Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 5, May 2017, pg.72 – 83

Data Mining and Gene Expression Analysis in Bioinformatics

Saurabh Sindhu, Divya Sindhu

Department of Computer Science, CRM Jat College, Hisar, Haryana, India

Abstract - Recent technological advances in computational biology and bioinformatics have generated a lot of biological information and voluminous biological data all over the world. This biological dat/information has necessitated the requirement of computerized databases to store, categorize and index the data as well as to view and analyze the data using specialized tools and algorithms. Feature selection, normalization and standardization of the data, visualization of the results and evaluation of the produced knowledge are equally important steps in the knowledge discovery process. Data mining aims to provide the analysts with novel and efficient computational tools to overcome the constraints posed by the traditional statistical methods. It is used to extract interesting, nontrivial, implicit, previously unknown and potentially useful information from data. Likewise, bioinformatics is the analysis of accurate and reliable biological information using computers and statistical techniques in order to gain new biological insights. Computers are used to gather, store, analyze and integrate biological and genetic information which can then be applied to gene-based drug discovery and development. This paper will focus on issues related to data mining and relevance of these data base sequences and gene expression in bioinformatics.

Keywords - Data mining, DNA sequences, Gene expression, Proteomics, Knowledge Discovery, Bioinformatics

I. INTRODUCTION

Rapid developments in genomics and proteomics in recent years have generated a large amount of biological data. Sophisticated computational analyses are required to draw conclusions from these data. Bioinformatics or computational biology is the interdisciplinary science of interpreting and analysis of biological data using information technology and computational techniques [1]. This area has evolved tremendously in recent years due to the explosive growth of biological information generated by the scientific community. Bioinformatics is the science of managing, mining, integrating and interpreting information from biological data at the genomic, proteomic, phylogenetic, cellular or whole organism levels [2, 3]. The need for bioinformatics tools and expertise has increased as genome sequencing projects have resulted in an exponential growth in complete and partial sequence databases. The importance of this new field of bioinformatics will grow as we continue to generate and integrate large quantities of genomic, proteomic and other data.

Data mining is the use of automated data analysis techniques to uncover previously undetected relationships among data items. Data mining often involves the analysis of data stored in a data warehouse. Data mining helps the scientists and the researches to extract the useful information from the

huge amount biological data at hand by providing sophisticated techniques [4]. It is a technique applied for the discovery of patterns hidden in large data sets, focusing on issues relating to their feasibility, usefulness, effectiveness and scalability. Combination of soft computing (which deals with information processing) and data mining in a constructive way can effectively be used for knowledge discovery in large databases.

A particular active area of research in bioinformatics is the application and development of data mining techniques to solve biological problems. Analyzing large biological datasets requires making sense of the data by inferring structure or generalizations from the data. Examples of this type of analysis include protein structure prediction, gene classification, cancer classification based on microarray data, clustering of gene expression data, statistical modeling of protein-protein interactions etc. Therefore, there is a great potential to increase the interaction between data mining and bioinformatics. The present article provides an overall understanding of data mining techniques and their application and usage in bioinformatics.

II. DATA MINING

Data mining (DM) refers to extracting or "mining" of knowledge from huge amounts of biological data records. It is defined as the process of discovering meaningful new correlations, patterns and trends by digging into large amounts of data stored in warehouses. Data mining is also sometimes called Knowledge Discovery in Databases (KDD) [5]. It requires sharp technologies and the compliance to discover the possibility of hidden knowledge that resides in the data. Data mining approaches are ideally suited for Bioinformatics, since it is data-rich, but it lacks a comprehensive theory of life's organization at the molecular level. The extensive databases of biological information craft both challenges and opportunities for development of novel KDD methods. Mining of biological data helps to extract useful knowledge from massive datasets gathered in biology and in other related life sciences areas such as medicine and neuroscience etc.

Knowledge discovery consists of an iterative sequence of the following steps (Fig. 1).

i. **Storage and selection:** Data obtained from various sources is stored in data warehouse and data of interest is selected.
ii. **Preprocessing and transformation:** The target data is cleansed and transformed to new and common format.
iii. **Data mining:** Processed data is used for data mining to obtain desired results.
iv. **Interpretation/Evaluation:** Results are presented in the form of model or theory to the user in a meaningful manner.
v. **Knowledge discovery:** The scope of data mining is the knowledge extraction from large datasets with the help of computers.

Data mining is an interdisciplinary area of research that has its roots in databases, machine learning and statistics, and has contributions from many other areas such as information retrieval, pattern recognition, visualization, parallel and distributed computing. The applications of data mining include customer relationship management, fraud detection, market and industry characterization, stock management, medicine, pharmacology and biology. The extensive databases of biological information create both challenges and opportunities for development of novel KDD methods. The field of bioinformatics has many applications in the modern day world, including molecular medicine, industry, agriculture, stock farming and comparative studies.

