Available online at www.elixirpublishers.com (Elixir International Journal)

Bio Technology

Elixir Bio Tech. 80 (2015) 31138-31142

ISSN: 2229-712X

# Data Mining in Bioinformatics

Rajesh Kumar. M[1] and R.Rajasekaran[2,*]

[1]Periyar Maniammai University, Chennai, Tamil nadu, India.

[2]Department of Biology, College of Science, Eritrea Institute of Technology, Mai Nefhi, P.O.Box no:-12676, Asmara, Eritrea, North East Africa.

**ARTICLE INFO**

**Article history:**

Received: 27 January 2015;

Received in revised form:

28 February 2015;

Accepted: 16 March 2015;

**Keywords**

Databases,

Bioinformatics,

Data mining,

Genomics,

Proteomics.

**ABSTRACT**

Bioinformatics is a science typically associated with databases in genomics and proteomics and structure and Function information for genes and proteins, of all forms of life on earth. In the past decade there has been a 'cyber-war', with the introduction of a number of biological databases on genomics and proteomics. The major aim here is to introduce data mining techniques as an automated means of reducing the complexity of biological data in large bioinformatics databases and of discovering meaningful, useful patterns and relationships in data. The main purpose of data mining in the field of bioinformatics is the mining of complex data which is fast growing and can be said to be outgrowing our processing power.



## Introduction

Bioinformatics is the science of data management system in genomics and proteomics of life forms. It is a comparatively young discipline in information technology and has progressed very fast in the last few years. Bioinformatics is practiced worldwide by biotechnologists to access various databases for research and to exchange information for comparison, confirmation, storage and analysis. The term Bioinformatics was coined by Paulien Hogeweg in 1979 for the study of informatics processes in biotic systems (www.wikipedia.org). As on date, there are a number of databases on specific genes and proteins pertaining to human, animals, plants, bacteria, and other life forms. These are being enriched and updated through research in modern biology with the practice of bioinformatics. It can be said that this scientific field deals with the computational management of all kinds of biological information, whether it may be about genes and their products, whole organisms or even ecological systems. Most of the bioinformatics work that is being done deals mainly with analyzing biological data, although a growing number of projects deal with the organization of biological information. There are various ways in which Bioinformatics may be defined. Most commonly it is defined as the interface between biological and computational sciences.

It may also be defined as the application of computer technology to biology; a combination of techniques and models in statistical, computational, and life sciences to understand the significance of biological data. In more formal terms Bioinformatics is the recording, annotation, storage, analysis, and searching/retrieval of nucleic acid sequence (genes and RNAs), protein sequence and structural information. This includes databases of the sequences and structural information as well methods to access. Search, visualize and retrieve the information[1]. It basically, is the field of science in which biology, computer science, and information technology merge into a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principle in biology can be discerned. The biotech revolution was ignited by decoding the human genome. It also created one of the biggest hurdles the Information Age has ever faced: which was to make sense of three billion base pairs of human DNA. It promised to cure every disease with a genetic component, including cancer, asthma, diabetes and mental illness. This grand, genomic dilemma gave birth to the newborn discipline of bioinformatics, the use of computers and information technology to tackle biology problems[2].

There are three important sub-disciplines within bioinformatics

- The development of new algorithms and statistics with which to assess relationships among members of large data sets.
- The analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures.
- The development and implementation of tools that enable efficient access and management of different types of information.

One of the tasks used in bioinformatics concerns the creation and maintenance of database of biological information. While the storage and organization of millions of nucleotides is far from trivial, designing a database and developing an interface whereby researches can both access existing information and submit new entries is only the beginning.

