RESEARCH ARTICLE OPEN ACCESS

Gene Expression Profiling of DNA Microarray Data using various Data Mining Methodologies

K.Ganeshbabu [1], N.Sevugapandi [2], C.P.Chandran [3]
Research Scholar [1] & [2],
Ph.D Part-Time, Category –B
Research and Development Centre,
Bharathiar University, Coimbatore
Associate Professor [3]
Department of Computer Science,
Ayya Nadar Janaki Ammal College,Sivakasi
Tamil Nadu – India

ABSTRACT
This paper aims to mine the gene expression profiling of DNA microarray data using various Data Mining methodologies with the biological vital sequence and to visualize the numerous data processing methodologies like classification, clump and association rule mining. DNA microarray technology has been extremely used in the field of bioinformatics for exploring genomic organization. It enables to analyze expression of many genes in a single reaction. The techniques currently employed to do analysis of microarray expression data is clustering and classification. In this paper, the cancer gene expression is analyzed using hierarchical clusterin g that identifies a group of genes sharing similar expression profiles and dendrograms are employed that provides an efficient means of prediction over the expression. Knowledge discretization is completed by clump the sequence in to two clusters i.e up-regulated and down-regulated.
Keywords :— Data Mining, DNA Micro array, Gene Ontology, KEGG pathway.

I. INTRODUCTION

A. Data Mining

Data mining is defined as the non-trivial process of searching and analyzing data in order to find implicit but potentially useful information. Let D = {d1... dn} be the dataset to be analyzed. The data mining process is described as the process of finding a subset D' of D and hypotheses HU(D',C) about D' that a user U considers useful in an application context C. D' have fewer data elements than D, but it also have a lower dimensionality (m'). In databases the data is partitioned into relations or object classes. D is considered as a union of relations R1... Rk each has its own dimensionality (m1... mk)[1].

B. Bioinformatics

Bioinformatics is the Science of integrating, managing, mining and interpreting information from biological datasets at genomic, metabalomic,
proteomics, phylogenetic and cellular or whole organism levels.

According to (National Institute of Health) NIH organization, the Bioinformatics and Computational Biology have been defined as "Bioinformatics is research and development or application of computational tools and approaches for expanding the use of biological, medical, health data including those to acquire store, organize, active, analyze or visualize such data" [2].

Genomics

DNA (Deoxyribonucleic Acid) is a molecule encoding the genetic instructions used in the development and functioning of all known living organisms many viruses. DNA is one of the three major macromolecules that are essential for all known forms of life.

Genetic information is encoded as a sequence of nucleotides (Guanine, Adenine, Thymine, and Cytosine) recorded using the letters G, A, T, and C. Most DNA molecules are double-stranded helices, consisting of two long polymers of simple units called nucleotides with the nucleobases (G, A, T, C) attached to the sugars. DNA is well-suited for biological information storage, since the DNA backbone is resistant to cleavage and the double-stranded structure provides the molecule with a built-in duplicate of the encoded information [3].

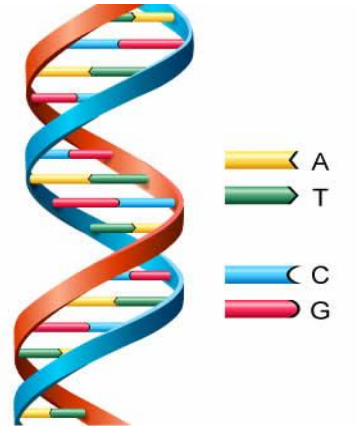

Fig. 1 Structure of DNA

Molecular Biology research evolves through the development of the technologies used for carrying them out. In the past , only genetic analyses on a few genes had been conducted and it is not possible to research on a large number of genes using traditional methods. DNA Microarray [4] is one such technology which enables the researchers to investigate how active thousands of genes at any given time and address issues which were once thought to be non traceable. One can analyze the expression of many genes in a single reaction quickly and in an efficient manner. DNA Microarray technology has empowered the scientific community to understand the fundamental aspects underlining the growth and development of life as well as to explore the genetic causes of DNA Microarray.

