A Survey on Computational Analysis of Gene Expression Pattern

K. Vimala
Research Scholar
Department of Computer Science
Mother Teresa Women's University

D. Usha, PhD
Assistant Professor
Department of Computer Science
Mother Teresa Women's University

ABSTRACT
A gene is a segment of DNA that contains all the information necessary to analysis the defects and genetic problems that evolves in an organism. A gene is also the unit of information that is transferred through transcription and translation. This paper discusses the changes that happens in the cell either internal or external environment can lead to changes in gene expression. Most human diseases manifest through a mis-regulation of gene expression. The outputs of DNA Microarray are processed by computation tools to take out biological significance which may help to detect human disease. Computation tools include a variety of algorithms of data mining, support vector machines, pattern recognition etc. Finding desired algorithm plays a major role in research to satisfy the requirements. Surveys on computational analysis of gene expression pattern are discussed here.

General Terms
Computational analysis on supervised, unsupervised and semi supervised classification are considered for survey.

Keywords
Gene, DNA, Microarray, Protein structure, Gene expression, Computational Analysis

1. INTRODUCTION
A major challenge to the medical sciences is the large number of disorders that are primarily genetic in origin. The most common diseases in which genetic abnormalities play a role is congenital heart malfunction and diabetics. These disorders include a wide variety of debilitating and fatal illnesses for which effective method of prevention can be provided. Gene Expression involves identification of differential expression of functionally related groups of genes and coordination in regulation of multiple genes. Protein processing, extracellular matrix remodeling and inflammation are three of the dominant processes in finding disorders that are simultaneously regulated in whole blood.

2. DATABASE
Gene databases plays major role in identifying biological occurrence, to detect various disease. These gene databases are very large in size and complex to work with, hence to store, access and manipulate these data efficiently is important deal. Gene databases are categorized as sequence databases, genome databases, microarray databases, protein structure databases and many more. Gene databases represent sequence information of the organisms. The few largest databases available are GeneBank, European Nucleotide Archive (EMBL), and DNA Data Bank of Japan (DDBJ). Microarray databases contain gene expression under various biological conditions. Microarray databases are ArrayExpress, Gene Expression Omnibus, and Genome databases collect organism genome sequences and fabricate the analysis for same. Xenbase, Corn, SEED, RGD are few other database available microarrays. These databases contain species genomes, or a single organism genome. Protein structure databases include inclusive domain of protein structure based on their similarities such as amino acid sequences and three-dimensional structure. Protein structure database includes PDB, SCOP and CATH. Enormous biological data are available in text format, with many databases like PubMed and OMIN. The database contents [1] represent two main challenges a) Hierarchies of Co-expressed Genes and Coherent Patterns. b) Address the High Connectivity of Gene Expression Data Sets.

3. COMPUTATIONAL TOOLS
Data mining, Pattern recognition, Machine learning are some of the computational tools required for to analysis the database. Data mining extract usable data from a larger set of any raw data. It implies analyzing data patterns in large batches of data using one or more software. The analyses undergone are cluster analysis and frequent pattern analysis to find hidden pattern in data samples [18][3][15]. In Pattern recognition systems, finds some pattern that are drawn from the available raw data by applying some computational action. Patterns are focused on two important methods, Supervised and Unsupervised learning [5][6][9]. While machine learning is an artificial intelligence part which focuses on complex pattern using statistics, probability theory, or artificial intelligence. The decisions are made on this identified data [8][7].

4. RESEARCH IN BIOINFORMATICS
4.1 Sequence alignment
Sequence alignment [11], is an element method of information management, it has all important sense for discovering biologic sequence function, structure and evolution. The sequence alignment is to find out similar relation and biologic character of two sequences or multi-sequence by certain special mathematic model or arithmetic. Sequence [10] subsets are identified using bisecting-kmeans algorithm where K-mer counts are considered as attributes for clustering. A score matrix was built for the sequences in the subset by obtaining pairwise alignments. The center sequences were identified by the sequence which maximizes the sum of pairwise score to the rest of the sequences. Finally, the sequences were merged based on pair wise alignments between the center sequence and other sequences. Progressive alignment process is followed in order to obtain the final alignment.