III. BIOINFORMATICS

The term bioinformatics was coined by Paulien Hogeweg in 1979 for the study of informatic processes in biotic systems. Bioinfomatics is the field of science in which biology, computer science and information technologies merge to form a single discipline. It was primarily used in genomics and genetics involving large-scale DNA sequencing. At the beginning of the "genomic revolution", a bioinformatics concern was the creation and maintenance of a database to store biological information such as nucleotide and amino acid sequences [6]. Bioinformatics is the science of storing, extracting, organizing, analyzing, interpreting

and utilizing information from biological sequences and molecules. The actual process of analyzing and interpreting various types of data including nucleotide and amino acid sequences, protein domains and protein structures in the field of bioinformatics is also referred to as computational biology. Computational biology encompasses the use of algorithmic tools to facilitate biological analysis. Recent rapid developments in genomic and other molecular research technologies and developments in information technologies have combined to produce a tremendous amount of information related to molecular biology. The aims of bioinformatics according to Luscombe et al. [7] are: (i) the organization of data in such a way that allows researchers to access existing information and to submit new entries as they are produced, (ii) the development of tools that facilitates the analysis and management of data, (iii) the use of these biological data and tools to analyze and interpret the individual systems in order to gain new biological insights.

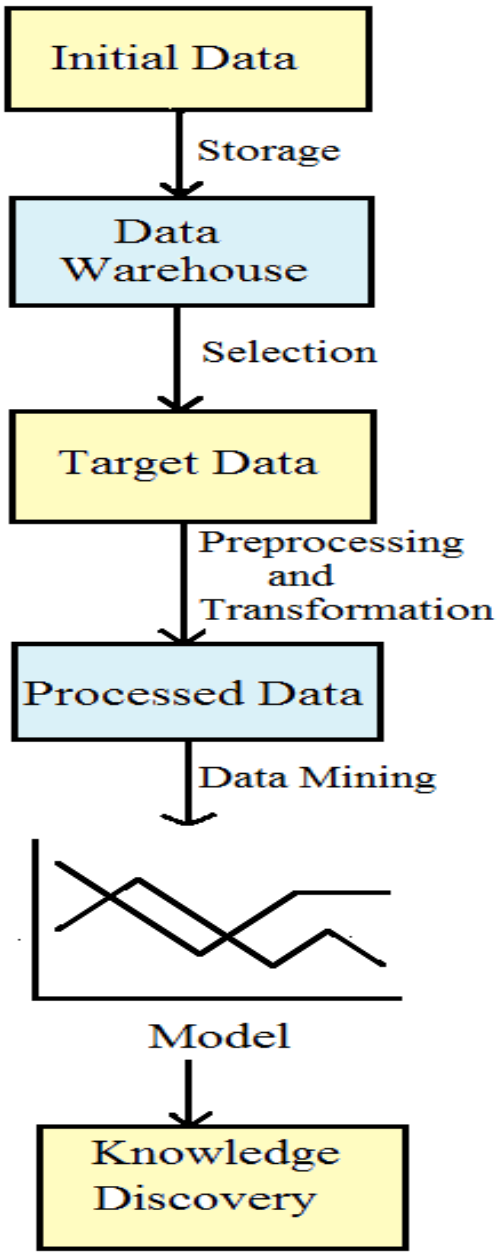

Figure 1. Diagrammatic representation of different sequence of steps foe knowledge discovery in biological data mining

There are three important sub-disciplines within bioinformatics: the development of new algorithms (mathematical formulas) and statistics with which to assess relationships among members of large datasets; the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains and protein structures, and the development and implementation of tools that enable efficient access and management of different types of information. A biological database is a large, organized body of persistent data, usually associated with computerized software designed to update, query and retrieve components of the data stored within the system. A simple database might be a single file containing many records, each of which includes the same set of information. For researchers to benefit from the data stored in a database, two additional requirements must be met: easy access to the information and a method for extracting only that information needed to answer a specific biological question.

Computation biology involves theoretical methods, mathematical modeling and computation simulation techniques to study the biological systems [8, 9]. It deals with the analysis, investigation, estimation and elucidation of dynamic interactions between genes and proteins in naturally occurring systems [10]. In this perceptive, the end goal is to finally engineer a system/organism to perform how we want it to perform [11]. The modeling component of synthetic biology allows the designing of biological circuits and the analysis of its expected behaviour. The experimental component merges models with real systems by providing quantitative data and sets of available biological units that can be used to construct circuits [12]. Therefore, the ability to create such systems address to innovative approaches for a wide range of applications such as bioremediation using biosensors, sustainable energy production and biomedical therapies [13].

The rationale for applying computational approaches to facilitate the understanding of various biological processes includes: a more global perspective in experimental designing of the ability to capitalize on the emerging technology of database mining. In this the process, testable hypotheses are generated regarding the function or structure of a gene or protein of interest by identifying similar sequences in better characterized organisms. Different biological problems considered within the scope of bioinformatics fall into the following main tasks.

- Alignment and comparison of DNA, RNA and protein sequences
- Gene finding and promoter identification from DNA sequences
- Interpretation of gene expression and microarray data
- Gene regulatory network identification
- Construction of phylogenetic trees for studying evolutionary relationship
- Protein structure prediction and classification
- Molecular design and molecular docking

IV. COMPUTATIONAL TOOLS USED FOR MINING OF BIOLOGICAL DATA

Data mining is the discovery of useful knowledge from databases and is the main step in the process known as Knowledge Discovery in Databases [5]. Other steps of the KDD process are the collection, selection and transformation of the data and the visualization and evaluation of the extracted knowledge. Data mining employs genetic algorithms [14] and techniques from statistics, machine learning, artificial intelligence, databases and data warehousing etc.

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. However, some biological data mining tools such as GeneSpring, Spot Fire, VectorNTI, COMPASS, Statistics for Microarray Analysis and Affymetrix Data Mining Tool have been developed [15]. Some biological data mining tools are also provided by National Center for Biotechnology Information and by European Bioinformatics Institute.

