MOLECULAR & CELLULAR TOXICOLOGY, Vol. 1, No. 4, 281-283, December 2005

BINGO: Biological Interpretation Through Statistically and Graph-theoretically Navigating Gene Ontology™

Sung Geun Lee[1], Jae Seong Yang[2], Il Kyung Chung[3] & Yang Seok Kim[1]

[1]Bioinformatics Division, ISTECH Inc., #506, Woongshin Art Plaza, 847 Janghang2-dong, Ilsan-gu, Goyang-si, Gyeonggi-do, 411-837, Korea
[2]Department of Life Science, Pohang University of Science and Technology, Pohang, Korea
[3]Department of Plant Genetic Engineering, Catholic University of Daegu, Gyeonsan 712-702, Korea
Correspondence and requests for materials should be addressed to Y.-S. Kim (yskim@istech21.com)

Accepted 23 November 2005

Abstract

Extraction of biologically meaningful data and their validation are very important for toxicogenomics study because it deals with huge amount of heterogeneous data. BINGO is an annotation mining tool for biological interpretation of gene groups. Several statistical modeling approaches using Gene Ontology (GO) have been employed in many programs for that purpose. The statistical methodologies are useful in investigating the most significant GO attributes in a gene group, but the coherence of the resultant GO attributes over the entire group is rarely assessed. BINGO complements the statistical methods with graph-theoretic measures using the GO directed acyclic graph (DAG) structure. In addition, BINGO visualizes the consistency of a gene group more intuitively with a group-based GO subgraph. The input group can be any interesting list of genes or gene products regardless of its generation process if the group is built under a functional congruency hypothesis such as gene clusters from DNA microarray analysis.

Keywords: Bioinformatics, Toxicoinformatics, Data mining

The high-throughput technologies in the post-genomic era have generated unprecedented amount of data and novel hypotheses about complex biologi-cal phenomena previously uncharacterized. Various data mining techniques such as clustering and classification have contributed to the pattern discovery of the massive data. The results are often some aggregates of genes or gene products, but their biologically consistent contexts are difficult to capture as the size becomes large. The underlying biological meanings need to be mined considering the interactive functional relationship between genes. This has been an interesting subject of current bioinformatics applications[6,9] and can be efficiently accomplished by use of ontologies. Ontologies are systematic rules and vocabularies that can be utilized effectively for annotating genes or gene products[2]. The relationship between objects (terms) defined in a bio-ontology is useful for assigning characteristic biological terms to a gene group of interest. Currently Gene Ontology (GO) is widely used for that purpose due to its lucid and systematic properties[12].

Recent approaches to extracting or assessing the enriched GO terms from a gene group include statistical methodologies[1,3,13] and graph-theoretic modeling[11]. The former methods emphasize the statistical significance of extracted GO terms, mostly evaluated by p-values, and the latter deals with the geometric configurations of a gene group on GO space. Despite the benefits of p-value approach, the excessive dependence on the *exact* p-value should be avoided since it may be biased rather from current *incomplete* annotation data than from theoretical procedures. Owing to the current incomplete knowledge of gene functions, i.e. the functional information of many genes are insufficient and unknown, the certainty of p-value should be carefully managed even for such statistical improvements as false discovery rate (FDR) and p-value adjustments for multiple comparisons. This is the point where graph-theoretic measure and well-designed visualization can be valuable as complementary information of statistical significance.

Results

BINGO has three distinct features that provide statistical and graph-structural evidences on functionally coherent groupings.

P-value Computation

This process is to extract significant GO terms from a gene group of interest. Hypergeometric distribution is adopted to compute p-values. In BINGO, p-value is

used for ranking candidate GO terms as well as approximating the significance of them. That is, p-value is primarily involved in the role of ranking GO terms. The highly ranked GO terms with the smallest p-values are selected as candidate characteristic GO terms.

Path Congruency and Proximity

These are simple and effective graph-theoretic indexes for measuring the consistency among group elements. These indexes are computed for each candidate GO term selected by p-value approach. After all the genes in an input group are mapped onto a GO DAG via their corresponding GO annotations, the graph-theoretic properties among them are examined. Path congruency is the average length of common paths between the root-input gene GO path and the root-candidate GO term GO path. Proximity is (path congruency) / (GO level of candidate GO term). The greater the path congruency and the closer to 1 the proximity, the more coherent the biological context of a gene group of interest.