Microarray technology will help researchers to learn more about many different diseases, including heart disease, mental illness and infectious diseases, to name only a few. One intense area of microarray research at the National Institutes of Health (NIH) [1] is the study of cancer. In the past, scientists have classified different types of cancers based on the organs in which the tumors develop. With the help of microarray technology, however, they will be able to further classify these types of cancers based on the patternsof gene activity in the tumor cells. Researchers will then be able to design treatment strategies targeted directly to each specific type of cancer. Additionally, by examining the differences in gene activity between untreated and treated tumor cell s - for example those that are radiated or oxygen-starved - scientists will understand exactly how different therapies affect tumors and be able to develop more effective treatments.

In addition, data mining clustering technique having an appealing property is employed, such that the nested sequence of clusters can be graphically represented with a tree, called a *dendogram.* It simplifies the identification of gene expression over the microarray thus provides an efficient means of prediction over the expression.

Nucleic Acids

Nucleic acids are large biological molecules essential for all known forms of life. They include DNA. Together with proteins, nucleic acids are the most important biological macromolecules; each is found in abundance in all living things, where they function in encoding, transmitting and expressing genetic information.

The nucleic acids Deoxyribonucleic acid DNA are polymers of nucleotides, arranged in a specific sequence. To form macromolecular polymers, nucleotides are joined between the 3’ and 5’ carbon atoms in their sugar moiety by a phosphodiester bond, giving rise to a nucleic acid with a sugar–phosphate ‘backbone’ to which is attached a series of bases in a specific order. Hydrogen bonding between pairs of bases can occur, leading to the formation of double-stranded polymers if the sequences are complementary [4].

(i) Base pairs

Base pairs are the building blocks of the DNA double helix, and contribute to the folded structure of both DNA. Dictated by specific hydrogen bonding patterns, Watson-Crick base pairs (Guanine-Cytosine and Adenine-Thymine) allow the DNA helix to maintain a regular helical structure that is independent of its nucleotide sequence. The complementary nature of this based-paired structure provides a backup copy of all genetic information encoded within double-stranded DNA. Fig. 2 shows the pairs of ATGC.

The regular structure and data redundancy provided by the DNA helix make DNA an optimal molecule for the storage of genetic information, while base-pairing between DNA and incoming nucleotides provide the mechanism through which DNA polymerase replicates DNA transcribes. Many DNA-binding proteins can recognize specific base pairing patterns that identify particular regulatory regions of genes.

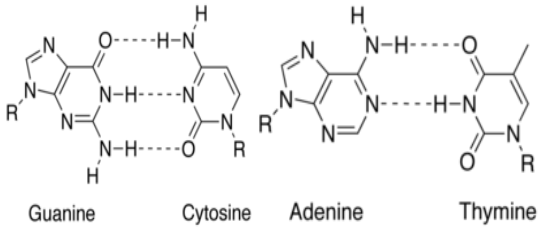

Fig. 2 Base pairs of ATGC

(ii) Gene Expression Data

Gene expression refers to a complex series of processes in which the information encoded in a gene is used to produce a functional product such as a protein that dictates cell function. It involves several different steps through which DNA is converted to an RNA which in turn is converted into a protein or in some cases RNA, for example, genes encoding the necessary information for transfer RNAs and ribosomal RNAs (tRNAs and rRNAs). The information flow from DNA to RNA to protein can be controlled at several points helping the cell to adjust the quality and quantity of resulting proteins and thus self-regulate its functions. Thus, regulation of gene expression is a critical step in determining what kind of proteins and how much of each protein is expressed in a cell.

(iii) DNA Motifs

Complex designs are often created using a relatively small set of common building blocks called motifs. DNA self-assembly can exploit this same design principle to hierarchically create more sophisticated a periodic structures.

