Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Views & Challenges**

Visualization of microarray gene expression data

Tangirala Venkateswara Prasad* and Syed Ismail Ahson

Department of Computer Science, Jamia Millia Islamia University, Jamia Nagar, New Delhi - 110025, India;
Tangirala Venkateswara Prasad* - Email: tvprasad2002@yahoo.com; * Corresponding author

received January 1, 2006; revised April 16, 2006; accepted April 28, 2006; published online May 03, 2006

Abstract:

Microarray gene expression data is used in various biological and medical investigations. Processing of gene expression data requires algorithms in data mining, process automation and knowledge discovery. Available data mining algorithms exploits various visualization techniques. Here, we describe the merits and demerits of various visualization parameters used in gene expression analysis.

Keywords: Bioinformatics; visualization; microarray gene expression data; data mining

Background:

Bioinformatics provides data mining tools for prediction, comparison and discovery. **[1]** The growing amount of sequence, expression and pathway data demands efficient storage and computing systems. Visualization of gene expression data is extremely important for biological knowledge discovery. **[2]** Therefore, it is essential to develop meaningful visualization tools for information extraction. Available tools require multi-layered analysis for knowledge extraction.

Visualization of extracted data represents results in a lucid and concise manner. However, improved visualization techniques are required for the representation of statistical information on genes in multiple dimensions (behavioral trend, historical details, internal relationships with other elements and external relationship with other organisms). Here, we describe different visualization tools and parameters used in gene expression analysis.

Current Developments:

Biological data are often in text and numerical formats unlike symbols and images. Hence, visualization of biological phenomena after data mining is critical. **[3]** This is particularly essential for gene expression and sequence data analysis. Comparative genomics also require visualization tools for the extraction of meaningful insights.

Gene expression data are stored in rows (genes) and columns (samples/observations). The use of ANN (artificial neural network) and other powerful techniques, especially SOM (self organizing map) and HC (hierarchical clustering) are well known for visualization. **[4, 6, 10, 11, 12, 13,14]** Pre-processing of data into binary form and then processing through ANN for visualization has been demonstrated. **[3]** However, efficient methods are required to communicate results to the user in simple and easy manner. Most widely used visualization techniques for visualizing microarray gene expression data are listed in Table 1. Data are of three broad types in gene expression analysis and they are described as follows.

Time series data:

Gene expression data of a single patient with multiple samples taken over regular time intervals is of this type. This type of data is easily analyzed and a comparative study of the gene expression of a set of genes over time is obtained. This is useful in identifying active and inactive genes over time. **[4]** Methods to compare two or more sets of time series data for different patients are important. Development of such tools requires high dimensional processing and mining for visualization of data.

Data with identical parameters

Samples from different patients are arranged in a 2D grid for easy comparison. This arrangement provides information about dormant and active genes. **[5, 6, 7, 8]** Such arrangement are useful for (1) establishing new classes of disorders **[6, 9]**, and (2) comparison of genes across samples. **[5, 6, 10]**

Data with different parameters

Data containing many parameters for different sets of observations come under this category. GEDAS (Gene Expression Data Analysis Suite) developed by our group **[15]** is primarily used for analyzing gene expression data of this nature using various visualization techniques.

Scope for improvements

There are other visualization techniques available for incorporation into the current developments. **[12, 13, 16]**

References:

[1] J. Vilo, *et al., ISMB2000,* (2000) [PMID: 10977099]
[2] A. Brazma, *et al., Advances in Biochemical Engineering/Biotechnology*, 77:113 (2002) [PMID: 12049748]
[3] A. Narayanan, *et al., Applied Bioinformatics*, 1:191 (2002) [PMID: 15130837]
[4] http://rana.lbl.gov/EisenSoftware.htm
[5] http://classify.stanford.edu/
[6] http://gepas.bioinfo.cnio.es
[7] http://www.csie.ntu.edu.tw/~cjlin/libsvm
[8] C. H. Chen, *Statistica Sinica,* 12:7 (2002)
[9] http://www.dbsr.duke.edu/research/softwaredev/applications/desktop/treemapclusterview/default.aspx
[10] http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster
[11] http://www.cis.hut.fi
[12] http://www.silicocyte.com
[13] http://bioinformatics.upmc.edu/GE2/GEDA.html
[14] H. Caron, *et al., Science*, 291:1289 (2001) [PMID: 11181992]
[15] T. V. Prasad, *et al.*, *Bioinformation,* 1:83 (2006)
[16] http://www.tomsawyer.com/gallery/gallery.php?printable=1

