leave implicit many important features relating to the structure and semantics of the data. Conceptual data modeling involves the development of implementation-independent models that capture and make explicit the principal structural properties of data. Entities such as a biopolymer or a reaction, and their relations, egg catalyses can be formalized using a conceptual data model. Conceptual models are implementation-independent and can be transformed in systematic ways for implementation using different platforms, eg traditional database management systems.

Data modeling is the formalization and documentation of existing processes and events that occur during application software design and development. Data modeling techniques and tools capture and translate complex system designs into easily understood representations of the data flows and processes, creating a blueprint for construction and/or re-engineering.

Data Model:

Managing large quantities of structured and unstructured data is a primary function of information systems. Data models describe structured data for storage in data management systems such as relational databases. They typically do not describe unstructured data, such as word processing documents, email messages, pictures, digital audio, and video. Early phases of many software development projects emphasize the design of a conceptual data model. Such a design can be detailed into a logical data model. In later stages, this model may be translated into physical data model.

A data model describes the structure of the data within a given domain and, by implication, the underlying structure of that domain itself. This means that a data model in fact specifies a dedicated 'grammar' for a dedicated artificial language for that domain. Because there is little standardization of data models, every data model is different. [5][6] This means that data that is structured according to one data model is difficult to integrate with data that is structured according to another data model. A data model may represent classes of entities (kinds of things) about which a company wishes to hold information, the attributes of that information, and relationships among those entities and (often implicit) relationships among those attributes. The model describes the organization of the data to some extent irrespective of how data might be represented in a computer system.

The entities represented by a data model can be the tangible entities, but models that include such concrete entity classes tend to change over time. Robust data models often identify abstractions of such entities. For example, a data model might include an entity class called "Person", representing all the people who interact with an organization. Such an abstract entity class is typically more appropriate than ones called "Vendor" or "Employee", which identify specific roles played by those people.

A proper conceptual data model describes the semantics of a subject area. It is a collection of assertions about the nature of the information that is used by one or more organizations. Proper entity classes are named with natural language words instead of technical jargon. Likewise, properly named relationships form concrete assertions about the subject area. For example, a relationship called "is composed of" that is defined to operate on entity classes "Order" and "Line item" forms the following concrete assertion definition: Each "Order" "is composed of" one or more "Line items." Note that this illustrates that often generic terms, such as "is composed of", are defined to be limited in their use for a relationship between specific kinds of things, such as an order and an order line. This constraint is eliminated in the generic data modeling methodologies.

***Data Organization:***

Another kind of data model describes how to organize data using a database management system or other data management technology. It describes, for example, relational tables and columns or object-oriented classes and attributes. Such a data model is sometimes referred to as the physical data model, but in the original ANSI three schema architecture, it is called "logical". In that architecture, the physical model describes the storage media (cylinders, tracks, and table spaces). Ideally, this model is derived from the more conceptual data model described above. It may differ, however, to account for constraints like processing capacity and usage patterns.

While data analysis is a common term for data modeling, the activity actually has more in common with the ideas and methods of synthesis (inferring general concepts from particular instances) than it does with analysis (identifying component concepts from more general ones). Data modeling strives to bring the data structures of interest together into a cohesive, inseparable, whole by eliminating unnecessary data redundancies and by relating data structures with relationships.

### *G. PATTERN DISCOVERY*

Biology has been transformed from a data poor to a data rich field, with massive accumulation of disparate types of data, for example huge databases of sequences (DNA, RNA, or protein). This data allows important biological insights to be made, partly by finding patterns and motifs that are conserved across many individuals or species; there is now a huge biological literature reporting on such conserved patterns and motifs that have been found in biological datasets. In contrast to the area of pattern matching, the patterns and motifs are generally not known ahead of time, but must be identified or discovered from the data; this task is often very subtle and difficult because the patterns and motifs may be short, may be highly degenerate (containing wildcards and variable length elements), may be ordered differently in different genomes, and are generally hidden in that they make up a small fraction of the data. 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 is most often concerned with the automatic classification of character sequences representative of the nucleotide bases or molecular structures,and of 3D protein structures.

The following diagram gives the gist of pattern recognition and discovery process.

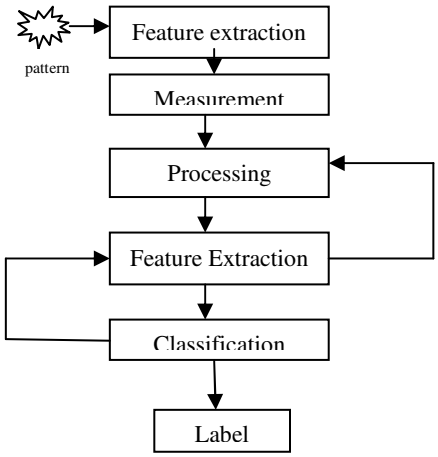


*Fig 4: pattern recognition and discovery process*

### *H. DATA VISUALIZATION*

Visualizing biological data is one of the most challenging part 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. [5] [6] 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.

