IOSR Journal of Pharmacy and Biological Sciences (IOSR-JPBS)
e-ISSN: 2278-3008, p-ISSN:2319-7676. Volume 9, Issue 2 Ver. III (Mar-Apr. 2014), PP 62-66
www.iosrjournals.org

In Silico Gene Expression Data Analysis of Head and Neck Paired Squamous Cell Carcinoma

Vijay Laxmi Saxena [1], Pragya Chaturvedi [2], Roohi Fatima [3]

[1]*(Coordinator, Bioinformatics infrastructure facility center, Department of zoology, D.G.P.G. College, Kanpur, India and Member of the Academic Council* of *KIIT University, Bhubaneswar,India*)

[2]*(Research scholar, School of Biotechnology, KIIT University, Bhubaneswar, India &*
Bioinformatics infrastructure facility center, Department of zoology, D.G.P.G. College, Kanpur, India)

[3]*(Department of Biotechnology, D.G. P.G. College, Kanpur, India)*

Abstract : *The recent explosion in the number of gene expression experiments and the need to extract useful biological information from the resulting volumes of data has presented a new challenge, that is, to identify the most significant changes in gene expression and to interpret the results in terms of biological relationships. There are a number of web resources that provide data annotation services for gene expression datasets. Microarray Analysis has become a huge widely used tool for the generation of gene expression data on functional genomic scale. The information gained offers an unprecedented opportunity to fully characterize biological processes. The global burden of cancer continues to increase largely because of the aging and growth of the world population alongside an increasing adoption of cancer-causing behaviours, particularly smoking, in economically developing countries. Mainly the Head and Neck squamous cell carcinoma, and is one of the leading causes of human death worldwide and the deaths from cancer are projected to continue rising. With the help of Bioinformatics techniques the analysis of gene expression data is done through BRB Array tool .Microarray data is analysed using PAM and Hierarchical clustering. The two genes i.e. Keratin 4 and Mal, T-cell differentiation protein are predicted through PAM. These genes can have important role in the Head and Neck squamous cell carcinoma.*

Keywords: *Clustering, Gene expression , Head and neck squamous cell carcinoma ,Microarray, PAM*

I. INTRODUCTION

MICROARRAY have become extremely useful for analysing gene expression phenomenon but establishing a relation between microarray analysis results (typically a list of genes) and their biological significance is often difficult[9]. Microarrays have revolutionised biology as a science and have ignited the bioinformatics revolution. The large amount of data produced by microarray experiments requires constant computational support in order to be of any use. The success of genome sequencing projects has led to the identification of almost all the genes responsible for the biological complexity of several organisms. Microarray are becoming increasingly more common laboratory for simultaneous change in expression across a large number of genes .Image data from the array lead to gene specific numerical intensities representing relative expression level, and these in turn form the input to computational analysis designed to access significance and relationship across biological samples[18]. This high-throughput technique can be used to predict the function of unknown genes, in medical diagnostics, in biomarker discovery, to infer networks from the regulatory interactions between genes, and to investigate the mechanisms by which a drug, disease, mutation and environmental condition affects gene expression and cell function. Large datasets are produced, particularly from whole-genome arrays, and public databases hold substantial quantities of gene-expression information [20].Microarrays facilitate the discovery of totally novel and unexpected functional roles of genes. The power of these tools has been applied to a range of applications, including discovering novel disease subtypes, developing new diagnostic tools, and identifying underlying mechanisms of disease or drug response [8].

Cancer is a potentially fatal disease caused mainly by environmental factors that mutate genes encoding critical cell-regulatory proteins. The resultant aberrant cell behaviour leads to expansive masses of abnormal cells that destroy surrounding normal tissue and can spread to vital organs resulting in disseminated disease, commonly a harbinger of imminent patient death [12]. Head and neck squamous cell carcinoma (HNSCC), which affects the oral cavity, the oropharynx, the larynx and the hypo pharynx, is the sixth most common cancer among men in the developed world. Well-known risk factors include tobacco and alcohol. Over the last decades, diagnosis and management have improved, but not long-term survival rates. The prognosis of HNSCC is influenced by many factors, such as TNM staging and pathological grading of differentiation. However, since these factors are not sufficient [25].We found that several major biologic systems or pathways were globally altered in HNSCC, at least in this set of studies. Among these are some previously implicated in HNSCC pathogenesis, such as cell cycle control [17] . This Bioinformatics based analysis of gene expression data of head and neck squamous cell carcinoma leads to identification of novel genes involved in HNSCC pathogenesis..