There are many possible DNA motifs and the focus here is on only a few in the context of the target nanostructure. Motifs include junctions that enable three or more double stranded helices of DNA to interact and thus form specific structures (e.g., a triangle, a corner, and so on). Another important motif is a single strand of DNA protruding from a double stranded helix called a sticky-end.

Two motifs with complementary sequences on their sticky-ends will bind to form a composite motif. Composite motifs may also have embedded sticky-end motifs and thus can also bind with other composite motifs to form another, larger, composite motif. This results in a hierarchical structure for motifs.

(iv) Various Gene Expression Techniques

Analytical methods may be used to examine mRNA expression levels or differential mRNA expression. Some examples of these techniques are listed below.

Serial Analysis of Gene Expression (SAGE): SAGE is a technique used to create a library of short sequence tags which can each be used to detect a transcript. The expression level of the transcript can be determined by assessing how many times each tag is detected. This technology enables comprehensive expression analysis across the genome [5].

DNA microarray: Also known of as biochip or DNA chip, a DNA microarray is a solid surface to which a collection of microscopic DNA spots are attached. The microarrays are used to determine expression levels across a large number of genes or to perform genotyping across different regions of a genome.

RNA Seq : This refers to methods used to measure the sequence of RNA molecules. Examples include shotgun sequencing of cDNA molecules acquired from RNA through reverse transcription and technologies used to sequences.

RNA molecules from a biological sample so that the primary sequence and abundance of each RNA molecule can be determined.

Tiling arrays: A tiling array is a type of microarray chip, with labelled DNA or RNA targets hybridized to probes attached to a solid surface. However, the probes used differ to those used with traditional microarrays. Rather than known sequences or predicted genes being probed, tiling arrays probe for sequences known to be present in a contiguous region.

II. PREVIOUS WORKS

S. Orlando et. al (2014) Ensemble techniques have been successfully applied in the context of supervised learning to increase the accuracy and stability of classification. Recently, similar techniques have been proposed for clustering algorithms. In this context, the potential of applying cluster ensemble techniques to gene expression microarray data has been analyzed. The experimental results show that there is often a significant improvement in the results obtained with the use of ensemble when compared to those based on the clustering techniques used individually [6].

F. Tao (2003) Data mining methods have been widely applied to discover patterns and relations in complex datasets. This work, particularly concerned with a type of taxonomy discovery called cluster analysis, the discovery of distinct and non-overlapping sub-population within a large population, the member items of each sub-population sharing some common features considered relevant in the problem domain of study [7].

S. Orlando et. al (2013) From oncology science, the uncontrolled growth of malignant/benign tumours refers to secreted reasons causing the formation of new blood vessels sprouting from pre-existing vessels. Consequently, scientists attribute this abnormal behaviour to intratumour factors, defined as tumour-derived factors. These factors are

guided through protein molecules that work on cellular signalling path [8].

Authors/ Year	Algorithms	Description
S. Orlando et. al (2014)	CA	Ensemble techniques have been successfully applied in the context of supervised learning to increase the accuracy and stability of classification.
F. Tao (2003)	Heuristic K-Means algorithm	Data mining methods have been widely applied to discover patterns and relations in complex datasets.
S. Orlando et. al (2013)	ANTCLUST	From oncology science, the uncontrolled growth of malignant/benign tumours refers to secreted reasons causing the formation of new blood vessels
C. F. Ahmed et. al (2008)	ASM	Various data mining techniques for analyzing Alzheimers disease Gene Expression Dataset using Clustering and Association Rule Mining.
S. Mallik (2013)	A4C	Gene expression data based cancer classification is of great importance to the computer aided diagnosis.

Accordingly, the deoxyribonucleic acid (DNA) is considered as the maestro of this process. Analysing changes on the gene expression may give rise for diagnosis enhancement of affected tissues in their early stages [9].