S. No.	Visualization technique	Description or interpretation	Special considerations or features	Advantages and drawbacks	Complexity	Application	References
1	Cluster view – (a) textual view	Clusters contain actual gene names, as generated in most of the text based ANN software. It is the most primitive of all and could be confusing for extremely large datasets	None	Output is impressive, but could be difficult to understand. It does not give an idea of overall gene expression	O(n) if the entire matrix has been sorted on cluster number; otherwise $O(n^2)$	SOM, LVQ, SVM and k-means; extended to HC and PCA	[3], [6], [10], [16]
2.	Cluster view – (b) temporal or wave graph	Clustered data is visualized in the form of a set of waves, in which each wave corresponds to a gene across samples on the X-axis. Also known as the temporal or wave graph view, this visualization technique can also be displayed as pie graph	An extra wave is plotted in black colour to indicate average value of each sample in the cluster, which gives a fair idea of the expression level. It is also essential to display zoomed view of each cluster to enable scientists to see the expression behaviour in enlarged form	Combined plot of all the genes can determine level of expression, overall as well as for specific genes. Very helpful for representing time series data. Requires a further GUI support to extract corresponding gene names. For a dataset containing discrete data (such as those containing numerous parametric values), this representation could render no meaningful use	O(n) if the entire matrix has been sorted on cluster number; otherwise $O(n^2)$	SOM, LVQ, SVM and k-means; extended to HC and PCA	[3], [6], [10], [16]
3.	Heat map	Introduced as one of the most elegant graphical representations of cluster contents, it is very helpful in large scale data mining applications	It can be further enhanced by use of coloring schemes to represent cluster of similar clusters. The coloring scheme conveys "logical classes" in the dataset or "knowledge" about certain hidden parameter common amongst all the clusters	When number of clusters is too large, there can be many null or empty clusters, which convey no meaning and thus be eliminated. It can be applied to datasets having a number of features in them. Coloured sections or groups of clusters give an idea of number of possible classes in the dataset. Requires data to be submitted in a specific format. When there are extremely large numbers of features (microarray gene expression does not have many), output could become cumbersome	$O(n^2)$	SOM, LVQ and k-means, SVM and HC	[11]
4.	Dendrograph view	Also called as checks view, it is very similar to dendrogram output generated by [16] or other graph visualization software. It can be	User can control selection of colours for representing low to high gene expression such as green to red or blue	Most effective form of visualizing trend of gene expression in many samples and genes in one shot. However, if the dataset is very large,	O(n) if the entire matrix has been sorted on cluster number; otherwise	SOM, LVQ and k-means, SVM and HC	[16]

		used for visual inspection of raw, preprocessed and clustered data. This representation alone is not a true dendrogram output as it does not accompany gene tree and array tree	to red or so on. Different colour codes can be assigned to represent null values or zero values. Shades represent intensity or magnitude of expression	it requires another GUI support to extract the gene names. Very helpful for studying trend in time series data and data of same parameter over different samples	$O(n^2)$		
5.	Proximity or distance map	It is a plot of distances between genes vs. genes similar to the distances table of various cities in the world as seen in the diaries. The gene expression matrix is sorted on cluster number, and then distance matrix is developed, which is a diagonal matrix. Each value is then displayed in the form of a coloured box. While white colour represents zero distance, black represents maximum distance. Diagonal line is always white indicating zero distance between same genes	Black and white proximity map can be given a coloured effect by displaying all bands of genes in a single colour shade within a cluster	With just a small plot, a fair view of cluster distances can be determined. It can provide better GUI so that a desired rectangular portion can be selected and corresponding genes listed out from database. One of the most powerful visualization techniques for analysis of gene expression data	$O(n^2)$. The biggest drawback is that it requires sorting of the entire gene expression data matrix	SOM, LVQ, k-means, HC, SVM and PCA	[8], [16]
6.	Microarray view	Very much similar to dendrograph view (or checks view) and is used for visual inspection of raw, preprocessed and clustered data	Same as dendrograph view	Same as dendrograph view	$O(n^2)$. Biggest drawback is that it requires sorting of the entire data matrix	SOM, LVQ, k-means, HC, SVM and PCA	[16]
7.	Tree or dendrogram view	One of the most effective and powerful representations of clustered gene expression data, consisting of three portions viz., gene tree, array tree and colour coded band of gene expression. Also known as matrix tree plot or 2-way dendrograms. In HC, it is the most common output. In order to standardize and provide common outputs to all data mining applications, output was converted into cluster view that further led to views such as microarray, textual,	Same as dendrograph view	Offers clustering of both genes and samples simultaneously. However, if the dataset is very large, it also requires another GUI support to extract the gene names. Very helpful for studying the trend in time series data and data of same parameter over different samples. Inter- date relationship will be lost by representing multi dimensional data in a 2D tree format	$O(n^3)$; two different algorithms work together to build up the gene tree on one side, and array tree and the colour coded band of expression values on the other	HC	[4], [6], [13], [16]

