Identifying Potential Prognostic Biomarkers by Analyzing Gene Expression for Different Cancers

Pragya Verma
Department of Mathematics, Bioinformatics and Computer Applications
Maulana Azad National Institute of Technology, Bhopal, Madhya Pradesh, India

C.K. Verma
Department of Mathematics, Bioinformatics and Computer Applications
Maulana Azad National Institute of Technology, Bhopal, Madhya Pradesh, India

Abstract- Cancer as everyone knows it is a very harmful and deadly disease which causes number of deaths world-wide and still there are no hundred percent cures for this disease. So for curing cancer scientist and researchers are working world-wide to find a probable cure for this deadly disease but still full diagnosis and recovery is not available. For providing some help and information for further analysis for scientists and researchers this work provides helpful hand. The affymetrix data which contained the differential expression values for different cancer were included for the analysis. The analysis includes normalization and filtration and retrieving the up regulated and down regulated genes from the datasets. After retrieving the genes responsible for different cancers common genes list were identified in multiple cancers. This would be helpful for identifying probable biomarkers for ceasing cancer progression. Our further analysis includes clustering using DPClus. The common genes obtained after normalization were clustered and this generated dense cluster from these cluster the gene that could be our probable biomarker were identified and hence could be used for further computation study.

Keywords: Microarray Data, GEO datasets, GENOWIZ Tool, DPClus, DAVID.

I. Introduction

Cancer is one of the most common diseases in the world. Cancerous cells are uncontrolled growth in the body of abnormal cells. Cancerous cells are also called as malignant cells. Cancer grows in the body of normal cells. The normal cell will grow up when our body requires cells, and will die accordingly when the body will not need these cells [1]. Cancer appears when the number of cells in the body is not in controlled manner and calls starts divide rapidly thus this situation will occur when cell looses it memory to die. Cancers are caused by abnormalities of the transformed cells in the genetic material [2]. These abnormalities may be due to the affects of carcinogens, such as chemicals, tobacco, radiation, smoke or infectious agents. Now the cancer promoting cell which causes genetic abnormalities which might have occurred randomly are transferred from parents to offspring or it may by through DNA replication, so it would be present in all of the cells from birth of offspring [3]. Thus the inheritance of character of cancer are generally affected by complex interaction of the two namely genome of host and oncogenic gene.

DNA microarray is (also commonly known as gene chip, DNA chip, biochip, or gene array) is a collection of microarray DNA spots attached to a solid surface, such as glass, or silicon chip forming an array [4]. It is made to solve the problem of expression and profiling of unknown gene, so it is helpful in recording expression level of thousands of genes one by one. So microarray analysis is very useful for the improved diagnosis and for new innovative treatment by early detection the different cancer [5].

Microarray Datasets are the gene expression data which are stored in a repository of microarray database [6]. Microarray datasets for different cancer are available at an integrated platform “Gene Expression Omnibus (GEO)” which is a public repository which collect and without any charge servers free information regarding the microarray data, Next-Generation Sequencing information and also different other form of high-throughput functional genomic data. Thus GEO database stores measured data and manages search engine, and provides insight for the users for interpretation.

Biomarkers short form for biological markers are biological measures of a biological state. Biomarkers are active biomolecule that is used for indicating processes, biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention. Several parameters are considered for biomarker study such as: sensitivity, specificity, robustness, accuracy, reproducibility [7]. Biomarkers are required for study of chronic diseases, whose treatment may require patients to take medications for years to get accurate diagnosis. So biomarkers are becoming more and more important, because they confirm a difficult diagnosis for number of diseases. Together with powerful bioinformatics tools, search for cancer biomarkers is conducted and are capable of performing parallel rather than serial analysis, and it helps to identify distinguishing patterns and multiple markers rather than just a single marker; such strategies represent a central component and a paradigm shift in the search for novel biomarkers.