C. F. Ahmed et. al (2008) worked on various data mining techniques for analyzing Alzheimers disease Gene Expression Dataset using Clustering and Association Rule Mining [10].

S. Mallik (2013) Gene expression data based cancer classification is of great importance to the computer aided diagnosis. In this paper, a novel cancer selection method, AR-SVM. In ARSVM, association rules are used as feature extraction approach to catch the non-linear relation among different genes, and support vector machine is used to classify the transformed gene expression data [11].

TABLE 1

PREVIOUS WORKS RELATED TO GENE EXPRESSION PROFILING

III. METHODOLOGY USED

Microarray data for a simple dataset having five samples and four genes, represented in dots of different color indicating the intensity of tumor have been interpreted. The different colors of the spots have to be converted to numbers before analysis in order to obtain the intensity of tumor. There are many approaches but here a simplified version of common techniques is employed.

Thousands of spotted samples known as probes (with known identity) are immobilized on a solid support (a microscope glass slides or silicon chips or nylon membrane). The spots can be DNA, cDNA, or oligonucleotides. These are used to determine complementary binding of the unknown sequences thus allowing parallel analysis for gene expression and gene discovery. An experiment with a single DNA chip can provide information on thousands of genes simultaneously. An orderly arrangement of the probes on the support is important as the location of each spot on the array is used for the identification of a gene.

A t-test is any applied math hypothesis check within which the check datum follows a Student's t-distribution underneath the null hypothesis. It will be accustomed confirm if two sets of knowledge area unit considerably completely different from one another, and is most typically applied once the check datum would follow a traditional distribution if the worth of a scaling term within the check datum were better-known. Once the scaling term is unknown associate degree is replaced by an estimate supported the info, the check datum (under bound conditions) follows a Student's distributions.

A one-sample location check of whether or not the mean of a population encompasses a worth per a null hypothesis. The p-value is "commonly used and misinterpreted", in step with the yankee applied math Association, that took the exceptional step of issue an announcement on the utilization of p-values. The use of bright-line rules as cutoffs, notably $p \leq$ zero.05, is especially criticized. The widespread use of applied math significance (generally taken as $p \leq$ zero.05) as a license for creating a claim of a scientific finding (or understood truth) ends up in substantial distortion of the scientific method. Whereas there's widespread agreement that p-values area unit used, there's no accord on alternatives, and misuse of p-values continues to be the topic of criticism.

A check of whether or not the slope of a regression curve differs considerably from zero. In frequents statistics, the p-value could be a operate of the ascertained sample results (a check statistic) relative to an applied math model, that measures however extreme the observation is. The p-value is that the likelihood that the ascertained result has nothing to try to with what one is truly testing for. Specifically, the p-value is outlined because the likelihood of getting a result adequate or "more extreme" than what was really ascertained, presumptuous that the model is true.

There are a unit completely different definitions of "extreme", in frequents abstract thought, the p-value is wide employed in applied math hypothesis testing, specifically in null hypothesis significance testing.

In this methodology, as a part of experimental style, before acting the experiment, one initial chooses a model (the null hypothesis) and a threshold worth for p, known as the importance level of the check, historically five-hitter or a hundred and twenty fifth and denoted.

If the p-value is a smaller amount than or adequate the chosen significance level, the check suggests that the ascertained knowledge is inconsistent with the null hypothesis, therefore the null hypothesis should be rejected. However, that doesn't prove that the tested hypothesis is true. Once the p-value is calculated properly, this check guarantees that the kind I error rate is at the most.

