International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014
ISSN 2091-2730

Analysis of Cancer Gene Expression Profiling in DNA Microarray Data using Clustering Technique

[1]C. Premalatha, [2] D. Devikanniga
[1, 2]Assistant Professor, Department of Information Technology
Sri Ramakrishna Engineering College, Coimbatore, Tamil Nadu
[1]premalathajck@gmail.com

Abstract— 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 clustering that identifies a group of genes sharing similar expression profiles and dendrograms are employed that provides an efficient means of prediction over the expression.

Keywords — Hierarchical clustering, Microarray data, Gene expression, Dendrograms.

INTRODUCTION

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 anomalies occurring in the functioning of the human body.

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 cells - 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.

GEO DATABASE

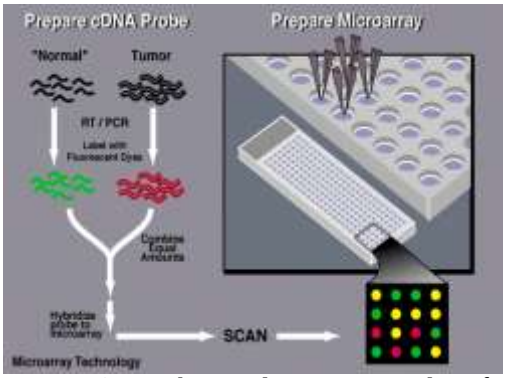
Microarray technology has been extensively used by the scientific community. Consequently, over the years, there has been a lot of generation of data related to gene expression. This data is scattered and is not easily available for public use. For easing the accessibility to this data, the **National Center for Biotechnology Information** (**NCBI**) has formulated the **Gene Expression Omnibus** or **GEO**. It is a data repository facility which includes data on gene expression [6] from varied sources. GEO currently stores approximately a billion individual gene expression measurements, derived from over 100 organisms, addressing a wide range of biological issues.

MICROARRAY TECHNIQUE

An array is an orderly arrangement of samples where matching of known and unknown DNA samples is done. 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.

WORKING OF DNA MICROARRAY TECHNOLOGY

DNA microarrays are created by robotic machines that arrange minuscule amounts of hundreds or thousands of gene sequences on a single microscope slide. When a gene is activated, cellular machinery begins to copy certain segments of that gene. The resulting product is known as messenger RNA (mRNA), which is the body's template for creating proteins. The mRNA produced by the cell is complementary, and therefore will bind to the original portion of the DNA strand from which it was copied. To determine which genes are turned on and which are turned off in a given cell, first the messenger RNA molecules present in that cell are collected then they are labelled by using a reverse transcriptase enzyme (RT) that generates a complementary cDNA to the mRNA. During that process fluorescent nucleotides are attached to the cDNA.



The tumor and the normal samples are labeled with different fluorescent dyes[7]. Next, the researcher places the labeled cDNAs onto a DNA microarray slide. The labeled cDNAs that represent mRNAs in the cell will then hybridize – or bind – to their synthetic complementary DNAs attached on the microarray slide, leaving its fluorescent tag. A special scanner is used to measure the fluorescent intensity for each spot/areas on the microarray slide.If a particular gene is very active, it produces many molecules of messenger RNA, thus, more labeled cDNAs, which hybridize to the DNA on the microarray slide and generate a very bright fluorescent area.

Fig.1. Microarray Technology

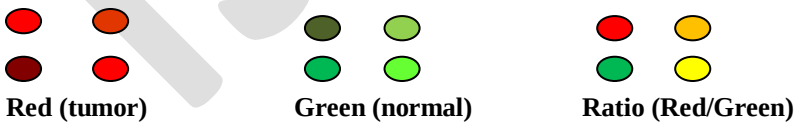
Genes that are less active produce fewer mRNAs, thus, less labeled cDNAs, which results in dimmer fluorescent spots. If there is no fluorescence, none of the messenger molecules have hybridized to the DNA, indicating that the gene is inactive. Researchers frequently use this technique to examine the activity of various genes at different times. When co-hybridizing Tumor samples (Red Dye) and Normal sample (Green dye) together, they will compete for the synthetic complementary DNAs on the microarray slide. As a result, if the spot is red, this means that that specific gene is more expressed in tumor than in normal (up-regulated in cancer). If a spot is Green, it means that the gene is more expressed in the Normal tissue (Down regulated in cancer). If a spot is yellow that means that the specific gene is equally expressed in normal and tumor.