Related work - Recent works has described statistical methods for the identification of differentially expressed genes in replicated cDNA microarray experiments. Dudoit *et al.* in 2002 [8] described new method for pre-processing steps of image analysis and normalization were proposed for dataset analysis. Differentially expressed genes were identified based on adjusted p-values by a multiple testing procedure and the dependence structure between the gene expression levels [9]. This method is applied to microarray data for the genes from multiple slides and compared to those identified by recently published single-slide methods for gene expression in the livers of mice with very low HDL cholesterol levels and this study was carried by Parmigiani, Giovanni *et al*. [10]. Marker genes study done by Jiang *et al.* [11] Identified Biomarkers among different case involving patient with different or same disease. So biomarkers will help in efficient diagnosis of disease by identifying the minimum number of genes necessary for accurate prediction of disease status between healthy and cancerous patients.

Roded Sharan *et al.* in 2005 used the conservation, and found statistically significant support for 4,645 previously un-described protein functions and 2,609 previously un-described protein interactions [12]. The functions and interactions could not be identified from sequence similarity alone, demonstrating that network comparisons provided essential biological information beyond what is obtained from the genome [13].

Eitan Hirsh *et al*. in 2007 [14] mentioned his work and developed a probabilistic model for protein complexes that are conserved across two species. His model describes the evolution of conserved protein complexes from an ancestral species by protein interaction attachment and detachment and gene duplication events. Then he applied his model to search for conserved protein complexes within the PPI networks of yeast and fly, which are the largest networks in public databases. His work detected 150 conserved complexes that match well-known complexes in yeast and are coherent in their functional annotations both in yeast and in fly. This model yields higher specificity and sensitivity levels in protein complex detection.

Jian-xin Wang *et al.* in 2008 [15] described about the known complexes in protein networks, his paper proposes a new topological structure for protein complexes, which is a combination of sub-graph diameter (or average vertex distance) and sub-graph density. By his approach of that of the previously proposed clustering algorithm DPClus expands clusters starting from seeded vertices it were easy and faster. In his work he applied the algorithm DPClus to the protein interaction network of *Sacchromyces cerevisiae*. DPClus algorithm is based on the new topological structure which makes it possible to identify dense sub-graphs in protein interaction networks. Miyako Kusano *et al.* in 2011 [16] said about the graph clustering to the constructed correlation networks to extract densely connected metabolites and evaluated the clusters by biochemical-pathway enrichment analysis. He also demonstrates that the graph-clustering approach identifies tissue- and/or genotype-dependent metabolomic clusters related to the biochemical pathway.

II. Proposed Methodology

Different platform were utilized in this proposed method. National Centre for Biotechnology Information (NCBI), GENOWIZ (tool for Normalization of the expression data), and Microsoft office excel, DAVID, DPClus.

GENOWIZ tool is a comprehensive microarray data analysis package. Starting from data upload through analysis and visualization, Genowiz analyze microarray data of different formats from various sources. The program allows different things such as viewing expression data, preprocessing, normalization and filtration of microarray data. It Ensure data quality by adequate pre-processing methods and generate automated work flows. It aims to perform functional classification and pathway analysis. By this tool targets were identified quickly and reliably. So comparisons of expression data across multiple platforms were performed. From gene list comparisons and finally our desired result were obtained and exported data and images in standard formats.

For every research work particular protocol should be followed to find out the solution of a particular problem. So for the Microarray Data analysis there are particular steps that should be kept in mind. So the basic steps that are followed are listed below in the figure1.

A. *Retrieval of Data Sets*

Stanford microarray database is open source for exploring and downloading gene expression profiles which were downloaded in tabular format. The Gene Expression Omnibus (GEO) as it's a public repository which collect and without any charge servers free information regarding the microarray data. The website for downloading datasets is

htpp://www.ncbi.nic.nih.gov/.

