International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Issue 2, Volume 2 (February 2015) www.ijirae.com

A Statistical Approach to Identify Cancer-Related Vital Genes

Jahiruddin
Department of Computer Science,
Jamia Millia Islamia (Central University),
New Delhi-110025, India

Abstract— To identify genes that play an important role in cancer treatment is an important task. In this paper we tried to construct gene regularity network from microarray gene expression profile using statistical approaches. The used microarray gene expression profile has 20 samples out of which 12 are cancerous and 8 are normal. We applied the statistical techniques to compare the samples by using GEO2R tool. We takes top 250 genes and filtered out the gene that does not have the gene names and the collect the expression values in different samples of remaining genes. We use the Singular Value decomposition (SVD) to get the modified expression values of the genes in different samples. The interaction between filtered genes is calculated using correlation and five different networks are analysed to identify vital genes related to cancer. Finally, validation of the result is done with biological literature.

Keywords— Cancer, Microarray gene expression, Gene expression omnibus, Gene regulatory network, Singular value decomposition, Latent semantic analysis

I. INTRODUCTION

Cancer also known as malignant tumor is a disease involving uncontrolled cell growth. The cancerous cells divide and grow uncontrollably and infect the nearby body cells. The mostly cases of cancer causes due to use of tobacco, alcohol, obesity, and poor diet. Some other factors of cancers are radioactive radiations, pollutions, and certain infections. The cancers are developed due to genetic changes in the cells by above causes. Very less number of cancer cases is due to genetic defects inherited from their ancestors. Initially cancers are diagnose by their symptom or screening, but finally they are confirms by medical tests. People with suspected cancer are confirms by blood test, CT scans, endoscopy, and biopsy. Generally the cancers are treated by surgery, chemotherapy, radiation therapy, targeted therapy etc. The identification of the gene responsible for cancer may play very import role in its treatment.

Due to rapid grow in microarray technique, the analysis of expression of the thousands genes can be done in a single experiment with small sample size [11]. This technology allows us to measure the expression levels of a large number of genes simultaneously in a microarray experiment [9]. Microarray technique may be used in many applications like disease diagnosis, drug discovery, gene discovery, and toxicogenomics. Recently microarray technology has been extensively used by the researchers. There is a huge amount of microarray data, but it is scattered and is not available for public use. The National Center for Biotechnology Information (NCBI) has the Gene Expression Omnibus (GEO) data repository to a large number of microarray data from various sources. It also has the GEO2R web tool for comparing two or more group of samples to identify genes that are differently expressed. It performs comparisons on original processed data using the GEO query and limma R package. It performs a number of statistical testing based on corrected p-value.

A microarray gene expression data is a matrix A of expression values (real numbers) of order m x n, where m is number of gene, n is number of samples, and a_{ij} is expression value of i^{th} gene in j^{th} sample. In microarray data there are problems of dimensionality (m >> n) and noise. There is no control on these problems as this is introduced due to experimental limitation.

Microarray based prediction of vital gene related to cancer is new and growing area of research. A gene regularity network (GRN) is a collection of DNA segments in a cell is used to govern gene expression levels of mRNA and proteins. It is used to model regulatory interaction in the cell and represent the gene regulation. Microarray data can be used to construct cancer specific gene regulatory network. The expression value of the gene in different in normal and cancer samples provide the information to get the relationship between gene pairs that may be used to construct the GRN. A central problem in system biology is mapping the topology of gene regulatory networks [8]. Accurate computational method needed for construction of GRN from gene expression profiles.

SVD is an important factorization of a rectangular real or complex matrix [4]. It is based on a theorem from linear algebra which states that a rectangular matrix A of order m × n can be decomposed into the product of three matrices – an orthogonal matrix U of order m × n, a diagonal matrix S of order n × n, and an orthogonal matrix V of order n × n. The columns of the matrix U are orthogonal eigenvectors of AA^T . Similarly, columns of V are orthogonal eigenvectors of A^TA, and S is a diagonal matrix containing the square roots of eigenvalues from U or V in descending order. SVD has the added benefit that the representation of genes that share samples become more similar to each other, and genes that were dissimilar may become more dissimilar. Presently, SVD is included in a number of software and java classes like MATLAB, JAMA java package etc. Caporaso et al. [2] explored Latent Semantic Analysis (LSA) that uses the SVD, for biomedical question answering system. They explained that LSA increases the number of phrases returned in response to a question.