Group-based GO Subgraph Structur

Visualization is useful for intuitive understanding of mathematical results. This module visualizes functional group dispersion in the form of GO subgraph. Only the paths between the root and the vertices (GO terms) with which a gene in an input group is annotated, are depicted from the entire GO DAG

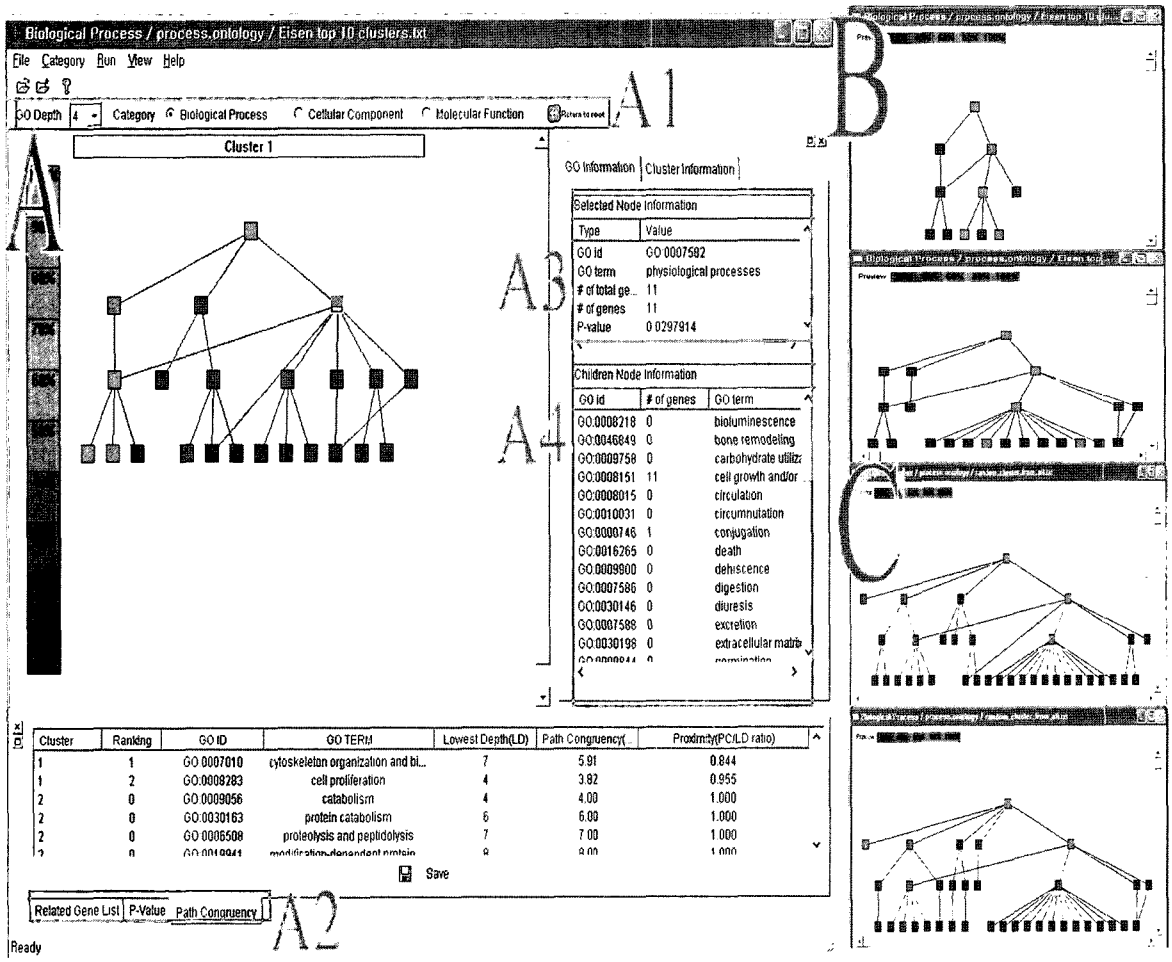

Fig. 1. The main interface of BINGO (A): entire GO DAG or its subgraph is displayed for each gene group/cluster. The color-bar indicates the extent of gene involvement. One of the three GO categories can be chosen and GO depth to be displayed is determined (A1). The computational results (p-value, path congruency) and related gene lists can be easily seen in the tab button (A2). For a node selected, all the related information is shown (A3, A4). The group-based GO subgraph structures are shown for Eisen clusters (B: two pictures above) and random groups of yeast genes (C: two pictures below). The Eisen clusters have 27, 138 GO-annotated genes and the random groups 50, 44, respectively. The former groups. exhibit tighter structure than the latter ones.

and form a subgraph of GO DAG. If the graph structure is tight, the gene group is assumed to be functionally well-clustered; if it has a scattered pattern, the grouping seems rather bad in the sense of functional commonality (Fig. 1).