There are various types of data involved in the fields of Bioinformatics. Some are

- Amino acid sequences[3].
- Protein domain cartoons.
- Different renderings of three-dimensional structures.
- Protein hydrophobicity data (www.fasta.bioch.virginia.edu/o_fasta/grease.htm)

Databases consisting of data derived experimentally, such as nucleotide sequences and three-dimensional structures are known as primary databases. Those data that are derived from the analysis or treatment of primary data such as secondary structures, hydrophobicity plots URL: http://fasta.bioch.Virginia.edu/o-fasta/grease.htm and domains are stored in secondary databases. A protein database

Tele:

E-mail addresses: rajasekharan.r@gmail.com

consisting of the conceptual translation of nucleotide sequences would also be considered a secondary database. There has been considerable mention of biological database in the above paragraphs. Let us look into what exactly does this biological database mean. A database is basically a collection of information that is structured, searchable, updated periodically and cross-referenced. The databases can be of many types:

- Flat file database
- Relational database
- Object oriented database

**Flat file database are simple text only files**

Relational databases have data stored in tables and relationships can be defined on a many-to-one basis or a many-to-many basis. The latter type requires a separate relationships table linking the two tables. All records in a given table have to necessary had identical features and any unique feature have to be stored in a separate table. For example, a sequence table might contain records with the attributes accession number and protein sequence while a function table might contain records with the attributes accession number and protein function. In Object-oriented databases, Data are defined as objects which have a class hierarchy, that is, they can be grouped into classes and subclasses etc. In a hierarchical manner. With the amount of data available, one of the most pressing tasks is the analysis and interpretation of data. This can be done in various ways.

- Individual entries in sequence and structure databases can be complied to reveal patterns and trends in biology.
- Common sequence features in sequence families can be identified in multiple alignments[4]. Clustering of sequences into trees reflects the degree of similarity between each sequence and all of the others in the family reveals evolutionary relationships.
- Finally, identification of homologs to each gene in well-characterized metabolic pathways provides information about the prevalence of that pathway in other organisms.

Getting the sequence the structure of data to molecular biology databases and the functional data in the online biomedical literature is complicated by the size and complexity of the database. Searching for raw data and performing the transformations and manipulations on the data through manual operations is often impractical. To avoid the computational constraints imposed by these large molecular biology databases, researchers turn to biological heuristics to avoid exhaustive searches. However, even with heuristics, user-directed discovery is inherently limited by the time required to manually search for new data.

Data Mining is one stage in the whole knowledge discovery process. This process involves selection and sampling of appropriate data from the databases, preprocessing and cleaning of the data to remove redundancies, errors and conflicts; transforming and reducing data to a format more suitable for data mining; evolution of mined data and visualization of the evolution results. Data Mining can be defined as the analysis of (often large) observational data which sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. It is basically the process of selecting, exploring and modeling large amounts of data to uncover previously unknown patterns for business advantage. (SAS Institute)

Data Mining has been used in a wide variety of applications including those in Marketing (sales model), Finance (loan decision), Insurance (risk analysis), Telecom (load predication), Web/Text Mining, Bioinformatics and many more. Let us take a look at some of the applications that data mining has in bioinformatics.

Data mining is used in the:

- Analysis of Micro array Data
- Mining free text
- Structural Genomics-Protein Crystallization
- Predicting structures from sequence

Data mining is a collection of techniques for extracting new, valid, interesting and surprising information from large databases. Traditional data analysis is assumption-driven, which basically means that a hypothesis is developed in advance and tested against data. The new techniques discussed here are discovery-driven. The extracted knowledge itself acts as a starting point for a new hypothesis. Researchers in areas such as machine learning, pattern recognition, and artificial intelligence first developed these techniques[5]. Most of these systems are able to import and export models in PMML (Predictive Model Markup Language) which provides a standard way to represent data mining models so that these can be shared between different statistical applications. PMML is an XML-based language developed by the Data Mining Group (DMG) an independent group composed of many data mining companies. PMML version 4.0 was released in June 2009 (www.wikipedia.org).

The data mining process typically involves the following

- Data cleansing – handling of noisy, missing or irrelevant data.
- Data integration – integration of multiple data sources.
- Data selection – retrieving relevant data.
- Data transformation – reducing the size of the data-set to be examined to a minimum.
- Data analysis – applying intelligent pattern extraction algorithms.
- Pattern evaluation – assign measures of "interestingness" to the patterns.
- Presenting the extracted knowledge to the user.

