

Using a Gene Ontology Grid

¹Liu, W.L. ¹Duan, Z.-H. ^{1*}Xiao, Y.C.
²DiDonato, J.A.

¹*Department of Computer Science*

University of Akron

Akron, Ohio 44325-4403

United States of America

²*Department of Cancer Biology*

Cleveland Clinic Foundation

Cleveland, Ohio 44195

United States of America

*Correspondence should be addressed to:

xiao@uakron.edu

☎ (330) 972-5809

ABSTRACT

A new visualization technique, gene ontology grid, is described in this paper. The gene ontology grid is designed to allow biologists to visually analyze the relationship between microarray gene expression data and gene functions. The grid is generated using a gene ontology dendrogram and a list of genes under study. Using this technique, genes are clustered, based on their expression values and biological functions, and can be subsequently visualized side-by-side and analyzed for meaningful patterns.

CATEGORY

-Microarray Data Analysis

-Software Application

Proc Virt Conf Genom and Bioinf (2):12-16

Print ISSN 1547-383X

Online ISSN 1547-7320

Copyright © 2003. All Rights Reserved

www.virtualgenomics.org

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not distributed for profit or commercial advantage.

Keywords: microarray, gene expression, gene ontology, gene functions, visualization.

1. INTRODUCTION

DNA Microarrays, also known as gene chips, provide an effective tool for monitoring and profiling gene expression patterns by measuring the expression levels of thousands of genes simultaneously [2,6,8,20]. Microarrays are widely used to identify genes of particular interest. For example, after studying the expression patterns of 16,000 genes using Affymetrix gene chips, Ramaswamy et al. [19] identified a gene expression signature and 17 genes in cancer cells that predict metastasis and early death.

To reveal patterns in large amount of gene expression data from microarrays, biologists use clustering methods to group genes with common features [18]. Among the frequently used clustering methods are hierarchical clustering [27], k-means [16] and self-organizing maps [14].

Gene ontologies (GO) [23] are vocabularies that describe the attributes of genes (for example, their biological functions). Each term in the vocabularies is called a GO term which represents a possible attribute value possessed by one or more genes. Gene ontologies are relational in the sense that GO terms are related and form a directed acyclic graph. The Gene Ontology Consortium [24] is creating three standard gene ontologies that describe the associated biological processes, cellular components and molecular functions for genes and their products (RNA or protein products encoded by genes). GO annotations are made to associate each gene or gene product with its related GO terms. Those GO terms reflect the normal function, process, and localization (component) of the gene or gene product. Therefore, they are of great importance in studying the behavior of the gene or gene product.

Many visualization techniques have been developed for visualizing gene expression data and gene ontologies. Heat-maps, introduced by Eisen et al. [9] use coded colors to display gene expression values so that genome-wide expression patterns can be revealed. Yang et al. [28] recently developed a 3D cluster display technique that plots clusters in a divided 3D grid. Each cube in the 3D grid contains one cluster of genes rendered as balls color-coded according to their cluster numbers. Biologist can interactively examine each cluster to find patterns. GO terms are usually displayed as directed acyclic graphs

(DAGs) or trees (after simplification) with each node of a DAG/tree showing the textual information of the terms [9, 23]. However, to make the textual information readable on the screen, the number of nodes that can be displayed simultaneously is limited. This type of display is usually used to reveal details of the genes after a pattern has already been identified. It is difficult to use such a display to find patterns in genes.

Many of the clustering and visualization techniques have been implemented in various gene analysis software. Some of the software are commercially available, such as GeneSpring [13], and many are free to the public [1-5,7,10-12,17,21,25,26].

In this research, we developed a new visualization technique: gene ontology grid (GOG). It represents GO terms as compact color-coded rectangles. Thousands of GO terms can be displayed on the screen simultaneously with color-coded and clustered gene expression values and optionally fold-change values. GOG can reveal what GO terms are significantly represented in clusters of genes.

In contrast to the textual-based presentation method, GOG provides a more intuitive graphical presentation of the correlations between gene expression and gene ontologies for biologists to find meaningful patterns.

2. MATERIALS AND METHODS

We have developed a software system, GOVis, to implement the GOG technique. The design and implementation details of the system are outlined below. The software will be freely available to the academic users by the end of 2003 (www.cs.uakron.edu/~xiao/gog).

2.1 System Design

The GOVis software system is composed of three components: data acquisition, data processing, and visual presentation (Figure 1). These three components form a visualization pipeline.