V. DATA MINING TASKS

Some of the most popular data mining tasks are classification, clustering, association, sequence analysis and regression. Depending on the nature of the data as well as the desired knowledge, there are a large number of algorithms for each task. All of these algorithms try to fit a model to the data [16]. Such a model can be either predictive or descriptive (Fig. 2). A predictive model makes a prediction about data using known examples, while a descriptive model identifies patterns or relationships in data. The main tasks well suited for data mining, which involves mining of meaningful new patterns from the data are:

(i) Classification: Classification is learning a function that maps (classifies) a data item into one of several redefined classes.

(ii) Estimation: With a given input data, it comes up with a value for some unknown continuous variable.

(iii) Prediction: Same as classification and estimation except that the records are classified according to some future behaviour or estimated future value.

(iv) Clustering: Segmenting a population into a number of subgroups or clusters.

(v) Association rules: It determines about the things, which go together and it is also called dependency modeling.

(vi) Description and visualization: It involves representation of the data using visualization techniques.

Learning from data falls into two categories: directed (supervised) and undirected (unsupervised) learning. The first three tasks i.e., classification, estimation and prediction are examples of supervised learning. The next three tasks i.e., association rules, clustering and description and visualization are examples of unsupervised learning. In unsupervised learning, no variable is singled out as the target and the goal is to establish some relationship among all the variables. Unsupervised learning attempts to find patterns without the use of a particular target field. The development of new data mining and knowledge discovery tools and their potential application is a subject of active research in modern biology.

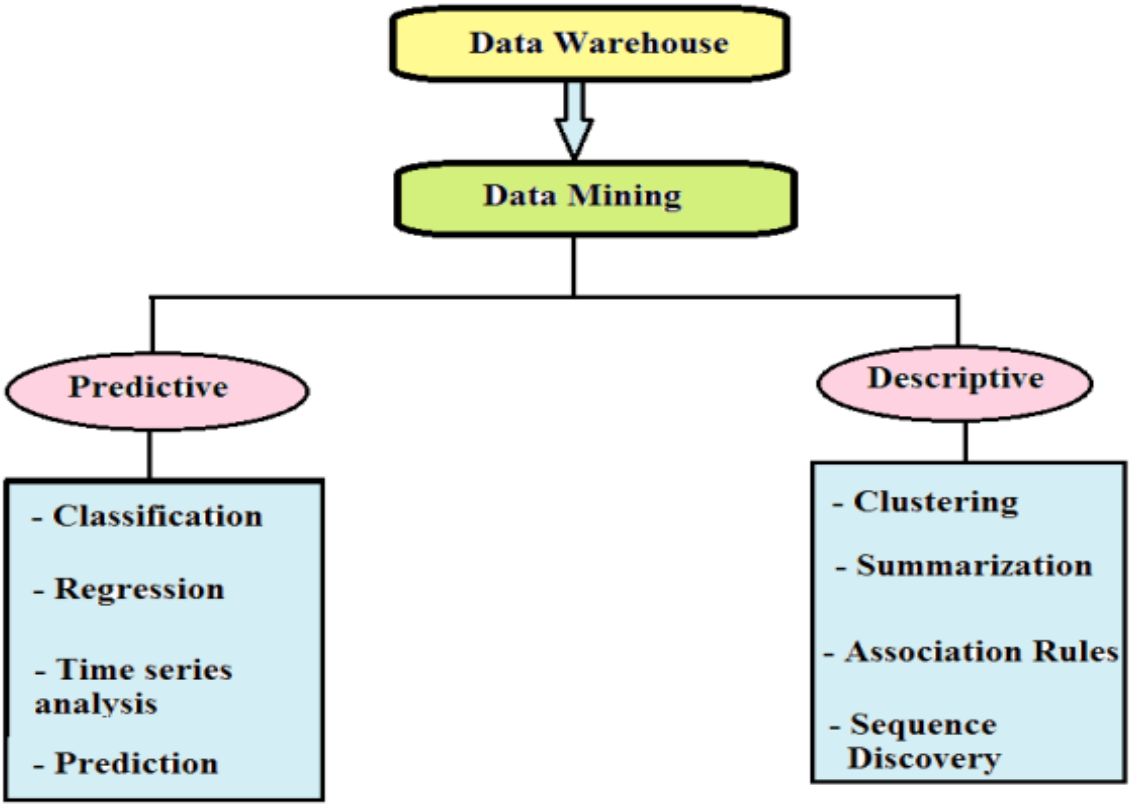

Figure 2. Common data mining tasks

VI. APPLICATIONS OF DATA MINING IN BIOINFORMATICS

Applications of data mining in bioinformatics include 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, and protein sub-cellular location prediction. For example, microarray technologies are used to predict a patient's outcome. On the basis of patients' genotypic microarray data, their survival time and risk of tumour metastasis or recurrence can be estimated. Machine learning can be used for peptide identification through mass spectroscopy [17]. Correlation among fragment ions in a tandem mass spectrum is crucial in reducing stochastic mismatches for peptide identification by database searching. An efficient scoring algorithm that considers the correlative information in a tuneable and comprehensive manner is highly desirable.

