



A Comparative Analysis on Visualization of Microarray Gene Expression Data

Poornima. S

PG Scholar, Department of Information Technology
Sona College of Technology
Salem, India

Dr. J. Jeba Emilyn

Associate Professor, Department of Information Technology
Sona College of Technology
Salem, India

ABSTRACT

Visualization technique helps in the easy analysis of data. But, Visualization of biological data is one of the most challenging processes and visualization of the computed clustered and biclustered data still remains an open issue. Clustering and Biclustering techniques were the popular methods for classifying the gene expression data. There is no standard visualization technique for the biclustered data. Visualization of multiple biclusters is very harder to implement because of the overlapping property. Here, we analyzed the merits and demerits of various visualization techniques and visualization tools. We Compared and provided a detailed study of each technique. Finally, a conclusion to overcome the common challenges in the visualization of microarray gene expression data is provided.

Keywords

Visualization, microarray gene expression data, clusters, biclusters.

1. INTRODUCTION

Visualization technique is the study of the visual representation of data. Visualizing gene expression data is a challenging one. The most common and efficient method for analyzing gene expression data is clustering that groups together genes with similar expression profiles. We have many standard visualization techniques for gene clustering. But it is not the case for the gene biclustering. In gene expression data, a bicluster is a subset of the genes exhibiting consistent patterns over a subset of the conditions. Biclustering techniques group the genes under a certain subset of conditions. At the same time, a gene or condition can be in more than one bicluster called overlapping, but in clustering a gene or condition is usually assigned



to a unique cluster. The outputs of biclustering algorithms provide the basis for better understanding of the biological process underlying the data. However, to provide a clear knowledge about the biclustered results a better and efficient visualization technique is required. Here, we have analyzed some visualization techniques of both clustering tools and biclustering tools.

This paper proceeds as follows. In the next section, the background study is described. Section 3 describes related works in this field, etc.

2. Study of Existing Visualization Techniques and Tools

2.1 Existing Visualization Techniques

2.1.1 Visualization of Multilayered Clustered Results

Leishi Zhang et al, proposed a paper [1] where 3D visualization of gene clusters can be effectively performed by using “Force directed placement spring model”. Detecting and allocating genes within the cluster to a local area (Infocube). Allocate all the infocubes within a global area. Here nodes are considered as physical bodies and edges are the spring force in between the gene (nodes). Genes are clustered within the infocube in the local area and then the infocubes are clustered in the global area. Therefore two layers are obtained. It can be extended to display as a multilayered graph. Spring force and repulsive force of each node is calculated and then they are positioned inside the respective infocube. Similarly all the infocubes’ spring and repulsive forces are calculated and then positioned in its global area.

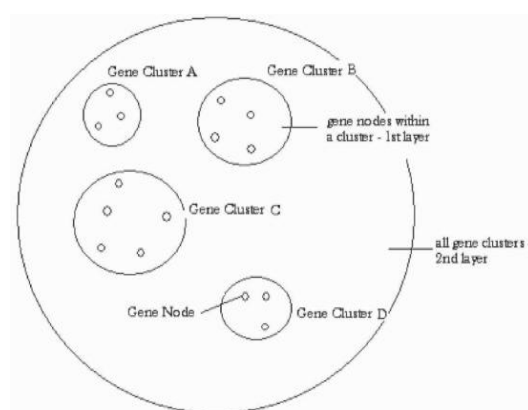


Figure 1. Representing clusters in two layers in [1]



2.1.2 *Integration of Clustering and Visualization*

Gunther H. Weber et al proposed this paper [12] where both clustering and visualization methods are involved for the analysis of gene expression. Brushing and linking is the visualization technique used. In physical views, colour and height are used for visualizing spatial gene expression pattern clusters. In abstract views, positions are ignored and expression levels for multiple genes in the cluster are plotted with respect to each other using scatter plots or parallel coordinates. Colour mapping plays the major part. It combines different visualization methods to overcome the shortcomings of single techniques. It provides very high Interactivity.