The data acquisition component imports gene expression data and GO terms into the system and stores the information in software objects. The data processing component filters the data so that only the desired portion of the data passes through for analysis. The selected data is then clustered to reveal potentially meaningful patterns. This component has a graphical user interface (GUI) that allows users to interactively specify selection criteria for

filtering and control parameters for clustering. The visual presentation component generates the final image graphically representing the processed data.

The image generated by the visual presentation component consists of four main graphical objects as shown in Figure 2: gene fold-change column (optional), heat-map, GO dendrogram and GOG. The GO dendrogram is a simplified tree structure based on the GO terms defined by the Gene Ontology Consortium [24]. Each node of the tree corresponds to a biological function of genes. The color of each node is coded differently among neighboring nodes so that the node can be easily identified. The gene fold-change column, heat-map, and GOG are color-coded matrices. Rows of the matrices represent individual genes. The column of gene fold change shows the color-coded average fold-change values of genes. The columns in the heat-map present individual samples. The colors of the cells in the heat-map code the gene expression intensity. The columns of GOG correspond to the biological functions of the genes as shown in the GO dendrogram above it. If a gene possesses a certain function, the related cell will be coded with the same color as the function node in the GO dendrogram. A gene with multiple functions will have multiple cells color-coded.

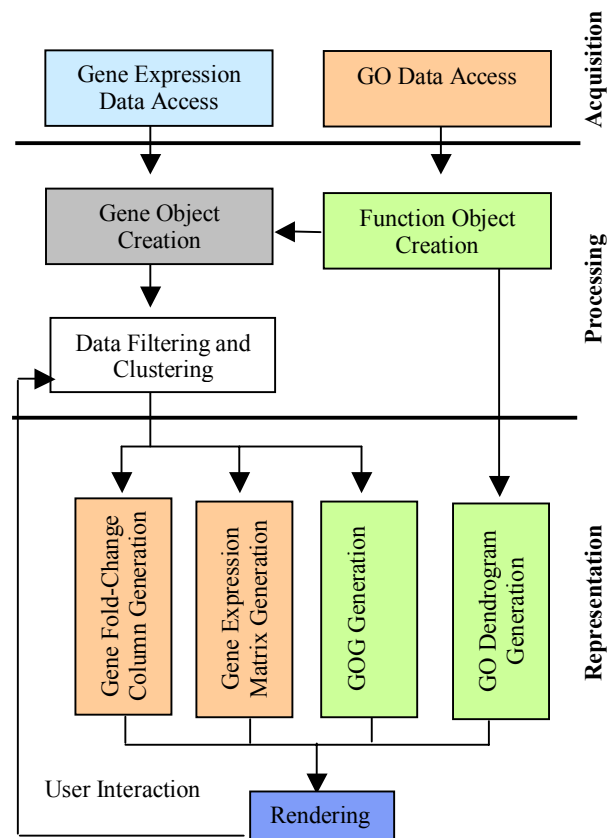


Figure 1. GOVis Visualization Pipeline.

Rows of genes in the graphic image can be ordered by average fold changes (Figure 2), by clusters generated from clustering algorithms (Figure 3), or by their biological functions (Figure 4). We note that a multiple-function gene appears in multiple rows in Figure 4, once for each functional group it is associated with.

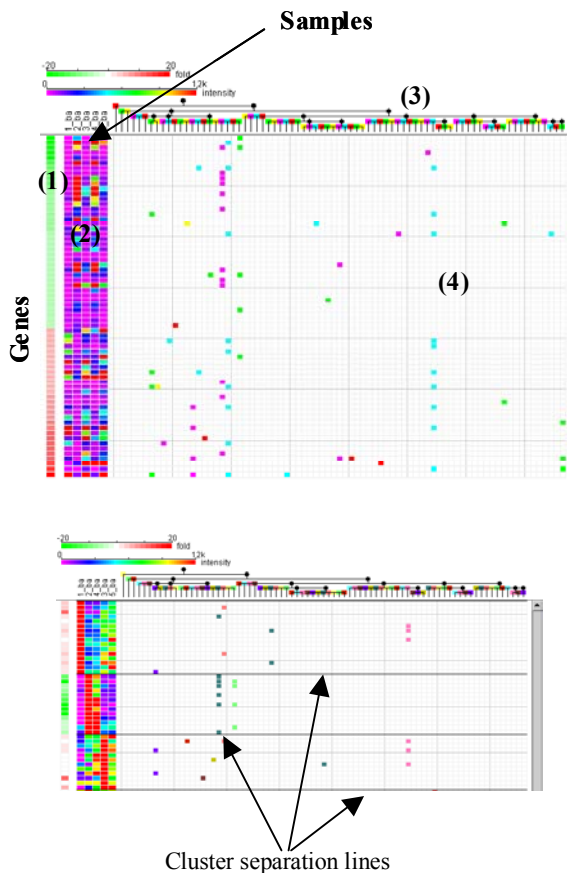


Figure 3. Clustered genes. Genes between two adjacent separation lines belong to the same cluster.