6.1. Genome analysis using data mining

DNA (deoxyribose nucleic acid) was recognized as the most important biological molecule of cellular organisms, wherein it serves to store genetic information DNA is located on the chromosomes. Its proper replication and transmission to next progeny is crucial for maintaining the genotype characteristics [18]. The ability to store billions of data as nucleotide base sequences is an important feature of the DNA. DNA is a polymer of nucleotides consisting of adenine, guanine, cytosine and thymine bases, commonly abbreviated to A, G, C and T, respectively. The DNA double helix is stabilized primarily by hydrogen bonds between nucleotides and base-stacking interactions among nucleotide bases. Both strands of the double-stranded structure store the same biological information. The chemical ties between the two base pairs are different i.e., C and G pair through three hydrogen bonds, while A and T pair through two hydrogen bonds. The different nucleotides i.e., ATP, GTP, CTP and TTP are polymerized by the DNA polymerase enzyme using a template of the parent strand. Each strand has, according to chemical convention, a 5' and a 3' end. Thus, a hypothetical DNA molecule having the following base sequences will have complementary bases as shown below:

AACGCGTACGTACAAGTGTCCGAATGGCCAATG
TTGCGCATGCATGTTCACAGGCTTACCGGTTAC

The arrangement of nucleotides varies from one organism to the other and different microbial species and organisms have different arrangement of nucleotides. The discrete segments of DNA coding for particular protein/polypeptide are called genes. Each person's genome has about three billion bases, having the capability to encode about 100,000 genes and these coding regions make up only about 10% of human's genome. The biological information encoded in genes is made available by gene expression [19, 20]. The information is transferred from DNA to mRNA by the process of transcription and this information is further translated into the protein by making use of the ribosomes. The hereditary information present in the nucleotide sequence is maintained intact by complex metabolism involving both replication and repair functions. Occasionally, mistakes may occur and it may result in alteration of a particular nucleotide in DNA. The change of even single base pair in particular gene of the DNA structure could lead to mutation and it may result in morphological and functional changes. These changes in DNA composition could result in evolution of new species.

The technique by which precise order of nucleotides in a piece of DNA either in full chromosome or entire genomes can be determined is termed as DNA sequencing [21]. Recently, automated DNA sequencing machines are capable of identifying 10,000 nucleotide base pairs per day and have become commercially available. The detection and analysis of sequencing reactions in second generation sequencing techniques 454 FLX and Solexa as well as in 3rd generation advanced sequencing techniques including PacBio SMRT and nanopore sequencing [22], is carried out by instruments controlled by computers. Different bioinformatics tools are used in these sequencing techniques like alignment tools, variation detection tools and assembly tools. The advent of rapid DNA sequencing methods has greatly accelerated biological and medical research and discovery. Knowledge of DNA sequences has become indispensable for basic biological research and in numerous applied fields such as diagnostic, biotechnology, forensic biology, virology and biological systematics.

(i) Sequence analysis

Many data mining techniques have been proposed to deal with the identification of specific DNA sequences. The most common techniques include neural networks, Bayesian classifiers, decision trees and Support vector machines (SVMs) [23, 24]. Sequence recognition algorithms exhibit performance tradeoffs between increasing sensitivity (ability to detect true positives) and decreasing selectivity (ability to exclude false positives) [25]. Another data mining application in genomic level is the use of clustering algorithms to group structurally related DNA sequences [26].

DNA sequence analysis is the most primitive operation in computational biology. It entails the prediction of genes in uncharacterized genomic sequences. The objective is to be able to take a newly sequenced uncharacterized genome and break it up into introns, exons, repetitive DNA sequences etc. and other elements. This operation consists of finding which part of the biological sequences are alike and which part differs during medical analysis and genome mapping processes. The sequence analysis implies subjecting a DNA or peptide sequence to sequence alignment, sequence databases, repeated sequence searches or other bioinformatics methods on a computer.

(ii) Genome annotation

In the context of genomics, annotation is the process of marking the genes and other biological features in a DNA sequence. Dr. Owen White designed the first genome annotation software system in 1995. The various components of genome analysis are:

- **Gene evaluation:** From a given DNA sequence, it is determined that what part of it codes for a protein and what part of it is junk DNA.
- **Genome classification:** The junk DNA is classified either as intron, untranslated region, transposons, dead genes and regulatory elements etc.
- **Gene prediction:** The coding regions in a newly sequenced genome is predicted into the genes (coding) and the non-coding regions

(iii) Comparative genomics

Comparative genomics is the study of the relationship of genome structure and function across different biological species. Gene finding is an important application of comparative genomics, as is discovery of new, non-coding functional elements of the genome. Comparative genomics exploits both similarities and differences in the proteins, RNA and regulatory regions of different organisms. Application of computational approaches in genome comparisons have recently become a common research topic in computer science.

6.2. Gene expression analysis

The technological advances in DNA synthesis and high-fidelity assembly of DNA fragments led to the developments and improvements of molecular biology and genetic engineering tools (27]. For metabolic engineering applications and the investigation of all biological functions alon with a finer control over the expression of genes in a given pathway could be designed and implemented [28]. Designed expression control for individual genes typically occurs either at the transcription or translation levels.

(i) Transcription: The first level of functional control for specific genes occurs during transcription. The transcriptional process involves binding of RNA polymerase on a DNA sequence (promoter) to initiate biosynthesis of mRNA. The analysis of naturally occurring promoter sequences showed the occurrence of conserved sequence motifs that physically bind to the sigma subunit of the RNA polymerase. Sequence variation in the promoter was found to affect transcriptional strength. In the early stages of the development of synthetic biology, transcriptional DNA parts were created and catalogued allowing the use of genetic variants for transcriptional control [29]. Controlled sequence-to-function relationships could then be extrapolated using mathematical correlation methods such as a position weight matrix (PWM). Recently, PWMs have been used to quantitatively describe the sequence-to-function relationship for promoters in *Escherichia coli* [30]. Moreover, synthetic promoters have been designed to produce a desired level of transcriptional strength with the recent modeling developments.