II. Tools and databases

A. Array express (database)

Array Express is a public database of microarray gene expression data at the EBI, which is a generic gene expression database designed to hold data from all microarray platforms[2].ARRAY EXPRESS relates to European Bioinformatics Institute (EBI), Europe, established 2002 .It stores microarray data only and its main aim is data repository as well as well-annotated data warehouse. [10]

B. Brb array tool (biometric research branch)

BRB-Array Tools is an integrated software system for the comprehensive analysis of DNA microarray experiments. The version of BRB Array used is 4.3.1. It was developed by professional biostatisticians by Dr. Richard Simon Biometrics Research Branch National Cancer Institute And Amy Lam The EMMES Corporation February 21, 2002 . The software is designed for use by biomedical scientists who wish to have access to state-of-the-art statistical methods for the analysis of gene expression data and to receive training in the statistical analysis of high dimensional data. [19]

C. System requirement

- PROCESSOR: Intel (R) core (TM) i3-2100 CPU
- PROCESSOR SPEEDS: @3.10 GHz 3.10GHz
- RAM: 2.00GB
- SYSTEM TYPE: 32 bit operating system
- OPERATING SYSTEM: WINDOWS 7

III. Methodology

Forty -four samples were being taken of patients suffering from head and neck squamous cell carcinoma from the array express database in which Twenty-two patients were normal and other twenty-two are in cancerous stage. These samples are analyzed through BRB Array tool technique which is an integrated software system for comparing the gene expression. The expression is compared between normal versus disease, after reiterating and normalization the actual genes are produced in the form of heat map. Prediction analysis of microarray is done to identify novel genes involved in HNSCC.

IV. Results and discussion

4.1. Cluster analysis of sample (Hierarchical)

Cluster analysis of samples is done with the help of BRB Array tool, Java tree viewer option. The current analysis is based on hierarchical clustering. According to which we can divide samples into two major groups. The clustering of sample explains the gene expression of each sample which is denoted with three different colors. Here the Green color Cluster analysis of samples is done with the help of BRB Array tool, Java tree viewer option. The current analysis is based on hierarchical clustering. According to which we can divide samples into two major groups. The clustering of sample explains the gene expression of each sample which is denoted with three different colors. Here the Green color represents the normal genes, red signify the diseased genes and black represents unexpressed genes. These samples are divided into two groups 1 and 2 and further subdivided into four subgroups a, b, c and d. These subgroups are further classified into various sub clusters. Samples are clustered according to distance between them.

We can see that in cluster 1 the sub cluster a show mainly the red images which are expressing as diseased genes and some of them are expressing normal too. In subgroup b mainly the black colour is been highlighted which are unexpressed genes and some are normal and disease too. Now in cluster 2 we can see under sub cluster c mainly the green colour images is been seen which means normal genes are highly expressed. In group d different images of all 3 colors is presents i.e. green, black, red means normal, diseased and unexpressed expression level is present .In a broad view we can say that cluster 1 is showing the diseased samples and the group 2 shows the normal samples .However , other expression can also be seen in both of clusters.