B. *Data Analysis*

The microarray datasets are in very huge and large so the analysis of these data will be influenced by many numbers of variables. So for normalization of data it is needed to perform statistical analysis this is carried out by removing background noise. It is suitable for some platform and on large scale commercial use it might just a proprietary. For microarray data analysis GENOWIZ tool was kept into consideration for data analysis. The data analysis was performed by uploading the dataset in microarray data analysis tool.

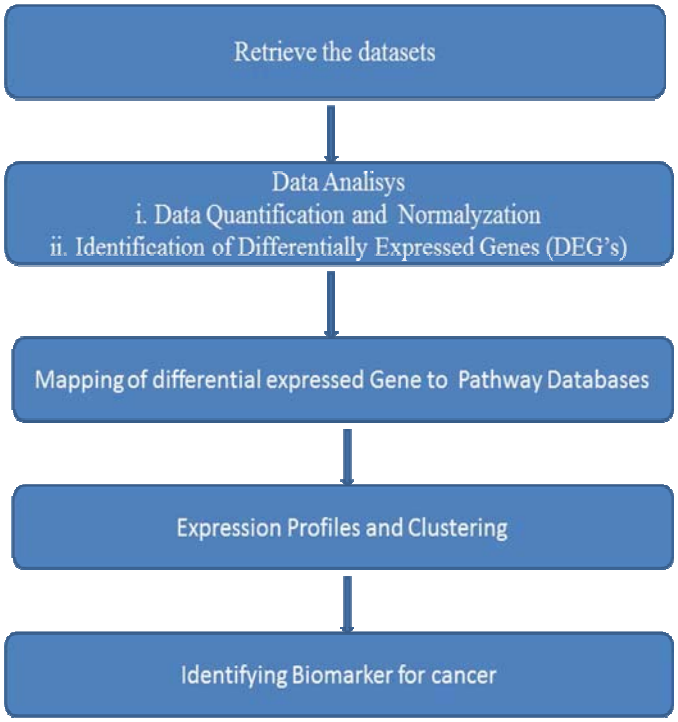
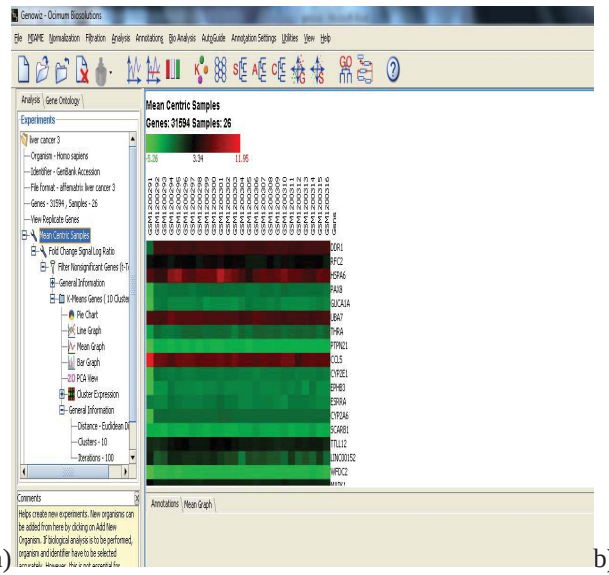


Figure 1 Methodology

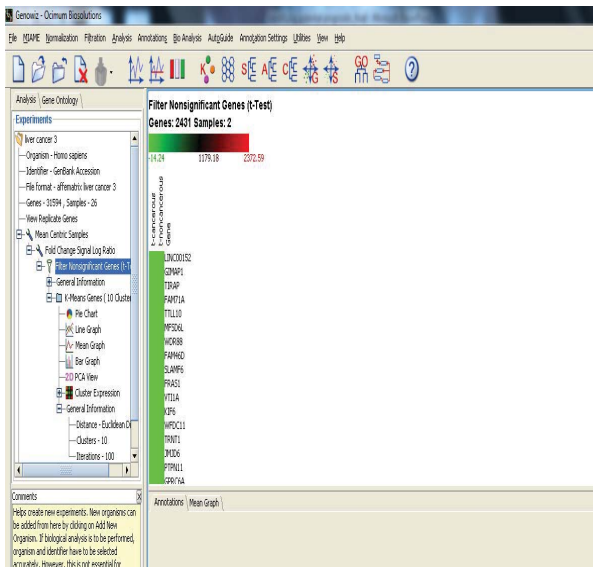
The two main following steps involved in data analysis are:

i. Data Quantification and Normalization

Data were normalized using GENOWIZ tool to make the samples comparable. GENOWIZ uses a stochastic model to estimate gene expression. This tool uses the probe data stored in Affymetrix .CEL files as input and converts the probe level expression data into gene level expression data. GENOWIZ preproscesses and normalizes the data to a certain extent in such a way that the distribution of expression values is comparable across the different samples. The following operations were performed on the raw datasets namely background correction, Local mean Normalization access Microarray surface, Logarithmic Transformation and Calculate Mean log Intensity and Log ratio.



a)



b)

Fig 2. Program window for Genowiz suite a) Importing Affemetrix data for normalization b) Gene list obtained after filtering and normalization.

ii. Identification of Differentially Expressed Genes (DEG's)

By GENOWIZ tool different experiments were conducted. The probes were distributed on values marked as 1 for over expressed and -1 for the under expressed. For normal a gene that is not differentially expressed 0 was assigned and blank field is left if not analyzed. Thus values were calculated for the probes. So in cancerous cell genes which were differentially expressed are selected. So more than one datasets were taken for different cancers namely for kidney, prostate liver and lung.

C. Mapping of Differentially Expressed Genes to Pathway Databases

The common differentially expressed genes obtained after Normalization were mapped to their pathway Databases. This gave the information about the genes and the pathway on which the gene acts. Some of the databases are KEGG Pathway, MetaCyc, Pathguide, BioCyc databases. The total differentially expressed genes; Unregulated and Down regulated were mapped to open source databases, this indexing provided us curated evidence and confirmation of these genes as differentially expressed. For our study the tool used is DAVID which is an online tool for annotation of common gene list which were obtained after normalization.