2.1.3 *Parallel Coordinate Plot Visualization For Biclusters*

K. O. Cheng *et al* proposed this paper [4], in which the parallel coordinate (PC) plot visualization technique is used for high dimensional data. A bicluster is a subset of rows which exhibit coherent patterns across a subset of columns. Generally there are two types of biclusters, namely additive-related biclusters and multiplicative-related biclusters. Here, all these axes were ordered in a parallel way on a 1-Dimensional plane. And the orthogonal property is destroyed. Also the geometric structure of the data is consistent. The additive-related bicluster shows a number of lines with the same slope across the conditions [4]. Biclusters detection is done by using the split and merge mechanism.

2.1.4 *R Package Biclust*

Sebastian Kaiser and Friedrich Leisch introduced the R package biclust in [7], which consists of multiple biclustering algorithms and multiple visualization techniques for different purposes. Bubbleplot is a graphical representation of biclusters which is shown in a 2 dimensional format. Bicluster within a bicluster is represented as small and big circles. Parallel coordinates is used to represent the similarity of rows over columns inside the bicluster. Heatmap is used to provide a representation of difference between the bicluster with the rows and columns.

2.1.5 *Visualization of Gene Clusters and Gene Expressions*

Ashraf S. Hussein (2008) proposed a web based framework [14] for the analysis and visualization of gene expression and protein structures. Two types of visualizations are used. They are sequence visualization and



IJCSBI.ORG

clustering visualization. In Clustering visualization we have hierarchical clustering where $N \times N$ matrix is the input and D is the pairwise distance. Depth first algorithm is implemented for displaying the results of hierarchical clustering, in Self organising maps the clusters are represented like Cartesian graph. In k-algorithm, clusters are represented as group of spheres with distinct colors.

3. Existing Visualization Tools

3.1.1 Gene 3D Miner

YONGGAO YANG, JIM X. CHEN and WOOSUNG KIM proposed in [9], where, clustering and 3D visualization takes place by using Self organizing maps, principal component analysis and 3D plotting techniques. Clusters are visualized in 3D space. Here, the genes are grouped into clusters using SOM algorithm and then they are reduced to 3 dimension data using PCA algorithm. And they are placed in the 2D and 3D space. Since all the genes are plotted in the same 3D space, the distance between the gene pairs represents the dissimilarities between the genes. The screen is divided into many panes to avoid overcrowding of genes for the comparison. Time serial values are provided in each pane. The size of the pane is directly proportional to the number of genes in the cluster. The entire 3D space is divided into smaller cubes. And each cube represents a cluster. Finally, the full 3D space will look like a big cube which encloses clusters as smaller cubes.

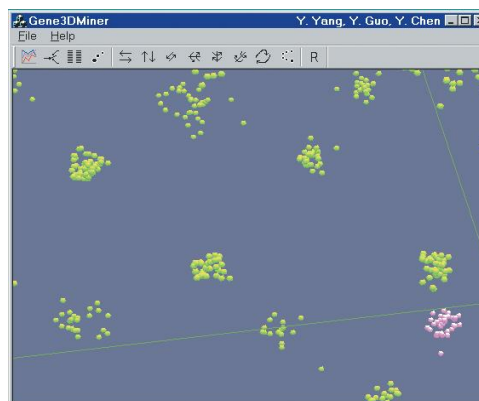


Figure 2. Representation of genes as cubes in gene 3D miner in [9]



3.1.2 Bicluster Viewer

Julian Heinrich deals with Bicluster Viewer in [2]. It supports contiguous representation of selected biclusters. To achieve this Bicluster viewer [2] allow row and column reordering as well as duplication. Here the representation of biclusters is done with several unique coloured rectangles inside the heat map. In Parallel-coordinate plots, each polyline represents the expression of a gene over all conditions [2]. Genes belonging to a bicluster are visualized using the same color as the corresponding bicluster in the heatmap [2]. The axes of the parallel-coordinates plot are arranged in the same order as the columns in the heatmap representation [2]. In order to visualize the conditions of genes belonging to a bicluster, it compute the average vertical position of all lines of a bicluster halfway between adjacent conditions if at least one of the conditions is part of the bicluster [2]. Then, the corresponding lines are forced to cross this point, which we call the centroid [2]. This uses the linking and brushing techniques to link heatmap and parallel coordinate plots.

