GO-Chart, A Toolkit for Visualization of Genes in the Context of Gene Ontology with Variable *P*-Values

Deepali Shah[1] Marc Ma[1, 2] Gerard Tromp[3]

[1]Department of Computer Science, [2]Center for Applied Mathematics and Statistics, New Jersey Institute Technology
[3]Center for Molecular Medicine and Genetics, Wayne State University-School of Medicine
Emails: {das23, qma}@njit.edu, gerard.tromp@sanger.med.wayne.edu

Abstract

The analysis of DNA microarray data presents significant hurdles to the researcher. A large number of probe sets may be identified as differentially regulated, yet it is difficult to determine the importance of these changes. Gene Ontology (GO) is a hierarchical categorization of gene functions that can be used to identify patterns in the data from gene expression studies. A number of software tools have been and continue to be developed to perform gene ontology analyses and identify functional categories among differentially expressed genes. A common problem is that each tool lacks some desirable feature or features that are present in other tools. We have developed a tool, "GO-Chart," for visualizing more conveniently the output from GOTM (GOTree Machine), while adjusting the *p*-values for multiple testing. GO-Chart allows users to analyze genes by GO functional categories, occurrence, multiple parents-children relationships. We present preliminary results in the evaluation of GO-Chart on analysis and visulization of Affymetrix® microarray data sets with variable *p*-values.

Keywords: DNA microarrays, gene ontology (GO), variable *p*-values, Affymetrix® microarray data.

1. Introduction

DNA microarrays typically consist of array of specific sequences, called probes, spotted or synthesized on a substrate that is usually rigid. Microarray technology is a powerful technology capable of detecting expression of many thousands of genes simultaneously. Using microarrays, biologists can study the differential expression of genes over time, between tissues and disease states; identify genes in complex diseases; discover new drugs and perform toxicology studies; detect gene mutation/polymorphism; and carry out pathogen analysis [1]. The amount of data produced by each experiment is overwhelmingly large. Meanwhile, though a large number of probe sets may be identified as differentially regulated, it is difficult to determine the importance of these changes. Thus the analysis of microarray data presents significant hurdles to the researcher

Early gene-expression experiments produced data for a limited set of genes based on prior hypotheses and the data were therefore relatively easily interpreted. The advance presented by assaying the expression all or most genes in a single experiment lies in the lack of bias in selection of genes, *i.e.* no prior hypothesis. Lacking a prior hypothesis, the data need to be analyzed for a pattern that can become the foundation for a hypothesis with which to interpret the biological meaning of the results.

The description of gene functions in terms of biological properties such as function and type is problematic. Until recently, there has been no common and controlled vocabulary built that described biological processes and properties, because every researcher described process in his own words. Not only did no common vocabulary exist, the problem was further complicated by a lack of formal relationship between individual terms in the vocabularies. A formal, agreed upon, ontology solves the problem [2, 3]. The ontology imposes a structure with formal rules that can be interpreted and used by computers. The Gene Ontology (GO) Consortium [1, 4] is establishing the necessary common vocabulary and hierarchical structure to describe gene functions.

GO has become increasingly important because of wide variety of biological information through microarrays and other high-throughput technologies. The tremendous amounts of biological information need to be described and classified in meaningful ways, which forms GO. GO is beginning to produce a common, structured, precisely defined, common, controlled vocabulary for describing the roles of genes and gene products in different species. The Gene Ontology Consortium has three extensive ontologies under development [5]: (a) Molecular function, which describes what a gene product does at the biochemical level. (b) Biological process, which describes a broad biological objective. (c) Cellular component, which describes the location of a gene product, within cellular structure and macromolecular complexes.

A number of web-based or stand-alone software tools have been and continue to be developed to perform gene ontology analyses and identify functional categories among differentially expressed genes. A common problem is that each tool lacks some desirable feature or features that are present in other tools. Our experience with EASE (Expression Analysis Systematic Explorer) [6], DAVID (Database for Annotation, Visualization and Integrated Discovery) [7, 8], Onto-Express [9], GOFigure [10] and GOTM (GOTree Machine) [11] shows that each of the tool performs functional profiling of genes from Affymetrix® probe sets, but there is a need of flexibility in desired the significant of enrichment (*p*-value) since researchers desire to choose their own threshold for visualization (bar chart) of the gene ontology data. GOTM allows Probe set IDs and visualizes bar chart based on given level number in GO hierarchy (depth) information from user at non-adjustable default *p*-values of less than 0.01 (multiple comparisons not allowed).