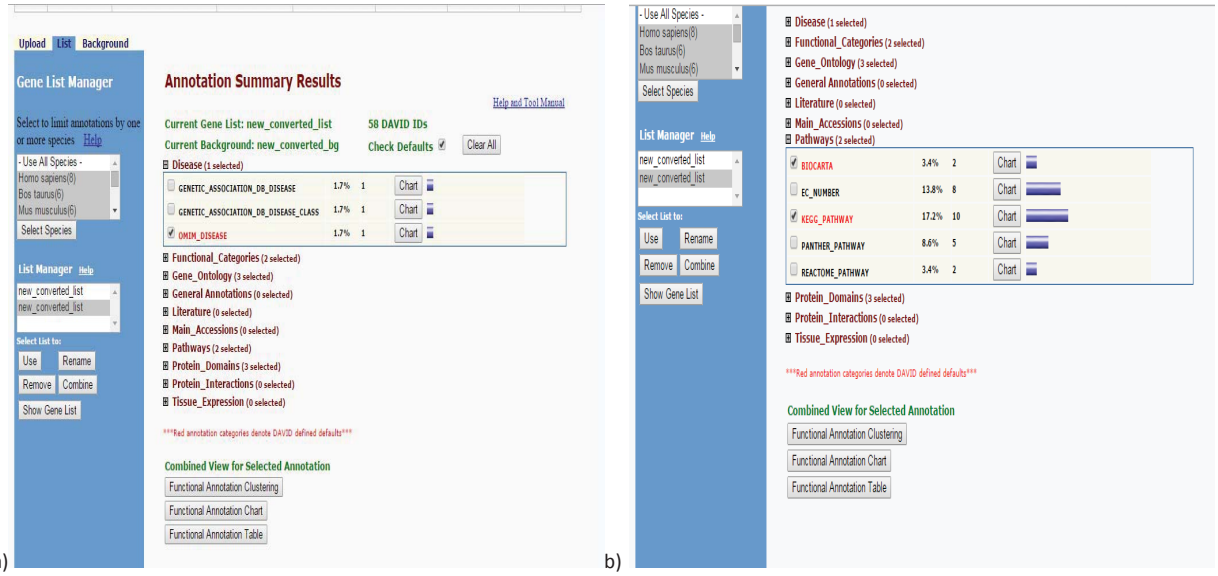


Fig.3 a) DAVID annotation results for disease ontology b) DAVID annotation results for pathway database

D. Expression Profiles and Clustering

Cluster analysis of gene expression data were commonly used to group gene expression measurements, longitudinally or cross-sectionally, into categories of genes that had similar patterns of expression. The objective is to cluster gene expression temporal profiles observed in one specific biological condition and here interest is in discovering the genes those were commonly expressed in different cancer disease. Simultaneously discovery of these genes expression level were identified and also group them.

- Gene expression profile for up and down regulated genes
- Clustering (Method: Complete Linkage clustering) -The complete linkage clustering (or the farthest neighbor method) is a method of calculating distance between clusters in Visualization hierarchical graph cluster analysis.
- Co-expression of genes in different diseases.

DPClus is a freely available tool for clustering of genes. The parameters were set for cluster property value, density values. DPClus receives two inputs:

1. Minimum density (din) $0 < d_{in} \leq 1$

2. Minimum value for cluster property (cp) $0 < cp_{in} \leq 1$

E. Identifying biomarker for cancer

The DPClus gave the clustering output in the form of Hierarchical graph. In the hierarchical graph a node represented a cluster and the edges represent the relations between the clusters The generated clusters were in such a way that intra cluster edges are green and inter cluster edges is red color. Some of the nodes were independent of clustering showing no interaction between the genes. The nodes which were interconnected with red and green lines were obtained and analyzed. Clusters with the high density were taken and nodes with the most number of edges interconnected were considered as the probable biomarker for the cancer disease.

III. RESULTS AND DISCUSSION

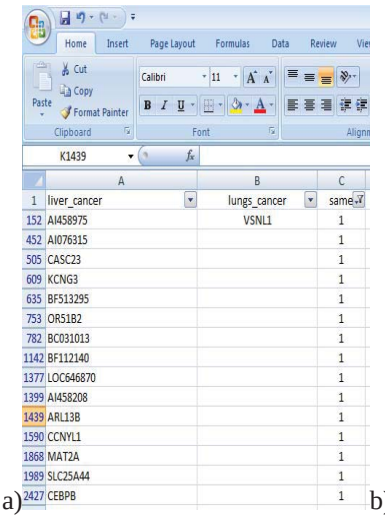
A set of genes were identified that were common and differentially expressed in multiple cancers those might be useful in the further analysis of fundamental signal transduction pathways that lead to carcinomas, so that these genes act as biomarkers for drug design.

Table1. List of common genes

Cancer name	**Common Genes**	**Common overexpressed Genes**	**Common Underexpressed Genes**
Lung cancer and liver cancer	15	8	7
Colon cancer and Kidney cancer	5	5	0
Colon cancer and liver cancer	2	1	1

