# GeneBox:visualizing gene expression data resulting from microarray experiments

Nameeta Shah[*] Dina A. St. Clair[†] Clark Dodsworth Jr.[‡] Bernd Hamann[*] Kenneth I. Joy[*]

## Abstract

DNA microarray technology is used in biology to study the simultaneous expression of hundreds to thousands of genes in an organism under specific experimental conditions (treatments, time, genotypes, etc.). Experiments using microarrays to study global gene expression often produce massive amounts of abstract multivariate, multidimensional (MDMV) data. Considering the size and biological complexity of these data, the task of analysis to extract biological meaning can be extremely challenging[Nadon and Shoemaker 2002; J. 2001]. There are multiple sources of 'noise' (error) in microarray data, including both biological and technical. Proper statistical techniques are required to measure and separate the real biological effects (the changes in gene expression in response to experimental conditions) from the sources of error. Therefore the challenge with global gene expression data sets is twofold: their nature (large, abstract MDMV) and separation of real biological effects from error. To begin to explore the biological meaning of these complex data sets, our group has developed an interactive data visualization technique in 3D that depends on a mapping considering the experimental variables time, genotype, and treatment.

**Keywords:** microarray experiments, visualization

## 1 Introduction

Currently, the generation and analysis of global gene expression data is the focus of many biological researchers[J. 2001; van Berkum and Holstege 2001]. Research on appropriate statistical techniques for the analysis of expression data to make possible precise comparisons of changes in gene expression is also being pursued by statisticians[Nadon and Shoemaker 2002]. One of the ways gene expression data can be examined is by identifying and grouping genes with similar patterns of expression changes. One hypothesis is that genes with similar expression patterns under certain experimental conditions may be involved in a common function or biological pathway. Various data clustering techniques like hierarchical clustering, k-means and SOM are being applied by biologists in an effort to identify commonalities of expression level over time, treatments or other experimental variables. The choice of a particular normalization method, error-measurement model[Rocke and Durbin 2001], and clustering technique with a specific distance metric all directly influence the identified clusters of genes that behave similarly[J. 2001; Nadon and Shoemaker 2002]. Thus, it is important in any data analysis and visualization tool that the researcher has the ability to make choices appropriate for their data set, in order to avoid arbitrary results in terms of the number of data clusters, size, density, etc.

## 2 Motivation

One "best" technique for pattern discovery in gene expression data does not currently exist[J. 2001; R et al. 2000]. Interactive visualization and data exploration can be important ways to complement statistics-based data analysis. A visualization tool should equip a biologist with a means to display the results of various data analysis techniques and reveal biological meaning. Commercial and freeware visualization tools are available for gene expression data analysis[R et al. 2000; J. 2001]. Many use only two dimensions, and often the third dimension is not exploited well. Consequently, we have begun to develop a three-dimensional visualization tool we hope will help visualize differentially expressed genes in a biologically meaningful way. We call our prototype tool "GeneBox."

## 3 GeneBox

One typical question a biologist may ask about gene expression data is: Which genes exhibit change in expression with respect to different treatments, genotypes and at different time points? GeneBox is a prototype visualization tool designed to help address such questions in a visual way. Visualization in GeneBox is based on a mapping that maps each gene to a point in the unit cube, whose coordinates are defined by a differential operator. This operator maps genes that do not change expression level with changing time/genotype/treatment to the origin of the cube, i.e., the point (0,0,0); genes that do change expression level maximally with changing time/genotype/treatment are mapped to the point (1,1,1). Each gene, by considering its differential expression level properties, is thus identified with a point in the unit cube and is visualized by a sphere primitive. The visual representation of an entire microarray experiment is a set of spheres rendered in the unit cube. In GeneBox, the user can select a subset of the genes, using a selection box, and apply various clustering techniques to the specified subset. The color of a sphere primitive can provide additional information about that gene's expression behavior. In preliminary tests, we have used GeneBox for the visualization of a relatively simple experiment described in the Results section. GeneBox will also support different normalization techniques and application of different distance measures.