We have developed a stand-alone tool, "GO-Chart," mainly for visualizing more conveniently the GOTM bar chart output while adjusting the *p*-values for multiple testing. GO-Chart allows users to analyze genes by Gene Ontology (GO) functional categories, occurrence, multiple parents-children relationships. Meanwhile, GO-Chart also cleans, matches with Gene Ontology for validation, stores and retrieves gene information obtained from GOTM effectively for multiple datasets without recalculation of the data.

In this paper, we will cover the design of GO-Chart and some preliminary results in the evaluation of GO-Chart on analysis and visualization of Affymetrix® microarray data sets with variable *p*-values in the context of GO.

2. GO-Chart Design

A schematic overview of the GO-Chart is shown in Fig. 1. The central database, Go-Chart DB, is at the heart of GO-Chart design. Multiple relational tables are constructed to store the following information:

(1) Updated GO files which are downloaded from Gene Ontology Consortium website [4].
(2) Functional profiled GO data from GOTM output, which are cleaned using Perl parser.
(3) Derived hierarchical relationship between gene and GO categories.
(4) Experimental information of each dataset.

Users query the GO-Chart DB using forms, which work with tables stored in the database to retrieve desired information. The query results can be shown graphically in the form of bar chart. In Fig. 2, the user is asked to choose desired data set to analyze and draw the bar chart for the same.

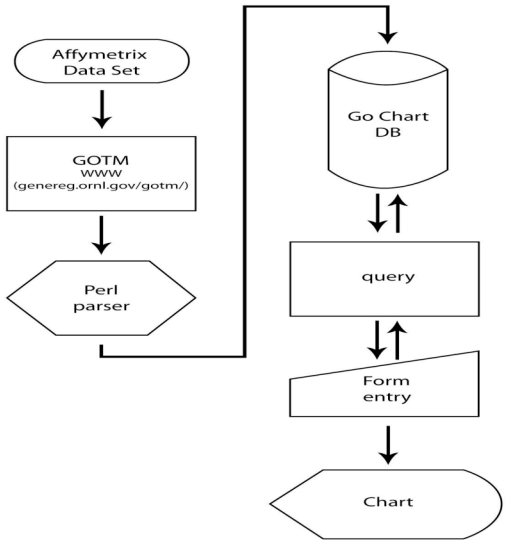

Fig. 1: Schematic view of the Go-Chart design.

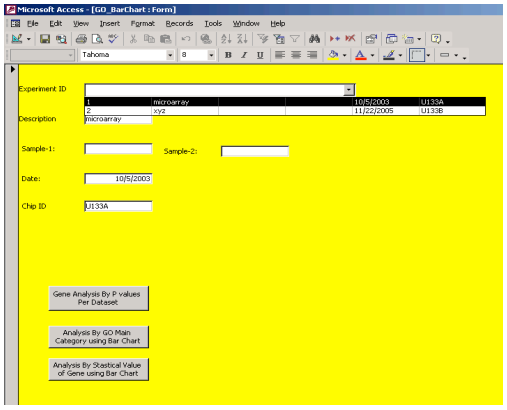

Fig. 2: GO Bar Chart provides user to select experiment (dataset) from dropdown list.

3. *P*-Value Adjustment

Identifying GO categories for biological process, cellular component, or molecular function with significantly enriched gene numbers (the *p*-value) in the interesting gene set compared to a reference gene set will allow the user to focus on biological areas that are most important for the interesting gene set [6]. Once we have *p*-value from GOTM data, it must be adjusted in order to perform meaningful comparisons among multiple data sets. The Sidak correction [12] is used to adjust the *p*-value per experiment using the following equation:

$$P_{adj} = 1 - (1 - P_{unadj})^n$$

in which P_{adj} is the adjusted *p*- value that is corrected for experiment-wide significance, P_{unadj} is the original unadjusted point-specific *p*- value, and n is the number of unique comparisons performed in the experiment.

4. Evaluation of GO-Chart

There are three options available after choosing desired experiment. The first option is to get information about all terms at specified *p*-value, which can establish the number of terms occurred that particular *p*-value. An example is shown in Fig. 3. Information listed includes date, chip ID, GO main category, depth, gene name, number of observed occurrence, expected occurrence, unadjusted and adjusted *p*-values.