4.2 Protein structure alignment and prediction
Protein Structure Prediction [13] is the process of predicting the three dimensional structure of a protein from its amino acid sequence. Proteins are large biological molecules that

contains large amount of amino acid sequence. A layered architecture [14] with two interacting levels has been defined for dealing with both primary and secondary-structure information of target protein sequences. Proteins that have low sequence identities, but whose structural and functional features suggest that common evolutionary origins are placed together in super families [12]

4.3 Gene expression

Gene expression is the most fundamental level at which genetic constitution of a cell of an organism rise to the physiological characteristics. A novel chromosome representation [4] involves each chromosome to be embedded with sub-functions, which can be deployed to construct the final solution. As part of the chromosome, the sub-functions are self-learned or self-evolved. Gene expression [16] data being high dimensional and redundant, dimensionality reduction is of prime concern; here Randomized search is employed to reduce the dimensionality of data. Genes [gene 17] have several distinctive roles in cellular processes; this is very difficult problem for classical clustering methods so mixture model is used to avoid this problem, with hidden Markov models (HMMs) as effective and flexible components

5. MICROARRAY ANALYSIS OF GENE EXPRESSION

5.1 Gene

A gene is the basic physical and functional unit of heredity. Genes are made up of DNA, act as instructions to make molecules called proteins. In humans, genes vary in size from a few hundred DNA bases to more than 2 million bases. Every person has two copies of each gene, one inherited from each parent. Most genes are the same in all people, but a small number of genes are slightly different between people. Alleles are forms of the same gene with small differences in their sequence of DNA bases. These small differences contribute to each person's unique physical features.

5.2 Gene analysis

Genetic analysis refers to experimental procedures designed to identify the genes influencing physical characteristics in organisms and their patterns of inheritance. Major research is concentrated for analysis of this interpreted data. Number of tools and techniques are available for analysis purpose like DNA Microarray, SAGE, Tiling array etc.

5.3 Microarray

An array is an orderly arrangement of samples where matching of known and unknown DNA samples is done based on base pairing rules. An array experiment makes use of common assay systems such as micro plates or standard blotting membranes. The sample spot sizes are typically less than 200 microns in diameter usually contain thousands of spots.

5.4 Microarray analysis of gene expression

Microarray analysis of gene expression involves the cDNA derived from the mRNA of known genes is immobilized. The sample has genes from both the normal as well as the diseased tissues. Spots with more intensity are obtained for diseased tissue gene if the gene is over expressed in the diseased condition. This expression pattern is then compared to the expression pattern of a gene responsible for a disease. Microarray technologies [2] have provided the means to monitor the expression levels of a large number of genes simultaneously. Gene clustering and gene ordering are important in analyzing a large body of microarray expression data.

6. COMPUTATIONAL ANALYSIS OF GENE EXPRESSION

Gene computational analyses are required categories the gene based on their behavior. Classifications of genes are given as, supervised, unsupervised and semi supervised methods. Supervised method includes all data labeled and the algorithms learn to predict the output from the input data. Unsupervised method includes all data unlabeled and the algorithms learn to inherent structure from the input data. Semi-supervised includes some data labeled but most of it is unlabeled and a mixture of supervised and unsupervised techniques can be used. K-means clustering algorithm, run fast and consume less memory compared to hierarchical clustering algorithms. Accordingly, Cai et al.[19] developed two Poisson-based measures and employed them into a K-means Clustering procedure to group tags with similar count profiles across libraries. An effective ensemble [20] approach is proposed. Ensemble classifiers increase not only the performance of the classification, but also the confidence of the results. The ensemble classifiers results are less dependent on peculiarities of a single training set. Semi-supervised classification [21] improved prediction accuracy with supervised method SVM, the performance increased with the number of unlabeled samples; the LDS method was robust with regard to the number of input features.

6.1 Analysis

The outcome of the computational analysis of gene data provides the study of different algorithms and their performance towards the desired results.

Table1.Computational Analysis

Classification Methods	Algorithms	Accuracy
Supervised	Ensemble classifiers	80%
	Back propagation	75%
Unsupervised	K-means clustering	82%
	Fuzzy C-means clustering	90%
Semi-Supervised	Support Vector Machine	97%

Semi-supervised algorithm proves to be the best in performance and accuracy, when compared to other classification methods.

7. CONCLUSION

An efficient system can be developed to handle gene expression changes based the disorders related to heart disease such as blood pressure, congenital heart disease and cardiac attack and diabetic for animals and human. The system involves two-level analysis. The first level includes preprocessing techniques, where the redundancies in the dataset are removed and summarizations of the data are collected. The second level includes statistical analysis, classification and development of ontology. The development in gene expression can provide an opportunity to study features of disorder in disease and provide path physiological context to handle complex disease.

8. REFERENCES

[1] Daxin Jiang, Jian Pei, "An Interactive Approach To Mining Gene Expression Data", IEEE Transactions On Knowledge And Data Engineering, Vol. 17, No. 10, Page.456,2005

[2] Huai-Kuang Tsai, Jinn-Moon Yang, Yuan-Fang Tsai, and Cheng-Yan Kao, "An Evolutionary Approach for Gene Expression Patterns", IEEE Transactions On Information Technology In Biomedicine, Vol.8, No.2, pages.69, 2004

[3] de Menezes RX, Boer JM, van Houwelingen HC. "Microarray Data Analysis". Applied Bioinformatics. Vol 3, Issue 4, Pages. 229,2009.