The process of data mining is concerned with extracting patterns from data by using techniques such as classification, regression, link analysis, segmentation and deviation detection. Classification involves mapping data into one of several predefined or newly discovered classes. It basically consists of assigning unknown data patterns to existing equivalence classes. Regression methods involve assigning data continuous numerical variables base on statistical methods. It uses a feed-forward neural network that can study sets of data and learn correlation between inputs and outputs. Link analysis involves evaluating apparent connections or links between data in the database. Segmentation identifies classes or groups of data that behave similarly, according to an established metric.

Deviation detection identifies data values that are outside of the normal. These methods of data mining are typically used in combination with each other, either in parallel or as part of sequential operation. Most of the various methods and techniques uses in data mining, originate from statistics, machine learning, and pattern recognition. Among the techniques used are Neural Networks, Bayesian Networks, Hidden Markov model etc (www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-fayyad.pdf).

Common data mining methods and tools are

- Clustering.
- Dependency modeling.

Given a number of observations, each described by a set of attributes, "clustering" aims to group the observations into equivalence classes, i.e. sub-sets of data that share attributes and are therefore "alike" in some sense. In "dependency modeling"

the degree to which one observation depends on another is described. So, if assertion "x" is true, then assertion "y" is also true to a certain probability. This allows for probabilistic reasoning to be carried out on a set of data.

Identifying and interpreting interesting patterns hidden within the immense list of bases that constitute a genome is a critical goal in molecular biology research. A number of genes in a new genome will be unlike anything previously known. The ability of an algorithm not only to find known sequences but also to recognize new sequences is crucial. Data mining algorithms that can find these patterns in polynomial running time are of great interest. Such an algorithm is the "SUBDUDE" algorithm proposed by researchers at the University of Arlington in Texas. SUBDUDE discovers interesting substructures in structural data without any prejudices about existing, known patterns. Thus it can find unexpected patterns that may have biological importance. SUBDUDE has been used successfully to identity patterns in three proteins: haemoglobin, myoglobin, and ribonuclease A. These patterns can now act as unique identifiers for the protein and provide a mechanism for classifying unknown proteins. SUBDUDE has also been used to find biologically important patterns in the DNA of baker's yeast, Saccharomyces cerevisiae. For example, the pattern AAGGG in this yeast is very probably involved in gene regulation[6].

Another example of data mining of DNA, the "peculiarity oriented mining" proposed by researchers at Maebashi Institute of Technology and Waseda University in Japan. This algorithm is based on the proposition that if a sequence is peculiar then it might also be interesting. Data is defined to be "peculiar" if it is very different from other objects in the data set. A peculiarity factors is assigned to each object by calculating the sum of the square root of the conceptual distance between that object and other objects in the set. The peculiar data is then studied by assigning a "relevance" to each piece of data. This technique has revealed "reasonable and interesting", albeit specialized and abstruse, relationships between amino-acid models and experimental results.

Data mining and knowledge discovery tools are not only being used for fundamental biological research but are also being used for medical research and for patient care and diagnosis. Glaucoma, a disease that affects the optic nerve, is one of the leading causes of blindness in the developed world. If diagnosed early suitable treatment can delay the progress of the disease. Indeed, determining if the disease is progressing and how quickly, is a crucial consideration in deciding on the method of treatment. Also, if a method of determining the progress of the disease is known, the effect of different drugs on the disease can be assessed. Measuring visual function in 76 locations of the central visual area is an important diagnostic tool in the treatment of glaucoma. Data mining techniques such as "support vector machines" and "IR decision stumps" have been applied to visual field measurement data collected from 113 patients each having at least eight visual fields measurements over a four-year period. These techniques detected progressing glaucoma more accurately than point wise univariate linear regression analysis – the most accurate traditional diagnostic tool[7].

