International Journal of Advanced Research and Development
ISSN: 2455-4030
Impact Factor: RJIF 5.24
www.advancedjournal.com
Volume 3; Issue 1; January 2018; Page No. 155-163

Comparative analysis of cancer gene using microarray gene expression data

Vaishali Gupta
Department of Statistics, School for Physical and Decision Sciences, Babasaheb Bhimrao Ambedkar University, Lucknow, Uttar Pradesh, India

Abstract
It is possible to simultaneously monitor the expression levels of huge number of genes during important biological processes and across collections of related samples by DNA microarray technology. But, the huge number of genes and the complicacy of biological networks increases the challenges of interpreting the results which includes large number of measurements. In this paper, we have analysed and reduced the high dimensional data by applying various clustering techniques to identify interesting and useful patterns. Then, heatmap is used for the visualisation of clustered and biclustered data. All the analysis is done using R software. The analysis educates us about correlation of various genes by identifying patterns which will surely help us to better understand the structure and function of human genes and thereby develop new strategies to take action to reduce human diseases.

Keywords: hierarchical clustering, K-means clustering, PAM clustering, model-based clustering, biclustering, heatmap

1. Introduction
A DNA Microarray is a latest technology which allows the simultaneous measurement of the expression levels of huge numbers of genes. Researchers, basically in the field of bioinformatics, have generated huge amounts of gene expression data, but there is a great desire to develop analytical methodology to analyse and to explicit this information (Lander, 1999). The volume of genetic data is growing faster than the rate of its analysis. Clustering techniques provide a possible solution for analysing and interpreting such speedily increasing genetic data. Success in applications has been communicated for many clustering algorithms, but still no single method has appear as the method of choice in the gene expression analysis. Most of the advanced clustering algorithms are largely heuristically motivated, and the issues of determining the "correct" number of clusters and choosing a "good" clustering algorithm are not yet rigorously solved. Visualizing gene expression data is also a typical task.
In particular, we would like to know whether *gene expression*, the process by which genetic information encoded in DNA is converted, first, into mRNA (messenger ribonucleic acid), and then into protein or any of several types of RNA, is any different for cancerous tissue as opposed to healthy tissue (Izenman, 2008) [5]. In exploratory data analysis and pattern discovery, *Clustering* is a ground laying technique aiming at extracting underlying clusters. When a new gene is allotted to a cluster, the biological function of this cluster can be attributed to this gene with high confidence. Yeung *et al*. (2001b) [14] suggested clustering the data set leaving out one experiment at a time and then comparing the performance of different clustering algorithms using the left-out experiment. A large number of clustering algorithms have been suggested for the analysis of gene expression data, including hierarchical clustering (Eisen *et al*., 1998) [1], k-means (Tavazoie *et al*., 1999) [12], PAM clustering and so on. Eisen *et al*. (1998) [1] and Tamayo *et al*. (1999) [11] used visual display to determine the number of clusters. Recently, Raza (2014) [8] used four different clustering algorithms, such as k-means, hierarchical, density-based and expectation maximization approaches to analyse five different types of cancerous gene expression data (lung, breast, colon, prostate, breast and ovarian cancer) and Gupta *et al*. (2016) [4] give the review of the miscellaneous research papers and journals on recent research done for cluster analysis of microarray gene expression data in brief.
Here, we use variables clustering (clustering of genes) instead of observations (patient samples), the main difference being the measure of distance between variables (Sarmah, 2010) [10]. We have many standard visualization techniques, the process by which the data is represented visually, for gene clustering like heatmap. For the analysis of gene data till now various clustering algorithms have been proposed and applied. Thus, the aspiration of this paper to provide a brief review of the various clustering algorithms, comparative study of clustering and biclustering, and also provide solutions for above described problems.

2. Materials and Methods
2.1 Microarray Gene Expression Dataset
DNA microarray gene expression data used in this paper is obtained from the website *www.cancergenome.nih.gov* consisting of 77 genes which are taken as variables and 121 patient samples which are taken as observations. Data is in the form of the table in which rows consist of gene identity and columns consist of patient's sample. Patients sample defined in terms of TCGA (The Cancer Genome Atlas) identity while gene identity defined in terms of numeric form. Some information about genes used in data is in the form of Table 1(here we specify starting 10 genes) and microarray gene expression data used a bit more here describe in Table 2.