In the literature a number of computational methods have been proposed to model GRN. It includes, Boolean networks, graph method, Bayesian networks, differential equations and so on [5], [6], [8], [10]. Madhamshettiwar et al. reconstructed cancer specific GRN [7]. They study application of gene regulatory network inference to ovarian cancer. An exhaustive state of art study for GRN modelling is done in [7]. They applied best method to infer GRN of ovarian cancer. In [1] information theoretic approach is used to construct gene regulatory network. They calculate the mutual information between genes to get the interaction between genes from gene expression profile. In this paper, statistical approach is used to reconstruct GRN from gene expression data and to identify important gene related to cancer. We applied this approach to generate the GRN of colorectal cancer from microarray data, which is a genetic disease.

II. PROPOSED SYSTEM

We now present the complete architecture of our system, which is designed to identify vital gene related to cancer from microarray gene expression profile. This may use for better diagnosis and treatment of collector cancer. The proposed system is shown in figure 1. This is characterized by following key functionalities – *Microarray data loading and analysis, Genes and their expression values extraction, relationship identifying between gene pairs and GRN generation,* and *vital gene identification* and *verification.* A brief description about these functionalities is given in the following paragraphs:

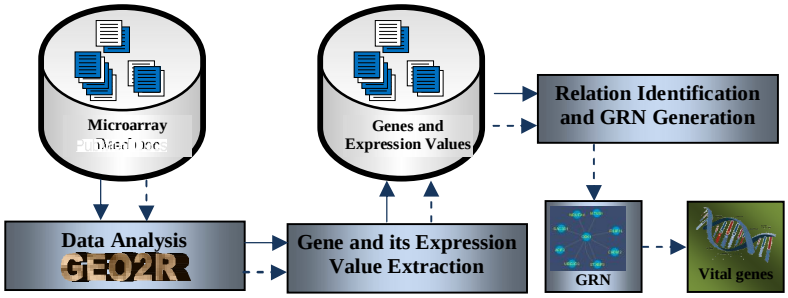

Fig. 1 System architecture

- Microarray data loading and analysis is focused to download the microarray data and analyse them. For this purpose GEO's accession viewer interface may be used. The GEO accession number of required microarray data is entered in the text field and go button is clicked to download the data. It returns the microarray data with its description. Thereafter, we can analyse the data using GEO2R tool. This tool does the ANOVA testing on the data and arranges the genes in ascending order of their p-values.
- Genes and their expression values extraction is focused on extraction of the efficient genes and downloading their expression values. The extraction of efficient genes is based on their p-values. The genes with p-value less than 0.01 are treated as an efficient gene and their expression value is to be downloaded. The gene that does not have gene name is required to filter out.
- Relationship identifying between gene pairs and GRN generation is focused on identifying relationship among gene pairs and generating the gene regularity networks (GRNs). First of all singular value decomposition (SVD) is used for factorization of gene expression matrix. It brings related gene more close and unrelated genes more far. Thereafter, interactions between gene pains are generated by calculating Pearson's correlation coefficient. Based on absolute value of correlation coefficients it is decided whether a relation between gene pair exist or not. We arrange the gene pairs in ascending order of absolute value of their correlation coefficients. The gene regularity networks (GRNs) are generated using these sorted gene pairs.
- Vital gene identification and verification focused to identify vital gene and verify them from existing literature. For this purpose we use the generated GRNs and identify the important genes that have maximum degree. Thereafter, the genes that present as an important gene in all the GRNs are declared as vital genes. It is required to verify these genes from existing literatures. An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

Further details about these functionalities along with the experimental results are presented in the following sections.

III. EXPERIMENTAL SETUP AND RESULTS

In this section, we present our experimental setup and results. We have used the GEO of NCBI to load microarray gene expression profile data using GEO accession number GSE4988 [3]. Gene expression profiling of circulating plasma RNA is a tool that is used to detect cancer and study tumor progression and therapy responsiveness. There are total 20 samples out of which 12 samples are taken from colorectal cancer patients and 8 from healthy donors. There were total 15552 genes expressions, but a large number of the genes do not have gene names. In general cases the expression value of genes in cancer samples are higher than that of in normal samples. Figure 2 show the expression value of gene 'DDX46' in normal and cancer samples. From figure 2 it is clear that many of the normal sample profile are down regulated.

After loading this data from GEO, we analyse it using GEO2R tool. GEO2R web tool perform the ANOVA testing using the GEO query and limma R package. It sort the gene in ascending order of p-values. There are options to select different columns and save the results. There are a number of options to choose different algorithms and to set the parameters. We use the default parameters and Benjamini & Hochberg algorithm for analysis.