(ii) Translation: During translation, mRNA is translated into proteins. Translation initiates when a ribosome interacts with a ribosome binding site (RBS) and facilitates the subsequent tRNA binding to mRNA codons to produce polypeptides by the addition of amino acids. Translation involves three steps: Initiation, elongation and termination. Of these three steps, translation initiation is the rate-limiting step and different rates of translation initiation are found due to variation in the DNA sequence of the RBS within each cell. Synthetic biology has now developed a complement of experimental and computational tools to design and control individual gene expression levels at both the transcriptional and translational levels [31]. These tools enable a finer level of design control for biological systems and can be implemented for metabolic engineering applications.

The expression of many genes can be determined by measuring mRNA levels with various techniques such as microarrays, expressed cDNA sequence tag (EST) sequencing, serial analysis of gene expression (SAGE) tag sequencing, massively parallel signature sequencing (MPSS) or various applications of multiplexed in-situ hybridization etc. The major research area involves developing statistical tools to separate signal from noise in high-throughput gene expression studies. The main types of microarray data analysis include gene selection, clustering and classification [32]. 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. This increases the likelihood of finding "false positives". In gene expression analysis, the features are the genes. Gene selection is a process of finding the genes most strongly related to a particular class. Moreover, a large number of genes are irrelevant when classification is applied. The danger of overshadowing the contribution of relevant genes is reduced when gene selection is applied.

Clustering is the most used method in gene expression analysis [33, 34]. Clustering methods were classified in two categories: one-way clustering and two-way clustering (Table 1) [35]. Methods of the first category are used to group either the genes with similar behaviour or samples with similar gene expressions. On the other hand, two-way clustering methods are used to simultaneously cluster genes and samples. Hierarchical clustering is currently the most frequently applied method in gene expression analysis. An important issue concerning the application of clustering methods in microarray data is the assessment of cluster quality. Many techniques such as bootstrap, repeated measurements, mixture model-based approaches and sub-sampling etc. have been proposed to deal with the cluster reliability assessment [36-38].

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. Detailed descriptions of different methods used in gene expression analysis include class discovery and class prediction [39], identification of genes with similar expression patterns [40] and gene expression data using Plaid models [41]. Most of the methods used to deal with microarray data analysis can be used for SAGE data analysis. Finally, machine learning and data mining can be applied in order to design microarray experiments [18].

Table 1. Popular microarray data mining methods

One-way clustering	**Two-way clustering**	**Classification**
Hierarchical clustering Self-organizing maps (SOMs) K-means singular value decomposition (SVD)	Block clustering Gene shaving Plaid models	SVMs K-nearest neighbors Classification/Decision trees Voted classification Weighted gene voting Bayesian classification

6.3. 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 and 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 [42]. Neural networks have also been used to predict protein properties such as stability, globularity and shape. The use of hierarchical clustering algorithms could be used for predicting protein domains. Data mining has been applied for the prediction of protein secondary structure and many techniques have been developed. Besides, initial statistical approaches, more accurate techniques based on information theory, Bayes theory, nearest neighbors and neural networks were developed. A density based clustering algorithm (GDBSCAN) can be used to deal with protein interactions [43]. This algorithm is able to cluster point and spatial objects according to both spatial and non-spatial attributes. Prediction of three dimensional structures of drug targets, design of biocatalysts and nano biomachines are a few of the multitude of foreseeable applications. A computational protocol is being developed for modeling and predicting protein structures at the atomic level.

(i) Analysis of protein expression

Gene expression is measured in many ways including mRNA and protein expression. However, protein expression is one of the best clues of actual gene activity since proteins are usually final catalysts of cell activity. Protein microarrays and high throughput (HT) mass spectrometry (MS) can provide a snapshot of the proteins present in a biological sample. Bioinformatics is involved in making sense of protein microarray and HT MS data.

(ii) Protein structure prediction

The amino acid sequence of a protein (primary structure) can be easily determined from the sequence on the gene that codes for it. In most of the cases, this primary structure uniquely determines a structure in its native environment. Knowledge of this structure is vital in understanding the function of the protein. Structural information is usually classified as secondary, tertiary and quaternary structure. Protein structure prediction is one of the most important for drug design and for the designing of novel enzymes.

(iii) Protein-protein docking

In the last two decades, three-dimensional structures of tens of thousands of proteins have been determined by X-ray crystallography and protein nuclear magnetic resonance spectroscopy (protein NMR). Biological scientists predict possible protein-protein interactions based on these 3D shapes, without doing protein-protein interaction experiments. A variety of methods have been developed to tackle the protein-protein docking problem [44].

6.4. Analysis of mutations in cancer

In cancer, the genomes of affected cells are rearranged in complex or even unpredictable ways. Massive sequencing efforts are used to identify previously unknown point mutations in a variety of genes in cancer. Bioinformaticians continue to produce specialized automated systems to manage the sheer volume of sequence data produced and they create new algorithms and software to compare the sequencing results to the growing collection of human genome sequences and germline polymorphisms. New physical detection technologies are employed such as oligonucleotide microarrays to identify chromosomal gains and losses and single-nucleotide polymorphism arrays to detect known point mutations. Another type of data that requires novel informatics development is the analysis of lesions found to be recurrent among many tumours.