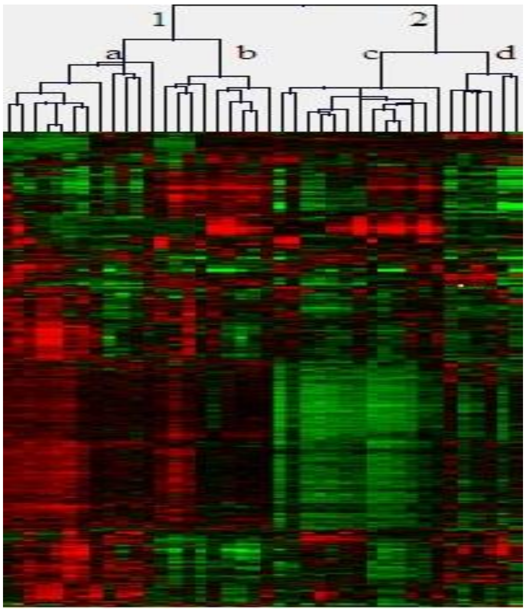

Fig 1- Hierarchical clustering of samples

4.2. Prediction analysis of microarray (PAM)

(PAM) Prediction analysis of microarray is done with the help of BRB Array tool. Total number of 1248 genes are used for PAM analysis .The above graph shows the misclassification error for 1248 genes. The minimum threshold is 5.07 is selected. The value of 5.07 is selected because the misclassification error is least in genes at this value which can be clearly seen in graph.

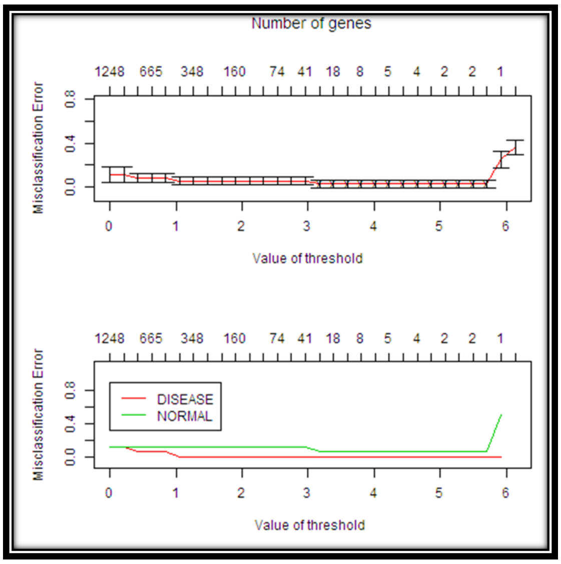

Fig 2 -Misclassification error of prediction analysis of microarray

This graph represents cross validated probability between the paired samples of normal versus disease which shows the variation of genes between the paired samples. Cross-validation mis-classification rate is 3 per cent. The classes are color-coded. Each gene is labelled by the unique ID and gene symbol. The green colour shows normal genes and red show diseased. The cross-validated sample probabilities from PAM classifier, stratified by the true classes. Threshold 5.07 was used.

	DISEASE	NORMAL
DISEASE	18	0
NORMAL	1	17

Table 1- a cross-tabulation of true (rows) versus predicted (columns) classes for the PAM fit

Sr.No.	**Array id**	**Class label**	**Prediction Correct?**	**Prediction**
1	GSM153814	DISEASE	YES	DISEASE
2	GSM153816	DISEASE	YES	DISEASE
3	GSM153818	DISEASE	YES	DISEASE
4	GSM153820	DISEASE	YES	DISEASE
5	GSM153822	DISEASE	YES	DISEASE
6	GSM153824	DISEASE	YES	DISEASE
7	GSM153826	DISEASE	YES	DISEASE
8	GSM153828	DISEASE	YES	DISEASE
9	GSM153830	DISEASE	YES	DISEASE
10	GSM153832	DISEASE	YES	DISEASE
11	GSM153834	DISEASE	YES	DISEASE
12	GSM153836	DISEASE	YES	DISEASE
13	GSM153838	DISEASE	YES	DISEASE
14	GSM153842	DISEASE	YES	DISEASE
15	GSM153846	DISEASE	YES	DISEASE
16	GSM153850	DISEASE	YES	DISEASE
17	GSM153852	DISEASE	YES	DISEASE
18	GSM153856	DISEASE	YES	DISEASE
19	GSM153813	NORMAL	YES	NORMAL
20	GSM153815	NORMAL	YES	NORMAL
21	GSM153817	NORMAL	YES	NORMAL
22	GSM153819	NORMAL	YES	NORMAL
23	GSM153821	NORMAL	YES	NORMAL
24	GSM153823	NORMAL	NO	DISEASE
25	GSM153825	NORMAL	YES	NORMAL
26	GSM153827	NORMAL	YES	NORMAL
27	GSM153829	NORMAL	YES	NORMAL
28	GSM153831	NORMAL	YES	NORMAL
29	GSM153833	NORMAL	YES	NORMAL
30	GSM153835	NORMAL	YES	NORMAL
31	GSM153837	NORMAL	YES	NORMAL
32	GSM153841	NORMAL	YES	NORMAL
33	GSM153845	NORMAL	YES	NORMAL
34	GSM153849	NORMAL	YES	NORMAL
35	GSM153851	NORMAL	YES	NORMAL
36	GSM153855	NORMAL	YES	NORMAL
Percent correctly classified:			**97**	

