BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 7 2008, pages 1026–1028
doi:10.1093/bioinformatics/btn068

Databases and ontologies

GOTreePlus: an interactive gene ontology browser

Bongshin Lee[1], Kristy Brown[2], Yetrib Hathout[2] and Jinwook Seo[2,*]

[1]Microsoft Research, One Microsoft Way, Redmond, WA 98052 and [2]Children's National Medical Center, 111 Michigan Ave, NW, Washington, DC 20010, USA

Received on December 13, 2007; revised on January 13, 2008; accepted on February 17, 2008

Advance Access publication February 21, 2008

Associate Editor: John Quackenbush

ABSTRACT

Summary: We developed an interactive gene ontology (GO) browser named *GOTreePlus* that superimposes annotation information over GO structures. It can facilitate the identification of important GO terms through interactive visualization of them in the GO structure. The interactive pie chart summarizing an annotation distribution for a selected GO term provides users with a succinct context-sensitive overview of their experimental results. We tested our *GOTreePlus* using a proteome profiling dataset obtained on differentiation of retinal pigment epithelial cells where 399 proteins were quantified.

Availability: http://bioinformatics.cnmcresearch.org/GOTreePlus/

Contact: jseo@cnmcresearch.org

1 INTRODUCTION

Hypothesis generation and testing in biology these days involve informatics tasks due to the heavy volume of data generated by cutting-edge techniques. Microarray techniques increased the data resolution to merely over tens of thousands of features on a chip. Recent single nucleotide polymorphism chips push the limit even further to one million features per chip. Mass spectrometry data in the proteomics field also stretches the limit. As datasets become larger, it becomes more challenging to extract global information on the underlying biochemical pathways and biological processes. To deal with such a large dataset, it is essential to aggregate or summarize the initial dataset in a universal language so that future testing could focus on more relevant parts of data to the aims of the project. This has led biomedical researchers to project the dataset over the gene ontology (GO) to reveal overall meaning of their data and to set a more focused hypothesis.

GO enrichment tools such as GOMiner (Zeeberg *et al*., 2003) and Database for annotation, visualization, and integrated discovery (DAVID; Dennis *et al*., 2003) systematically sort the massive amount of GO data in a more meaningful and focused format based on various enrichment algorithms. However, it is still challenging to effectively explore the enrichment results over the GO structure. It is due to the size and complexity of GO data structure and the lack of visualization tools to efficiently browse and search the structure with experimental data combined.

*To whom correspondence should be addressed.

GO contains ontology terms organized as a Directed Acyclic Graph (DAG), which is more complex than a tree structure because of the cross links in it. There are various GO browsers to help interpret microarray and proteomics data using the GO structure (see www.geneontology.org/GO.tools.browsers.shtml). Most of them are text-based tools showing the structure using a simple tree control that has '+' or '−' sign in front of each term and indents terms to show different levels. There are some graphical browsers, but they still do not support effective user interactions. The lack of effective navigation in these tools ignited the use of tools equipped with interactive visualization techniques. Baehrecke *et al*. (2004) incorporated a famous 2D space filling visualization called 'Treemap' to make an intuitive graphical overview of the raw dataset. Treemap improved the way biologists interpret their dataset. However, since it is not trivial in Treemap to show the hierarchical structure, users often struggle to grasp the underlying structure of the GO.

A different approach can be used to show expression/proteome profiling data with the GO annotation in a generalized Venn diagram (Kestler *et al*., 2005). Each GO term is represented as a circle whose size is proportional to the number of proteins/genes mapped to that term. It shows intersections between GO terms as well as a graphical overview of the GO association with the input dataset. But again the hierarchical structure of the GO was not incorporated in this visualization.

To address these problems, we developed an interactive GO browser called GOTreePlus (Fig. 1) by improving TreePlus (Lee *et al*., 2006). It visualizes the GO DAG as a tree and provides interactive zoom and pan. GOTreePlus maps proteome profiling data to the GO terms and visualizes them over the GO structure. It also provides succinct context-sensitive overviews of annotation distribution over children nodes of a selected node in a pie chart (Fig. 2). We believe GOTreePlus could help users better understand their genomic or proteomic datasets.