[*] Center for Image Processing and Integrated Computing, Department of Computer Science, University of California, Davis, CA 95616-8562, {shahn, hamann, joy}@cs.ucdavis.edu

[†] Department of Vegetable Crops, University of California, Davis, CA 95616-8746, stclair@vegmail.ucdavis.edu

[‡] Osage Associates, 3420 Pierce St., SF CA 94123, clark@dodsworth.com

## 4 Results

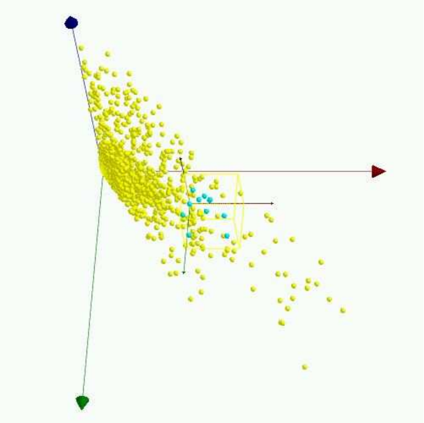

Figure 1: Typical visualization of all genes in the GeneBox. Expression values are not "log-transformed." Genes are closely clustered near the origin.

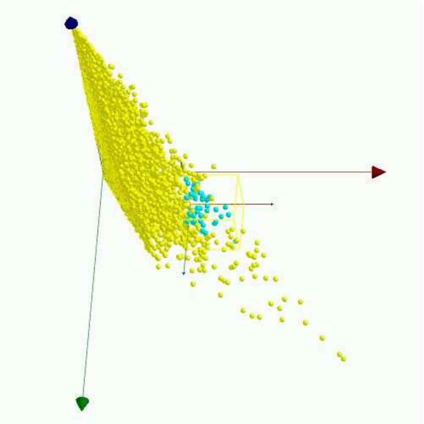

Figure 2: Effect of transforming expression data. Same genes as shown in Figure 1, expression values "log-transformed." Genes are distributed more uniformly.

GeneBox was used to visualize the data from an experiment to screen for differences in gene expression changes in Arabidopsis thaliana (a small flowering plant widely used as a model organism in plant biology). Salicylic acid (SA) was applied to plants of two different genotypes of Arabidopsis to induce a subset of genes to respond, and replicated biological samples were taken 0, 4, and 28 hours after either application of SA or water (a control treatment). One objective of this experiment was to identify those genes which respond to SA induction differently in one genotype compared to the other. The gene expression data set was displayed in GeneBox without and with prior log-transformation[Nadon and Shoemaker 2002] (Figures 1 and 2, respectively), illustrating how data transformation can affect the display.

## 5 Conclusion

GeneBox is a potentially useful tool for the visualization of gene expression data to help provide valuable insight into the complex biological behavior of gene expression. More work is planned to extend the functionality and utility of GeneBox to biologists. Further testing is required with additional gene expression data sets to determine the biological validity of the mapping results from GeneBox and conceptual utility to research biologists. Mapping abstract data to visual representations may provide answers to questions such as: (1) Considering a specific gene, how sensitively does it change its expression level in response to changing time/genotype/treatment? (2) Considering all genes simultaneously, which genes exhibit highly similar behavior, as evidenced by primitives closely clustered together in the unit cube?

## References

GILBERT, D. R., SCHROEDER, M., AND VAN HELDEN J. 2000. Interactive visualization and exploration of relationships between biological objects. *Trends in Biotechnology, Vol.18 No.12*.

J., Q. 2001. Computational analysis of microarray data. *Nature Reviews Genetics, Vol. 2*, 418–427.

NADON, R., AND SHOEMAKER, J. 2002. Statistical issues with microarrays: processing and analysis. *TRENDS in Genetics, Vol.18 No.5*, 265–271.

ROCKE, D. M., AND DURBIN, B. 2001. A model for measurement error for gene expression arrays. *Journal of Computational Biology, Vol.8 No.6*, 557–569.

VAN BERKUM, N. L., AND HOLSTEGE, F. C. 2001. DNA microarrays: raising the profile. *Current Opinion in Biotechnology, Vol.12*, 48–52.