Metabolic engineering has the potential to engineer novel diagnostic and therapeutic strategies for relatively intractable medical conditions such as cancer and infectious diseases. A critical shortcoming in cancer treatments has been their inability to distinguish between cancerous and normal cells, but one reliable signature of tumor growth is hypoxia. Anderson et al. [45] engineered *Escherichia coli* to invade mammalian cells selectively in hypoxic environments. Recently, Wright et al. [46] employed similar principles to link enzymatic activity to a cancer marker of hypoxia. HIF-1a is a hypoxia-inducible factor selectively found in cancer cells. The activity of a segment from p300, a binding partner of HIF-1a, was coupled to the activity of cytosine deaminase, an enzyme that converts the relatively benign prodrug 5-fluorocytosine to the chemotherapeutic 5-fluorouracil. This enabled selective activity of the drug within cancer cells, which could result in significant improvement in the side effects typical of chemotherapy.

6.5. Modeling of biological systems

Modeling of biological systems is a significant task of systems biology and mathematical biology. Computational systems biology aims to develop and use efficient algorithms, data structures, visualization and communication tools for the integration of large quantities of biological data with the goal of computer modeling. It involves the use of computer simulations of biological systems like cellular subsystems such as the networks of metabolites and enzymes, signal transduction pathways and gene regulatory networks to both analyze and visualize the complex connections of these cellular processes. Artificial life is an attempt to understand evolutionary processes via the computer simulation of simple life forms within the laboratory [47] and a broader challenge in synthetic biology is to engineer existing genomes for bio-manufacturing or to decipher the principles that govern the operation of biological systems [48, 49].

The rapid improvements in DNA synthesis and enhanced assembly techniques have enabled the construction of entire genomes. Recently, synthesis capabilities have progressed from a *Mycoplasma* genome of 582,970 base pairs [50] to a 1.08-mega-base-pair *Mycoplasma* genome transplanted into a recipient cell lacking a genome [51]. Dymond et al. [52] reported the remarkable synthesis of the right arm of chromosome IX in yeast and a portion of chromosome VI. These genomes were integrated into yeast cells with minimal phenotypic variation in growth and gene expression. This work provides a valuable method of studying the yeast genome and adapting yeast to specific applications such as biosynthesis.

Computational technologies are used to accelerate or fully automate the processing, quantification and analysis of large amounts of high-information-content biomedical images. A fully developed image analysis system may completely replace the observer. Therefore, biomedical imaging is becoming more important for both diagnostics and research. Some of the examples of research in this area include clinical image analysis and visualization, inferring clone overlaps in DNA mapping and Bioimage informatics etc.

6.6. Drug design

Metabolic engineering utilizes biological information to genetically modify the cellular function for production of a targeted chemical, drug or protein product. Successful metabolic engineering results are based on directed designs built upon biological databases or combinatorial screening that uses high-throughput experimental techniques. Computational tools have been developed to utilize biological data in the analysis and design of microbial strains for metabolic engineering [53].

Designing of a novel drug is one of the biggest challenges faced by the pharmaceutical industry. The use of computers accelerates the process of drug design which is a time intensive process and also reduces the cost of whole process. Computational methods are used in various forms of drug discovery like QSAR, virtual screening and structure-based drug designing methods. Among these, structure based drug design is gaining importance due to rapid growth in structural data (available in nucleic acid data bank). This structural data can be used in molecular modeling to design lead molecules based on the structural features of the active site.

VII. CONCLUSIONS

Bioinformatics and data mining are developing as interdisciplinary science. Bioinformatics provides the opportunities for development of novel and improved data mining methods. Active and collaborative research by the academia as well as the industry is needed because of the special characteristics of biological data and the extremely high importance of bioinformatics research. In metabolic engineering, major components of metabolism could be completely redesigned for more efficient utilization of resource pools to minimize material drains. It often incorporates the most recent biological database and tools at all levels of biological organization within a cell. Using genome-scale models and optimization of algorithms, metabolic network analysis and designs can be achieved [54]. Designed tools coupled with rapid DNA synthesis and assembly technologies have accelerated the prototyping, tuning and deployment of synthetic biological systems for various applications [9].

However, data mining in bioinformatics is hampered by many facets of biological databases including their size, number, diversity and the lack of a standard ontology to aid the querying of

heterogeneous data and the information they contain. Other problems in bioinformatics are the accurate prediction of protein structure and analysis of gene behaviour in microarrays. Recent technologies will enable the prediction of protein function in the context of regulation of gene expression, metabolic pathways involving a series of chemical reactions catalyzed by enzymes within a cell and signaling cascades consisting of reactions which occur as a result of a single stimulus. Novel genetic circuits with useful applications have been constructed through rational design and forward engineering by the synthetic biologists and efficient strategies have been described for fine-tuning of genetic circuit characteristics [55]. In future, predictive computational models need to be developed that could be validated by experimentation and applicable across many host organisms.

REFERENCES

[1] Mount, D. W., "Bioinformatics: Sequence and genome analysis". Spring Harbor Press, (2002).

[2] Cristianini, N. and Hahn M., "Introduction to computational genomics", Cambridge University Press, 2006. ISBN 0-5216-7191-4.

[3] Sindhu, S. and Sindhu, D., "Development of computational tools for metabolic engineering". International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, issue 5, pp. 9208-9217, 2016a.

