A Literature Survey on Data Mining in the Field of Bioinformatics

[1]Lakshmana Kumar. R, [2]M.S. Irfan Ahmed and [3]M.Amala Jayanthi

[1]Department of Computer Applications, Hindusthan College of Engineering and Technology, Coimbatore, India
E-Mail:research.laksha@gmail.com
Mobile:+91-9894433116
[2]Department of Computer Applications, Nehru Institute of Engineering and Technology, Coimbatore, India
E-Mail:msirfan@gmail.com
Mobile:+91-9003750009
[3]Department of Computer Applications, Kumaraguru College of Technology, Coimbatore, India
E-Mail:amalawithcontact@gmail.com
Mobile:+91-9940858116

Abstract- In the field of Biological research the bioinformatics plays a vital role in enriching the results for developing computer databases and algorithms. As various Bioinformatics tools are available in the market still it requires skills for their correct usage. We need trained people that can use their skills autonomously. NCBI (National Centre for Biotechnology Information) is a resource for molecular biology information. NCBI creates and maintains public databases, conducts research in computational biology, develop software tools for analyzing genome data, and disseminates biomedical information. This article identifies: 1) Tools used in the bioinformatics; 2) Research approaches and methods of delivery for conveying bioinformatics problems; and 3) Future enhancements on the impact of these programs, approaches, and methods in Bio informatics. Based on these findings, it is our goal to describe the landscape of scholarly work in this area and, as a result, identify opportunities and challenges in bioinformatics paradigm.

Keywords: Data mining, bioinformatics and NCBI tools

1. MOTIVATION:

Bioinformatics deals with large amounts of information with Biological aspects. Mainly, it focuses on molecules like DNA. Bioinformatics mixes computer science, statistics, mathematics, and engineering to analyze and interpret biological data [12]. 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 [5]. Mining biological databases imposes challenges which knowledge discovery process and data mining have to address. Methods have led to the emergence of a promising new field: Bioinformatics. On the other hand, in past few years' progress in data mining researches to the development of numerous efficient and scalable methods for mining biological data i.e. interesting patterns and knowledge in large databases, ranging[13]. There are from efficient classification methods to clustering, outlier analysis, Sequential, frequent and structured pattern analysis methods, and visualization and spatial/temporal data analysis tools.

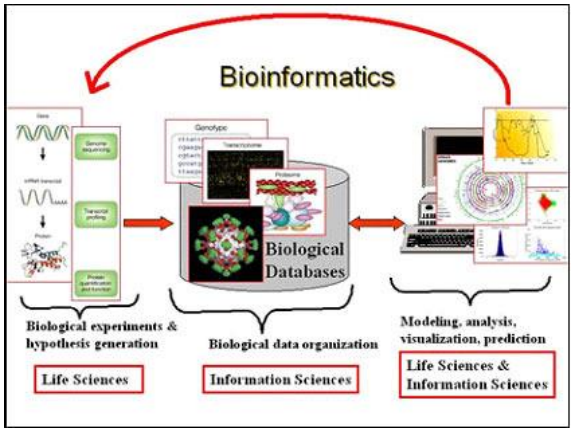

Figure-1.Archiecture of Bio Informatics

II. LITERATURE REVIEW

Bioinformatics derives knowledge from computer analysis of biological data. the collection, classification, storage, and analysis of biochemical and biological information using computers especially as applied to molecular genetics and genomics [1]. Bioinformatics is a rapidly developing branch of biology and is highly interdisciplinary, using techniques and concepts from informatics, statistics, mathematics, chemistry, biochemistry, physics, and linguistics. Review articles are the summary of current state of understanding on a particular research topic [2]. They analyze or discuss research previously published by scientist and academicians rather than reporting novel research results. Review article comes in the form of systematic reviews and literature reviews and are a form of secondary literature. Systematic reviews determine an objective list of criteria, and find all previously published original research papers that meet the criteria [16]. They then compare the results presented in these papers. Literature reviews, by contrast, provide a summary of what the authors believe are the best and most relevant prior publications. The concept of "review article" is separate from the concept of peer-reviewed literature. It is possible for a review to be peer-reviewed, and it is possible for a review to be non-peer-reviewed.

Data mining [3] 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 [14]. 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 [15]. 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.