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.

INTERPRETING MICROARRAY DATA

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.

- First, each spot is converted to a number that represents the intensity of the red dye and green dye. In this example, arbitrary light units are used.
- Next, we calculate the ratio of red to green (red/green).



Red (tumor)

400	200
00	400

Green (normal)

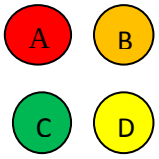
100	200
300	400

Ratio (Red/Green)

4	1
0.33	1

Fig.2. Ratio between tumor and normal gene

TABLE I CELL INTENSITY RATIO



Genes / *Cell intensity*	**Gene A**	**Gene B**	**Gene C**	**Gene D**
Red(tumor)	400	200	100	400
Green(normal)	100	200	300	400
Ratio	4	1	0.33	1

TABLE II INTENSITY RATIO FOR FIVE SAMPLES

Genes / *Samples*	**Gene A**	**Gene B**	**Gene C**	**Gene D**
Sample1	4	1	0.33	1
Sample2	2	0.8	1	1.3
Sample3	3.5	2	0.25	3
Sample4	1.5	0.5	0.25	1
Sample5	0.8	1	1.2	0.8

When, Ratio >1: Indicates the gene was induced by tumor formation.
Gene A induced four fold.
Ratio <1: Indicates the gene was repressed by tumor formation.
Gene C repressed 3 fold.
Gene B and D not affected by tumor formation.

PROCESSING MICROARRAY DATA

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.

TABLE III PROCESSING MICROARRAY DATA

Genes / *Samples*		**Gene A**	**Gene B**	**Gene C**	**Gene D**
Sample1	**Ratio**	4	1	0.33	1
	Log 2	2	0	-1.599	0
	Log 10	0.602	0	-0.481	0
Sample2	**Ratio**	2	0.8	1	1.3
	Log 2	1	-0.321	0	0.378
	Log 10	0.301	-0.096	0	0.114
Sample3	**Ratio**	3.5	2	0.25	3
	Log 2	1.807	1	-2	1.584
	Log 10	0.544	0.301	-0.602	0.477
Sample4	**Ratio**	1.5	0.5	0.25	1
	Log 2	0.584	-1	-2	0
	Log 10	0.176	-0.301	-0.602	0
Sample5	**Ratio**	0.8	1	1.2	0.8
	Log 2	-0.321	0	0.263	-0.321
	Log 10	-0.096	0	0.079	-0.096

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.

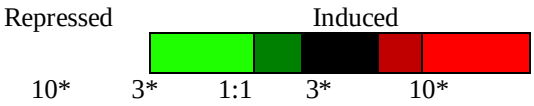


Fig.3. Color conversion for repressed and induced cell

TABLE IV CELL PATTERNS (TUMOR vs. NORMAL)

Genes / *Samples*	Gene A	Gene B	Gene C	Gene D
Sample1	0.602	0	-0.481	0
Sample2	0.301	-0.096	0	0.114
Sample3	0.544	0.301	-0.602	0.477
Sample4	0.176	-0.301	-0.602	0
Sample5	-0.096	0	0.0792	-0.096

CALCULATING SIMILARITIES

To calculate similarity score, mean and standard deviation of each gene's expression values are calculated.

TABLE V SIMILARITY CALCULATION USING MEAN AND STANDARD DEVIATION

Genes / *Samples*	Gene A	Gene B	Gene C	Gene D
Sample1	0.602	0	-0.481	0
Sample2	0.301	-0.096	0	0.114
Sample3	0.544	0.301	-0.602	0.477
Sample4	0.176	-0.301	-0.602	0
Sample5	-0.096	0	0.0792	-0.096
Mean	*0.305*	*-0.0194*	*-0.321*	*0.0988*
Std dev	*0.254*	*0.194*	*0.299*	*0.200*

Next, normalize the values subtracting the mean for that gene and dividing it by standard deviation.
For Ex, Gene A - sample1 becomes
(0.602-0.305)/0.254=1.167