Another application is in detection of the onset of ventricular fibrillation (VF). VF is an irregular heart rhythm that causes the heart to be incapable of pumping blood. It is fatal within minutes unless terminated by the passage of a large electric current through the heart muscle. The ability to predict the onset of VF in time to deliver a less serve shock or to administer a preventative drug is obviously a safer and more managed solution. Data mining techniques have been applied to short data intervals of ECG recordings. The algorithm can classify ECG rhythms in only 2.5 seconds. Other tools available require at least 10 seconds doing the same. Although this is a great achievement, improved accuracy would be necessary to avoid delivering unnecessary shocks.

Drug discovery is another area of application for data mining. This is of major interest to the fundamental active "ingredient" in a compound is, a study of various compounds is carried out with the same functionally and attempt to discern their common characteristics. Data mining techniques that are functional programming algorithms have been applied at Plymouth University and at GlaxoSmithKline. These algorithms mine the structural data of organic compounds by making use of a functional programming language called "Gofer". It was possible to classify accurately 64 test organic compounds according to the key common characteristics provided the length of the compound was 1024 bytes or less. Such results indicate that these classification techniques will become a valuable decision support tool in the pharmaceutical industry[8]. Other examples of applying data mining technique applications are biomedical data facilitated by domain ontology, mining clinical trial data[9].

Life science produce vast databases of information, as living organisms are the most sophisticated and complex systems encountered. Life science research is becoming increasingly database driven and many databases consisting of biological signals have been complied. Some examples include: the MIT-BIH arrhythmia database, the MIT-BIH polysomnographic database of recordings of multiple physiologic signals during sleep and the collection of animal sounds at the Borror Laboratory of bioacoustics at The Ohio State University. The ultimate goal of mining DNA sequence is to discover all of the genes, how they are regulated and how the genes work together to produce higher function and behavior levels. Relatively short sequence of DNA in the region surrounding the gene control gene expression. Finding these short sequences is a fundamental problem to molecular biologists and computer scientists. To get an idea of the difficulty, imagine you are given 30 DNA sequence, each of length 800. Now find a common pattern of length 8. In the simplest case where the pattern occurs exactly once in each sequence, there are a possible 80030 potential locations. When problems such as these are solved, and as more knowledge is mined out of DNA sequences, this will bring a revolution in understanding of human health, genetics and the functioning of living organisms.

**Information Retrieval from Databases**

With the amount of data available today in the growing field of bioinformatics, one of the most pressing tasks is the retrieval of the data from the databases and the interpretation of the data. These are various tools used together with the data mining techniques to effectively mine and interpret the vast amount of data available. Some of them are discussed below.

BLAST sequence-similarity searching which is standard tool used in sequence analysis. Basically used for comparing gene and protein sequences against others in public databases[10,11]. Another tool used in Sequence Analysis is the LocusLink which covers information on Official nomenclature, aliases, sequences, phenotypes etc. UniGene[12] is another tool used in Sequence Analysis. It can assist in gene discovery, gene mapping projects, and large-scale expression analysis. Another tool Electronic PCR allows searching DNA sequence tagged sites STSs and EST[13,14] which have been uses as landmarks in

various types of genomic maps. Spidey aligns one or more mRNA sequences to a single genomic sequence.

Clusters of Orthologous Groups (COGs) currently covers 21 complete genomes from 17 major phylogenetic lineages. It is a tool used in Clustering which is a technique aimed to group observations into equivalence classes[15]. Another tool used in biologically-oriented cluster analysis of DNA microarray data, automated functional annotation via clustering, and functional categorization of biological objects is GOODIES. It is used for various purposes which can be classified as below:

- In cluster analysis of DNA microarrays, biologists primarily want to know how well clusters of expression profiles are associated with known functional categories and cellular processes. GOODIES performs such a task in terms of Gene. Ontology (GO) that it is complementary to statistical clustering methods.
- The function of unknown genes can be putatively predicted through the clustering interpretation of GOODIES. After the biological relationships for each cluster is quantitatively measured by a newly-defined metric 'AverPd' on GO terms, the clusters whose AverPd score is sufficiently low can be used for functional assignment of unknown genes in those clusters.
- GOODIES can accomplish a large-scale functional categorization of biological entities – e.g. ESTs, genes, and proteins – according to the GO annotations of each entity that are extracted from reliable, curated databases.

