

# GENEGOBI : VISUAL DATA ANALYSIS AID TOOLS FOR MICROARRAY DATA

Eun-kyung Lee, Dianne Cook, Eve Wurtele, Dongshin Kim, Jihong Kim, and Hogeun An

*Key words:* Data visualization, Gene Expression data, GGobi, Metabolic networks, Multivariate data analysis.

*COMPSTAT 2004 section:* Statistical software.

**Abstract:** GeneGobi is a software for the exploratory analysis of microarray data and metabolic networks. It helps biologists analyze the connections between microarray data and metabolic pathways visually and interactively. This software is built on the best of open-source statistical analysis software, R, and the best of open-source data visualization software, GGobi. GeneGobi combines the advantages of these two softwares and provides a “user-friendly” interface to the analysis and graphical tools available in these powerful packages.

## 1 Introduction

Microarray experiment generate huge data sets. Current software provides only limited way to look at microarray data and there is no available software that combines visualization and analysis of metabolic networks. GeneGobi is developed for exploratory analysis of microarray data and metabolic networks. It helps biologists explore patterns in gene expression data. With this software, biologists also can explore the regulatory and metabolic pathways. This software is originally designed for Arabidopsis and analyze the connections between microarray data and metabolic pathway of Arabidopsis visually and interactively. However it can be used for general gene expression data analysis connected to pathways.

The software is built on the best of open-source statistical analysis software, R, and the best of open-source data visualization software, GGobi. R is the package that the major statistical analysis packages are being developed in today, for example, BioConductor, which is used heavily for gene expression data analysis. GeneGobi provides a “user-friendly” interface to the analysis and graphical tools available in these powerful packages.

## 2 GeneGobi

GeneGobi is a visual data analysis aid tool for microarray data and metabolic networks. It is based on R and GGobi. R is an open source statistical analysis software and there are many contributed statistical analysis packages in R. Also packages for gene expression data analysis are available (e.g. Bioconductor).

GGobi is an interactive and dynamic data visualization system for multivariate data. It is able to use in R (Rggobi package). GGobi allows the user to explore multivariate data using bar charts, scatterplots, 3D rotating plots and higher dimensional rotations (unique to GGobi), profile plots. The user can interact with each plot by brushing points and lines using different symbols and line types, and these actions simultaneously change elements of other plots as appropriate. The linking between plots is fast and can be sophisticated, using selected columns in the data as the key to graph elements in different plots. This is different from virtually all other packages. Elements of the plots can also be identified with user-provided labels, or by selected columns in the data.

GGobi can also display metabolic networks. Users can read in a layout from another package such as FCModeler, and interact with the network by brushing nodes and edges and identifying nodes or edges. Elements of the network can be linked to other types of data displayed in other plots. For example, where we have identified the LocusID of the gene responsible for a particular entity in the network this provides a key to link to gene expression data.

There are a couple of drawbacks in R and GGobi. To analyze gene expression data or develop new methods, we need to write code in R and it is not easy for a naive user. Another one comes from huge data. Usually gene expression data has thousands of genes. In stand-alone ggobi with thousands of genes, it is not easy to brush or identify a few specific genes. In Rggobi, we can control gene by gene and it also needs to use code in command line.

The main purpose of GeneGobi is to help biologist explore patterns in gene expression data and the regulatory and metabolic pathway. It helps analyze the connections between microarray data and metabolic pathway visually and interactively. It also can combine statistical analysis results with interactive plots to improve the analysis.

GeneGobi provides a user-interface to both GGobi and R. It adds a spreadsheet to more information about each gene, links to literature from the Web, menus of analysis and visualization options and an interface to lists of selected genes created from expert knowledge or previous analyses. We can brush selected genes from this spread sheet using color buttons.

We give an overview of GeneGobi by summarizing main GUI, menu bar and buttons. For demonstration, we use gene expression data of Arabidopsis. This experiment uses two different genotype(WT and ACL) at 5 different time points. We also considered a hypothetical jasmonic acid network to show how gene expression data analysis and network analysis are combined with GeneGobi.

## 2.1 Main GUI

Main GeneGobi GUI has a lot of features. **Gene Information** provides a lot of information about genes. It should be provided by user. In this **Gene**

**Information**, user can select genes that need to be analyzed. **Chip List** provides a list of experiments. We can choose interesting experiments for future data analysis. **Exp.info** button shows the experimental information, such as genotype, time, treatments, etc. for each experiment. This information is useful for advanced statistical analysis. **Gene List** show a list of cluster that can be well-known one or user-defined one. When one cluster in the gene list is selected, genes in this cluster are changed to “Red” and the other genes are changed to “Gray” in GGobi plot. In the Gene Information spreadsheet, genes in this cluster come up to the top list and the name of this cluster shows in the Gene List column.