Figure 4. Genes grouped by functions.

Figure 2. Visualization of gene expression data and gene functions using GOG: (1) gene fold-change column; (2) heatmap; (3) GO dendrogram; (4) GOG.

To effectively use the limited screen space to visualize patterns, no detailed textual information (e.g. the names of the genes and their functions) is displayed. But the graphical display is made interactive so that such textual information can be displayed as needed. We term this method as DoD (details on demand). When a user right-clicks on a cell in the expression matrix, the system displays the textual information about the corresponding gene in a pop-up box. A user can also right-click on a node in the GO dendrogram to reveal the name of the function that the node represents. When a user right-clicks in a cell in the GOG, the textual information related to the gene on the row and the name of the function corresponding to the column are displayed (Figure 5).

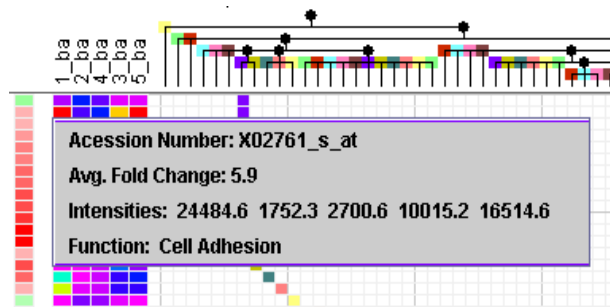


Figure 5. Details-on-Demand (DoD). When a grid cell in GOG is right-clicked, a text box pops up to display the detail textual information of the gene.

2.2 Software Implementation

The GOVis software system is implemented using the JAVA programming language [22], which is platform-independent. Software programs written in Java can run on any platform that supports the Java Virtual Machine.

The GUI of GOVis is implemented using Java Swing and Java Beans [22]. The DoD feature of GOVis is implemented using the event-driven programming paradigm, in which a mouse click is treated as an event and the program routine that processes the mouse click is coded as an event handler. The even handler displays different detail information of the gene depending on the location of the mouse click.

GOVis is a stand-alone software application and it requires the sample data to reside on the same system to be visualized. The advantage of this approach is that run-time data acquisition can be relatively fast since the program does not need to fetch the data across a network.

Users can upload data from other programs into GOVis as long as the data exported from the other programs are first adapted to the GOVis input file format.

3. TEST AND RESULTS

To test GOVis, we applied the system to a case study, in which we visualize and analyze a set of microarray gene expression data [15] of renal cell carcinoma (RCC). The five samples in the data obtained from HuFL gene chips (manufactured by Affymetrix) include one from the RCC cell lines, two from pooled clear cell RCC tissues and two from the pooled, patient-matched normal kidney tissues.

The GO data were acquired using GeneSpring software from Silicon Genetics. The gene ontology annotations for the gene chips are parsed by GeneSpring's "Build Simplified Ontology" tool to group genes hierarchically into meaningful biological categories based on the Gene Ontology Consortium classifications [24]. The result is a simplified GO tree.

The microarray data along with the GO tree data were loaded into the system and filtered by their signal intensity and fold change. Using the visualization system, we observed that most of the genes with greater than or equal to five fold changes in the functional group cell adhesion are on the up-regulated pathway while the ones in transport are on the down-regulated pathway. This result agrees with the findings reported in [15]. Figure 2 displays a portion of the genes and their corresponding biological functions on GOG. The genes were sorted by the average fold change of the gene expression values of RCC and normal kidney tissue.

Three clusters of genes obtained using the k-means method [16] and their biological functions are shown in Figure 3. As we can see, the second cluster includes a group of down regulated genes involved in metabolism and transport. This result demonstrates one of the unique features of the visualization system — correlating the clustering results based on the gene expression values with biological functional groups.

4. CONCLUSIONS AND FUTURE WORK

The GOG visualization technique presented in this paper displays the relationship between the expression values of the genes and their functions. This technique allows us to analyze and visualize the clusters of genes based on their expression values and biological functions together. It

provides a quick and intuitive way to detect useful biological patterns and their potential meanings.

A case study of microarray gene expression data of renal cell carcinoma was performed and the findings of the visualization are consistent with the reported results [15]. In addition, some potentially meaningful patterns were observed. Currently, the k-means clustering algorithm is implemented. Future work will make other clustering methods available in the software. In this first version, GOVis uses a simplified GO structure and displays the structure as a dendrogram. We are looking into ways to import the general GO structure from the Gene Ontology Consortium and to visualize it along with gene expression data.

