IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 6, Ver. V (Nov.-Dec. 2016), PP 01-06
www.iosrjournals.org

Fundamentals of Molecular Biology for Gene Analysis

V.Sujatha [(1)], Dr.Shaheda Akthar[(2)]
[(1)]*Research Scholar, Department Of CSE, Acharya Nagarjuna University, Guntur and Assistant Professor,Vignan's Nirula Institute Of Technology and science for women,Pedapalakaluru.*
[(2)]*Head Of The Department, Dept of Computer Science, Govt. College for Women(A),Guntur*

Abstract: *Biological data analysis and data management is very useful, interesting and mostly needed dynamic research area. Biological data details are usually represented by means of data sequences for ease of analysis of bio relationships. State-of-the-art, automated, efficient, effective, scalable, and interoperable bio data analysis and data management techniques are needed. Accurate and automated data mining techniques are needed to analyze, process, compare biological data sequences of data in order to discover relationships among the patterns of bio data sequences. Specific and convenient indexing techniques are needed for efficient, effective and scalable management of very large number of biological data sequences because time complexity of biological data sequences is exponential. Many dynamic programming based analysis techniques of biological data sequences are available in the literature of bio information data management. Present study is the foundation for developing powerful indexing techniques as well as cancer data type classification techniques.*

Keywords: *DNA, RNA, SAGE, NCBI, gene expression, biological data, DNA microarray,*

I. Introduction

The **National Center for Biotechnology Information** (**NCBI**) has formulated the **Gene Expression Omnibus (GEO)**. It is a data repository facility which includes data on gene expression from different sources. It is therefore appreciable that the knowledge obtained from microarray gene expression analysis will probably increase our basic understanding of the cause and consequences of diseases (pathogenesis). The rapid increase in the quantity and quality of available biological data causes for the introduction of distributed and massively parallel methods for gene expression measurements analysis and state of the art methods and techniques are needed for applying efficient computational techniques for gene data analysis. Evolutionary algorithms (EAs) are needed for classification of gene data. EAs may constitute an efficient method for optimal gene selection, and can also help in reducing the size (number of features used) of classifiers. In many real life applications, the classification accuracy obtained using evolutionary algorithms, often in conjunction with other state of the art methods represents a significant improvement over the results obtained without the use of evolutionary algorithms.

Data mining techniques including many advanced data mining techniques allow many provisions for finding interactive and interrelationships present in the complex data of gene expression profiles. Data mining offers correct, precise and accurate learning and discovery of many useful patterns in decision making. Many machine learning and statistical techniques are available for selecting set of potential genes called discriminatory genes, which are used for creating relationships between types of samples versus gene expression values.

The collection of deoxyribonucleic acid sequences that are needed to represent complete details of any living creature is called a genome. Genome is divided into a set of genes and each gene is represented as one sequence of small components. First, gene expression data set has very unique characteristics which are very different from all the previous data used for classification. Most publicly available gene expression data has the following properties[2]:

- high dimensionality: up to tens of thousands of genes,
- very small data set size: less than 100, and
- most genes are not related to cancer classification.

Gene expression data set has very unique characteristics which are very different from all the previous data used for classification [2]. Various performance measures that must be considered are – classification accuracy, computational time, selecting best subset of genes. Since the amount of gene expression data available is small, the classification accuracy of various algorithms cannot be compared extensively [2]. There are several scalable decision tree algorithms available, as the date size increases [2]. In the future decision tree classifier models will be considered in greater detail for obtaining in depth analysis of cancer classification subtypes with reasonably accurate classification results, performance issues, robust and efficient, effective results.