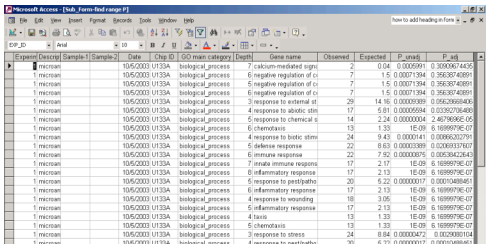

Fig. 3: Query results showing genes (GO information) within a range of *p*-values.

The second option is to analyze GO main categories to establish the number of observed and expected terms occur in that particular experiment for multiple testing. An example is shown in Fig. 4 in which we see the experimentally observed numbers of genes are much larger than the expected ones.

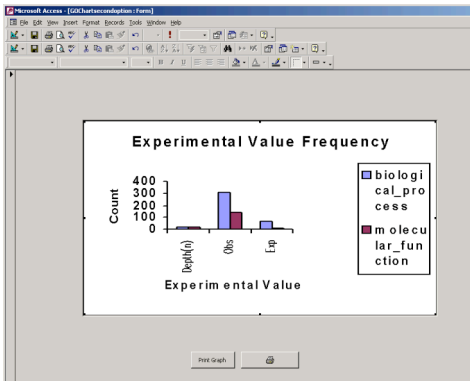

Fig. 4: The bar chart for GO main category analysis.

The third option is to analyze each term with GO main category by multiple testing with different *p*-values and visualize the results with nice graphics. When researchers present their interpretation of the biological significance of their data, they should focus on how the bar chart visualization is affected by changing the threshold for significance of enrichment, the *p*-value. Fig. 5 shows the bar chart for GO depth 4 with *p*-values less than 0.01. Fig. 6 shows the same as in Fig. 5 except the p values are now adjusted for multiple testing. Figs. 5 and 6 illustrate the effect of correction for multiple testing for a data set. Fig. 7 shows the bar chart for GO depth 5 with *p*-values less than 0.0001. Fig. 8 shows the same as in Fig. 7 except the p values are now adjusted for multiple testing. Fig. 7 and Fig. 8 illustrate the effect of selection of α, the appropriate choice of which is a balance between the presence of false-positives and false-negatives, for the same data set as in Figs. 5 and 6. To some degree it is the researcher's responsibility to decide on the threshold of α, *i.e.,* how many false positives the researcher is willing to pursue in order to ensure few false negatives. It will be very useful for presentation and publication to focus on appropriate biological relevance of the data.

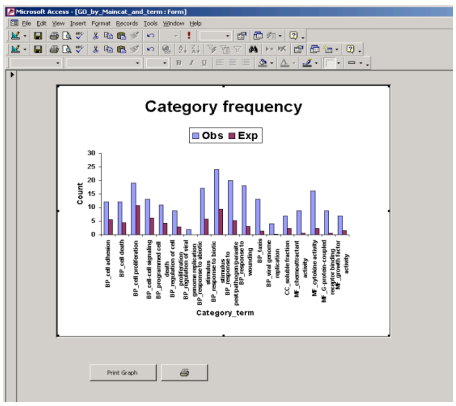

Fig. 5: The bar chart for GO depth 4 and P_{unadj} <0.01.

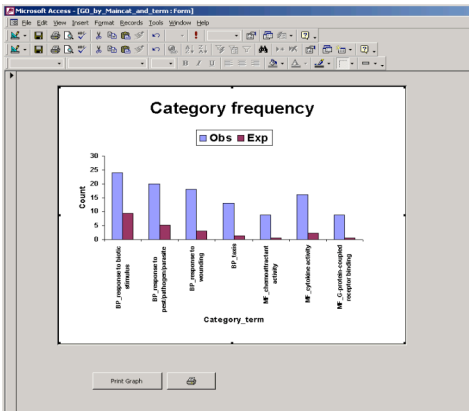

Fig. 6: The bar chart for GO depth 4 and P_{adj} <0.01 for multiple testing.

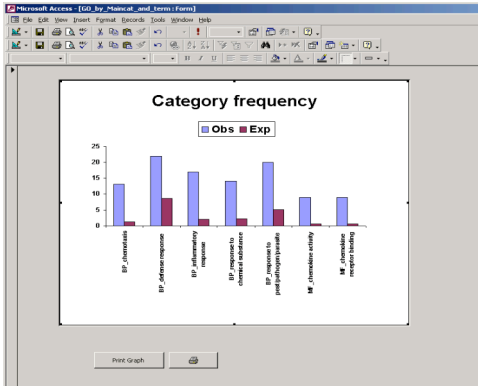

Fig. 7: The bar chart for GO depth 5 and P_{unadj} <0.0001.