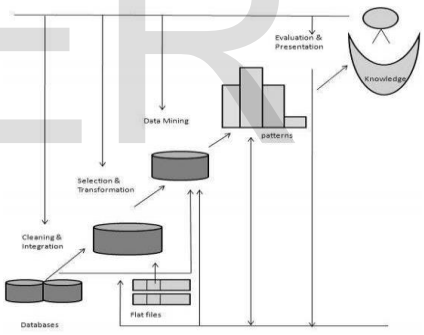

Figure-2.Data Mining Process

III. METHODS

Data mining is conducted against data accumulated in OLTP repositories, data warehouses, data marts and archived data [9]. The steps for data mining follow the following pattern:

- Data Extraction
- Data Cleansing
- Data Transformation /Reduction
- Data Mining Methods
- Applying Data Mining Algorithm

• Modelling 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. 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 [17]. 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.

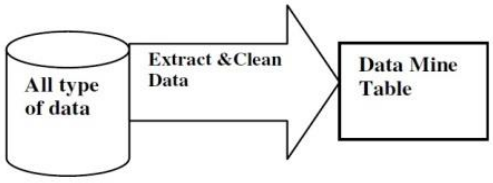

Figure-3.Data Extraction

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 pre-processed and cleaned. This is done in following steps:

Data categorization: It basically deals with documentation of data in an appropriate and meaningful manner, so that any person could understand and interpret the data comfortably. 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 [18]. 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. 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 [19]. 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 pre-processing 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.

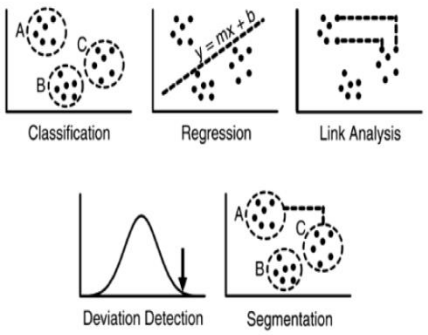

Figure-4.Techniques in Data Mining

E. APPLYING DATA MINING ALGORITHM

It 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 modelling, adaptive control theory, psychology, and artificial intelligence (AI). 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.

F. DATA MODELING

Data modelling 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 [20]. Conceptual data modelling 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, e.g. traditional database management systems. Data modelling is the formalization and documentation of existing processes and events that occur during application software design and development. Data modelling 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 reengineering.

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.

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. 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 is types of data visualization techniques.

Sequence Visualization 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 [4]. 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.

1. Structure Visualization

Structural bioinformatics is the branch of bioinformatics which is related to the analysis and prediction of the three-dimensional structure of biological macromolecules such as proteins, RNA, and DNA. It deals with generalizations about macromolecular 3D structure such as comparisons of overall folds and local motifs, principles of molecular folding, evolution, and binding interactions, and structure/function relationships, working both from experimentally solved structures and from computational models. The term structural has the same meaning as in structural biology, and structural bioinformatics can be seen as a part of computational structural biology.

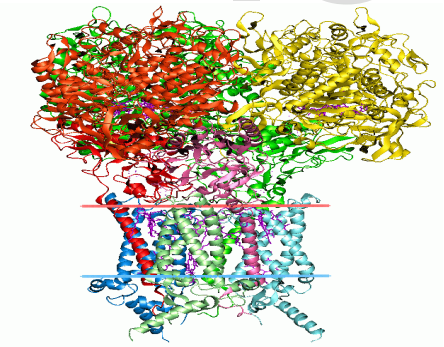

Figure-5-Structure Visualization

2. Protein structure prediction

It is the inference of the three-dimensional structure of a protein from its amino acid sequence—that is, the prediction of its folding and its secondary and tertiary structure from its primary structure [7]. Structure prediction is fundamentally different from the inverse problem of protein design [8]. Protein structure prediction is one of the most important goals pursued by bioinformatics and theoretical chemistry; it is highly important in medicine (for example, in drug design) and biotechnology (for example, in the design of novel enzymes). Every two years, the performance of current methods is assessed in the CASP experiment (Critical Assessment of Techniques for Protein Structure Prediction). A continuous evaluation of protein structure prediction web servers is performed by the community project CAMEO3D.