II. Terminologies of Molecular Biology

DNA stands for deoxyribonucleic acid.
RNA stands for ribonucleic acid
SAGE stands for serial analysis of gene expression

Genes are categorized into three types:
1) Protein-coding genes
2) RNA- specifying genes
3) Un-transcribed genes

RNAs are classified as
1) Messenger RNA (mRNA)
2) Transfer RNA (tRNA)
3) Ribosomal RNA (rRNA)

Human body contains collection of cells. A subgroup of cells is called a cell population. A cell population is represented as expression profiles of genes. Expression profiles of genes are divided into small together for efficient and effective sequencing[3]. Gene tag sequences in the gene expression profile are frequently used in many human studies and standard libraries of gene tags gene libraries have been developed from different cells or tissues such as
1) Dendric cell
2) lling fibroblast cells
3) Oocytes
4) Thyroid tissue
5) B-cell lymphoma
6) Cultured keratinocytes
7) Muscles
8) Brain tissues
9) Sciatic nerve
10) Retina
11) Macula
12) Skin cells
13) Retinal pigment epithelial cells
14) Cord blood-derived mast cells

Human body is treated as collection of cells. Cells are the basic building block of any living creation. D**NA** is responsible for completing cell related tasks by executing a set of needed instructions. Components for DNA are called nucleotides. Collection of DNA area arranged in a particular fashion is called a DNA sequence. Sometimes DNAs are called Fundamental units of any living organism. The collection DNA sequences that are needed to represent complete details of any living creature are called a genome. The size of genome varies greatly from one living creature to other living creature.

Protein molecules are generated by protein coding genes. RNA-specifying genes are responsible for chemical reactions. After chemical reactions different types of RNA-specifying molecules are generated. Un-transcribed genes are responsible for performing a set of functions. DNA has self replicated property and also generates additionally other molecules called RNAs. RNA stands for ribonucleic acid. Different types of RNA molecules are generated by genome.

III. DNA Microarray Technique

A DNA microarray is also known as chip or biochip. It is a collection of microscopic DNA spots and at least DNA spots are attached to a solid surface. DNA microarray allows researchers to analyze and study multiple genes simultaneously. Each DNA spot is a collection of a specific DNA sequences. A sequence is a collection of probes. Fundamental principal of micro array is hybridization between two DNA spots. DNA spots are called features and each feature is a collection of probes. Each location in DNA array contains thousands of features. The process of measuring gene expression via CDNA is called gene expression profiling or gene expression analysis. Expression levels of thousands of genes are analyzed simultaneously in gene expression profiling for finding different types of diseases effects of certain treatments and so on. Microarray based gene expression profiling allows researchers to identify gene whose expressions are changed in response to changes in organisms by comparing infected and uninfected gene cells or tissues[4].

DNA Microarray technology helps many biological researchers to learn more about different diseases such as heart diseases, mental disorders or illness, infectious diseases and particularly the study of cancer type classifications. Until recently, different types of cancer disease types have been classified on the basis of the organs in which the tumors develop. Now, with the evolution of latest microarray technology, it will be possible for the many biological researchers to further classify the types of cancers on the basis of the patterns of gene activity in the tumor cells. This will tremendously help the pharmaceutical community to develop more effective drugs as the treatment strategies will be targeted directly to the specific type of cancer[5].

Classification techniques can be used in microarray analysis to predict class label of the sample. Many researchers provide practical comparison methods for the classification of tumors using gene expression data. Many tools are available for classifying the gene expression data. Some of the tools are:

1) Discriminant analysis
2) Linear technique,
3) Logistic technique
4) Discrimination techniques
5) Classification and regression trees (CART) (Tree-based algorithms)
6) Generalized additive models and
7) Neural networks

The classification of different tumor types is of great importance in cancer diagnosis and drug discovery. DNA microarray technique allows analyzing and processing thousands of gene expressions simultaneously[6]. For efficient and effective analysis and processing of different types of cancer related data certainly it is needed different types of gene selection methods. Gene selection methods are considered as an integral preprocessing step for cancer classification. No doubt, cancer research is one of the most important areas in medical field. Traditional cancer classification methods are not suitable for many real life applications. In order to know and apply best latest methods for the problem of cancer classification, systematic approaches based on global gene expression analysis are necessarily needed.

Different types of classification and clustering methods from statistical and machine learning fields have been applied to cancer classification. Present study proposes latest methods and tools for cancer classification. The gene expression data is very different from any of the traditional data methods.