2 FEATURES AND FUNCTIONALITIES

GOTreePlus (Fig. 1) consists of the GO terms list, the proteins/genes list and the TreePlus control. When users open a file containing a list of proteins/genes and an annotation file from the GO annotations download page (www.geneontology.org/GO.current.annotations.shtml), the number of annotations for each GO term is computed and shown in the GO structure using the TreePlus control. Each node representing a GO term

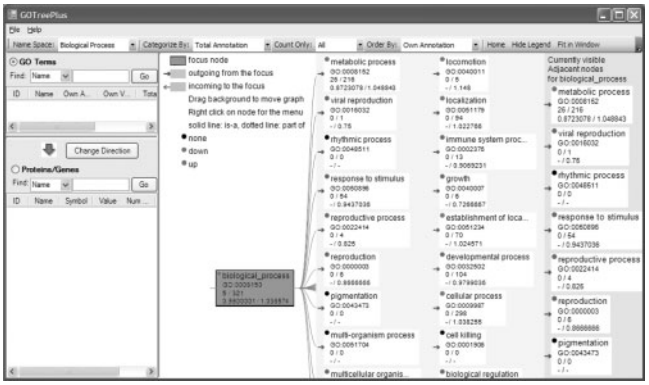
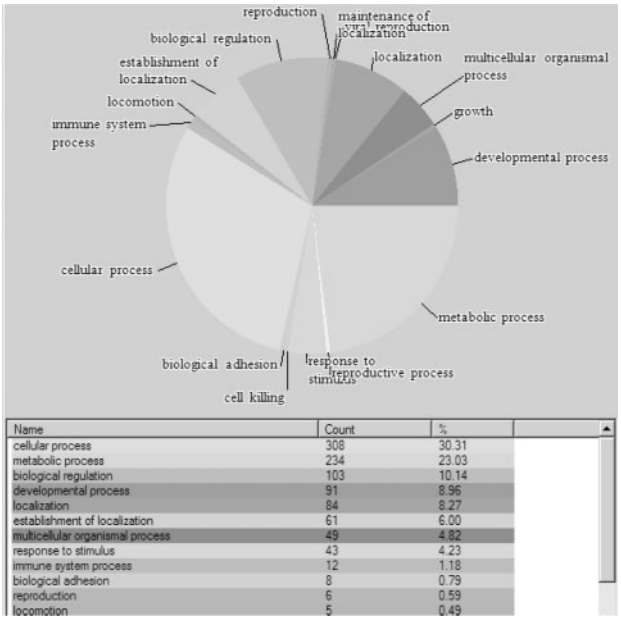


Fig. 1. GOTreePlus consists of two lists (GO terms list and Proteins/Genes list) on the *left* and the TreePlus control on the *right*. Online user manual with high-resolution figures are available at http://bioinformatics.cnmcresearch.org/GOTreePlus/.

has six attributes: name, ID, number of its own annotations, sum of the number of its descendents' and its own annotations, average value of the proteins mapped to this node and average value of the proteins mapped to it or its descendents. Since the nodes in the TreePlus control, by default, are sorted by the sum of the number of its own annotations, users can easily see which GO term is most relevant in their data. Each node has a colored dot that shows up- or down-regulation of the proteins/genes mapped to that node. For example, red color indicates up-regulation and green color indicates down-regulation (Fig. 1).

Since the ontologies are structured as DAGs, it is common for a node to have more than one parent. GOTreePlus can visualize multiple parent nodes in a node-link diagram with color-coded edges. As users click on a GO term node, all children nodes are shown as outgoing edge with blue arrows and all parent nodes are shown as incoming edge with red arrows. Any structural change required to show DAG as a tree is smoothly animated to help users follow the change. Unlike other GO browsers, GOTreePlus enables users to select any GO term and make it the root node to initiate a focused exploration from the node. Users can also select any node to see a localized overview of annotation distribution over its children nodes in a standard pie chart with a coordinated list view (Fig. 2).