[4] Jinghui Zhong, Yew-Soon Ong, and Wentong Cai, 'Self-Learning Gene Expression Programming", IEEE Transactions On Evolutionary Computation, Vol.2, No.3,Pages.23, 2015

[5] Chen L-F, Su C-T, Chen K-H, Wang P-C. "Particle swarm optimization for feature selection with application in obstructive sleep apnea diagnosis". Neural Computing and Application, Vol 21, Issues 8 , Page 2087,2012.

[6] Shuanhu Wu, Alan Wee-Chung Liew, "Cluster Analysis of Gene Expression Data Based on Self-Splitting and Merging Competitive Learning", IEEE Transactions On Information Technology In Biomedicine, Vol. 8, No. 1, Page.234,2004.

[7] Safiye Celik, Benjamin A. Logsdon, Scott M. Lundberg "A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia", Nature Communications Vol. 9, Article number: 42, 2018.

[8] Lingyun Gao , Mingquan Ye , Xiaojie Lu , Daobin Huang, "Hybrid Method Based on Information Gain and Support Vector Machine for Gene Selection in Cancer Classification", Genomics Proteomics Bioinformatics, Vol 15, No.1, Pages 389, 2017

[9] Chen Y, Miao D, and Wang R. "A rough set approach to feature selection based on ant colony optimization". Pattern Recogniton Letters, Vol 31, No.3, Pages 226, 2010

[10] Kokila K. Perera, C. Thusangi Wannige, "A Hybrid Algorithm for Multiple DNA Sequence Alignment", International Conference on Advances in ICT for Emerging Regions, Vol.3, No.5,Page 323,2016

[11] LIU Chao, LIU Shuai, "The research on DNA Multiple Sequence Alignment Based on Adaptive Immune Genetic algorithm", International Conference on Electronics and Optoelectronics,2011

[12] Sankar K. Pal, "Evolutionary Computation in Bioinformatics: A Review", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 36, No. 5, page. 601, 2006

[13] Subhendu Bhusan Rout, Satchidananda Dehury, Bhabani Sankar Prasad Mishra, "Protein Structure Prediction using Genetic Algorithm", International Journal of Computer Science and Mobile Computing, Vol.2, Issue.6, pages 187, 2013

[14] Giuliano Armano, Luciano Milanesi, and Alessandro Orro, "Multiple Alignment Through Protein Secondary-Structure Information", IEEE Transactions On Nanobioscience, Vol.4, No.3, Page.34, 2005

[15] Latkowski T, Osowski S. "Data mining for feature selection in gene expression autism data." Expert Systems with Applications, Vol 42, Issues 2, Page 864,2015

[16] Sushmita Mitra,"Feature Selection and Clustering of Gene Expression Profiles Using Biological Knowledge", IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 42, No. 6, Page.1590, 2012

[17] Alexander Schliep, Ivan G. Costa, Christine Steinhoff, and Alexander Schonhuth, "Analyzing Gene Expression Time-Courses", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 2, No. 3, pages. 179, 2005

[18] Francisco Torres Aviles," Data mining and influential analysis of gene expression data for plant resistance gene identification in tomato (Solanum lycopersicum)", Electronic Journal of Biotechnology Vol 17, Issue 2, Page. 79, 2014.

[19] Cai, L, Huang, H, Blackshaw, S, Liu, J. S, Cepko, C. L. and Wong, W. H. "Clustering Analysis of SAGE Data Using a Poisson Approach," Genome Biology, Vol.5, No.51, Page.56, 2004

[20] Sara Tarek , Reda Abd Elwahab, Mahmoud Shoman, "Gene expression based cancer classification", Egyptian Informatics Journal, Vol.4, No.3, Pages.234, 2016

[21] Mingguang Shi and Bing Zhang,"Semi-supervised learning improves gene expression-based prediction of cancer recurrence", Bioinformatics Vol. 27, No. 21 , pages 3017, 2011

[22] David Seo, MD, Geoffrey S. Ginsburg, "Gene Expression Analysis of Cardiovascular Diseases Novel Insights Into Biology and Clinical Applications", Cardiovascular Genomic Medicine, Vol. 48, No. 2, Page 227,2006

[23] Michelle M. Kittleson, Khalid M. Minhas, "Gene expression analysis of ischemic and nonischemic cardiomyopathy: shared and distinct genes in the development of heart failure", Physiol Genomics, Vol. 21, No.4, Page. 299, 2005

[24] Haifang Wang, Yuting Liu, Marko Briesemann, and Jun Yan, "Computational analysis of gene regulation in animal sleep deprivation", Physiol Genomics,Vol.42,No.2, Page. 427, 2010.