Table 1: Table for information about genes included in the data

S No.	Gene symbol	Gene official name	Gene identity	Gene type
1	DFFB	DNA fragmentation factor, 40kDa, beta polypeptide	1677	Protein coding
2	PIK3CD	phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit delta	5293	Protein coding
3	DFFA	DNA fragmentation factor, 45kDa, alpha polypeptide	1676	Protein coding
4	CASP9	caspase 9, apoptosisrelated cysteine peptidase	842	Protein coding
5	PIK3R3	phosphoinositide-3-kinase, regulatory subunit 3 (gamma)	8503	Protein coding
6	PRKACB	protein kinase, cAMPdependent, catalytic, beta	5567	Protein coding
7	NGF	nerve growth factor (beta polypeptide)	4803	Protein coding
8	NTRK1	neurotrophic tyrosine kinase, receptor, type 1	4914	Protein coding
9	FASLG	Fas ligand (TNF superfamily, member 6)	356	Protein coding
10	CAPN2	calpain 2, (m/II) large subunit	824	Protein coding

Table 2: Table of microarray gene expression data of order 4x5

Gene	Patient Sample				
	TCGA. BL.A0C8	TCGA. BL.A13I	TCGA. BL.A13J	TCGA. BL.A3JM	TCGA. BT.A0S7
69	-0.111	-0.002	-0.035	-0.009	-0.184
109	-0.111	-0.002	-0.035	-0.009	-0.184
120	-0.111	-0.002	-0.035	-0.009	-0.184
157	-0.111	-0.002	-0.035	-0.009	-0.184
573					

2.2 Underlying Theory

Cluster analysis, the most well-known example of *unsupervised learning* (*Problems in which no prior information available to define an appropriate output variable*), is a very popular and powerful tool for analyzing unstructured multivariate data. In gene clustering, the ($m \times n$) data matrix $X = (X_{ij})$ represents the gene expression data, where i indexes the row (gene), j indexes the column (tissue/patient sample), and X_{ij} is a measurement of how strongly the i^{th} gene is expressed in the j^{th} sample (Everitt *et al*., 2011). There are many packages in R to analyze clustering, (see R, 2014, for more discussion) [7].

2.2.1 Hierarchical Clustering

There are two types of hierarchical clustering method named *agglomerative* and *divisive*. We use agglomerative hierarchical clustering method to cluster the gene expression data, (see Izenman, 2008, p. 411) [5].

2.2.2 Nonhierarchical or Partitioning Methods

Nonhierarchical clustering methods (also known as partitioning methods) simply divide the data into a predefined number K of groups or clusters, where there is no hierarchical relationship between the K-cluster and the (K + 1)-cluster, (see Izenman, 2008, p. 422) [5].

2.2.3 Biclustering

Let Y be a $m \times n$ matrix. The goal of *biclustering* now is to find subgroups of rows and columns which are as similar as possible to each other and as different as possible to the rest (Kaiser, 2008) [6].

3. Results and Discussion

The following 11 plots have been adopted for the analysis of DNA microarray gene expression data used in this paper by using R software.

First, we illustrate unsupervised clustering of the dataset using three methods (complete linkage, single linkage, and average linkage). R and CRAN have a variety of agglomerative hierarchical clustering algorithms. First, start with the most commonly used procedure, function *hclust* in base-R. The procedure runs on the matrix of pairwise distances between points constructed using the function *dist*. The result of *hclust* is a dendrogram which is displayed using the function *plot*. To see clear merging of the clusters, we use the function *str.dendrogram*.

In Figure 1, we investigate the tree by trial-and-error, and find that 'cutting the tree' at a height k = 4 clusters provides a useful result. It is clear from all these three plots is that among all 77 variables the smallest distance is between 8218 and 8219, so it merges first at the height 0 and becomes two member cluster and this merges to 2999 and 8131 at the height 3.98 and proceeding in this way member of cluster increases and forms a group of cluster. Lastly all the group of clusters merge 5987 at different heights in dendrogram. Thus, it is seen that the distance between first two member cluster (8218 and 8219) is minimum, so it represents most similarity and as height increases similarity decreases and we get most dissimilar cluster 5987. So the *gene* 8218 *and* 8219 *have almost same expression level whereas gene 5987 have very different expression level* from all other 76 genes. It is also clear from the graphs is that 5987 is the *outlier*. By using all three methods, we cannot say clearly which method is best and also which gene is high expressed, low expressed or completely expressed. Next method gives the answer of this particular question. For more details about dendrogram, (see Izenman, 2008, p. 412) [5].

3.1 Graphical visualization of different clustering methods

There are different methods to visualize different clustering techniques. Some of them are described in this paper.