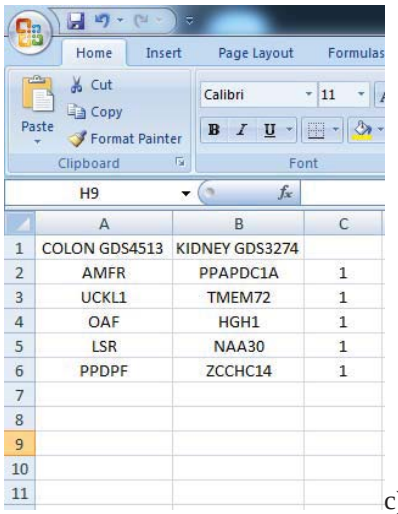
Table1 shows the common list of genes and these were analyzed by clustering. The common gene lists were annotated. The annotated gene lists were used and it gave vital information regarding disease pathway. By knowing about the disease progression in pathway the active biomolecule were identified. Each active molecule may be active biomolecule target. Similar work would be followed for other different cancer and thus disease progression can be given a halt.

a)



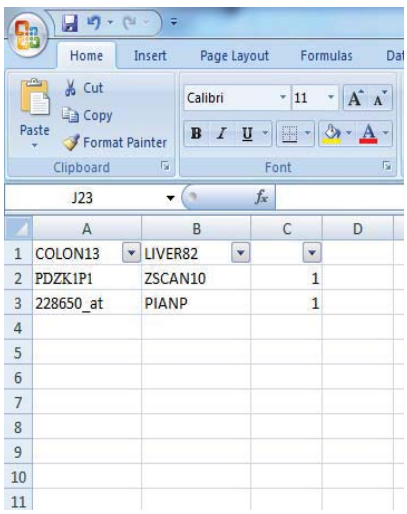
	A	B	C
1	liver_cancer	lungs_cancer	same
152	AI458975	VSNL1	1
452	AI076315		1
505	CASC23		1
609	KCNG3		1
635	BF513295		1
753	OR51B2		1
782	BC031013		1
1142	BF112140		1
1377	LOC646870		1
1399	AI458208		1
1439	ARL13B		1
1590	CCNYL1		1
1868	MAT2A		1
1989	SLC25A44		1
2427	CEBPB		1

b)



	A	B	C
1	COLON GDS4513	KIDNEY GDS3274	
2	AMFR	PPAPDC1A	1
3	UCKL1	TMEM72	1
4	OAF	HGH1	1
5	LSR	NAA30	1
6	PPDPF	ZCCHC14	1
7			
8			
9			
10			
11			

c)



	A	B	C	D
1	COLON13	LIVER82		
2	PDZK1P1	ZSCAN10	1	
3	228650_at	PIANP	1	
4				
5				
6				
7				
8				
9				
10				
11				

Fig.3 a) list of common gene in lung cancer and liver cancer (shows the name of list of common gene in lung cancer and liver cancer. 15 genes were found common and they contain both overexpressed and underexpressed genes in both the cancer.) b) List of common gene in Colon cancer and Kidney cancer (shows the name of list of common gene in Colon cancer and Kidney cancer. 5 genes were found common and they are overexpressed in both the cancer.) C) List of common gene in Colon cancer and Liver cancer (shows the name of list of common gene in colon cancer and liver cancer. 2 genes were found common and they contain both overexpressed and down regulated genes in both the cancer.)

So after analyzing data for different cancerous sample (e.g. diseased vs. normal) clustering in the previous steps obtained genes that were common and have some interaction and were defined in following group distribution. After careful data study, a handful of marker genes were selected and classified. For the data sets for lungs and liver cancer 15 genes were common in both the cancer. Among these 15 common genes, 3 are tumor-suppressor genes (*CEBPB, ARL13B, KCNG3),* 2 Oncogenes (*MAT2A, CASC23*), and 10 cancer-related genes (*SLC25A44, CCNYL1, AI458208, LOC646870, BF112140, BC031013,* OR51B2, BF513295, AI076315, AI458975). After the clustering analysis it is confirmed that "*MAT2A"* gene would be used as marker gene and it is associated with different metabolic pathways. Similar analysis shows that "*KCNG3"* which is common gene will be a marker gene and would act as an active bimolecular target. Five genes were found common in colon and kidney cancer. Among the five common genes *AMFR* is oncogenic gene while 4 are cancer-related genes (*UCKL1, OAF, LSR* and *PPDPF*). The clustering result the gene "*AMFR*" is considered as marker gene for the active biomolecule target. Only two genes were found in common and differentially expressed in colon and liver. The *PDZK1P1* gene is a cancer-related gene and after clustering analysis of this gene it showed that this would act as a biomarker.