5. REFERENCES

- [1] AmiGO: <http://www.godatabase.org/dev/>
- [2] Baldi, P.; Hatfield, G.W. 2002. DNA microarrays and gene expression from experiments to data analysis and modeling. Cambridge University Press, Cambridge, UK.
- [3] Berrar, D.P.; Dubitzky, W.; Granzow, M. 2003. A practical approach to microarray data analysis. Kluwer Academic Publishers, Boston, MA.
- [4] Bouton, C.M.; Pevsner, J. 2002. DRAGON View: Information visualization for annotated microarray data. *Bioinformatics*, 18:323-324.
- [5] Causton, H.C.; Quackenbush, J; Brazma, A. 2003. A beginner's guide microarray gene expression data analysis. Blackwell Publishing, Malden, MA.
- [6] Chee, M.; Yang, R.; Hubbell, E.; Berno, A.; Huang, X.C.; Stern, D.; Winkler, J.; Lockhart, D.J.; Morris, M.S.; Fodor, S.P. 1996. Accessing genetic information with high-density DNA arrays. *Science*, 274:610-614.
- [7] Dahlquist, K.D.; Salomonis, N.; Vranizan, K.; Lawlor, S.C.; Conklin, B.R. 2002. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nature Genetics*, 31(1):19-20.
- [8] Duggan, D.J.; Bittner, M.; Chen, Y.; Meltzer, P.; Trent, J.M. 1999. Expression profiling using cDNA microarrays. *Nature Genetics*, 21:10-14.
- [9] Eisen, M.B.; Spellman, P.T.; Brown, P.O.; Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863-14868.

- [10] Expander: www.cs.tau.ac.il/~rshamir/expander/expander.html
- [11] FreeView: <http://magix.fri.uni-lj.si/freeview/>
- [12] GeneCluster: www-genome.wi.mit.edu/cancer/software/software.html
- [13] GeneSpring: <http://www.silicongenetics.com>
- [14] Kohonen, T. 2001. Self-organizing maps. Springer Series in Information Sciences, Vol. 30. 3rd ed. Springer-Verlag, Berlin.
- [15] Liou, L.S.; Shi, T.; Duan, Z.-H.; Sadhukhan, P.; Der, S.D.; Novick, A.A.; Hissong, J.; Almasan, A.; DiDonato, J.A. 2003. Microarray gene expression profiling and analysis in renal cell carcinoma (submitted).
- [16] MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 281-297.
- [17] Parmigiani, G.; Garrett, E.S.; Irizarry, R.A.; Zeger, S.L. 2003. The analysis of gene expression data methods and software. Springer, NY.
- [18] Quackenbush, J. 2001. Computational analysis of microarray data. Nature Reviews - Genetics, 2(6):418-427.
- [19] Ramaswamy, S.; Ross, K.N.; Lander, E.S.; Golub, TR.; 2003. A molecular signature of metastasis in primary solid tumors. Nature Genetics, 33(1):49-54.
- [20] Schena, M.; Shalon, D.; Davis, R.W.; Brown, P.O. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 270:467-470.
- [21] Slcview: <http://slcview.sourceforge.net>
- [22] Sun's Java web site: <http://java.sun.com>
- [23] The Gene Ontology Consortium. 2000. Gene Ontology: tool for the unification of biology. Nature Genetics, 25:25-29.
- [24] The Gene Ontology Consortium. <http://www.geneontology.org/>
- [25] TreeView: <http://rana.lbl.gov/EisenSoftware.htm>
- [26] Treeps: <http://monod.uwaterloo.ca/software/>
- [27] Weinstein, J.N.; Myers, T.G.; O'Connor, P.M.; Friend, S.H.; Fornace, A.J., Jr; Kohn, K.W.; Fojo, T.; Bates, S.E.; Rubinstein, L.V.; Anderson, N.L.; Buolamwini, J.K.; van Osdol, W.W.; Monks, A.P.; Scudiero, D.A.; Sausville, E.A.; Zaharevitz, D.W.; Bunow, B.; Viswanadhan, V.N.; Johnson, G.S.; Wittes, R.E.; Paull, K.D. 1997. An information-intensive approach to the molecular pharmacology of cancer. Science, 275:343-349.
- [28] Yang, Y.; Chen, J.; Kim, W. 2003. Gene expression clustering and 3D visualization. Computing in Science and Engineering, 5(5):37-43.