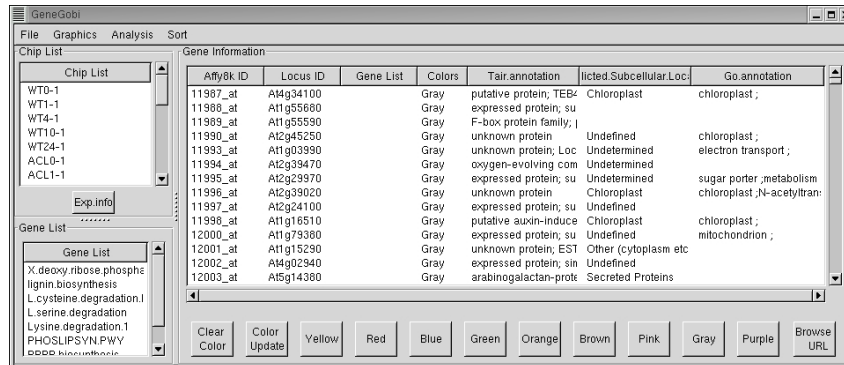


Figure 1. Main Window of GeneGobi : Menu bar is on the top. Chip List(experiment name) and Gene List(user defined) is in the left side. All the information about genes are in the right side. There are also color control buttons at the bottom of the right side.

## 2.2 Menu

### 2.2.1 File

- Read Files : read files that need for data analysis
  - Gene Expression : read gene expression data. It should be the normalized data in text mode(tab limited or comma separated)
  - Gene Information : read gene information file that contains affy id, locus id, tail.annotation, go.annotation, etc. At least, you need to have locus id.
  - Network : use SBML
  - Gene List : read user defined cluster file. It should have cluster name and a list of genes in text mode.
  - Experimental Information : read experimental information file. It contains genotype, replicate, time, treatments, etc.
- Exit

## 2.2.2 Graphics

- Open GGOBI : Before you choose this menu, you need to load data(at least gene expression and gene information data).
  - with Gene Expression : Only if you want to analyze gene expression data alone.
  - with Gene Expression + Network :

## 2.2.3 Analysis

- Compare Genotype : Before you choose this menu, you need to select at least two chips with different genotypes from Chip List.
  - Correlation : Calculate the correlation between two genotypes. In the Gene Information spreadsheet, genes are sorted by this correlation in descending order.
  - Covariance : Calculate the covariance between two genotypes. In the Gene Information spreadsheet, genes are sorted by this covariance in descending order.

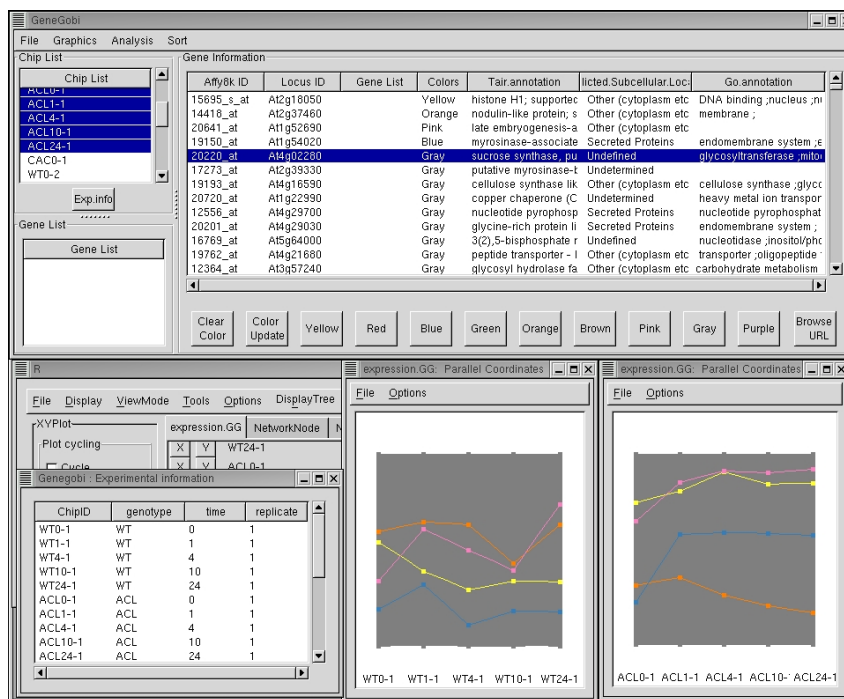


Figure 2. How to use “Compare Genotype” menu : 1) Choose the first replicates of two genotypes, WT and ACL from the Chip List. 2)