Entrez is a search and retrieval system that integrated information from databases. These database include nucleotide sequences, protein sequence, macromolecular structures, whole genomes, and MEDLINE. Entrez provides access to several linked databases, such as:

- PubMed: The biomedical literature[16,17].
- Nucleotide sequence database Genbank[13].
- Protein sequence database[18].
- Structure: three-dimensional macromolecular structures[19].
- Genome: complete genome assemblies.
- PopSet: Population study data sets.
- Taxonomy: organisms in GenBank.
- OMIM: Online Mendelian Inheritance in Man.

**Sequence Retrieval System (SRS)**

The Sequence Retrieval System (SRS) at the European Bioinformatics Institute (EBI) allows both simple and complex concurrent searches of one or more sequence databases. It is open-source software. The SRS system may be installed and used on a local machine to assist in the preparation of local sequence databases.

Structural Analysis – MSDfold or DALI can be used to query the protein structure and compare it to those in the Protein Data Bank (PDB) www.rcsb.org/pdb

**Protein Functional Analysis**

There are various tools used in Protein Functional Analysis[20]. These are:

- CluSTr Search: searches UniProt by accession numbers.
- FingerPRINTScan: Prints protein fingerprint searches.
- GeneQuiz: Highly automated analysis of biological sequences.
- InterProScan: Search protein sequences against InterPro member databases.
- PPSearch: A Protein motif searches.
- ScanProsite: Scans a sequence against PROSITE.
- ProtParam: Physico-chemical parameters of a protein sequence.
- PredictProtein: PHDsec, PHDacc, PHDhtm, PHDtopology from Columbia University.
- REP: Searches a protein sequence for repeats.
- Radar: Protein repeat detection.

**The Future of Bioinformatics**

Keeping in view the present pace of investment and developments, it is impossible to predict the future of bioinformatics. The searching of biological databases via the WWW is becoming increasingly difficult. Differences in database structures and nomenclature hinder research efforts where standardizations have met with much resistance. However, researchers are optimistic that web tools developed for other purposes may help bioinformatics[21]. New software is being produced for different applications in biotechnology. Such developments are important for the future of bioinformatics and development of biotechnology. Further, it has been apprehended that such fast developments in the internet and use of information available in databases may prove to be 'library killers' as subscriptions to science journals will reduce[22]. The future challenges of sequence analysis are pushing bioinformatics in a time when the demand of bioinformaticians outnumbers supply. Thus, in future more biotechnologists with computer knowledge are required. This can be made possible through various training programmes. Various network systems like EMBLnet and ICCBnet, initiated by UNESCO are playing a vital role in this regard imparting training to biotechnologists in bioinformatics through their training programmes[1].

With the increase of sequencing projects, bioinformatics continues to make considerable progress in biology by providing scientists with access to the genomic information. This progress is especially contributed by the Human Genome Project. The information obtained with the help of Bioinformatics tools furthers the understanding of various genetic and other diseases and helps identify new drug targets. With technological developments of the Internet, scientists are now able to freely access volumes to such biological information, which enables the advancement of scientific discoveries in biomedicine. In spite of being young, the science of Bioinformatics exhibits tremendous potential for playing a major role in the future development of science and technology. This is evident from the fact that modern biology and related science are increasingly becoming dependent on this new technology. It is expected that Bioinformatics will especially contribute in the future as the leading edge in biomedicine to pharmaceutical companies by expediently yielding a greater quantity of lead drugs for therapy.

**References**

1. Edelman, M., 1999. The ICCBnet Bioinformatics Training Workshop, International Centre for Co-operation in Bioinformatics Network, Weizmann Institute of Science, Israel, pp. 11-20.
2. Tripathi, K. K., 2000. Bioinformatics: The foundation of present and future biotechnology. Current science., 79(5), 570-575.
3. Bairoch, A., and Apweiler, R., 1999. The Swiss-Prot protein sequence data bank and its supplement TrEMBL in 1999. Nucl. Acids Res., 27, 49-54.
4. Thompson, J., Desmond, G., Higgins. and Toby, Gibson, J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl. Acids Res., 22, 4673-4680.
5. Kantardzic, Mehmed., 2003. Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons.