3.1.3 Bicoverlapper

Rodrigo Santamaria proposed a tool Bicoverlapper [3] to visualize biclusters from gene expression matrices. It is used to compare the biclusters, and to highlight relevant genes and conditions. Biclusters are visualized using the technique called “**Force directed graph**”. Each pair of nodes are held by two types of forces called spring force and expansive force. A glyph is used to convey multiple data values and glyphs on gene and conditions nodes improves our understanding the instances of overlapping when the representation becomes complex. Edge cluttering is avoided by wrapping the bicluster in a rounded shape (transparent hull) built by splines that take the positions of the outermost nodes in the bicluster as anchor points. A gene or a condition node can be in more than one bicluster, reflecting overlapping between biclusters [3], which can usually affect more than a node.



IJCSBI.ORG

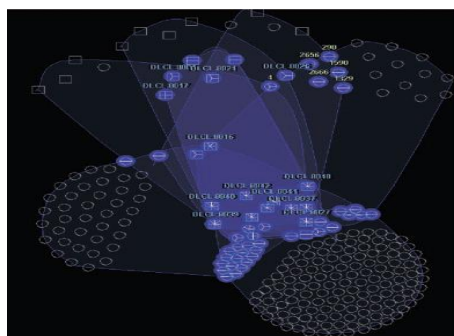


Figure 4. Representation of Biclusters in Bicoverlapper [3]

3.1.4 Biolayout Express 3D

Athanasios The ocharidis et al designed a tool [13] for various analysis and visualization of biological process. It supports both unweighted and weighted graphs together with edge annotation of pair wise relationships [13]. Fruchterman -Rheingold layout algorithm is used for the visualization. Various colour schemes make the network more informative and the clusters can be easily visualized. Markovian chain clustering algorithm is designed specifically for the clustering of simple or weighted graphs [13]. It has a very high user interaction. The UI functions supported here were zooming, scaling, rotation, translation, and selection of the graph. Landscape plot is used as an alternative representation for dendrogram. The subtrees at a given internal node are modified by reordering it, so that the larger tree is obtained in the left. And finally a histogram is created based on the joining events corresponding with the subsequent gene pairs.



Figure 5. Representation of Clusters in Biolayout express 3D in [13]

3.1.5 Bicat

Simon Barkow et al proposed in [5] a software tool where both clusterin and biclustering techniques are used for the visualization of gene expression data. Many biclustering and clustering algorithms were used. Higer user



interaction is provided in this tool. The expression matrix is displayed as a heatmap. Biclusters can be visualized as heatmap and gene expression profiles. The heatmap is rearranged in such a way that those genes and conditions that define the bicluster under consideration appear in the upper left corner of the map. Alternatively, the expression view of a bicluster, displays the profiles of those genes that are grouped within a bicluster. For different conditions, different colour lines are represented between the gene pairs.