Choose Compare Genotype  $\rightarrow$  Covariance 3) gene 15695\_s\_at(yellow) is the most different gene between WT and ACL according to this covariance

- Find Interesting Genes : Before you choose this menu, you need to select two chips from Chip List. After calculating measures, new measure is added to GGobi and genes in the Gene Information spreadsheet are sorted by this calculated measure.
  - Difference : Fit  $Y = X$  for two selected experiments from Chip List and calculate the residuals. Genes are sorted by the absolute values of these residuals in descending order. This measure is used for the log-scaled data to represent fold changes.
  - Regression : Fit a simple linear regression model( $Y = a \cdot X + b$ ) for two selected experiments from Chip List and calculate the residuals. Genes are sorted by the absolute values of these residuals in descending order.
  - Angle : Calculate the angle from  $Y=X$  line for two selected experiments from Chip List. Genes are sorted by these angles in descending order. This measure is used for the raw data(without log transformation) to represent fold changes.
  - Mahalanobis : Calculate the Mahalanobis distance from means of two selected experiments from Chip List. Genes are sorted by these distances in descending order.  $d_M(\mathbf{x}) = (\mathbf{x} - \bar{\mathbf{x}})^T \Sigma^{-1} (\mathbf{x} - \bar{\mathbf{x}})$  In here,  $\Sigma$  is estimated from two selected experiments from Chip List.)

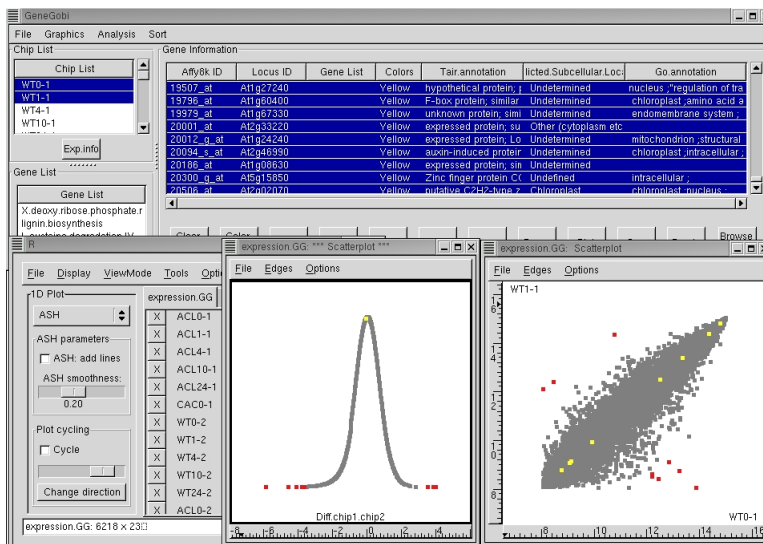


Figure 3. How to use “Find Interesting Genes” menu : 1) Choose two chips from Chip List, WT-0, and WT-1. 2) Choose Find Interesting Genes → Difference 3) genes with red colors are most different genes between two chips and genes with yellow colors are most similar genes between two chips.

- Find Similar Genes : Before you choose this menu, you need to select one gene from Gene Information spreadsheet and select chips that you want to consider. After calculating measures, new measure is added to GGobi and genes in the Gene Information spreadsheet are sorted by this calculated measure.
  - Euclidean : Calculate the Euclidean distance from the selected gene( $\mathbf{x}^*$ ). This distance variable adds to GGobi. In the Gene Information spreadsheet, genes are sorted by these distances in ascending order.
  - Corr : Calculate the correlation distance from the selected gene. This distance variable adds to GGobi. In the Gene Information spreadsheet, genes are sorted by these correlations in descending order.
  - Zerocorr : Calculate the zero correlation distance from the selected gene. This distance variable adds to GGobi. In the the Gene Information spreadsheet, genes are sorted by these zero correlations in descending order.