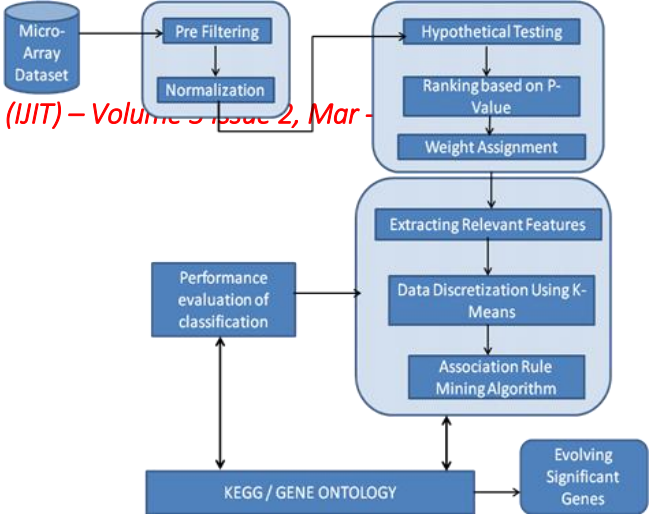

Fig. 3 Methodology Used for Classification of Biologically Significant genes

IV. RESULTS AND DISCUSSION

To analyze large amount of expression data, it's necessary to use statistical analysis. Unfortunately, fractions are not suitable for statistics. For this reason, the expression ratios are usually transformed by log2 function, in which, for every increase or decrease of 1, there are 2 fold changes. In our example, log10 is used, since it is easier for efficient outcome. In log10, for every increase or decrease of 1, there are 10 fold changes. The table below shows the relationships between log2 and log10. Numbers are often converted to colored scale i.e., red and green fluorescents, to make it easier to see the patterns. Results are often reported in this way of representation.

Data is made in such how, wherever the whole column represents the sample identity and also the row represents the cistron identity. Ultimately the info contains 27579x215 tagged with cistron id and column id on an individual basis. Every organic phenomenon values area unit delineate by its column and cistron id.

The filtered knowledge ought to be normalized gene-wise as standardization converts the info from totally different scales into a standard scale. During this dataset, zero-mean standardization is applied that converts the info into such a type wherever mean of every cistron becomes zero.

From the ensuing price of the t-statistic, corresponding p-value is calculated from t-table or accumulative distributions perform (cdf). If p-value of a cistron is a smaller amount than 0.05, then the cistron is termed DE/DM, otherwise not. The genes DE/DM area unit then hierarchal with relation to their p-values.

According to this knowledge matrix, one denotes up-regulated cistron and 0 denotes down-regulated cistron. As in ARM, one and 0 signify presence and absence of some item (gene) in some group action (sample), severally for binary knowledge, therefore here we will show solely the presence of by victimization one, and presence of by victimization 0.

When, Similarity score = + ve, two genes behave similarly i.e., when one is induced, so is the other. number, Similarity score = 1, two genes behave identically. Gene A obviously behaves exactly like Gene A. Similarity score = 0, two genes behave in unrelated manner. Similarity score = –ve, two genes behave in opposite ways i.e., when one is induced other is suppressed. By casual inspection, we could summarize that: Gene C's behavior is opposite to that of Gene A, B, and D Gene B and Gene D have the most similar behaviors.

TABLE 2
DATASET CONSTRUCTED FROM THE REPOSITORY

Genes Samples	Gene A	Gene B	Gene C	Gene D
Sample1	0.60	0	-	0
Sample2	0.30	-0.096	0	0.114
Sample3	0.54	0.3	-	0.477
Sample4	0.17	-0.301	-0.602	0
Sample5	-0.096	0	0.0792	-0.096

TABLE 3
NORMALIZED DATASET

Gene/Sample	Sample 1	Sample 2	.....	Sample 215	P-Value	Gene Rank	Weight
G9522	1.854133	0.73256	.....	0.049975	0.049975	1	1
G11289	1.374848	0.806773	.....	1.068918	0.049901	2	0.92212
G4802	-1.72562	-1.79166	.....	-1.25521	0.049891	3	0.8124
. .			.....		. .	. .	. .
G9421	-0.91353	-1.01116	.....	1.254453	3.91E-32	14557	0.14262
G19976	0.168014	-1.31806	.....	2.052473	2.94E-32	14558	0.141385