TABLE VI NORMALIZED DATA

Genes / *Samples*	Gene A	Gene B	Gene C	Gene D
Sample1	1.166744	0.099756	-0.53475	-0.49276
Sample2	-0.01659	-0.39902	1.074453	0.075694
Sample3	0.938726	1.649106	-0.93956	1.885779
Sample4	-0.50801	-1.4496	-0.93956	-0.49276
Sample5	-1.58087	0.099756	1.339419	-0.97595

Next for each pair of gene, multiply the values from each sample, add up the products and divide by the number of sample (5). The result is similarity score. For example, for Gene A and Gene B,

TABLE VII SIMILARITY SCORE FOR GENE (AB) AND (CD) USING DOT PRODUCT

Genes / *Samples*	Gene A	Gene B	Product (A and B)	Gene C	Gene D	Product (C and D)
Sample1	1.166744	0.099756	0.116389	-0.53475	-0.49276	0.26350341
Sample2	-0.01659	-0.39902	0.00662	1.074453	0.075694	0.081329645
Sample3	0.938726	1.649106	1.548059	-0.93956	1.885779	-1.771802517
Sample4	-0.50801	-1.4496	0.736406	-0.93956	-0.49276	0.462977586
Sample5	-1.58087	0.099756	-0.1577	1.339419	-0.97595	-1.307205973
SUM			2.249774	**SUM**		-2.271197849

SIMILARITY SCORE	**0.449955**	**SIMILARITY SCORE**	**-0.45423957**

TABLE VIII SIMILARITY SCORE FOR GENE (AC) AND (AD) USING DOT PRODUCT

Genes / *Samples*	**Gene A**	**Gene C**	**Product (A and C)**	**Gene A**	**Gene D**	**Product (A and D)**
Sample1	1.166744	-0.53475	-0.62391635	1.166744	-0.49276	-0.574924773
Sample2	-0.01659	1.074453	-0.01782518	-0.01659	0.075694	-0.001255763
Sample3	0.938726	-0.93956	-0.8819894	0.938726	1.885779	1.770229778
Sample4	-0.50801	-0.93956	0.477305876	-0.50801	-0.49276	0.250327008
Sample5	-1.58087	1.339419	-2.11744731	-1.58087	-0.97595	1.542850077
SUM			-3.16387237	**SUM**		2.987226325
SIMILARITY SCORE			**-0.63277447**	**SIMILARITY SCORE**		**0.597445265**

TABLE IX SIMILARITY SCORE FOR GENE (BC) AND (BD) USING DOT PRODUCT

Genes / *Samples*	**Gene B**	**Gene C**	**Product (B and C)**	**Gene B**	**Gene D**	**Product (B and D)**
Sample1	0.099756	-0.53475	-0.05334452	0.099756	-0.49276	-0.04915577
Sample2	-0.39902	1.074453	-0.42872824	-0.39902	0.075694	-0.03020342
Sample3	1.649106	-0.93956	-1.54943403	1.649106	1.885779	3.109849464
Sample4	-1.4496	-0.93956	1.361986176	-1.4496	-0.49276	0.714304896
Sample5	0.099756	1.339419	0.133615082	0.099756	-0.97595	-0.09735687
SUM			-0.53590553	**SUM**		3.647438305
SIMILARITY SCORE			**-0.10718111**	**SIMILARITY SCORE**		**0.729487661**

TABLE X SIMILARITY SCORE FOR ALL 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	-0.454
Gene D	0.597	0.729	0.454	1

When, Similarity score = + ve, two genes behave similarly i.e., when one is induced, so is the other. Larger the number, the more similar they are.

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.

HIERARCHICAL CLUSTERING

To analyze 1000 of genes, hierarchical clustering [3] is used which works by taking the most similar genes and joining them in a cluster. The nested sequence of clusters produced by hierarchical methods makes them appealing, when different levels of detail are of interest, because small clusters are nested inside larger ones. In microarray applications, interest may focus on both small groups of similar observations and a few large clusters. The former might occur when individuals provide multiple samples or a few samples have special meaning, such as the four samples in the carcinoma example that are normal tissue. The latter would occur when larger groups exist, such as samples from two different sources, or different stages of carcinoma, or from different experiments.