To illustrate the effectiveness of our approach, we applied BINGO to the gene clusters obtained from clustering DNA microarray data of yeast and human[8,10]. The biological implications of both results concurred with the descriptions about the gene clusters in the original articles. As expected, the p-values of the characteristic GO terms of Eisen clusters are much smaller than those of the randomly generated gene clusters that are considered functionally scattered. The proximity values of Eisen clusters are mostly around 0.8 with great path congruency values; On the other hand, random gene groups show low proximity or small path congruency. The results of human data are not as impressive as those of yeast, nonetheless some gene clusters showed favorable outcomes (see Supplementary data). This discrepancy may be attributed to the quantity and quality of the annotation information of the two species.

BINGO requires three kinds of input files; gene group file, gene ontology files, and GO annotation file. The input groups can be any interesting lists of genes or gene products, but BINGO is more appropriate to apply to the groups generated under the hypothesis of functional congruency such as gene clusters from DNA microarray data analysis. To facilitate the study associated with DNA microarray analysis, BINGO automatically recognizes the result format file of K-means and hierarchical clustering included in the widely used software 'Cluster'[8]. Gene ontology files can be downloaded from http://www.geneontology.org/ and need not be pre-processed separately to submit into BINGO that interprets the format as it is. GO annotation files can be obtained from species-specific public databases[4,7] or GO-specialized integrated databases[5]. Although all-genes annotation set is a usual superset for p-value computation, a superset restriction may be useful in some cases and this is easily done just by replacing GO annotation files. The format of these input files is described in detail in the manual of BINGO.

Discussion

BINGO is a stand-alone GUI program implemented in C++ and hence users can easily operate it under Windows environment with fast computation. BINGO combines statistical methods with graph-theoretic measures for reliable biological interpretation of gene groups, visualized in an intuitive and informative graphical setting. This integrative approach will be more beneficial in the situation that the functions of all genes are not revealed. The multiple hypothesis test issue is not addressed in this version since we prefer ranking by p-values, but it will be incorporated in the coming versions.

References

1. Al-Shahrour, F., Diaz-Uriarte, R. & Dopazo, J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578-580 (2004).
2. Bard, J. Ontologies: Formalizing biological knowledge for bioinformatics. *Bioessays*, **25**, 501-506 (2003).
3. Berriz, G.F., King, O.D., Bryant, B., Sander, C. & Roth, F.P. Characterizing gene sets with FuncAssociate. *Bioinformatics*, **19**, 2502-2504 (2003).
4. Bult, C.J. *et al.* The Mouse Genome Database (MGD): integrating biology with the genome. *Nucleic Acids Res.*, **32**, D476-D481 (2004).
5. Camon, E. *et al.* The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.* **32**, D262-D266 (2004).
6. Doniger, S.W. *et al.* MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol.*, **4**, R7 (2003).
7. Dwight, S.S. *et al.* Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.*, **30**, 69-72 (2002).
8. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863-14868 (1998).
9. Hosack, D.A., Dennis Jr, G., Sherman, B.T., Lane, H.C. & Lempicki, R.A. Identifying biological themes within lists of genes with EASE. *Genome Biol.*, **4**, R70 (2003).
10. Iyer, V.R. *et al.* The transcriptional program in the response of human fibroblasts to serum. *Science*, **283**, 83-87. (1999).
11. Lee, S.G., Hur, J.U. & Kim,Y.S. A graph-theoretic modeling on GO space for biological interpretation of gene clusters. *Bioinformatics*, **20**, 381-388 (2004).
12. The Gene Ontology Consortium Creating the Gene Ontology resource: design and implementation. *Genome Res.*, **11**, 1425-1433 (2001).
13. Zeeberg, B.R. *et al.* GoMiner : a resource for biological interpretation of genomic and proteomic data. *Genome Biol.*, **4**, R28 (2003).