Table.2. Performance of classifiers during cross validation

This table presents the results of the K-fold cross-validation procedure used to characterize accuracy of the classifiers This is result of prediction verification which means how many samples are been predicted as in correct status of class labeling. For each array (sample), the table presents whether the sample is belonging to the class in which it has been placed. PAM was able to predict the class of the sample correctly. It is seen that out of total 36 samples 35 represent the correct prediction and only one sample i.e. GSM153823 show the incorrect prediction which is actually in normal state.

The second table presents sensitivity and selectivity of the PAM predictor for each class. There are two classes namely normal and disease .Their sensitivity, specificity and positive and negative predicted value is been specified .In the case of normal their sensitivity in cross validation is less as 0.944 whereas its specificity is more i.e. 1. Their positive predicted value is 1 and its negative value is 0.947.But in case of disease their sensitivity is more as 1 and specificity is 0.944.and the cross- validation of positive predicted value is less as 0.947 and the negative value is 1.It means that disease class is more sensitive than normal one and normal class is more specific than disease one.

Sr. No.	Geom mean of intensities in class 1	Geom mean of intensities in class 2	Fold-change	Shrunken centroid in class 1	Shrunken centroid in class 2	Standard deviation si+s0	Name	EntrezID
1	700.68	11230.2	0.062	11.11	11.8	1.96	Keratin 4	3851
2	504.88	6615.54	0.076	10.6	11.07	1.92	Mal, T-cell differentiation protein	4118

Table 3- Two genes used for class prediction

The table contains genes that are used for class prediction. For each gene, Geometric mean of gene expressions (ratios): For each class, geometric mean of gene expression (ratios) values are listed. The excluding of genes can be under any of the following condition as if it Less than 20 % of expression data have at least a 2 -fold change in either direction from gene's median value. The p-value of the log-ratio variation in greater than 0.01 and the Per cent of data missing or filtered out exceeds 50 %.The two genes which passes above criteria are Keratin 4 and Mal, T-cell differentiation protein genes.

V. Conclusion

Through hierarchical clustering we conclude that cluster 1 is showing the diseased samples and the group 2 shows the normal samples .However , other expression can also be seen in both of clusters.

In PAM Cross-validation mis-classification rate is 3 per cent and total numbers of 1248 genes are used for PAM analysis. It is seen that out of total 36 samples 35 represent the correct prediction and only one sample i.e. GSM153823 show the incorrect prediction which is actually in normal state. The two genes which pass criteria are Keratin 4 and Mal, T-cell differentiation protein genes.

As these genes are already known to involve in cancer progression these genes can be future target for cancer treatment.

Reference

[1]. A. Brazma, P. Hingam , J. Quackenbush, Gavin Sherlock,Chris Stoeckert, John Aach, Wilhelm Ansorge, Catherine A. Ball, Paul Spellman, Helen C. Causton,Terry Gaasterland, Patrick Glenisso, Frank C.P. Holstege, Irene F. Kim, VictorMarkowitz, John C. Matese, Helen Parkinson, Alan Robinson, Ugis Sarkans, Steffen Schulze-Kremer, Jason Stewart, Ronald Taylor, Jaak Vilo& Martin Vingron , Minimum information about a microarray experiment (MIAME)—toward standards for microarray data (2001), Nature genetics, volume 29 december .

