GOfox: Semantics-based simplified hierarchical classification and interactive visualization to support GO enrichment analysis

Edison Ong, Yongqun He
University of Michigan, Ann Arbor, Michigan, USA

ABSTRACT

Gene Ontology (GO)-based statistical enrichment analysis is a popular approach to identify statistically enriched biological processes, molecular functions, and cellular components that are associated with a list of genes. However, such GO enrichment analysis often generates a large number of enriched GO terms that are difficult to interpret and analyze. To address this issue, we developed GOfox, a web tool that utilizes OWL-based ontology semantics and RDF triple store SPARQL queries to generate full or simplified hierarchical GO subsets to classify and display enriched GO terms. GOfox integrates and extends features from OntoFox and Ontobee, two ontology tools developed in the laboratory. GOFox also includes a newly developed algorithm for generating simplified hierarchical classification by considering the multiple inheritance of GO. Furthermore, GOfox provides an interactive visualization that supports GO subset tree exploration and term editing. GOfox is freely available at the website: http://gofox.hegroup.org/.

1 INTRODUCTION

A biological/biomedical ontology is a set of computer and human-interpretable terms and relations that represents entities in a biological/biomedical domain and how they relate to each other. Hundreds of biological ontologies have been developed. The most widely used biological ontology is the Gene Ontology (GO), which systematically and semantically represents three major attributed associated with gene products: Biological Processes (BP), Molecular Function (MF), and Cellular Components (CC) (Ashburner et al., 2000). One major GO application is GO-based statistical enrichment analyses. The rationale of such an enrichment analysis is that given a group of genes, the co-functioning genes should have a higher or enriched potential to be identified as a relevant group using high throughput technologies (*e.g.*, microarrays and RNA-Seq). Since often hundreds (or even more) of enriched terms are detected, the linear output of enriched terms can be very large and overwhelming, resulting in diluted focus on the analysis of related terms.

To address the ever increasing number of enriched GO terms resulting from high throughput studies, we developed GOfox to support GO enrichment analysis through integrating and extending the features of OntoFox (Xiang et al., 2010) and Ontobee (Xiang et al., 2011). OntoFox is able to fetch ontology terms and axioms. OntoFox includes several semantics algorithms for extracting different levels of intermediate layer terms between user-selected terms and a top level term of the ontology (Xiang et al., 2010). Ontobee is the default OBO ontology linked data server that facilitates ontology data sharing, visualization, query, integration, and analysis (Xiang et al., 2011). Ontobee also supports ontology visualization including the hierarchy, definition and annotations. By integrating and extending the features of OntoFox and Ontobee, GOfox is able to represent the enriched GO terms in an interactive hierarchical layout along with term-related information, and it allows users to manually modify the summarized enrichment result. Considering the multiple inheritance strategy used in GO development, GOfox developed a new algorithm to trim down the size of the enriched subset tree of GO. In addition, GOfox retrieves and displays related information such as definition, database cross references and comments, etc. of the selected GO term from Ontobee. This report provides the first time introduction of the GOfox to help researchers better visualize and analyze the results of GO gene enrichment studies.

2 GOFOX SYSTEM OVERALL DESIGN

The overall design and workflow is displayed in Fig. 1. Using a web form shown in Fig. 2, a user can input enriched or interested GO terms along with the p-values. Then the user can define a P-values cutoff (or another cutoff) and how intermediates are treated. After receiving the user's request, the GOfox server will extract a subset of GO that contains the input terms and related GO terms using PHP, Java and SPARQL. Specifically, the server queries against He Group's RDF triple store using SPARQL and retrieves a subset of GO. The query results will be in RDF/XML format and will be reformatted to the OWL format using OWL API (http://owlapi.sourceforge.net/). Then, based on the user's preference, GOfox will run simplification algorithm and generate results for downloading, visualization, and editing (Fig. 1). The results will be temporarily stored in He group RDF triple store and destroyed in a regular basis.

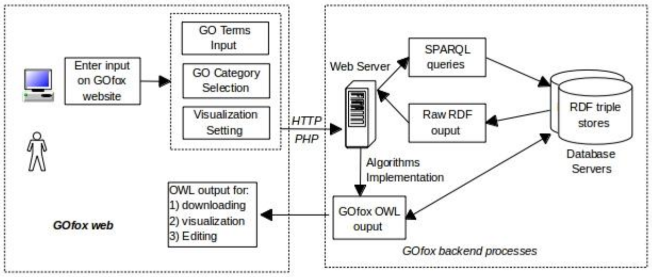


Fig. 1. GOfox program architecture and workflow design.

[*] To whom correspondence should be addressed: yongqunhe@umich.edu

3 GOFOX NEW ALGORITHM (SIM) FOR PROCESSING ENRICHED GO TERMS

The new GOfox algorithm “Include Computed Simplified Intermediates” (SIM) is developed on the basis of OntoFox “Include Computed Intermediates” (COM). The COM basically removes all intermediate GO terms that match the following rules: 1) the intermediate GO term is not included in the user’s input; 2) the intermediate GO term has only one parent and one children GO term (Xiang et al., 2010). Although COM works well for most ontologies, it often does not generate ideal results for ontologies (e.g., GO) that have multiple inheritance. SIM is developed to resolve this issue.