**Sequence Visualization**

It deals with analyzing the nucleotide sequence represented in various forms. Many drastic changes took place in programming the biological data right form machine code type where in the sequence was represented in terms of 0s and 1s,but this encountered many problems like heavy time consumption, errors which forced to shift to next level of programming called assembly level programming , which allows programmers to use mnemonics such as "CLR" to clear a buffer and "ADD" to add two values. To still go ahead next level would be using programming languages such as C++, BASIC, and HTML that insulate programmers from the underlying computational hardware infrastructure and allow them to work at a level nearer the application purpose. Still higher level would be the flow diagrams or storyboards—maps of sorts—that provide a graphic overview of the application that can be understood and critiqued by nonprogrammers.

Getting back to to nucleotide sequence work, the parallel to these storyboards are gene maps—high-level graphic representations of where specific sequences reside on a chromosome, many above said languages and techniques can be used to visualize the data.

**Structure Visualization**

Visualizing the biological data is one of the biggest challenges. There is a need to properly visualize the data in order to find the various properties of the data and also to examine if any defects present.

One such activity in proteomics R&D is determining and visualizing the 3D structure of proteins in order to find where drugs might modulate their activity. Several other activities include identifying all of the proteins produced by a given cell or tissue and determining how these proteins interact. In order to carry out all these activities the currently used methods include protein purification and X-ray crystallography. Both activities take significant time, even with robotic automation. As such, it's generally understood by the molecular biology research community that the sequencing of the human genome, which will likely take several more years to complete, is relatively trivial compared to definitively characterizing the proteome. Barring the introduction of some new technology, cataloging, interpreting, and dissecting the proteome will take many years. Unlike a nucleotide sequence, which is a relatively static structure, proteins are dynamic entities that change their shape and association with other molecules as a function of temperature, chemical interactions, pH, and other changes in the environment. Grasping the static structure of the approximately 30,000 proteins of the human proteome is difficult enough for many researchers, much less their potentially unlimited variation. In contrast to visualizing the sequence of nucleotides on a strand of DNA, visualizing the primary structure of a protein adds little to the knowledge of protein function. More interesting and relevant are the higher-order structures. For example, understanding the docking of two proteins is greatly facilitated by visualizing the two 3D structures interacting in 3D space. Visualizing a protein's tertiary structure is valuable in comparing protein structure predictions.

## IV. CONCLUSIONS

Both data mining and bioinformatics are fast-expanding and closely related research frontiers. It is important to examine the important research issues in bioinformatics and develop new data mining methods for scalable and effective bio data analysis. Here the basics of data mining have provided a short overview of bio data analysis from a data mining perspective. Although a comprehensive survey of all kinds of data mining methods and their potential or effectiveness in bio data analysis is well beyond the task of this short survey, the selective data

Presented here may give readers an impression that a lot of interesting work has been done and still more can be. It is believed that active interactions and collaborations between

these two fields have just started. It is a highly demanding and promising direction, and a lot of exciting results will appear in the near future.

## REFERENCES

[1] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipmann. Basic local alignment search tool. *Journalof Molecular Biology*, pages 403–410, 1990.
[2] M. Andrade and P. Bork. Automated extraction of Information in molecular biology. *FEBS Letters*,476:12–17, 2000.
[3] T. Attwood and D. Parry-Smith. *Introduction to Bioinformatics*. Longman Higher Education, 1999.
[4] A. Bairoch and R. Apweiler. The SWISS-PROT Protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, 28:45–48, 2000.
[5] P. Baldi and S. Brunak. *Bioinformatics: The MachineLearning Approach, Second Edition*. MIT Press, 2001.
[6] Bioinformatics Computing By Bryan Bergeron

## AUTHOR'S PROFILE

**Prof .Sapna V M**

Mrs .Sapna V M (Assistant professor )date of birth is 21st October 1982, received the Bachelors in IT and Masters of Technology in Comp science,from VTU University, from Belgaum ,state.Karnataka , country India. Currently she is working as AP in MIT College of Engineering, Pune. Her research interests are data mining,Wireless sensor network Embedded systems.She is life time member in CSI chapter Bangalore,She has published papers on wireless sensor networks.

**Prof . Khushboo Satpute**

*Miss Khushboo Satpute (Assistant professor )* Date of birth is 11th April 1988, received the Bachelors in CSE and Masters of Technology in IT ,from RGTU bhopal,Madhya pradesh, country India. Currently she is working as AP in MIT College of Engineering, Pune. Publish many research paper in data mining, and currently working on optimization techniques.