[2]. Alvis Brazma, Helen Parkinson, Ugis Sarkans, Mohammadreza Shojatalab, Jaak Vilo, Niran Abeygunawardena, Ele Holloway, Misha Kapushesky, Patrick Kemmeren, Gonzalo Garcia Lara, Ahmet Oezcimen, Philippe Rocca-Serra and Susanna-Assunta Sansonel , Array Express—a public repository for microarray gene expression data at the EBI (2003) Nucleic Acids Research, , Vol. 31, No. 1. DOI: 10.1093/nar/gkg091.

[3]. David SP Tan, Maryou BK Lambros, Rachael Natrajan and Jorge S Reis-Filhol, Getting it right: designing microarray (and not 'microarray') comparative genomic hybridization studies for cancer research ,(2007).

[4]. Donna K. Slonim, Itai Yanai ,Getting Started in Gene Expression Microarray Analysis (October 2009) | Volume 5 | Issue 10 | e1000543.

[5]. Franck Rapaport, Andrei Zinovyev , Marie Dutreix, Emmanuel Barillot and Jean-Philippe Vert, Classification of microarray data using gene networks (2007) BMC Bioinformatics. doi:10.1186/1471-2105-8-35.

[6]. Helen Parkinson, Misha Kapushesk , Nikolay Kolesnikov ,Gabriella Rustici, Mohammad Shojatalab , Niran Abeygunawardena Miroslaw Dylag, Ibrahim Emam, Anna Farne, Ele Holloway, Hugo Berube , Margus Lukk ,James Malone, Roby Mani , Ekaterina Pilicheva , Tim F. Rayner, Faisal Rezwan ,Anjan Sharma, Eleanor Williams, Xiangqun Zheng Bradley , Tomasz Adamusiak ,Marco Brandizi, Tony Burdett , Richard Coulson , Maria Krestyaninova ,Pavel Kurnosov, Eamonn Maguire, Sudeshna Guha Neogi ,Philippe Rocca-Serra, Susanna-Assunta Sansone, Nataliya Sklyar ,Mengyao Zhao, Ugis Sarkans and Alvis Brazma, Array Express update—from an archive of functional genomics experiments to the atlas of gene expression. (2009), Vol. 37, doi:10.1093/nar/gkn889.

[7]. Malcom R Alison , CANCER(2001) ENCYCLOPEDIA OF LIFE SCIENCES ,(2001) , Nature Publishing Group / www.els.net.

[8]. Peter Choi, Chu Chen, Genetic Expression Profiles and Biologic Pathway Alterations in Head and Neck Squamous Cell Carcinoma.(2005) CANCER September 15, / Volume 104 / Number 6. DOI 10.1002/cncr.21293.

[9]. R. Keira Curtis , Matej Oresic, and Antonio Vidal-Puig, Pathways to the analysis of microarray data, (2005), TRENDS in Biotechnology Vol.23 No.8 August doi:10.1016/j.tibtech.2005.05.011.

[10]. Richard Simon et.al, Analysis of gene expression data using BRB Array tool,(2007) ,Journal cancer information .

[11]. Russell ,Greg ,Elizabeth, Lee, Greg, Hisham, Cynthia Elizabeth, Pierre , Assessing gene significance from Cdna microarray expression data via mixed model . (2001), Volume 8, Pp. 625–637.

[12]. Anne Cromer Annaı ck Carles Re gine Millon Gitali Ganguli Fre De ric Chalmel Fre De Ric Lemaire Julia Young Doulaye Dembe Le Christelle Thibault, Danie le Muller Joseph Abecassis and Bohdan Wasylyk, Olivier Poch , Identification of genes associated with tumorigenesis and metastatic potential of hypopharyngeal cancer by microarray analysis .(2004) www.nature.com/onc