In Figure 2, plots give the standard visualization of hierarchical clustering method for gene expression data. The expression level are continuously showed on the color scale which represent three colors red, green and black which are given by using the function *color Ramp Palette* in R. Here, red represents over expression of gene, green represents low expression and black represents that gene are fully expressed.

It also contain two dendrograms which represent grouping of genes (left) and patients sample (top). Heatmap is displayed by using the function *heatmap.2* in R. Each gene is represented by a single color box for each patient but using complete dataset this cannot be shown clearly. So we extract starting 5 values of the data so that we can see clearly that each gene is represented by a single color box for each patient. Thus, we see that gene *69* is low expressed for first two and last two patient while it is fully expressed for third patient whereas gene *120* is over expressed for all the patients. Also gene *157* is fully expressed for first and last patient, low expressed for second and fourth patient and over expressed for third patient. Similarly, remaining two genes can be interpreted.

In Figure 3, after selecting the list of differentially expressed genes, we want to find out the relationship between these genes and the well-separated clusters which will be done using K-means and PAM clustering. The silhouette plot is quite useful for deciding the number of clusters. From the k-means plot, the first cluster contain 19 gene, second cluster contain 17 gene, third cluster contain 34 gene and fourth contain 7 gene represented by red, blue, green and yellow color respectively. Clearly, this plot shows that most of the genes in blue and green cluster have large average silhouette value lying between 0.3 to 0.5, indicating that the cluster is somewhat separated from neighboring clusters. However, red cluster contains many genes with low silhouette values, and yellow cluster contains a few genes with negative values, indicating that these two clusters are not well separated. So, this plot gives a good structure because most of genes within the cluster that they are in and also the genes in green and blue cluster can be used for further analysis but the drawback is that we cannot select the "best" number of clusters. This problem is solved by PAM clustering. One can run PAM several times, each times for different values of k and then compare the resulting silhouette plots. The average silhouette width can be used to select the "best" number of clusters, by choosing that k which have the maximum silhouette width. For, the microarray gene expression data used in this paper, average silhouette width is highest (0.15) for $k = 2$. So we select number of cluster 2 for the PAM clustering. The microarray gene expression data used for analysis consist of 121 patient samples which is being partitioned into 2 clusters, decision to the first cluster contain 82 samples, second cluster contain 39 samples. PAM work on observations directly so PAM work on samples rather than genes. First, we see, for each observation, that to how well it fits into the cluster that it has been assigned to. This is seen by comparing how close the sample is to other samples in its own cluster with how close it is to samples in other clusters. A value close to 1 means that the sample is well placed in its cluster and a value close to 0 means that a sample belong in some other cluster. Clearly, this plot shown that except 7 samples, for all the samples this value near 1 mean that the patients sample is well placed in its cluster. Thus, this plot gives better result for the cluster (patient) samples and also plot indicates that there is a good structure to the clusters, with most observations appear to belong to the cluster that they are in. Also, in both plots genes/patients sample which are in the same cluster means they represent similarity with each other. For more details about silhouettes, (see Rousseeuw, 1987) [9].

3.2 Model-Based clustering with the key merit of select the number of cluster and an appropriate model

There are several clustering methods so the problem is to choose a "good" clustering method and find out "correct" number of clusters. This problem is reduced to model selection problem in the framework of probability. We use model-based clustering only to fit an appropriate model according to BIC for EM algorithm by hierarchical clustering for parametrized gaussian mixture model for the microarray gene expression data used in this paper so that we can apply "Biclustering" to reduce the complexity of the data. Here, we use BIC criteria from "mclust" package in R for the 10 available model parametrization and up to 9 clusters for the data used in this paper. We select that model whose BIC value is minimum. So, the best fitted model among all multivariate mixture model is EII (Spherical equal volume with 9 components). With the help of this model we apply biclustering in the data, (see Yeung *et al*., 2001a) [13].

The first plot in Figure 4, we visualize the gene expression data matrix as a heatmap. Here, green and red indicate down and up regulation of genes. The gene expression data used for the analysis consist of 77 genes and 121 patient samples but after applying biclustering we get only 6 genes and 37 patient samples instead of 77 genes and 121 patient samples. So using this plot we reduce the complexity of the data by reducing size of the data. Remaining genes and samples are used for further analysis.

The second plot in Figure 4 is the parallel coordinate plot which represents expression levels through gene and patients sample in a bicluster as polylines. By applying the plot only row, the expression level for the genes in the selected bicluster will be drawn. This means each line is a gene profile with the genes on the x-axis. By applying plot only column, a second plot will be added to this plot with the patient samples on the x-axis. In this plot, each polyline represents the expression of a gene over all conditions in vertical axes. Genes belonging to a bicluster are captured using the same color as the corresponding bicluster in the heatmap. The axes of the parallel-coordinates plot are arranged in the same order as the columns in the heatmap. Here, red color shows gene/patients sample are from the bicluster while gray color shows they are from original data set.