TABLE 4

RANKING AND WEIGHTING OF GENES

Gene / Samples		GSM763063	GSM763064	GSM763064	.....	GSM763282
		Sample 1	Sample 2	Sample 3		Sample 215
cg00000292	G1	1.759552	1.541954	0.840594	.....	-0.97909
cg00002426	G2	1.855155	0.365548	2.060242	.....	-0.30238
cg00003994	G3	-1.20374	-1.0904	-0.95906	.....	1.566123
. . .					. . .	. . .
cg27665659	G22998	-1.56763	-1.28845	-1.48949		-0.19109

TABLE 5

SIMILARITY SCORE FOR ALL GENES

Genes	Gene A	Gene B	Gene C	Gene D
Gene A	1	0.450	-0.633	0.597
Gene B	0.450	1	0.107	0.729
Gene C	-0.633	-0.107	1	-
Gene D	0.597	0.729	0.454	1

V. CONCLUSION

Microarray technology has been extensively used by the scientific community. Advances in computer technology have made powerful analytical tools readily available. Even modest PC can analyze a dataset of 3,000 genes with 100 samples in minutes. In real situations, additional complications need to be taken into account, such as making sure that comparing fluorescence from different microarrays does not introduces additional variability. Other microarray may use different detection and analytical techniques that don't use fluorescence.

The clustering techniques have been widely used to identify group of genes sharing similar expression profiles and the results obtained so far have been extremely valuable. However, the metrics adopted in these clustering techniques have discovered only a subset of relationships among gene expression. Clustering can work well when there is already a wealth of knowledge about the pathway in question, but it works less well when this knowledge is sparse. The inherent nature of clustering and classification methodologies makes it less suited for mining previously unknown rules and pathways.

REFERENCES

[1] M. H. Dunham, Data Mining: Introductory and Advanced Topics, Pearson Education, Third Edition, New Delhi, India, ISBN 81-7758-880-X, 2012.

[2] Yi-Ping P. Chen (Ed.), Bioinformatics technologies, Springer International edition, Hiedelberg, ISBN 3-540-20613-2, 2008.

[3] http://en.ncbi.org//DNA,_RNA_and_proteins:_The_three_essential_macromolecules_of_life#DNA.2C_RNA_and_proteins.

[4] Seung-Hyun Lee and Chengde Mao,DNA Nanotechnology, Techniques essay, University, USA, 2004.

[5] S.Bandyopadhyay, Austin.H, Chenwei W 2014 'A survey and comparative study of statistical tests for identifying differential expression from microarray data', in IEEE/ACM Trans. Comput. Biol. Bioinformat., vol. 11, no. 1, pp.95115.

[6] R.Agrawal, T.Imielinski and A.Swami 2008 'Mining Association Rules between Sets of Items in large Databases', in ACM SIGMOD ACM, New York, vol. 58, no. 2, pp. 468-493.

[7] S.Orlando, M.Aouf, L.Lyanage and S.Hansen 2013 'Enhancing the apriori algorithm for frequent set counting', Data Warehousing and Knowledge Discovery pp. 7182.

[8] S.Mallik 2013 'Integrated analysis of gene expression and genomewide DNA methylation for tumor prediction: An association rule miningbased approach', Intell. Bioinformat . Comput. Biol. (CIBCB), Singapore, pp. 120127.

[9] F.Tao 2003 'Weighted association rule mining using weighted support and significance framework' in ACM SIGKDD, Washington, D.C., USA, pp.661666.

[10] C.F.Ahmed, Benoit Le Quau, Omair Shafiq and Reda Alhajj(2008) 'Mining weighted frequent patterns in incremental databases', Trends in Arti ficial Intelligence. vol. 5351,Lecture Notes in Computer Science, pp. 933938.

[11] Gene Set: ION HOMEOSTASIS Gene Ontology Reference Availableat http://software.broadinstitute.org/gsea/msigdb/cards/ION_ HOMEOSTASIS [January 2016].