1) The gene expression data is high dimensional
2) Gene expression data size may vary from thousands to ten thousands.
3) Publically available data sizes are very small
4) In many cases the available data is irrelevant

Previously existing methods may not have the capability to handle such cases. Most important point to be considered is how to select the most important genes for obtaining accurate and efficient cancer classifications.
Requirements of cancer classification methods

1) High classification accuracy (Less misclassification accuracy)
2) Selecting the best data mining technique
3) Ability to manage large data (provision is needed for data scalability test)
4) Intelligent, automated and online based results of cancer and tumor classification methods are needed
5) Cheap and accurate drug discovery techniques are needed
6) Results must be completely based on gene expression details
7) Gene selections, subsets of gene selections, gene pruning, interrelationships between and among genes
8) Alternate methods are needed for cancer and tumor classification
9) Thorough understanding of existing methods for cancer classification is needed[7].

Many data mining and statistical data analysis techniques are available for accurate, efficient, effective, precise and functionally robust analysis of gene expression data. Many statistical impurity measures such as entropy, Gini Index, Max Minority, and the Twoing rule are useful for selecting, analyzing and then deriving desired patterns from various gene libraries of human beings. Jaccard and vector cosine similarity measuring techniques are available for human gene analysis. Many multiclass cancer gene expression data sets are available publically for analysis. Some of the points to be considered during analysis of cancer data are:

1) Traditional cancer disease type classification techniques are computationally expensive.
2) Advanced cancer disease type classification methods are needed and these methods must be efficient, effective, robust, scalable and high accurate.

Following details are needed for analyzing microarray gene expression data for cancer disease classification:

1) Cellular mechanisms of genes

2) The regulatory functions of genes
3) The functions of genes and proteins
4) The structure of gene networks and pathways

The output of the analysis of gene expression profiles is the risk of being affected by cancer disease. Microarrays can be manufactured in different ways based on the number of probes under examination, array costs, customized requirements, types of diseases and the types of queries that will be asked[8]. The microarray size may vary from minimum size ten probes to maximum of 2.1 million probes in commercial microarray. Different techniques are available for fabricating microarrays. A microarray is a grid of probes.

Advanced microarray technology is a popular tool for gene expression profiling in predicting the different types of cancer disease types. Microarray cancer data is organized intelligently between sets of sample data and genes types and then tissue samples are classified into different types of tissue classes. Microarray cancer data is particularly useful for finding potential gene markers that cause different cancer subtypes. Different cancer types are diagnosed and determined using potential gene markers. Small training data sets are reasonably not good in deriving best decision classifiers. Traditional supervised decision classifiers are constructed using labeled training data sets. Microarray cancer data contains both labeled and unlabeled data. To process microarray cancer data, both supervised learning (for labeled data) and unsupervised learning (for unlabeled data) are needed and sometimes these two techniques must be executed in an interleaved fashion or separately[9]. Main problem with cancer disease type analysis is how to find potential gene markers from the microarray cancer data once potential gene markers are identified then the next step is how to classify cancer disease type based on these potential gene markers.

Details of genes present in the human body are represented by means of expression called gene expressions. These gene expressions are carefully analyzed for predicting cancer subtype using different types of data analysis techniques emerged from diversified fields such as machines clustering, decision tree classification, fuzzy logic, k-nearest neighbors, support vector machines, rough set approach, Bayesian, associative classification, different types of clustering methods, and so on.

Internal data details of human body are represented using terabytes of memory that contains string encoded data[10]. Many software tools are available for string data similarity searches and some of the software tools are

1) Sim rank: It is a rapid and sensitive general-purposes k-mer search tool.
2) SAGE tool: serial analysis of gene expression is a powerful tool, which provides quantitative and comprehensive expression profile of genes in a given cell population.

Sim rank provides molecular ecologists with a high-throughput, open source choice for comparing large sequence sets to find similarity. Simrank stand alone utility tools rapidly searches and identifies database string which are similar to the given query strings.