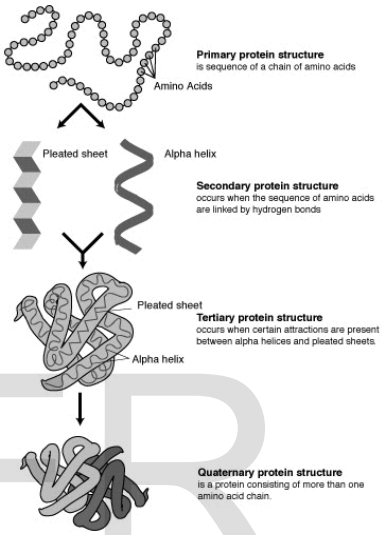

Figure-6.Protein Structure

IV. DISCUSSION

In this section we have discussed several tools which are readily available to address the difference issues in bio informatics

Table-1. Various tools for Bio Informatics

Tools	Description

1.BLAST	The **Basic Local Alignment Search Tool** for comparing gene and protein sequences against others in public databases now comes in several types including PSI-BLAST, PHI-BLAST, and BLAST 2 sequences. Specialized BLASTs are also available for human, microbial, malaria, and other genomes, as well as for vector contamination, immunoglobulins, and tentative human consensus sequences.
2.Electronic PCR	It allows you to search your DNA sequence for sequence tagged sites (STSs) that have been used as landmarks in various types of genomic maps. It compares the query sequence against data in NCBI's UniSTS, a unified, non-redundant view of STSs from a wide range of sources.

3.Entrez Gene:	Each Entrez Gene record encapsulates a wide range of information for a given gene and organism. When possible, the information includes results of analyses that have been done on the sequence data. The amount and type of information presented depend on what is available for a particular gene and organism and can include: (1) graphic summary of the genomic context, intron/exon structure, and flanking genes, (2) link to a graphic view of the mRNA sequence, which in turn shows biological features such as CDS, SNPs, etc., (3) links to gene ontology and phenotypic information, (4) links to corresponding protein sequence data and conserved domains, (5)links to related resources, such as mutation databases. Entrez Gene is a successor to LocusLink.

4.Model Maker	It allows you to view the evidence (mRNAs, ESTs, and gene predictions) that was aligned to assembled genomic sequence to build a gene model and to edit the model by selecting or removing putative exons. You can then view the mRNA sequence and potential ORFs for the edited model and save the mRNA sequence data for use in other programs. Model Maker is accessible from sequence maps that were analyzed at NCBI and displayed in Map Viewer.
5.ORF Finder	ORF Finder identifies all possible ORFs in a DNA sequence by locating the standard and alternative stop and start codons. The deduced amino acid sequences can then be used to BLAST against GenBank. ORF finder is also packaged in the sequence submission software Sequin.

6.SAGEMAP	It is a tool for performing statistical tests designed specifically for differential-type analyses of SAGE (Serial Analysis of Gene Expression) data. The data include SAGE libraries generated by individual labs as well as those generated by the Cancer Genome Anatomy Project (CGAP), which have been submitted to Gene Expression Omnibus (GEO). Gene expression profiles that compare the expression in different SAGE libraries are also available on the Entrez GEO Profiles pages. It is possible to enter a query sequence in the SAGEmap resource to determine what SAGE tags are in the sequence, then map to associated SAGEtag records and view the expression of those tags in different CGAP SAGE libraries.
7.Spidey	It aligns one or more mRNA sequences to a single genomic sequence. Spidey will try to determine the exon/intron structure, returning one or more models of the genomic structure, including the genomic/mRNA alignments for each exon.

8.VecScreen	It is a tool for identifying segments of a nucleic acid sequence that may be of vector, linker, or adapter origin prior to sequence analysis or submission. VecScreen was developed to combat the problem of vector contamination in public sequence databases.

Table-2.Comparision of tools with various parameters

Tools	General parameter	Scoring parameter	Filters and masking	Search set	Query sequence
1.BLAST	yes	yes	yes	yes	yes
2.ORF Finder	no	no	no	yes	yes

V. CONCLUSION

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.