6. Attwood, Teresa, K., and Parry-Smith, David, J., 2001. An Introduction to Bioinformatics; Pearson Education (Singapore) Pvt. Ltd, New Delhi.
7. Kretschmann, E., Fleischmann, W., and Apweiler, R., 2001. Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. Bioinformatics., 920-926.
8. Bate, A., Lindquist, M., Edwards, I. R., Olsson, S., Orre, R., Lansner, A., De Freitas, R. M. A., 1998. Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol., 54 (4), 315-21.
9. Xingquan Zhu, Ian Davidson., 2007. Knowledge Discovery and Data Mining: Challenges and Realities. Hershey, New York, pp. 163-189. .
10. Altschul, S. F, T. L Madden, A. A., Schaffer, J., Zhang, Z., Zhang, W., Miller., and Lipman, D. J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res., 25, 3389-3402.
11. Zhang, Z, A. A., Schaffer, W., Miller, T. L., Madden, D. J., Lipman, Koonin, E. V., and Altschul, S. F., 1998. Protein sequence similarity searches using patterns as seeds. Nucl. Acids Res., 26, 3986-3990.
12. Wheeler, L. D., Colombe Chappey, Alex, E., Lash, Detlef, D., Leipe, Thomas, L., Madden, Gregory, D., Schuler, Tatiana, A., Tatusova, and Barbara, Rapp, A., 2000. Database resources of the National Center for Biotechnology Information. Nucl. Acids Res., 28, 10-14.
13. Dennis, A., Benson, Ilene Karsch-Mizrachi, David, J., Lipman, James Ostell, Barbara, A., and David, Wheeler, L., 2000. Gen Bank. Nucl. Acids Res., 28 (1), 15-18.
14. Boguski, Mark, S., Todd, M. J., Lowe, and Carolyn, Tolstoshev, M., 1993. dbEST-database for "expressed sequence tags". Nature genetics. 4, 332-333.
15. Hennig, W., 1950. Grundzüge einer Theorie der phylogenetischen Systematik. (reprint 1994 by Koeltz Scientific Books,pp. 370.
16. Baasiri, R. A., Glasser, S. R., Steffen, D. L. and Wheeler, D. A., The breast cancer gene database: a collaborative information resource. Oncogene., 1999, 18, 7958-7965.
17. Wheeler, L. D., Tanya Barrett, Dennis, A., Benson, Stephen, H., Bryant, Kathi Canese, Deanna, M., Church, Michael DiCuccio, Ron Edgar, Scott Federhen, Wolfgang Helmberg, David, L., Kenton, Oleg Khovayko, David, J., Lipman, Thomas, L., Madden, Donna, R., Maglott, James Ostell, Joan, U., Pontius, Kim, D., Pruitt, Gregory, D., Schuler, Lynn, M., Schriml, Edwin Sequeira, Steven, T., Sherry, Karl Sirotkin, Grigory Starchenko, Tugba, O., Suzek, Roman Tatusov, Tatiana, A., Tatusova, Lukas Wagner. and Eugene Yaschenko., Database resources of the National Center for Biotechnology Information. Nucleic Acids Res., 2005, 33, 39-45.
18. Bairoch, A., and Apweiler, R., The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucl. Acids Res., 2000, 28, 45-48.
19. Helen, M., Borman, John, Westbrook, Zukang, Feng, Gary Gilli, and T.N., Bhat, Helge Weissig, Ilya, Shindyalov, N., and Philip, Bourne, E., The protein data bank. Nucl. Acids Res., 2000, 28, 235-242.
20. Mani, K., and Vijayaraj, N., Bioinformatics a practical approach, Aparna Publication, 2004, pp. 36-42.
21. Sobral, B. W. S., Common language of bioinformatics. Nature., 1997, 389, 418-420.
22. Butler, D., The writing is on the web for science journals in print (News). Nature., 1999, 397, 195-200.