In Figure 5, the first plot is a bicluster barchart, showing the patients sample. Each block shows one bicluster. Thus there are 3 bicluster and the bars inside of each bicluster display the means of bicluster values for the corresponding sample. In this plot, red dot represents the mean of each bicluster.

The barplot of biclusters (second plot) is used to compare the values inside a bicluster with the values outside of the bicluster. For each bicluster, three bars are drawn per sample of the bicluster. The darkest bar represents the values inside of the bicluster and the other two are the mean and median of the values outside of the bicluster and also calculate the number of biclusters which should be drawn with the box.

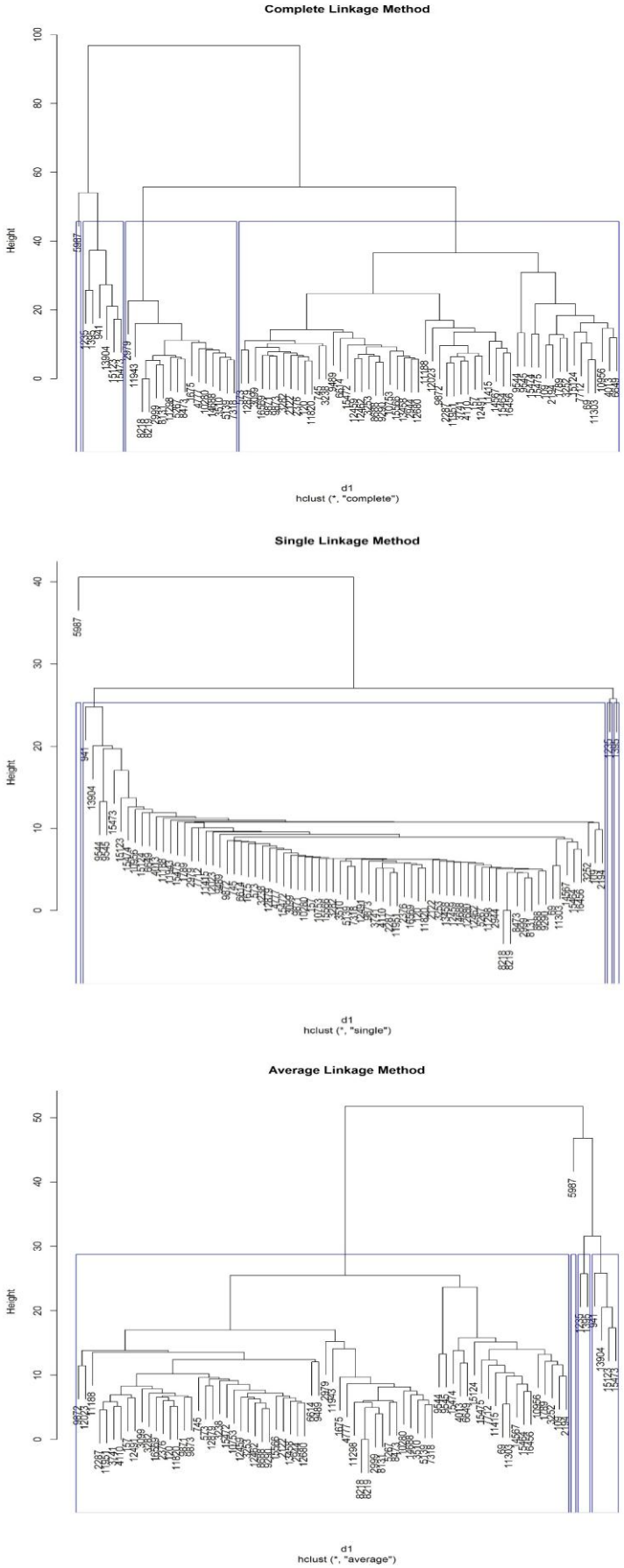

Fig 1: Dendrogram by using complete, single and average linkage method:

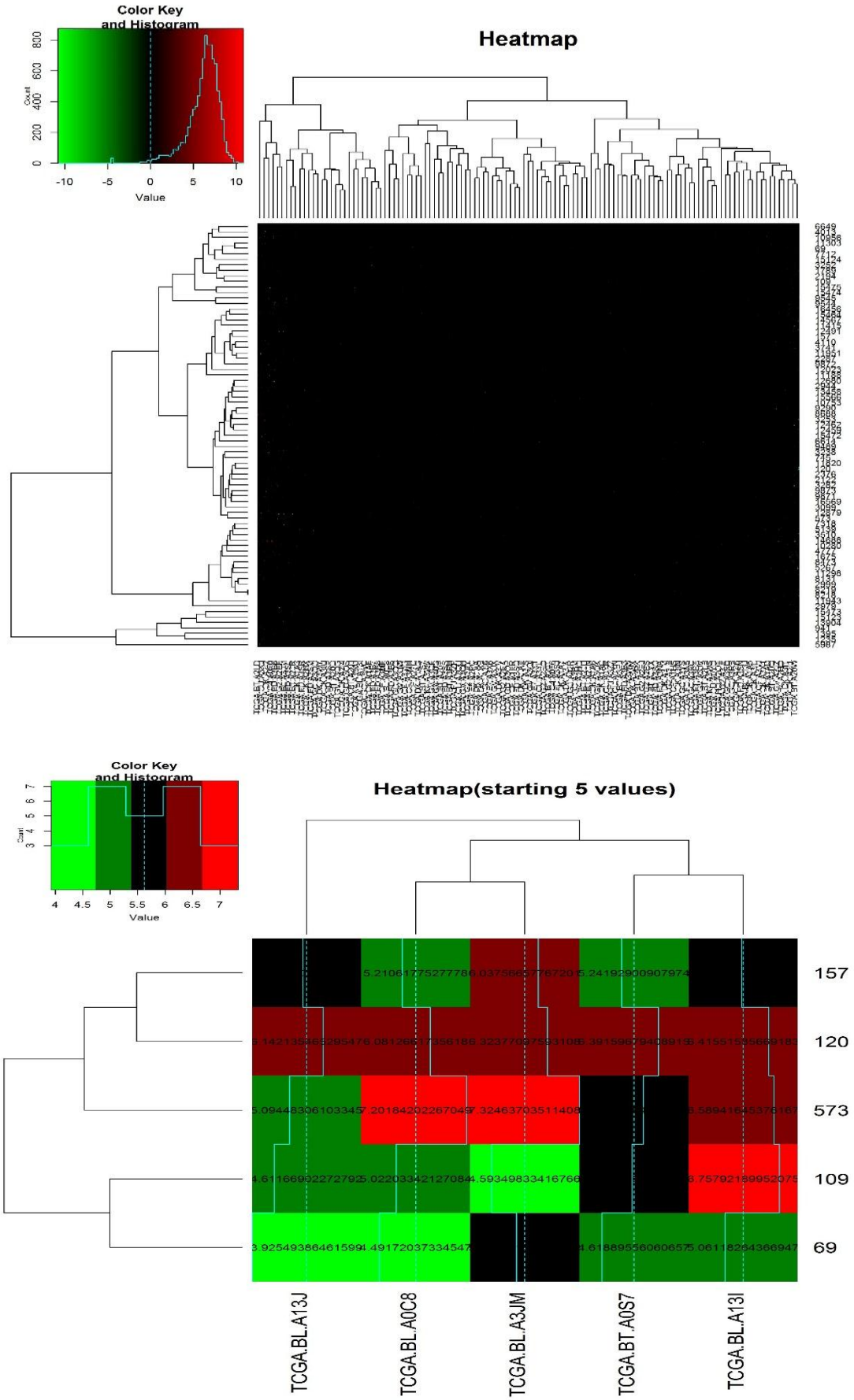

Fig 2: Heatmap for the graphical visualization of hierarchical clustering method using complete dataset and starting 5 values of the dataset