An efficient, effective, general purpose, open-source rapid, stands alone and flexible tools are needed but currently such tools are not available. Information details of genes are represented as a set of expressions. DNA molecules are responsible for generating mRNAs and mRNAs are translated into amino acid sequences of proteins that execute different types of cellular functions. Transforming DNA sequences for genes into RNA is called gene expression. Expression level of a gene increases as the number of RNAs increases. Different biological states are represented by using different patterns of gene expressions. Cells and tissues are responsible for generating different gene expressions. DNA microarrays are used for finding the effectiveness of gene expression patterns for identifying different gene functions and cancer diagnosis.

DNA microarrays and serial analysis of gene expressions are two latest technologies that are used for measuring the thousands of genome-wide expression values in parallel. The new molecular biology method called cDNA microarray analysis expands probe hybridization methods and as a result thousands of genes are accessed at once. An array called cDNA microarray that contains thousands of DNA sequences printed on a high density glass microscope.

SAGE stands for serial analysis of gene expression and it is a technique that is used for obtaining profile details of cellular gene expression. SAGE quantifies the gene details using a special tag. SAGE manipulates a list of tags using digital representation of cellular gene expression. Many real life applications of cancer classification methods use DNA microarray expression data.

IV. Cancer Classification Details

Cancer is a life threatening disease that needs effective diagnosis such as predicting accurate symptoms at an early stage, correctly and distinctly. Analyzing and then designing sophisticated set of algorithms is a crucial part for deriving different types of pattern from microarray technology it is possible to analyze and

process thousands of gene expressions parallel. The exponential growth rate of micro array data size poses problems for many computational scientists in understanding biologically significant cellular mechanisms.
Data clustering is an unsupervised learning technique in data mining. Clustering techniques are useful in understanding the biological process details of human beings by understanding the relationships between genes and different types of disease states.

Clustering technique is preformed on the genes or data samples to identify clusters of genes that have similar expression patterns or clusters of samples that have similar expression profiles. These exists different types of clustering techniques on similar functionalities of genes. Gene clustering algorithm creates clusters of correlated genes. Sometimes it is necessary to know how genes collectively react under certain conditions.

The data sets that are available for gene analysis must be preprocessed using standardization and normalization. In general, normalization is performed using the z-score method and stand and deviational. Normalized data is said to be standardized data.

Usually gene expression data sets contain less number of samples and more number of features one best way is rank all the features in the data set before processing gene data. In machine learning and statistical learning many technical measures are available for ranking the features of genes. Popular statistical learning many technical measures is available for ranking the features of genes. Popular statistical impurity based measures – Gini Index(GI), Max minority(mm), and the towing rule(TR) are frequently used to extract the relevant features. Very simplest idea is that before actually using the original data set convert or translates the original data set into reduced feature data set. Main advantage of gene ranking is tha if a particular gene is highly ranked then the other genes which are correlated with this gene are also likely to have high ranks. Useful decisions are taken based on high ranks. Useful decisions are taken based on these correlations between highly ranked genes.

Biological relationships are analyzed in terms of association rule mining techniques and particularly using rank based methods. Biological relationships are very difficult to analyze because of their density and genes exhibit high complex relationships. Gene correlations are very useful in the process of understanding of biologically significant cellar mechanisms.

Comparative genomic hybridization is a method used of assessing genome content in different cells or closely related organisms. Cell organisms are divided into small genes and each gene is identified by a gene-ID. Within cell structure gene-IDs are used for disease detection using microarray technology. A protein is a collection DNA sequences and these DNA sequences can be separated by immune precipitating process.
SNP detection is used in many applications of microarray. Some of the micro array applications are

1. Genotyping
2. Identifying drug consumed candidates
3. Determination of cancer type
4. Genetic linkage analysis
5. Forensic analysis
6. Assessing loss of heterozygosity

Potential parts of the genes are analyzed by using exon junction array. Exon junction array is somewhere between gene expressions array consisting of one or two or three probes per g3ne and a genomic tiling array consisting of hundreds or thousands of probes per gene. Exon arrays have been designed separately and differently for analyzing only predicted genes