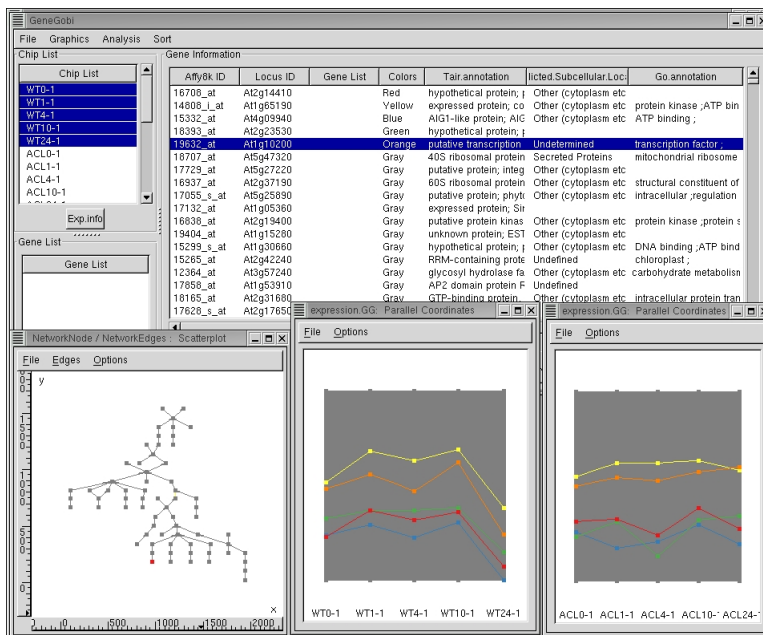


Figure 4. How to use “Find Similar Genes” menu : The plots at the bottom

include a view of a hypothetical jasmonic acid network and two profile plots of gene expression data

- Find Cluster :
  - hclust : with subset of data(selected genes and selected chips), find cluster using hclust function in R and draw an interactive dendrogram. This dendrogram is connected to GeneGobi and the other plots in GGobi.

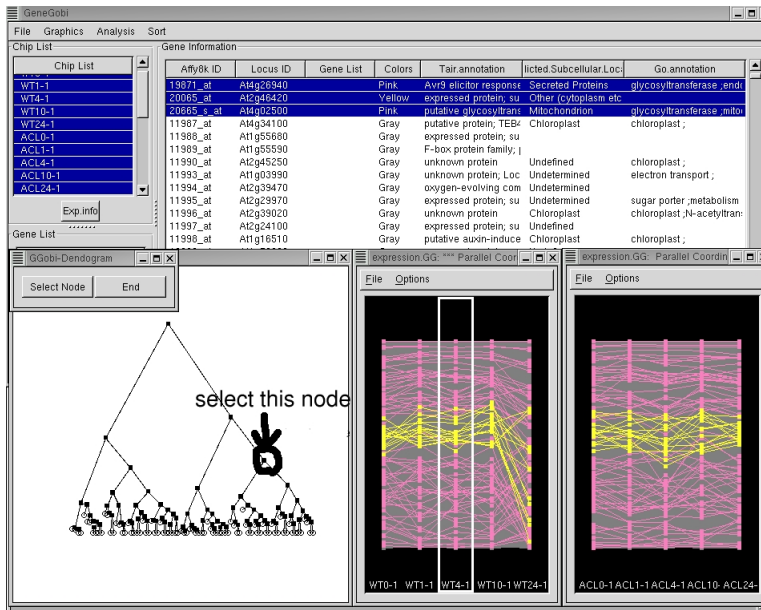


Figure 5. How to use “Find Cluster” menu : Whenever select “select node” button, this dendrogram changes to interactive mode and the right sides of selected node are denoted by color “Yellow” and the left sides are “Pink”

## 2.2.4 Sort menu : sort genes in the Gene Information spreadsheet

- by AffyID/LocusID/GeneList/Colors

## 2.3 Buttons

- Clear Color : change all colors in GGobi plot to “Gray” which is used as a base color
- Color Update : when user changes colors in GGobi directly using “BRUSH”, use this button to update color information in the gene information spreadsheet.

- Yellow/Red/Blue/.../Purple : change colors for selected genes in the gene information spreadsheet.
- Browse URL : for selected genes, link to literature from the web

### 3 Conclusion

GeneGobi provides a user-interface to both GGobi and R. It adds a spreadsheet to more information about each gene, links to literature from the Web, menus of analysis and visualization options and an interface to lists of selected genes created from expert knowledge or previous analyses. It is flexible to add new method. Therefore GeneGobi can be a new platform of the visualization of gene expression data and the analysis of functional relationship between genes, proteins and metabolites.

### References

- [1] Parmigiani, G., Garrett, E. S., Irizarry, R. A., and Zeger, S. L. (2003). *The Analysis of Gene Expression Data*. Springer, New York, NY.
- [2] Swayne, D. F., Temple Lang, D., Buja, A. and Cook, D. (2002). *GGobi: Evolving from XGobi into an Extensible Framework for Interactive Data Visualization*. *Journal of Computational Statistics and Data Analysis* **43**, 423–444.

*Address:* Iowa State University, Ames, IA, USA

*E-mail:* kyung@iastate.edu