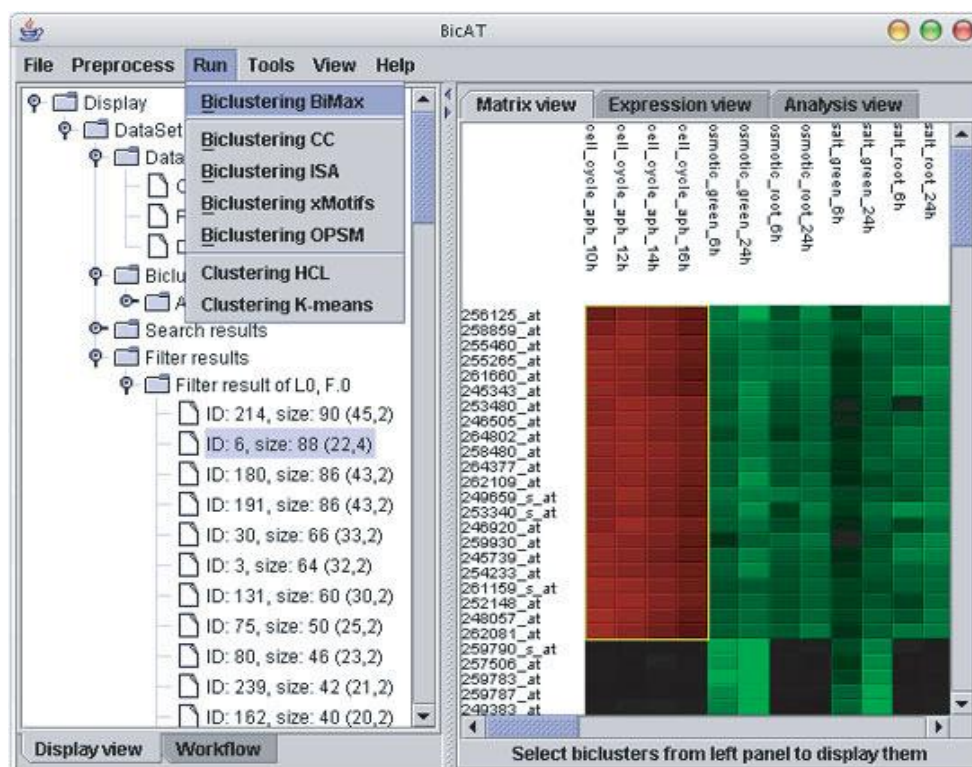


Figure 6. Representation of Biclusters in Heatmap in [5]

3.1.6 Bivisu

K.O. Cheng*, N.F. Law, W.C. Siu and T.H. Lau proposes a tool [6] called BiVisu, an open-source tool for detecting and visualizing biclusters. The biclustering results are then visualized in a two-dimensional set up for the easy analysis. In parallel co-ordinates all the axes in the plane are drawn parallel to each other. The axes are arranged properly to visualize the biclusters clearly. The tool allows the navigation of biclusters individually. The genes inside and/or outside a bicluster can be drawn for comparison purpose. Heat map is also provided. Detection of biclusters of constant, constant rows, constant columns, additive and multiplicative types are also possible. split-and-merge algorithm is used for bicluster detection.



3.1.7 *Jexpress*

B. Dysvik and I. Jonnasen designed [10], a JAVA tool with Multidimensional scaling, clustering and visualization methods to visualize the gene clusters. Gene graph viewer is interactive and when the internal nodes are clicked it shows the expression profiles of the gene in the subtree. Clusters are visualized in the form of gene graphs as an output. To get a 2D or 3D representation of clusters multidimensional scaling method is employed. Visualization of clusters are not upto the expected level, it is acceptable, since this is the primary tool in this field which leads to the invention of powerful visualization techniques. It is interactive and when the internal nodes are clicked it shows the expression profiles of the gene in the subtree

3.1.8 *clusterMaker*

John H Morris et al designed [11] clusterMaker, a plugin for the software cytoscape, which provides a variety of clustering algorithms and visualization techniques for the analysis of gene expression data. The three types of visualizations are heatmap, treeview and network view. Heatmaps are used for visualizing clusters from k-means, k-mediod and AutoSome clustering algorithms and also all the numeric attributes within the network are also visualized using heatmap. The clusters obtained from hierarchical clustering algorithm are visualized as dendrograms. Heat map and Treeview (dendrogram) visualizations are used from Java Treeview software. Network view, visualization is provided using force directed layout. It represents only the intra- cluster edges all the inter cluster edges are dropped.

3.1.9 *EUREKA-DMA*

Sagi Abelson in [16] designed a software called EUREKA-DMA. It is a GUI used for the fast classification and analysis of microarray gene expression data. It integrates many primary tools and forms a common interface. It provides a very efficient framework for the visualization and interpretation of gene expression data. Many visualization techniques are use for various purposes. They are volcano plot, heat map, gene ontology biograph, bar plots and box plots [16]. Volcano plot is used to show the genes with differential expression [16]. Heatmap is for the representation of hierarchial clustering analysis and the other techniques are for other purposes like finding protein (KEGG) pathways [16] and to analyse the



difference between the computed results. This heat map representation is helpful to the users with high degree of knowledge about the microarray gene expression data.