Data mining is a broad area where a set of data analysis techniques are available for obtaining useful knowledge. Classification is one of the most important data analysis techniques in data mining. Various classification methods are available for data analysis. Most important data classification techniques are:

1) Decision Tree Classification technique
2) Support Vector Machines (SVM)
3) Artificial Neural Networks
4) Bayesian Classification
5) Linear Discriminant Analysis (LDA)
6) Naïve Bayesian Classification
7) K-nearest Neighbor Classification (KNN)
8) Fuzzy Classification
9) Rough Set Based Classification

Cancer disease types are classified by using gene expressions. There exists a direct relationship between the gene expression changes and cancer disease types. Many cancer classification methods are proposed form two important fields – machine learning and statistics. There is no single best classifier for all types of data classification applications. Classification of cancer disease types must handle the following details:

1) Classifier accuracy
2) Classifier interpretability
3) Robustness of the classifier
4) Scalability of the classifier
5) Computational complexity of the classifier
6) Relationships among the attributes
7) Range of attribute values
8) Testability of the classifier
9) Dimensionality of the training data set
10) Attribute splitting function

V. Conclusions

State of the art cancer classification techniques are needed for better cancer treatment and better drug discovery. Most of the existing cancer classification methods are facing many problems:

1) Limited cancer related availability
2) Limited research development in cancer related areas
3) Availability of drugs for cancer disease removal or decreasing of problems caused by cancer
4) Existing methods are not suitable for applying large cancer data sets
5) Sometimes larger cancer related data sets are not available

In the future efficient, effective and scalable cancer diagnostic methods are needed. Also potential drug discovery methods and latest exactly suitable drugs are needed. Cancer research institutes and people who are doing cancer researchers must take necessary steps for strengthening cancer research developments by adopting latest and other medical and nonmedical related things. Gene expression must be thoroughly analyzed for finding cancer related symptoms and the related decisions must be taken for both cancer diagnoses and drug discovery. Automated, knowledge based and intelligent simulated techniques in a variety of directions and angles. In addition to the recent microarray technology, the researchers must concentrate other state of the art techniques also.

In the future existing cancer classification methods will be thoroughly analyzed for finding the performance differences among the various cancer classification methods. Efficient types of indexing techniques will be analyzed for selecting the best and scalable indexing technique for through gene analysis. Various classification errors will be analyzed for correct classification of cancer subtypes.

References

[1]. Data mining Jiawei Han, Kamber, Second edition Cancer Classification Using Gene Expression Data Ying
[2]. Lu Jiawei Han Department of Computer Science University of Illinois at Urbana-Champaign Urbana, IL 61801, USA
[3]. Li, W.; Han, J.; and Pei, J. 2001. CMAR: Accurate and efficient classification based on multiple class-association rules. In Proc. of the Int. Conf. on Data Mining (ICDM-01).
[4]. Madhavan, J.; Bernstein, P.; Chen, K.; Halevy, A.; and Shenoy, P. 2003. Matching schemas by learning from a schema corpus. In Proc. of the IJCAI-03 Workshop on Information Integration on the Web.
[5]. McCallum, A.; Nigam, K.; and Ungar, L. 2000. Efficient clustering of high-dimensional data sets with application to reference matching. In Proc. 6th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining.
[6]. Monge, A., and Elkan, C. 1996. The field matching problem: Algorithms and applications. In Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining.
[7]. Raman V., and Hellerstein, J. 2001. Potter's wheel: An interactive data cleaning system. In The VLDB Journal, 381–390.
[8]. Rosenthal, A.; Renner, S.; Seligman, L.; and Manola, F. 2001. Data integration needs an industrial revolution. In Proceedings of the Workshop on Foundations of Data Integration.
[9]. Sarawagi, S., and Bhamidipaty, A. 2002. Interactive deduplication using active learning. In Proc. of 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining.
[10]. Tejada, S.; Knoblock, C.; and Minton, S. 2002. Learning domainindependent string transformation weights for high accuracy object identification. In Proc. of the 8th SIGKDD Int. Conf. (KDD- 2002).