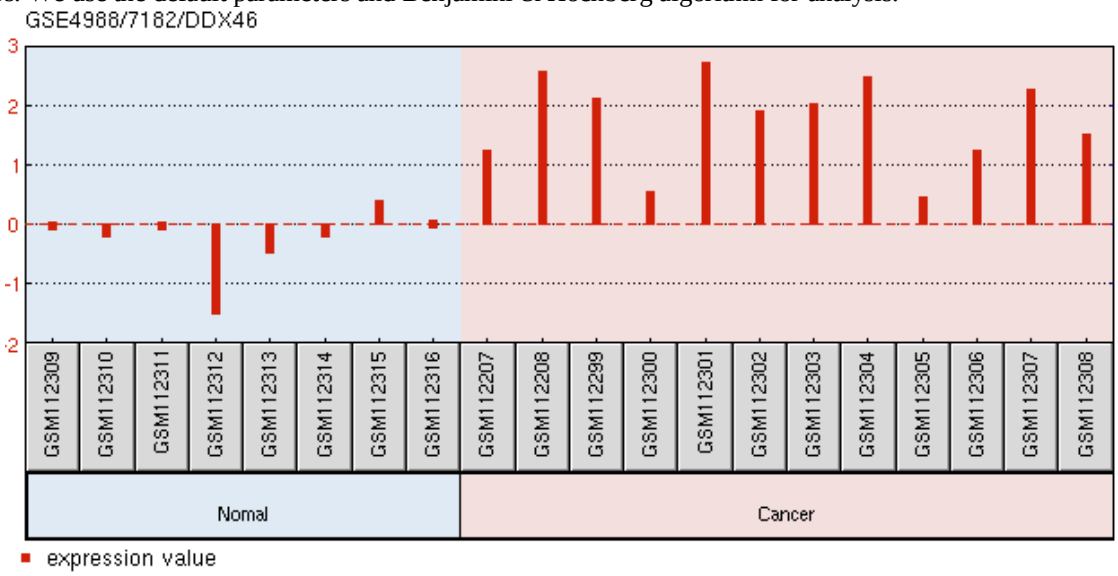

Fig. 2 Expression values of gene 'DDX46'in normal and cancer samples

There is a large number of genes, majority of them are irrelevant. After analysis, based on p-value we take the gene for further analysis. We select the gene from the analysed data whose p-value is less than 0.01. Thereafter, we eliminate the gene whose name is missing and get the total of 105 genes. Then we download the expression values of these genes for all samples. The table 1 show the partial list of genes and their expression values in the samples.

TABLE I
PARTIAL LIST of GENE and their EXPRESSION VALUES in the SAMPLES

	Sample ID↓ Gene→	DDX46	HNRNPH3	ARMC1	BAMBI	EPAS1
Normal	GSM112309	-0.03	-0.55	-1.26	-0.1	-2.16
	GSM112310	-0.24	-0.63	-4.26	0.52	0.47
	GSM112311	-0.03	-0.19	-7.41	0.74	-3.08
	GSM112312	-1.52	-3.57	-3.85	2.14	1.71
	GSM112313	-0.51	-1.22	-1.2	0.98	-1.09
	GSM112314	-0.23	-1.34	0.12	2.41	-0.66
	GSM112315	0.42	-0.9	-0.89	0.8	-1.34
	GSM112316	0.01	-1.19	-0.13	1.45	-1.86
Cancer	GSM112207	1.27	1.05	0.16	-0.84	3.1
	GSM112208	2.6	-0.63	-1.03	-0.65	-1.65
	GSM112299	2.14	0.87	1.17	-1.1	0.56
	GSM112300	0.58	-0.61	2.41	-2.15	-1.63
	GSM112301	2.73	2.35	2.66	-1.35	2.81
	GSM112302	1.93	1.68	1.07	0.77	3.15
	GSM112303	2.04	0.67	1.4	-0.09	2.17
	GSM112304	2.49	1.92	1.37	0.1	3.62
	GSM112305	0.47	1.83	0.4	-2.05	3.64
	GSM112306	1.28	1.2	0.89	-3.78	2.56
	GSM112307	2.3	0.89	0.31	-0.84	2.85
	GSM112308	1.53	1.39	1.05	-1.26	2.64

From the above data we get a gene expression matrix A. As we selected total 105 genes and there are 20 samples therefore the order of this matrix will be 105 × 20. Where a_{ij} element of this matrix represent the expression value of i^{th} gene in j^{th} sample. Now we apply the singular value decomposition (SVD) to factorize this matrix into – an orthogonal matrix U of order 105 × 20, a diagonal matrix S of order 20 × 20, and an orthogonal matrix V of order 20 × 20. The rows of the matrix U represent the gene and columns represent the sample. Here each row vector of matrix U is corresponds to a gene. We apply the Pearson's correlation on rows of matrix U to get the relation between pair of genes. As there are total of 105 genes therefore total 5460 correlation coefficient are calculated. We assume an interaction between a pair of genes if absolute value of correlation coefficient between those gene pair is greater than or equal to 0.50 as rounding of this gives 1. Then we sort gene pairs in descending order of absolute value of their correlation coefficient.