[4] Li, J., Wong, L. and Yang, Q., "Data mining in bioinformatics", IEEE Intelligent System, IEEE Computer Society, 2005.

[5] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R., "Advances in knowledge discovery and data mining". AAAI Press/MIT Press, Menlo Park, California, USA, 1996.

[6] Sindhu, D. and Sindhu, S., "Biological computers: Their application in gene mining and protein engineering", International Journal of Technical Research, Vol. 4, issue 3, pp. 15-21, 2015.

[7] Luscombe, N.M., Greenbaum, D. and Gerstein, M., "What is bioinformatics? A proposed definition and overview of the field". Methods of Information in Medicine, Vol. 40, issue 4, pp. 346-358, 2001.

[8] Elizabeth P., Isaac L. and Kevin T., "Computational modeling approaches for studying of synthetic biological networks", Current Bioinformatics, Vol. 3, pp. 1-12, 2008.

[9] Sindhu, D. and Sindhu, S., "Computational programming for product designing in synthetic biology", International Journal of Innovative Rsearch in Science, Engineering and Technology, Vol. 5, issue 5, pp. 8095-8103, 2016b.

[10] Tanaka R.J., Okano H. and Kimura H., "Mathematical description of gene regulatory units", Biophysics Journal, Vol. 91, pp. 1235-1247, 2006.

[11] Cameron D.E., Bashor C.J. and Collins J.J., "A brief history of synthetic biology", Natural Review Microbiology, Vol. 12, pp. 381-390, 2014.

[12] Chandran D., Bergmann F.T. and Sauro, H.M., "Tinker Cell: modular CAD tool for synthetic biology, Journal Biological Engineering, Vol. 3, pp. 19-36, 2009.

[13] Purnick P.E. and Weiss R., "The second wave of synthetic biology: from modules to systems. Natural Review of Molecular and Cellular Biology, Vol. 10, pp. 410-422, 2009.

[14] Chadha, P. and Singh, G.N., "Classification rules and genetic algorithms in data mining", Global Journal of Computer Science and Technology Software and Data Engineering, Vol. 12, issue 15, pp. 1-5, 2012.

[15] Han, J., "How can data mining help bio-data analysis"? In: Zaki, M.J., Wang, J.T.L. and Toivonen, H.T.T. (Eds). Proceedings of the 2nd ACM SIGKDD Workshop on data mining in bioinformatics, Vol. 1-2, 2002.

[16] Dunham, M.H., "Data mining: Introductory and advanced topics". Prentice Hall, Upper Saddle River, New Jersey, USA. 2002.

[17] Molla, M., Waddell, M., Page, D. and Shavlik, J., "Using machine learning to design and interpret gene-expression microarrays", AI Magazine, Vol. 25, issue 1, pp. 23-44, 2004.

[18] Krieger, M., Scott, M.P, Matsudaira, P.T., Lodish, H.F., Darnell, J.E., Lawrence, Z., Kaiser, C. and Berk, A., "Structure of nucleic acids, Section 4.1", Molecular Cell Biology. New York: W.H. Freeman and Co. 2004.

[19] Jacob, F. and Monod, J., "Genetic regulatory mechanisms in the synthesis of proteins", Journal of Molecular Biology, Vol. 3, pp. 318–356, 1961.

[20] Sindhu, S. and Sindhu, D., "Engineering of gene circuits for cellular computation", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 5, issue 7, pp. 12833- 12840, 2016c.

[21] Sanger, F., "Sequences, sequences, and sequences", Annual Review of Biochemistry, Vol. 57, pp. 1–28, 1988.

[22] Shendure, J. and Hanlee, J., "Next generation DNA sequencing", Nature Biotechnology, Vol. 26, pp. 1135–1145, 2008.

[23] Ma, Q. and Wang, J.T.L., "Biological data mining using Bayesian neural networks: A case study", International Journal on Artificial Intelligence Tools, Special Issue on Biocomputing, Vol. 8, issue 4, pp. 433-451, 1999

[24] Zien, A., Ratsch, G., Mika, S., Scholkopf, B., Lengauer, T. and Muller, R.-K., "Engineering support vector machine kernels that recognize translation initiation sites. Bioinformatics, Vol. 16, issue 9, pp. 799-807, 2000.

[25] 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. Available: http://www.biodatabases.com/whitepaper.html

[26] Oyelade, J., Isewon, I., Oladipupo, F., Aromolaran, O., Uwoghiren, E., Ameh, F., Achas, M. and Adebiyi, E., "Clustering algorithms: Their application to gene expression data", Bioinformatics and Biology Insights, Vol. 10, pp. 237-253, 2016.