		whole genome, etc					
8.	Principal component view	PC view is a line graph drawn as sum of principal components (Eigen value) and individual expression values. Though, all components are displayed, the first two or three PCs play important role in dimensionality reduction	A good PC line graph always falls down on X axis exponentially. Facility provided to change colours of PCs. Graph can be redrawn with required number of PCs	First two or three PCs are usually sufficient to generate the entire dataset.	O(n)	PCA	[10], [16]
9.	Scatter plot	In addition to plot genes after PCA, it can be used for MA plots, preprocessing, etc. For PCA display of two PCs, one vertical and other horizontal. All data points are projected on these two lines	Different colours for samples may be considered	For PCA, output is processed for second time to project all data points on the PCs	O(n)	PCA	[10], [13], [16]
10.	Whole microarray graph view	A highly versatile representation of gene expression data with each band (or line graph) corresponding to each sample. The portion between two horizontal lines contains expression values of 100 genes	The last band is used to represent median of all the samples. Facility to change background colour and colour of bands provided. Visualization can be zoomed or reduced as per need	While behaviour of all individual samples can be visually matched with neighbouring samples, it requires more time for representing more samples (when zoomed out)	$O(n^3)$	Raw data and pre-processed data; extended to SOM, LVQ, k-means, HC, SVM and PCA	[14], [16]
11.	Decision-space or search-space view	Used for classification, either SVM or LVQ, and this kind of representation of gene expression data come very handy with each decision space corresponding to each class of genes. The figure was extracted from a demo version of SVM [7], it could be applied to any multi-class classification	Coloured decision spaces give very impressive look and better understanding of the data grouping	Requires data to be in 2D form; higher dimensional representation could be very complex	Not available	LVQ and SVM. Could be extended to SOM, k-means, HC, PCA, etc. if it is possible to represent their output in 2D	[7]
12.	Tree-map view	Applied to the results of gene expression data from hierarchical clustering [4]. There are a large number of Tree-map visualization variants for representing hierarchical relationships	Colour and depth variation could be effectively used to form clusters, which further exhibit information such as cluster size, overall expression, etc.	Visually very attractive for smaller datasets; can be very confusing or scrammed for larger datasets	Not available	HC, could be extended to other clustering and/or classification techniques	[4]
13.	Box-	Very handy in dealing with raw and pre-processed gene	Colour variation could	Mean, median as well as upper and lower	O	Raw data	[13], [16]

Whisker plot	expression data. The plot provides information on overall over and under expression along with mean, upper and lower quartile together in one plot. It can, in many cases, be applied to reduced or transformed datasets with large number of insignificant genes and samples pruned	be introduced for different samples, upper and lower quartile variation	quartile can be viewed simultaneously; useful for preliminary analysis and when a number of genes and/or samples have been eliminated, causing change in the mean and other parameters	(n^2)	and pre-processed data

Table 1: Description of various visualization techniques for microarray gene expression data

Edited by P. Kangueane

Citation: Prasad & Ahson, Bioinformation 1(4): 141-145 (2006)