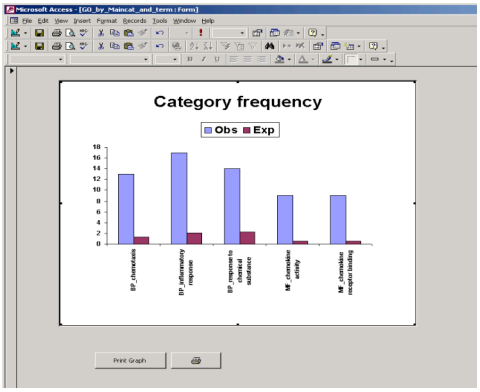

Fig. 8: The bar chart for GO depth 5 and P_{adj} <0.0001 for multiple testing.

5. Discussions and Future Work

There is significant difference in bar chart visualization after correcting *p*-value (P_{adj}) for multiple comparisons. Such significant difference implies the importance of performing *p*-value adjustment in mining useful gene information from multiple comparisons.

The advantages of the GO-Chart toolkit are obvious. First, the GO-Chart is well organized and easy to use. Second, it can be easily extended due to its modular design. Third, the entire project is done with commodity computer hardware and software technology, which is cheaply and readily available. The benefits of using GO-Chart for research are as follows. First, capital cost computing equipment and software is low. Second, it gives user a lot of flexibility in finding patterns of interesting genes and their biological relevance. Third, it provides powerful way to ensure correct statistical analysis by correcting *p*-values for multiple comparisons. Fourth, it provides solutions to many questions the biologists are interested to ask about Affymetrix® microarray data sets. Fifth, the GO-Chart database provides convenient access to GOTM data that are cleaned and synchronized with the curate GO database. Sixth, it reduces significantly the need for manual searching for gene information, retrieves information of an interesting gene, and provides multiple bar charts, while the user can specify the desired experiment. Last, GO-Chart is not restricted to the human species.

There are problems associated with GO-Chart, too. GOTM is very slow with large sets of data and is updated at a different cycle from the GO database, which may result in many unmatched GO terms with Gene Ontology database. This is a fundamental problem for bioinformatics in general. It is indeed extremely difficult to synchronize disparate data sources. Because GO-Chart make use of GOTM's output, the performance problem of GOTM will directly cause performance problems in GO-Chart.

Our future work includes building a Web interface and allowing the use of programmatic queries for faster tasks. Another extension will be to enable the research community to take advantage of the flexibility of user input such as the interesting GO main category and depth to find all gene associated with interesting GO category at specified depth.

References

[1] M. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, "The Gene Ontology (GO) database and informatics resource," *Nucleic Acids Res.* 32: D258-261, 2004.

[2] J. Bard, "Ontologies: Formalising biological knowledge for bioinformatics," *Bioessays*, 25(5): 501-506, 2003.

[3] J. Bard and S. Rhee, "Ontologies in biology: design, applications and future challenges," *Nat Rev Genet* 5: 213-222, 2004.

[4] M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, J. Eppig, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet* 25: 25-29, 2000.

[5] Gene Ontology Consortium, "The gene ontology resource: design and implementation," *Genome Res*. 11(8): 1425-1433, 2001.

[6] http://david.niaid.nih.gov/david/ease.htm

[7] http://david.niaid.nih.gov/david/version2/

[8] G. Dennis Jr, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, R. A. Lempicki, "DAVID: Database for Annotation, Visualization, and Integrated Discovery," *Genome Biol* 4(5): P3, 2003.

[9] P. Khatri, P. Bhavsar, G. Bawa and S. Draghici, "Onto-Tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments," *Nucleic Acids Res*. 32: W449-W456, 2004.

[10] S. Khan, G. Situ, K. Decker, C. Schmidt, "GoFigure: automated Gene Ontology annotation," *Bioinformatics* 19(18): 2484-2485, 2003.

[11] B. Zhang, D. Schmoyer, S. Kirovand, J. Snoddy, "GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies," *BMC Bioinformatics* 5(1): 16, 2004.

[12] Z. Sidak, "Rectangular confidence regions for the means of multivariate normal distribution," *J Am Stat Assoc* 62: 626 – 633, 1967.