Table 1. Comparison of Visualization techniques

#	Techniques	Cluster/ Bicluster	Methods	Merits	Demerits
1	Multilayered clusters in[1]	Clusters In 3D	Force directed placement spring model	Cluster visualization is simple, effective, easily adapted and extended	No optimal representation and takes very high running time
2	Integration of clustering and visualization in [12]	Cluster In 2 D	Scatter plots and parallel coordinate	Clear visual representation of the results.	possibility of missing data when there is a large and complex pattern of gene expression data
3	Gene cluster and gene expression visualization	Cluster in 2D	DFS tree, Cartesian graphs, spheres	Visually attractive patterns are obtained	Works good only for small datasets
4	Parallel coordinates plots in [4]	Bicluster in 2D	Parallel coordinate	Biclusters are detected and visualized. User interaction is possible	Contiguous biclusters are not obtained
5	Heatmap in [8]	Bicluster in 2D	Rearranging rows and columns in the 2D matrix	It visually highlights similarities and differences between the biclusters	only one occurrence of each bicluster can be highlighted
6	R package Biclust in [7]	Bicluster in 2D	Bubble plot, Heatmap, Parallel coordinate	Better analysis and understanding of both clusters and biclusters	Simultaneous visualization is not possible



Table 2. Comparison of Visualization Tools

#	Tools	Cluster/ Bicluster	Methods	User Interacti on	Merits	Demerits
1	Gene 3D miner	Cluster	PCA to reduce the dimension	Low	Overcrowding of genes are avoided in this tool	user interaction is poor
2	Cluster maker in [11]	Cluster	Java treeview software for tree view, heatmap,Fore directed layout for network view	Low	Increased efficiency	bicluster visualization is not available
3	Jexpres s in [10]	Cluster	PCA ,tree view	medium	First visualization tool.	Very basic visualization
4	Biolayo ut express 3D in [13]	Cluster	Fruchterman- Rheingold layout algorithm	High. zooming, scaling, rotation, translatio n, and selection of the graph	High-end visualization technologies, Parallelization, allowing the utilization of all available cores simultaneousl y and thus speeding up the operating time.	Possible of producing poor quality of images because of rendering large graphs.
5	Bicat in [5]	Cluster	Heatmap, parallel coordinates are available.	High. Selection of clusterin g algorithm is possible.	GUI level is very high	Selection and comparison of gene/condition inside the heatmap and expression profiles are not available



6	Bivisu in [6]	Bicluster	Heatmap and parallel coordinates.	High. Zooming, selection and comparison features are available.	Selection and comparison features are available.	Simultaneous navigation is not possible
7	Bicluster viewer in [2]	Bicluster	Heatmap and parallel coordinates	Medium	avoids visual clutter, zooming is possible	Continuous duplication and reordering of rows and columns may lead to the loss of originality.
8	Bicoverlapper in [13]	Bicluster in 2D	Force directed graph for overlapper, Heatmap, parallel coordinates	Very high.	Simultaneous visualization of all the biclusters in a single window. High level of user interaction is possible.	Heatmap and parallel coordinates are not good enough in UI
9	EUREKA-DMA	Cluster In 2D	Heat map	High	Apart from clustering, visualization techniques are available for other purposes.	Very basic tool for cluster visualization

4. Conclusion

According to the study, visualization of biclustered gene expression data in 3D provides an efficient way of analyzing the genes and condition present in all the biclusters. But it is a challenging process. Heatmap and parallel coordinates were efficient and often followed for the visualization of cluster genes. But for gene biclusters there is no standard visualization technique. Because, biclusters present in a micro array gene expression data cannot be visualized simultaneously, but can be analysed one by one. BicOverlapper provides an efficient way of visualizing the gene biclusters. Therefore it is better to visualize the gene biclusters in 3D so that all the biclusters can be visualized and analyzed in a single plane with higher level of user interaction.