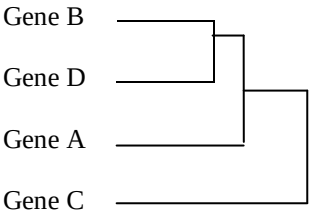
Gene B and D are the most similar at 0.729, so they are joined to become [BD]. Next, the log-transformed expression levels are averaged for the clustered genes and the similarity scores are recalculated:

TABLE XI SIMILARITY SCORE FOR CLUSTERED GENE

	Gene A	**Gene C**	**Gene [BD]**
Gene A	1	-0.633	0.564
Gene C	-0.633	1	-0.305
Gene [BD]	0.564	-0.305	1

Next highest score is chosen i.e., A and [BD], to form the cluster [ABD]. Since only 4 genes are processed it terminates with simple process. But this is an iterative process until a single pair is formed. The end product is a dendrogram, a graphic representation of clusters**:**

TABLE XII DENDROGRAM REPRESENTATION OF GENE



Genes / ***Samples***	**Gene B**	**Gene D**	**Gene A**	**Gene C**
Sample1	0	0	0.602	-0.481
Sample2	-0.096	0.114	0.301	0
Sample3	0.301	0.477	0.544	-0.602
Sample4	-0.301	0	0.176	-0.602
Sample5	0	-0.096	-0.096	0.0792

Many hierarchical clustering methods have an appealing property that the nested sequence of clusters can be graphically represented with a tree, called a dendogram[7]. Usually, each join in a dendogram is plotted at a height equal to the dissimilarity between the two clusters which are joined. Selection of *K* clusters from a hierarchical clustering corresponds to cutting the dendogram with a horizontal line at an appropriate height. Each branch cut by the horizontal line corresponds to a cluster. The result of the expression level analysis is usually presented as a dendrogram with an accompanying expression level table that has been reordered according to the clusters.

The picture emerging from our analysis is not simple one such as “Gene A is active in the tumor”. The true power of microarray technology is the possibility of assaying thousands of genes and hundreds of samples in the same experiment. Consider a slightly more complicated example. This experiment has 13 genes with 12 samples. They are ready to be stored by hierarchical cluster. The tree’s two major branches reveal two major groups of genes and samples as well:

Genes: C, J, H, F, B and G
Genes: L, K, E, M, D, I and A
Samples: 9, 8, 12, 4, and 10
Samples: 2, 7, 5, 6, 3, 11, and 1

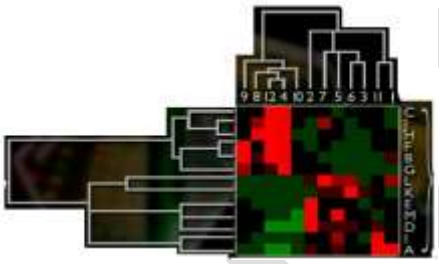


Fig.4.Microarray data with 13 genes and 12 samples

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] Holter, N.S., Mitra, M., Maritan, A. 2000 Fundamental patterns underlying gene expression profiles: Simplicity from Complexity, Proc.Natl.Acad. Sci., 97(15): 8409-8414

[2] Alizadeh, A.A., Eisen, M.B., Davis, R.E. 2000 Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling, Nature, 403:503-511

[3] C.B., Spellman, P.T., Brown, P.O., Botstein, D. 1998 Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci., 95(25): 14863-14868

[4] Schena, M., Shalon, D., Davis, R.W., Brown, P.O. 1995 Quantitative monitoring of gene expression patterns with a complementary DNA microarray, Science, 270:467-470

[5] Lashkari, D.A., DeRisi, J.L., McCusker, J.H., Namath, A.F., Gentile, C., Hwang, S.Y., Brown, P.O., Davis, R.W. 1997 "Yeast microarrays for genome wide parallel genetic and gene expression analysis", Proc. Natl. Acad Sci USA. 94(24):13057-13062

[6] Brazma, A., Vilo, J. 2000 Gene expression data analysis, FEBS Letters, 480: 17-24

[7] Alon, U., Barakai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J. 1999 Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proc. Natl Acad Sci USA. 96(12): 6745-6750