Thereafter we generate five gene regularity networks by taking top 30, 40, 50, 60, and 70 gene pairs. For these tasks we have written a java program.

In gene regulatory network the nodes are corresponds to gene and edges are corresponding to their interaction. The gene regulatory network-3 is shown in figure 3. From figure 3 it is clear that gene 'FILIP1L'and 'KCNQ2' are highly connected genes with a degree of 4. From network-1 we taken the genes with maximum degree as important genes and from other four networks the genes with maximum and second maximum degree are considered as important genes. The observation of highly connected gene in all five networks is shown in table 2.

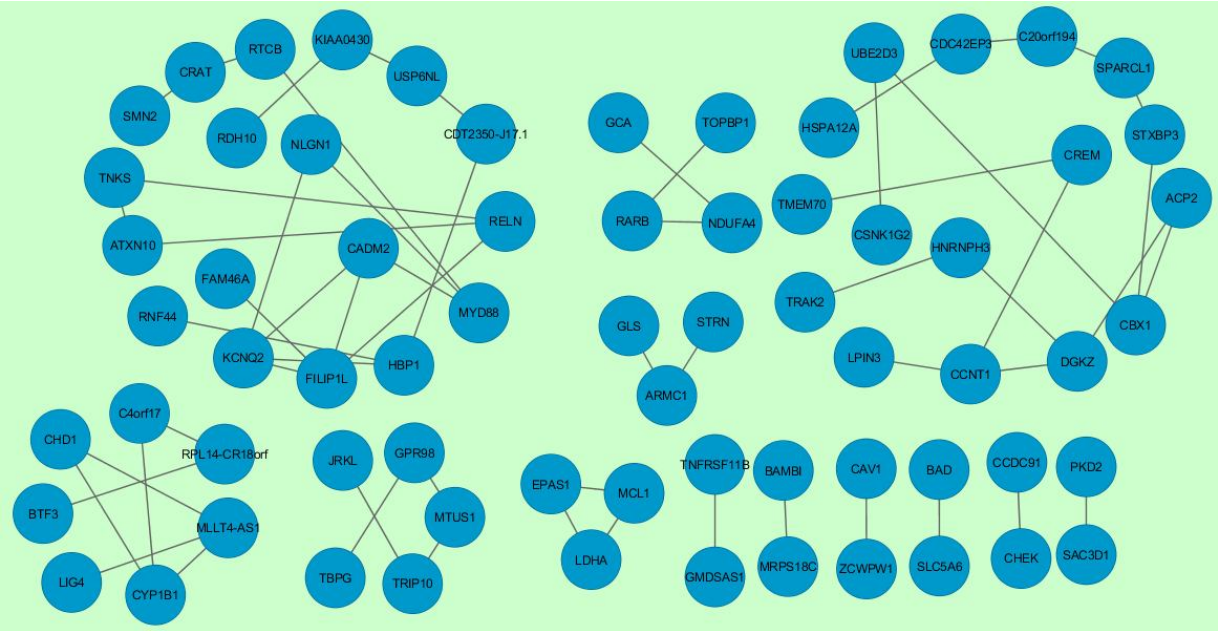

Fig. 3 Gene regulatory network-3

From the table it is clear that the gene 'CBX1' and 'KCNQ2' exist in all the networks as important gene therefore these two genes are more important for the diagnosis and treatment of the cancer and is a vital gene. After these two genes, the genes FILIP1L, CYP1B1, and MLLT4-AS1 exist in four networks except the network-1. Therefore, we can say that these genes are also important but not as the previous two genes for cancer diagnosis and treatments. These five genes are validated with the biomedical literatures. All other genes may be needed for validation in vet-labs if it does not exist in the literatures. This method also works on simulated dataset.