REFERENCES

- [1] Leishi Zhang, Xiaohui Liu and Weiguo Sheng.:3D Visualization of gene clusters, Computer Vision and Graphics Computational Imaging and Vision Volume 32, pp 349-354,2006.
- [2] Julian Heinrich, Robert Seifert, Michael Burch, Daniel Weiskopf: BiCluster Viewer: A Visualization Tool for Analyzing Gene Expression Data, Advances in Visual Computing Lecture Notes in Computer Science Volume 6938, pp 641-652, 2011.
- [3] Rodrigo Santamaría Roberto Thero and Luis Quintales: A Visual analytics approach for understanding biclustering results from microarray data”, Bioinformatics 9 (247) 2008.
- [4] K.O. Cheng¹, N.F. Law, W.C. Siu and A.W.C. Liew, “Biclusters Visualization and Detection Using Parallel Coordinate Plots”, Proceedings of the International Symposium on Computational Models for Life Sciences, American Institute of Physics, 2007.
- [5] Simon Barkow¹, Stefan Bleuler¹, Amela Prelic², Philip Zimmerman and Eckart Zitzler, “BicAT: a biclustering analysis toolbox”, Bioinformatics, Vol. 22, No. 10: 1282-1283 March 21, 2006.
- [6] K.O. Cheng*, N.F. Law, W.C. Siu and T. H, “BiVisu: Software Tool for Bicluster Detection and Visualization”, Bioinformatics, Vol. 23, No.17: 2342- 2344 June 22, 2007.
- [7] Kaiser Sebastian, Leisch Friedrich “A Toolbox for Bicluster Analysis in R.” In Compstat 2008: Proceedings iComputational Statistics, 2008.
- [8] Gregory A Grothau, Adeel Mufti¹ and TM Murali, “Automatic layout and visualization of biclusters”, Algorithms for Molecular Biology September 2006.
- [9] Yonggao Yang, Prairie View A & M, Niversity, Jim X. Chen “gene expression clustering and 3d visualization”, Computing in science and engineering, volume 5, issue 5, September 2003.
- [10] B. Dysvik and I. Jonnasen, “Jexpress: exploring gene expression data using java”, Bioinformatics, volume 17, issue 4, pp. 369-370, 2001.
- [11] John H Morris, Leonard Apeltsin, Aaron M Newman, Jan Baumbach, Tobias Wittkop⁴, Gang Su, Gary D Bader and Thomas E Ferrin, “clusterMaker: a multi-algorithm clustering plugin for Cytoscape”, BMC Bioinformatics Nov 9 2011.
- [12] Oliver Ru¹ bel, Gunther H. Weber, Min-Yu Huang, E. Wes Bethel, Mark D. Biggin, Charless C. Fowlkes, Cris L. Luengo Hendriks, Soile V.E. Kera² nen, Michael B. Eisen, David W. Knowles, Jitendra Malik, Hans Hagen, and Bernd Hamann “Integrating Data Clustering and Visualization for the Analysis of 3D Gene Expression Data”, IEEE/ACM Trans Comput Biol Bioinform, pp.64-79. 2010
- [13] Athanasios Theocharidis, Stjin van Dongen, Anton J Enright & Tom C Freeman “Network visualization and analysis of gene expression data using BioLayout Express 3D”, Nature Protocols 4, pp.1535 - 1550, Oct 2009.
- [14] Ashraf S. Hussein, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, 11566, Egypt “Analysis and Visualization of Gene Expressions and Protein Structures,” 2008.
- [15] O. R¹ ubel¹, G.H. Weber³, M.-Y. Huang, E. W. Bethel, S. V. E. Ker² anen, C .C. Fowlkes⁵, C. L. Luengo Hendriks, Angela H. De Pace, L. Simirenko, M. B. Eisen⁴, M.D. Biggin, H. Hagen, J. Malik, D. W. Knowles and B. Hamann, “PointCloudXplore 2: isual Exploration of 3D Gene Expression, Journal software, Vol. 3, 2008.
- [16] Sagi Abelson, “Eureka-DMA: an easy-to-operate graphical user interface for fast comprehensive investigation and analysis of DNA microarray data”, BMC Binformatics, Feb 24 2014.



IJCSBI.ORG

- [17] Georgios A Pavlopoulos, Anna-Lynn Wegener and Reinhard Schneider, “A survey of visualization tools for biological network analysis”, Bio Datamining, 28 November 2008.
- [18] Tangirala Venkateswara Prasad* and Syed Ismail Ahson, “A survey of Visualization of microarray gene expression data”, Bioinformation, pp. 141-145. May 3 2006.

This paper may be cited as:

Poornima, S. and Emilyn, J. J., 2014. A Comparative Analysis on Visualization of Microarray Gene Expression Data. *International Journal of Computer Science and Business Informatics*, Vol. 12, No. 1, pp. 87-99.