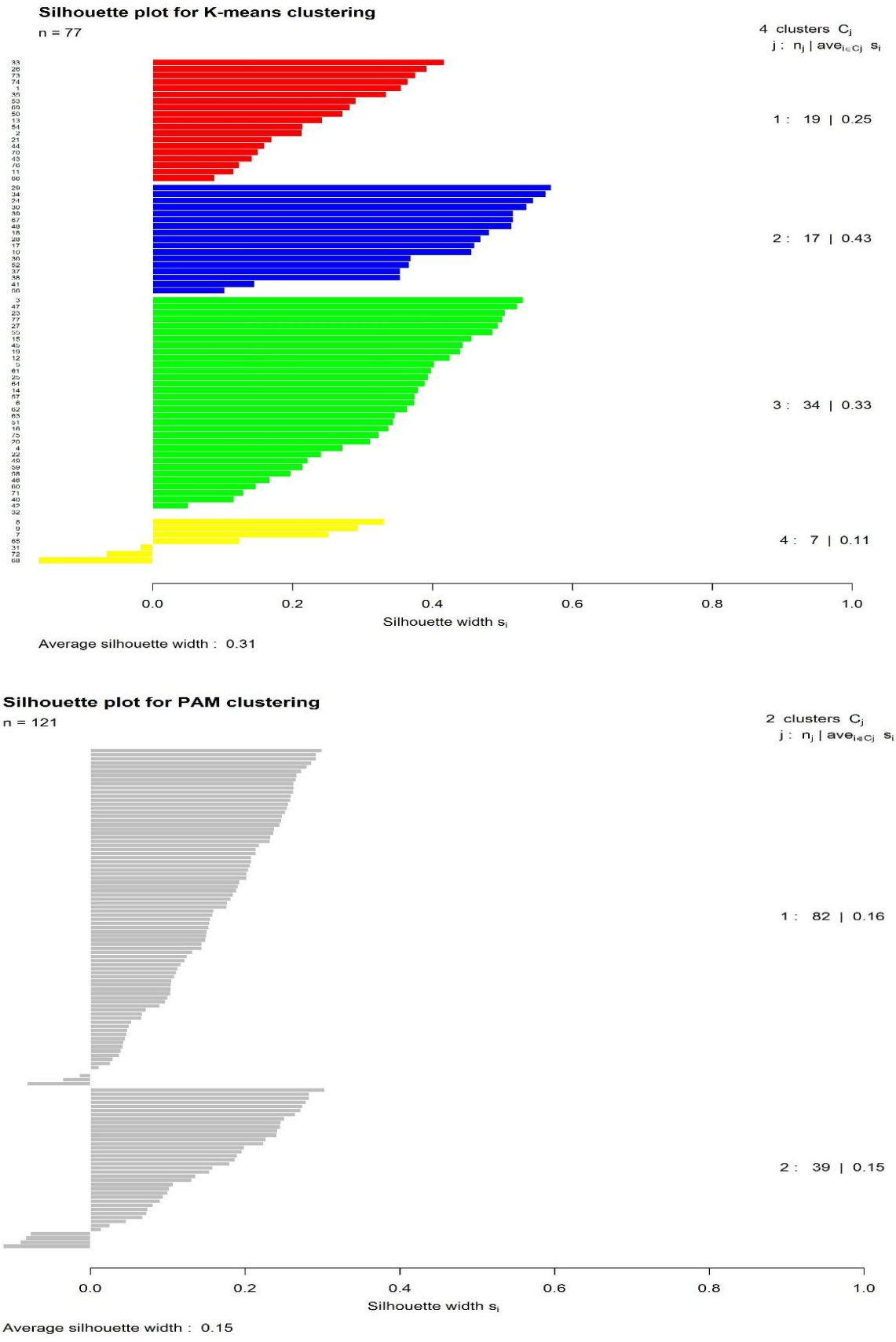

Fig 3: Silhouette plot for the graphical visualization of K-means and PAM clustering:

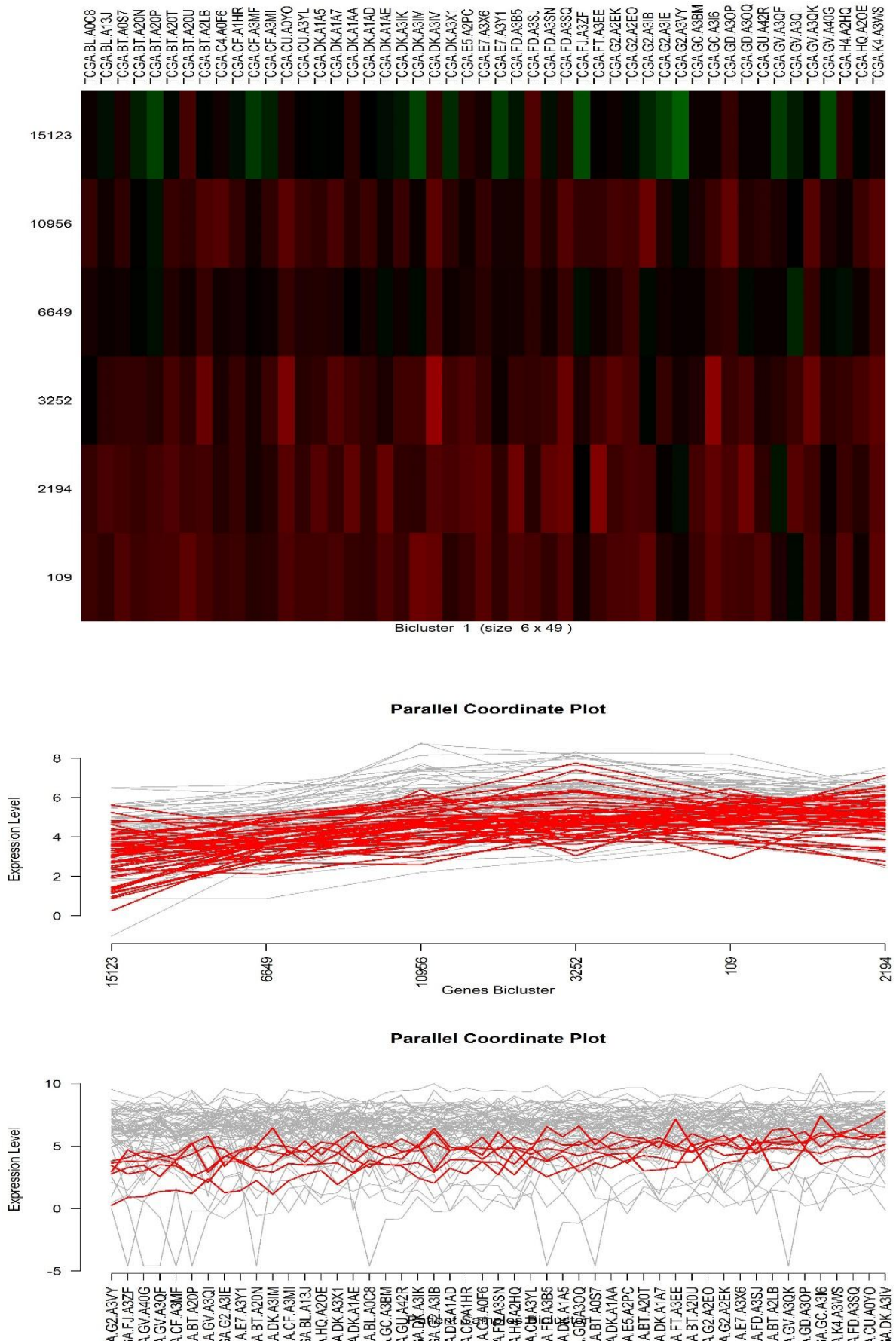

Fig 4: Graphical visualization of Biclustering by Heatmap and Parallel Coordinate Plot:

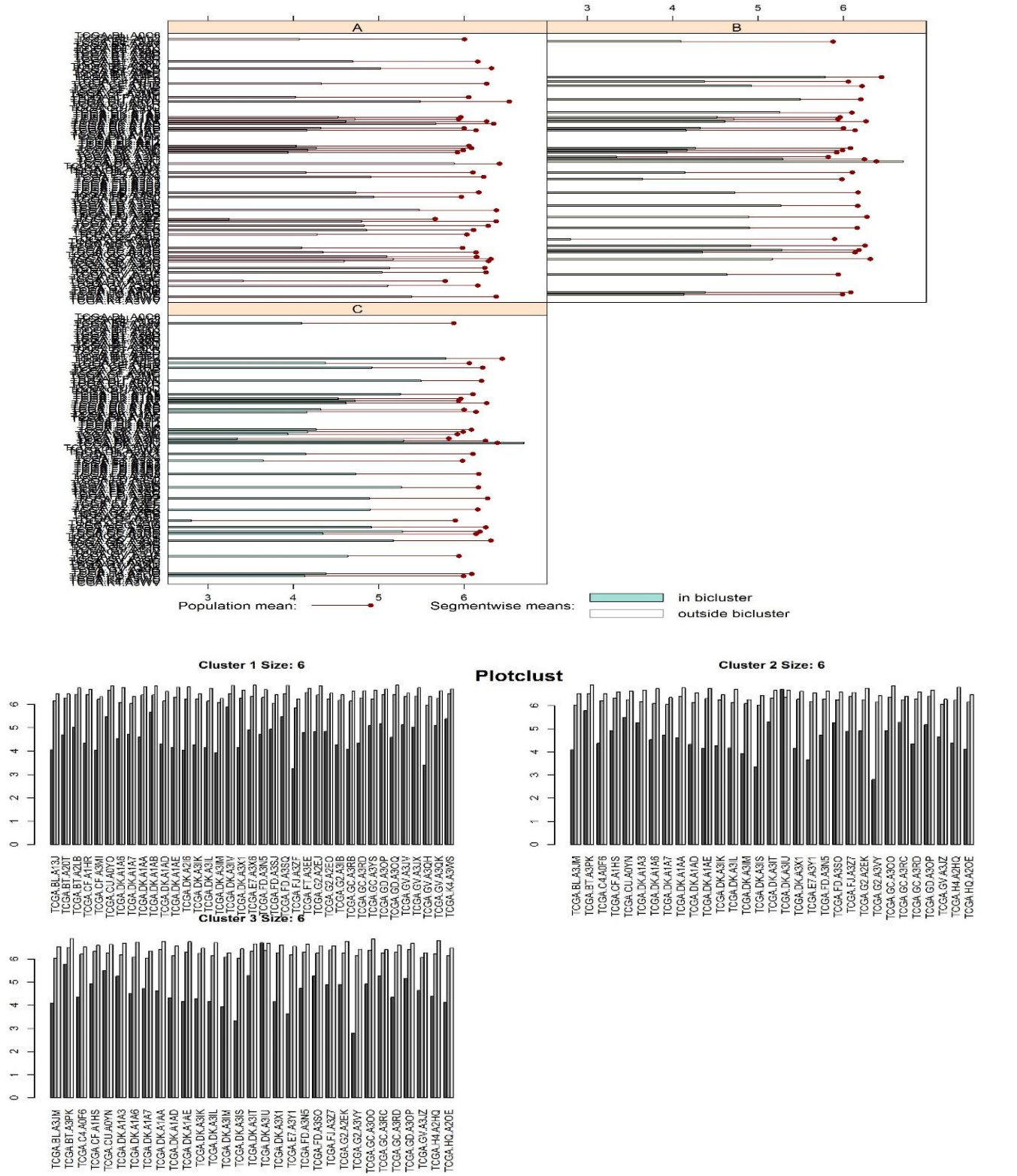

Fig 5: Graphical visualization of Biclustering by Bicluster Barchart and Barplot:

4. Conclusion

In this paper, microarray gene expression data set is analysed over different clustering techniques. First, with the hierarchical clustering technique, we see the expression level of genes by considering their similarity and also we see the standard visualization of this technique. Then, we find out the relationship between genes and clusters with the help of k-means technique and also correct number of clusters with the help of PAM clustering. After this, we reduce the complexity of the data by biclustering technique and see the graphical

visualization of this technique. Finally, we get the low dimensional data which is not difficult to analyse and interpret, mainly for the researchers in the field of bioinformatics.

5. Acknowledgement
I would like to thank Prof. Anoop Chaturvedi and Dr. Amit Kumar Misra for their fruitful comments and suggestions during this research.

6. References

1. Eisen *et al*. Cluster Analysis and Display of Genome-wide Expression Patterns. Proceedings of the National Academy of Science, USA. 1998; 95:14863-14868.
2. Everitt *et al*. Cluster Analysis, Wiley Series in Probability and Statistics, 5th Edition, 2011.
3. Golub *et al*. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, www.sciencemag.org, 1999; 286.
4. Gupta *et al*. Clustering Methods Applied for Gene Expression Data: A Study, Second International Conference on Computational Intelligence & Communication Technology, 2016.
5. Izenman AJ. Modern Multivariate Statistical Technique, Springer Texts in Statistics, 2008.
6. Kaiser S, Leisch F. A Toolbox for Bicluster Analysis in R, 2008. http://www.stat.unimuenchen.de.
7. R Core Team. R:A Language and Environment for Statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2014.
8. Raza K. Clustering Analysis of Cancerous Microarray Data. Journal of Chemical and Pharmaceutical Research. 2014; 6(9):488-493.
9. Rousseeuw PJ. Silhouettes: A Graphical Aidto the Interpretation and Validation of Cluster Analysis. Journal of Computational and Applied Mathematics, North-Holland. 1987; 20:53-65.
10. Sarmah S, Bhattacharyya DK. An Effective technique for Clustering Incremental Gene Expression data. International Journal of Computer Science Issues. 2010; 7(3):3.
11. Tamayo *et al*. Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and Applications to Hematopoiteic Differentiation, Proceedings of the National Academy of Science, USA, 1999; 96:2907-2912.
12. Tavazoie *et al*. Systematic Determination of Genetic Network Architecture, Nature Genete, 1999; 22:281-285.
13. Yeung *et al*. Model-Based Clustering and Data Transformations for Gene Expression Data, Bioinformatics, 2001a; 17(10):977-987.
14. Yeung *et al*. Validating Clustering for Gene Expression Data, Bioinformatics, 2001b; 17(4):309-318.