GOTreePlus provides a way to search for a specific GO term—a simple substring match either by term or by ID. Search results are shown in the GO terms list. When users select a GO term from the list, the selected term is shown in the TreePlus control. Furthermore, with the 'GO Terms' radio button selected, the proteins/genes list is updated with the proteins/genes associated with the selected GO term. The number of proteins/genes in the proteins/genes list is also updated and shown by the 'Proteins/Genes' radio button. Similarly, users can also search proteins by name or by symbol. When users select a protein/gene from the list, all GO terms related to the selected protein/gene are shown in the GO structure in the TreePlus control. If the 'Proteins/Genes' radio button is selected, the GO terms list is updated with the GO terms associated with the selected protein/gene.



Name	Count	%
cellular process	308	30.31
metabolic process	234	23.03
biological regulation	103	10.14
developmental process	91	8.96
localization	84	8.27
establishment of localization	61	6.00
multicellular organismal process	49	4.82
response to stimulus	43	4.23
immune system process	12	1.18
biological adhesion	8	0.79
reproduction	6	0.59
locomotion	5	0.49

Fig. 2. Annotation distribution for 'biological process' node. By selecting 'Show annotation distribution' from the pop up menu, users can see the overall annotation distribution of all child nodes of the selected node using a pie chart.

In summary, GOTreePlus has the following distinctive features:

- Visualize GO DAG structure as a tree with smooth animations
- Explore GO with any selected GO node as a root node
- Provide a context-sensitive overview of annotation distribution of children nodes of an interactively selected GO node in a pie chart
- Search for a GO term in GO and its associated proteins
- Search for a protein in user dataset and its associated GO terms in GO

3 APPLICATIONS

We used proteome profiling data obtained on dividing versus confluent human retinal pigment epithelial (RPE) cells to show the utility of GOTreePlus. Proteome profiling of dividing versus resting human RPE cells was obtained using stable isotope labeling by amino acid in cell culture (SILAC) strategy in combination with shotgun proteome profiling (Hathout *et al.*, 2005). Labeled dividing cells were mixed at 1:1 ratio with unlabeled resting cells. Total cytosolic proteins were extracted and digested with trypsin, and the resulting peptides analyzed by 2D chromatography coupled to an liquid trap quadrupole ion trap mass spectrometer. A total of 399 proteins were identified and quantified in this study.

The bottleneck is how to look at this data and extract useful information based on the GO term and differential expression. One can rank proteins in a group as up-regulated and

down-regulated proteins and look at their function and subcellular localization one by one. However, this is time consuming and the overall underlying biological process may be overlooked. By mapping this dataset to GO using GOTreePlus, we could easily follow the overall function and biological process underlying the global proteome profiling data obtained on dividing versus resting RPE cells. Most of the up-regulated proteins in dividing cells were found directly involved in cell cycle, synthesis and biogenesis of cell components, while the down-regulated proteins in resting cells were involved in actin remodeling and stress management.

The ontology terms that we can extract from this application directly reflects the biological status of the system studied. One can select all the proteins that are up-regulated in dividing cells and check if they have common ontology terms. The exploratory nature of using the GO structure makes the interactive graph visualization methods in GOTreePlus most useful. Users can delve into the sublevels of a GO term with annotated protein information to get a deeper insight into their datasets.

ACKNOWLEDGEMENTS

This work was supported by NIH 5R24HD050846-02 Integrated molecular core for rehabilitation medicine, and NIH 1P30HD40677-01 (MRDDRC Genetics Core).

Conflict of Interest: none declared.

REFERENCES

Baehrecke,E.H. *et al.* (2004) Visualization and analysis of microarray and gene ontology data with treemaps. *BMC Bioinformatics*, **5**, 84.

Dennis,G. *et al.* (2003) DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol.*, **4**, R60.

Hathout,Y. *et al.* (2005) Metabolic labeling of human primary retinal pigment epithelial cells for accurate comparative pro-teomics. *J. Proteome Res.*, **4**, 620–627.

Kestler,H.A. *et al.* (2005) Generalized venn diagrams: a new method of visualizing complex genetic set relations. *Bioinformatics*, **21**, 1592–1595.

Lee,B. *et al.* (2006) TreePlus: interactive exploration of networks with enhanced tree layouts. *IEEE TVCG*, **12**, 1414–1426.

Zeeberg,B.R. *et al.* (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.*, **4**, R28.