IV.CONCLUSION

Common DEGs were found out in multiple cancers, and classified into two groups the over expressed and under expressed genes. By finding out the DEGs and performing clustering, different gene were interacting with each other and which shows evidences supporting significant value for Tumor progression. Potential molecular Biomarker for Genes that are responsible for the Tumor progression were identified also predicted their underlying functions by Computational biology and clustering method. With our analysis genes that would be responsible and act as active biomolecule and used as target for drug response for curing particular disease were found. The densed plots were obtained from the cluster analysis. The common lists identified were used for clustering. These genes were annotated. The annotated gene list gave vital information regarding disease pathway. By knowing about the disease progression in pathway active biomolecules were studied. Each active molecule might be biomolecule target and but the exact prediction were done by using modeling and final analysis would be carried in wet lab and validation with experiments. So for the researchers and scientists this work would be really helpful and provide an insight to study regarding cure of cancer. Designing an inhibitor requires active site which is given by the experiment conducted. Similar work would be followed for other different cancer and disease progression would be given a halt.

REFERENCES

[1]. Strachey, Ray, and Florence Nightingale. *"The Cause": A Short History of the Women's Movement in Great Britain*. ICON Group International, 1928.

[2]. Maurano, Matthew T. "Systematic localization of common disease-associated variation in regulatory DNA." *Science* 337.6099 (2012): 1190-1195.

[3]. Lu, Jun. "MicroRNA expression profiles classify human cancers." *nature*435.7043 (2005): 834-838.

[4]. M. Schena, D. Shalon, R.W. Davis, P. O. Brown, "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray", *Science*, Volume 270, No. 5235, 1995, pp. 467-70.

[5]. Cheng, Jill, et al. "NetAffx Gene Ontology Mining Tool: a visual approach for microarray data analysis." *Bioinformatics* 20.9 (2004): 1462-1463.

[6]. Wang, Xiaofi. "miRDB: a microRNA target prediction and functional annotation database with a wiki interface." *Rna* 14.6 (2008): 1012-1017.

[7]. Strimbu, Kyle, and Jorge A. Tavel. "What are biomarkers?." *Current opinion in HIV and AIDS* 5.6 (2010): 463.

[8]. Dudoit, Sandrine, "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments." *Statistica sinica*12.1 (2002): 111-140.

[9]. Bhattacharjee, Arindam, "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses."*Proceedings of the National Academy of Sciences* 98.24 (2001): 13790-13795.

[10]. Parmigiani, Giovanni, *The analysis of gene expression data: an overview of methods and software*. Springer New York, 2003.

[11]. Jiang, Hongying, "Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes." *BMC bioinformatics* 5.1 (2004): 81.

[12]. Sharan, Roded, "Conserved patterns of protein interaction in multiple species." *Proceedings of the National Academy of Sciences of the United States of America* 102.6 (2005): 1974-1979.
[13]. Bandyopadhyay, Sourav, Roded Sharan, and Trey Ideker. "Systematic identification of functional orthologs based on protein network comparison."*Genome research* 16.3 (2006): 428-435.
[14]. Hirsh, Eitan, and Roded Sharan. "Identification of conserved protein complexes based on a model of protein network evolution." *Bioinformatics* 23.2 (2007): e170-e176.
[15]. Wang, Ling, "Conditional clustering of temporal expression profiles." *BMC bioinformatics* 9.1 (2008): 147.
[16]. Fukushima, Atsushi, "Metabolomic correlation-network modules in Arabidopsis based on a graph-clustering approach." *BMC systems biology* 5.1 (2011): 1.