VI. FUTURE ENHANCEMENT

Bioinformatics has become an essential interdisciplinary scientific field to the life science helping to “omics(Genomics)” field and technologies and mainly handling and analyzing “omics” data. Accumulation of high-throughput biological data due to the technological advances in “omics” fields required and prioritized the use of bioinformatics resources, and research and application for the analysis of complex and even further enlarging “Big Data” volumes, which would be impractical and useless without bioinformatics. Therefore, as highlighted herein, there is a critical need for the preparation of well-qualified, new generation scientists with integrated knowledge, multilingual ability, and cross-field experience who are capable of using sophisticated operating systems, software and algorithms, and database/networking technologies to handle, analyze, and interpret high-throughput and increasing volume of complex biological data.

Community resources and a globally coordinated foundation of bioinformatics training and education platforms as well as research conferences, workshops, short online training, and web-based educational courses and materials are available to accomplish toward this goal. However, there is an urgent need for the development of bioinformatics education and training, in particular in developing countries, which requires innovative platforms, training techniques, better funding, web and network access, and high-performance computing systems.

REFERENCES

1. Li, Jinyan, Limsoon Wong, and Qiang Yang. "Guest Editors' Introduction: Data Mining in Bioinformatics." *IEEE intelligent systems* 20.6 (2005): 16-18.

2. Abdurakhmonov, Ibrokhim Y. "Bioinformatics: Basics, Development, and Future." *BIOINFORMATICS-UPDATED FEATURES AND APPLICATIONS* (2016): 1.

3. Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

4. Lu, Dihui. *BIOLOGISTS'INFORMATION SEEKING BEHAVIOR WITH ONLINE BIOINFORMATICS RESOURCES FOR GENOME RESEARCH*. Diss. University of North Carolina at Chapel Hill, 2003.

5. Sapna, V. M., and Khushboo Satpute. "Data Mining in Bioinformatics: Study & Survey of Data Mining and its Operations in Mining Biological Data." *International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE)* 2.9 (2013): 1.

6. Raza, Khalid. "Application of data mining in bioinformatics." arXiv preprint arXiv:1205.1125 (2012).

7. Couto, Francisco M., Mario J. Silva, and Pedro Coutinho. "ProFAL: PROtein Functional Annotation through Literature." *JISBD*. 2003.

8. Tang, Haixu, and Sun Kim. "Bioinformatics: Mining the Massive Data from High Throughput Genomics Experiments." (2007): 3-24.

9. Hu, Xiaohua. "Data mining and its applications in bioinformatics: techniques and methods." *Granular Computing (GrC), 2011 IEEE International Conference on*. IEEE, 2011.

10. Pfaltz, John L., and Christopher M. Taylor. "Closed set mining of biological data." *Proceedings of the 2nd International Conference on Data Mining in Bioinformatics*. Springer-Verlag, 2002.

11. Han, Jiawei. "How can data mining help bio-data analysis?[extended abstract]." *Proceedings of the 2nd International Conference on Data Mining in Bioinformatics*. Springer-Verlag, 2002.

12. Luscombe, Nicholas M., Dov Greenbaum, and Mark Gerstein. "What is bioinformatics? An introduction and overview." *Yearbook of Medical Informatics* 1.83-100 (2001): 2.

13. Sapna, V. M., and Khushboo Satpute. "Data Mining in Bioinformatics: Study & Survey of Data Mining and its Operations in Mining Biological Data." *International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE)* 2.9 (2013): 1.

14. Romero, Cristóbal, and Sebastián Ventura. "Educational data mining: a review of the state of the art." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 40.6 (2010): 601-618.

15. Kitts, Brendan, Gabor Melli, and Karl Rexer. "Data Mining Case Studies." *The First International Workshop on Data Mining Case Studies, 2005 IEEE International Conference on Data Mining, Huston, USA*. 2005.

16. Altschul, Stephen F., et al. "Basic local alignment search tool." *Journal of molecular biology* 215.3 (1990): 403-410.

17. Andrade, Miguel A., and Peer Bork. "Automated extraction of information in molecular biology." *FEBS letters* 476.1-2 (2000): 12-17.

18. Counsell, Damian. "A review of bioinformatics education in the UK." *Briefings in Bioinformatics* 4.1 (2003): 7-21.

19. Bairoch, Amos, and Rolf Apweiler. "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000." *Nucleic acids research* 28.1 (2000): 45-48.

20. Baldi, Pierre, and Søren Brunak. *Bioinformatics: the machine learning approach*. MIT press, 2001.