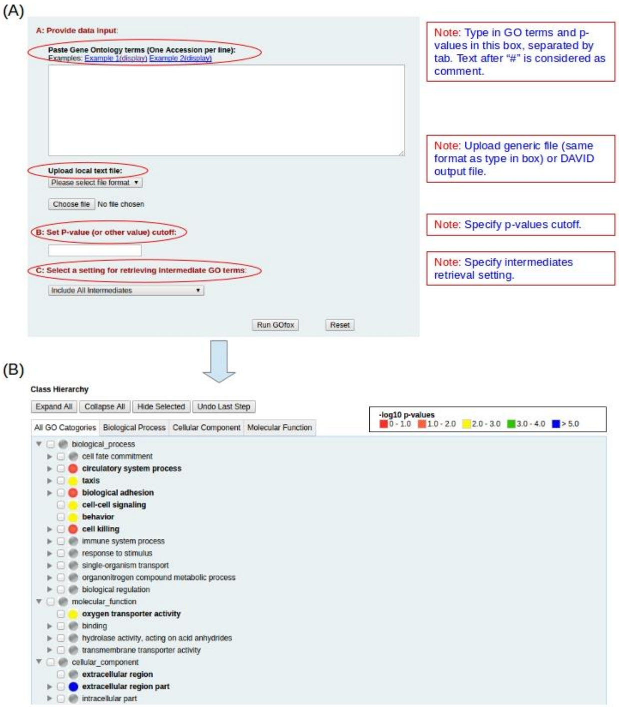
SIM first goes through the COM steps, and the COM results are further simplified by selectively removing some intermediate terms that have multiple parents (e.g., multiple inheritance) based on the following 3 steps. First of all, SIM reformats the OWL-formatted results by removing indirect subclass relationships. For example, the subclass axiom: *(regulates) some (transcription, DNA-templated)* will be removed because the parent-children relationship is not a direct *‘is a’* relationship. Second, SIM removes intermediate GO terms that match the following rules: 1) the intermediate GO term is not included in the user’s input; 2) if the intermediate GO term has less than two child GO terms within the user’s input list (*Note*: here we do not consider one parent condition as COM does). Third, SIM will further trim down the list by removing the subclass relationships between the GO terms and three GO top level terms of BP, CC, and MF. The requirements of the removal are: 1) the term is a direct subclass of BP, CC or MF; 2) there exists another direct subclass relationship between the GO terms and terms other than the three GO top level terms.

While GOfox still keeps the COM algorithm for users to choose, the SIM algorithm provides an extra way of shortening the GO terms in display.

4 GOFOX FEATURES AND WEB INTERFACE

GO provides many features for generating hierarchical classification given a list of user-provided enriched GO terms. Fig. 2 provides a demo on how GOfox works. Specifically, a user can choose to type in GO terms or upload a text file as input. The user can provide a standard P-value or other P-values such as false discovery rate adjusted P-value. A different value cutoff can also be used. The user can then select an intermediates retrieval setting, including COM, SIM, or all intermediates. GOfox will run after “Run GOfox” is clicked (Fig. 2A).

After the results are generated, GOfox provides an Ontobee-like term visualization interface (Fig. 2B). This feature is good for biologists who are not familiar with using the Protégé OWL editor to display output files. The user can interactively explore the hierarchy of retrieved GO terms and also hide unwanted GO terms from the web page.



A: Provide data input:

Paste Gene Ontology terms (One Accession per line):

Upload local text file:

Please select file format ▾

No file chosen

B: Set P-value (or other value) cutoff:

C: Select a setting for retrieving intermediate GO terms:

Include All Intermediates ▾

Note: Type in GO terms and p-values in this box, separated by tab. Text after “#” is considered as comment.

Note: Upload generic file (same format as type in box) or DAVID output file.

Note: Specify p-values cutoff.

Note: Specify intermediates retrieval setting.

Class Hierarchy

-log10 p-values
🟥 0 - 1.0 🟧 1.0 - 2.0 🟨 2.0 - 3.0 🟩 3.0 - 4.0 🟦 > 5.0

- ▼ □ ● biological_process
  - ▶ □ ● cell fate commitment
  - ▶ □ ● **circulatory system process**
  - ▶ □ ● **taxis**
  - ▶ □ ● **biological adhesion**
  - □ ● **cell-cell signaling**
  - □ ● **behavior**
  - ▶ □ ● **cell killing**
  - ▶ □ ● immune system process
  - ▶ □ ● response to stimulus
  - ▶ □ ● single-organism transport
  - ▶ □ ● organonitrogen compound metabolic process
  - ▶ □ ● biological regulation
- ▼ □ ● molecular_function
  - □ ● **oxygen transporter activity**
  - ▶ □ ● binding
  - ▶ □ ● hydrolase activity, acting on acid anhydrides
  - ▶ □ ● transmembrane transporter activity
- ▼ □ ● cellular_component
  - □ ● **extracellular region**
  - ▶ □ ● **extracellular region part**
  - ▶ □ ● intracellular part

Fig. 2. GOfox website interface and output. (A) GOfox web-interface input form. (B) Standard GOfox SIM algorithm output.

5 AVAILABILITY AND LICENSE

GOfox is freely available on: http://gofox.hegroup.org/. With the license of Apache License 2.0, the source code is released on Github: https://github.com/ontoden/gofox.

6 SUMMARY

GOfox is a simplified hierarchical classification tool to help user interpret the results of GO enrichment analysis. GOfox addresses a critical issue. *i.e.*, the difficulty to visualize, select and further analyze the increased number of enriched GO terms from the popular GO enrichment analysis studies.

REFERENCES

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25**,** 25-29.

Xiang, Z., Courtot, M., Brinkman, R.R., Ruttenberg, A., and He, Y. (2010). OntoFox: web-based support for ontology reuse. *BMC Res Notes* 3:175**,** 1-12.

Xiang, Z., Mungall, C., Ruttenberg, A., and He, Y. (Year). "Ontobee: A linked data server and browser for ontology terms", in: *The 2nd International Conference on Biomedical Ontologies (ICBO)*: CEUR Workshop Proceedings), Pages 279-281 [http://ceur-ws.org/Vol-833/paper248.pdf].