[27] Annaluru, N., Muller, H., Mitchell, L.A., Ramalingam, S. and Stracquadanio, G., "Total synthesis of a functional designer eukaryotic chromosome", Science, Vol. 344, pp. 55-58, 2014.
[28] Covert, M.W. and Palsson, B.O., "Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*", Journal of Biological Chemistry, Vol. 277, pp. 28058-28064, 2002.
[29] Kelly, J.R., Rubin, A.J., Davis, J.H., Ajo-Franklin, C.M. and Cumbers, J., "Measuring the activity of BioBrick promoters using an *in vivo* reference standard", Journal of Biological Engineering, Vol. 3, pp. 4-9, 2009.
[30] Rhodius, V.A. and Mutalik, V.K., "Predicting strength and function for promoters of the *Escherichia coli* alternative sigma factor σE", Proceedings of National Academy of Sciences, USA, Vol. 107, pp. 2854-2859, 2010.
[31] Mutalik, V.K., Guimaraes, J.C., Cambray, G., Lam, C. and Christoffersen, M.J., "Precise and reliable gene expression via standard transcription and translation initiation elements", Natural Methods, Vol. 10, pp. 354-360, 2013.
[32] Piatetsky-Shapiro, G. and Tamayo, P., "Microarray data mining: Facing the challenges", SIGKDD Explorations, Vol. 5, issue 2, pp. 1-5. 2003
[33] Pirim H, Ekşioğlu B, Perkins AD, Yüceer Ç., "Clustering of high throughput gene expression data", Computer Operation Research, Vol. 39, issue 12, pp. 3046–3061, 2012.
[34] Chandrasekhar, T., Thangavel, K. and Elayaraja E., "Effective clustering algorithms for gene expression data", International Journal of Computer Applications, Vol. 32, issue 4, pp. 25–29, 2011.
[35] Tibshirani, R., Hastie, T., Eisen, M., Ross, D., Botstein, D. and Brown, P., "Clustering methods for the analysis of DNA microarray data (Tech. Rep.)", Department of Statistics, Stanford University, Stanford, California, USA. 1999.
[36] Kerr, M.K. and Churchill, G.A., "Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments", Proceedings of the National Academy of Sciences, Vol. 98, pp. 8961-8965. 2001.
[37] Smolkin, M. and Ghosh, D., "Cluster stability scores for microarray data in cancer studies", BMC Bioinformatics, Vol. 4, pp. 36-42, 2003.
[38] Yeung, Y.K., Medvedovic, M. and Bumgarner, R.E., "Clustering gene- expression data with repeated measurements", Genome Biology, Vol. 4, issue 5, pp. R34, 2003.
[39] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S., "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring", Science, Vol. 286, issue 5439, pp. 531-537, 1999.
[40] Hastie, T., Tibshirani, R., Eisen, M.B., Alizadeh, A., Levy, R., Staudt, L., Chan, W.C., Botstein, D. and Brown, P., "Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns", Genome Biology, Vol. 1, issue 2, pp. research0003, 2000.
[41] Lazzeroni, L. and Owen, A., "Plaid models for gene expression data", Statistica Sinica, Vol. 12, pp. 61-86, 2002.
[42] Whishart, D.S., "Tools for protein technologies", In: Sensen, C.W. (Ed.), Biotechnology, Vol 5b, Genomics and Bioinformatics, Wiley-VCH. pp. 325-344, 2002.
[43] Sander, J., Ester, M., Kriegel, P.-H. and Xu, X., "Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications", Data Mining and Knowledge Discovery, Vol. 2, issue 2, pp. 169-194, 1998.
[44] Lee, K., "Computational study for protein-protein docking using global optimization and empirical potentials", International Journal of Molecular Science, Vol. 9, pp. 65-77, 2008.
[45] Anderson, J.C., Clarke, E.J., Arkin, A.P. and Voigt, C.A., "Environmentally controlled invasion of cancer cells by engineered bacteria", Journal of Molecular Biology, Vol. 355, pp. 619-627, 2006.
[46] Wright, C.M., Wright, R.C., Eshleman, J.R. and Ostermeier, M., "A protein therapeutic modality founded on molecular regulation", Proceedings of National Academy of Sciences, USA, Vol. 108, Issue, 39, pp. 16206-16211, 2011.
[47] Elowitz, M. and Lim, W.A., "Build life to understand it", Nature, Vol. 468, pp. 889–890, 2010.
[48] Nandagopal, N. and Elowitz, M.B., "Synthetic biology: integrated gene circuits", Science, Vol. 33, pp. 1244-1248, 2011.
[49] Duschak, V.G., "Synthetic biology: Computational modeling bridging the gap between *in vitro* and *in vivo* reactions", Current Synthetic and Systems Biology, Vol. 3, Issue 2, pp. 127-142, 2015.
[50] Gibson, D.G., Benders, G.A., Andrews-Pfannkoch, C., Denisova, E.A. and Baden-Tillson H, et al., "Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome", Science, Vol. 319, pp. 1215–1220, 2008.
[51] Gibson, D.G., Glass, J.I., Lartigue, C., Noskov, V.N. and Chuang, R.Y. et al., "Creation of a bacterial cell controlled by a chemically synthesized genome", Science, Vol. 329, pp. 52–56, 2010.
[52] Dymond, J.S., Richardson, S.M., Coombes, C.E., Babatz, T. and Muller, H. et al., "Synthetic chromosome arms function in yeast and generate phenotypic diversity by design", Nature, Vol. 477, pp. 471-476, 2011.
[53] Ma, K.C., Perli, S.D. and Lu, T.K., "Foundations and emerging paradigms for computing in living cells", Journal of Molecular Biology, Vol. 428, pp. 893-895, 2016.
[54] Farasat, I.. Kushwaha, M., Collens, J., Easterbrook. M. and Guido, M., "Efficient search, mapping, and optimization of multi-protein genetic systems in diverse bacteria", Molecular System Biology, Vol. 10, pp. 731-736, 2014.
[55] Bradley, R.W., Buck, M. and Wang, B., "Tools and principles for microbial circuit engineering, Journal of Molecular Biology, Vol. 428, Issue, 5, pp. 862-888, 2016.