TABLE II
FIVE GENE REGULARITY NETWORKS and IMPORTANT GENES

Network No.	No. of edges	No. of nodes	Important genes
Netwok-1	30	45	CBX1(2), UBE2D3(2), STXBP3(2), RARB(2), NDUFA4(2), MYD88(2), LDHA(2), MCL1(2), EPAS1(2), ATXN10(2), MLLT4-AS1(2), CDC42EP3(2), KCNQ2(2)
Netwok-2	40	54	CBX1(3), FILIP1L(3), UBE2D3(2), STXBP3(2), RARB(2), NDUFA4(2), MYD88(2), LDHA(2), MCL1(2), EPAS1(2), ATXN10(2), MLLT4-AS1(2), CDC42EP3(2), KCNQ2(2), ACP2(2), DGKZ(2), CCNT1(2), CHD1(2), CYP1B1(2), C4orf17(2), CADM2(2), RELN(2), RTCB(2), CRAT(2)
Netwok-3	50	61	FILIP1L(4), KCNQ2(4), CBX1(3), CYP1B1(3), MLLT4-AS1(3)
Netwok-4	60	68	FILIP1L(4), KCNQ2(4), CBX1(3), CYP1B1(3), MLLT4-AS1(3), RELN(3), HBP1(3), MYD88(3), CADM2(3), DGKZ(3), CCNT1(3)
Netwok-5	70	71	FILIP1L(4), KCNQ2(4), RELN(4), CBX1(3), CYP1B1(3), MLLT4-AS1(3), HBP1(3), MYD88(3), CADM2(3), DGKZ(3), CCNT1(3), RARB(3), CREM(3), EPAS1(3), LDHA(3), ARMC1(3)

IV. CONCLUSION AND FUTURE WORK

In this paper, we have presented an application of statistical approach to identify vital cancer-related genes. To this end, we have generated five different gene regulatory networks from microarray gene expression profile, and used ANOVA to compare various genes based on their expression values. Genes having p-values less than 0.01 are considered and genes having no gene names were filtered out. As a result, total number of 105 genes are identified and their expression values for all the samples are generated. Thereafter, singular value decomposition for matrix factorization is used which brings related gene more closer and unrelated genes more farther.

The Pearson's correlation is used to determine interaction between gene pairs, and an edge between a pair of genes is created if the absolute value of correlation coefficient is greater than or equal to 0.5. The generated network is analyzed and the genes with maximum and second maximum degree are considered as important genes. Finally, 'CBX1' and 'KCNQ2' genes are declared as vital genes as thay are present in all networks, and on analysis it is found that these two genes play important role in cancer diagnosis. In future, we have planned to extend our experiment on a larger

REFERENCES

[1] A.J. Butte, and I.S. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," in Proceedings of *Pacific Symposium on Biocomputing 2000*, p. 418–429.

[2] J.G. Caporaso, W. Baumgartner, H. Kim, Z. Lu, H.L. Johnson, O. Medvedeva, A. Lindemann, L.M. Fox, E.K. White, K.B. Cohen, and L. Hunter, "Concept recognition, information retrieval, and machine learning in genomics question-answering," in Proceedings of the *15th Text Retrieval Conference (TREC'2006)*, Gaithersburg, Maryland.

[3] M. Collado, V. Garcia, J.M. Garcia, I. Alonso, et al., "Genomic profiling of circulating plasma RNA for the analysis of cancer," *Clin Chem 2007 Oct;* vol. 53(10), pp. 1860-1862, 2007.
http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4988

[4] G.E. Forsythe, M.A. Malcolm, and C.B. Moler, *Computer Methods for Mathematical Computations*, Prentice Hall Professional Technical Reference, ISBN:0131653326, 1977.

[5] H.D. Jong, "Modeling and simulation of genetic regulatory systems: a literature review," *Journal of computational biology*, vol. 9(1), pp. 67-103, 2002.

[6] G. Karlebach, and R. Shamir, "Modelling and analysis of gene regulatory networks," *Nature reviews. Molecular cell biology*, vol. 9(10), pp. 770-780, 2008.

[7] P.B. Madhamshettiwar, et al., "Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets," *Genome Medicine*, vol. 4(41), 2012.

[8] S.R. Maetschke, P.B. Madhamshettiwar, M.J. Davis, and M.A. Ragan, M. A, "Supervised, semi-supervised and unsupervised inference of gene regulatory networks," *arXiv:1301.1083v1*, 2013.

[9] S.K. Mohapatra, and A. Krishnan, "Microarray data analysis," *Methods in Molecular Biology*, Vol. 678, pp. 27-43, 2011.

[10] K. Raza, and R. Parveen, "Evolutionary Algorithm in Genetic Regulatory Netowrks Model," *Journal of Advanced Bioinformatics Applications and Research*, vol. 3, no. 1, pp. 271–280, 2012.

[11] X. Wang and O. Gotoh, "Microarray-based cancer prediction using soft computing approach," *Cancer informatics*, vol. 7, pp. 123-139, 2009.

