BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 10 2011, pages 1444–1446
doi:10.1093/bioinformatics/btr142

Databases and ontologies

Advance Access publication March 17, 2011

GO-Module: functional synthesis and improved interpretation of Gene Ontology patterns

Xinan Yang[1], Jianrong Li[1], Younghee Lee[1] and Yves A. Lussier[1,2,*]

[1]Department of Medicine, Section of Genetic Medicine and Center for Biomedical Informatics and [2]Comprehensive Cancer Center, Ludwig Center for Metastasis Research, Computation Institute, Institute for Translational Medicine and Institute for Genomics and Systems Biology, the University of Chicago, Chicago, IL, USA

Associate Editor: Jonathan Wren

ABSTRACT

Summary: *GO-Module* is a web-accessible synthesis and visualization tool developed for end-user biologists to greatly simplify the interpretation of prioritized Gene Ontology (GO) terms. *GO-Module* radically reduces the complexity of raw GO results into compact biomodules in two distinct ways, by (i) constructing biomodules from significant GO terms based on hierarchical knowledge, and (ii) refining the GO terms in each biomodule to contain only true positive results. Altogether, the features (biomodules) of *GO-Module* outputs are better organized and on average four times smaller than the input GO terms list ($P = 0.0005$, $n = 16$).

Availability: http://lussierlab.org/GO-Module

Contact: ylussier@bsd.uchicago.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 6, 2010; revised on December 28, 2010; accepted on March 13, 2011

1 INTRODUCTION

Gene Ontology (GO; Ashburner *et al*., 2000) is a standardized representation of biological gene attributes organized in a directed acyclic graph structure, where nodes are GO terms and edges are the hierarchical relationships between them. Prioritized GO terms are routinely computed in functional analyses of genomic studies such as gene expression profiles and high-throughput sequencing. Enrichment and gene set analyses have been implemented to identify significant GO terms from large prioritized gene lists, such as *DAVID* (Dennis *et al*., 2003), *Onto-express* (Draghici *et al*., 2003a), *Global test* (Draghici *et al.*, 2003b), *GOstats* (Falcon and Gentleman, 2007), etc. However, when presented with the resulting extensive list of GO terms, biologists cannot readily synthesize their relationships and rigorously prioritize hypotheses.

Two approaches have generally been used for refining overabundant GO terms resulting from genomic studies. The first approach reduces the number of GO terms studied thus decreasing the total prioritized ones. For instance, *GOSlim* (Harris *et al.*, 2004) contains a small list of 127 key terms (2010 August version) curated a priori by expert biologists, while *GOPaD* dynamically calculates the most informative subset of GO terms from a prioritized gene list by disregarding other statistically significant GO terms (Alterovitz *et al*., 2007). These approaches are designed to eliminate a massive number of GO terms deemed as significant by statistics, thus severely limiting the expressiveness of the resulting list of GO terms. A second group of approaches removes falsely prioritized GO terms that are computational artifacts attributable to the inheritance of genes (signal) in the GO hierarchy occurring in enrichment studies (Rhee *et al*., 2008). For example, a bottom-up conditional hypergeometric test can conservatively remove inherited genes annotated to statistically significant descendants nodes before testing the parent GO nodes (Falcon and Gentleman, 2007; Grossmann *et al*., 2007). However, this test also eliminates true-positive GO nodes such as those whose child nodes are all significant. We developed an approach that only prunes a parent node if its inherited signal is equal or worse to that of its most statistically significant child (Lee *et al*., 2010). This approach retains key parent GO terms whose increased statistical significance is amplified by inheriting multiple children's signals indicating an overarching systems property, a key GO term, emerging from subsumed GO terms.

We hypothesized that grouping prioritized GO terms in biomodules could further improve their interpretation by biologists, and would thus allow for the synthesis and reduction of the overall number of biological features while explicitly encapsulating subsumed GO terms. This approach focuses on biomodules of GO terms rather than of individual genes. While prioritized gene lists have been shown to aggregate in biomodules associated to GO terms in co-expression (Wolfe *et al*., 2005) and protein interaction studies (Cho *et al*., 2007), software designed to group gene lists into gene biomodules using GO knowledge produce large numbers of GO terms for interpretation without any coherent organization.

2 GO BIOMODULE ONLINE TOOL: *GO-MODULE*

Here we described *GO-Module*, a user-friendly, parameter-free and web-accessible synthesis and visualization tool developed for end-user biologists to improve their analysis of prioritized GO terms by removing false-positive GO terms and organizing the remaining ones as GO biomodules. We previously described the algorithm as a hierarchical refinement of a prioritized list of GO terms (Lee *et al.*, 2010). Three characters are used to annotate the results where 'K' refers to the key terms of GO biomodules, 'T' refers to the truly significant hierarchical descendents of the key terms and 'F' refers to the false positive prioritized GO terms among the input. This

*To whom correspondence should be addressed.

© The Author 2011. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

algorithm follows the steps described below and is illustrated in more detail in Supplementary Fig. S1.

Inputs: GO-Module requires a table containing: (i) a list of prioritized GO IDs and (ii) their corresponding rankable qualifiers, such as *P*-values or simply ordinal numbers.

(1) Annotates a node as 'K' (key GO term) if every one of its children or parents is less prioritized than itself. To allow for straightforward interpretation of the key GO terms as 'local minima in a directed acyclic graph', these key terms are locally prioritized within contiguous relationships (e.g. parent and child).

(2) As long as contiguous descendents of a 'K' node are not themselves new 'K' nodes, *GO-Module* annotates them as 'T' (subsumed true-positive GO terms as a member of the biomodule).

(3) Assigns 'F' (false positive) to the remaining GO terms.

(4) Assigns a unique numerical label to each 'K' GO term and to all its 'T' descendants. Note that this assignment is a number that has no bearing on the rank and that a 'T' node may have more than one biomodule label.

Output: *GO-Module* provides three alternative output formats of identified terms and their relationships in biomodules: (i) an online table, (ii) a downloadable text file with five columns containing two input columns following by labels of significance, full names of the GO IDs and identified GO biomodules or (iii) a fixed graphic vector network visualization of resultant IDs or their terms (pdf format) preferably for lists of up to 500 GO terms.

3 IMPLEMENTATION

Example data: To demonstrate the efficacy of *GO-Module*, we examined prioritized GO lists reported in the literature (Table 1).

Results: *GO-Module* significantly reduces the features as measured by the ratio of 16 literature reported GO lists to their number of resultant biomodules (Table 1, *#K/#GO* ranging from 28 to 91%; $P = 0.0005$, $n = 16$, Wilcoxon's signed rank test compared with theoretical median of 100%). Specifically, we report that on average 28% of input terms were found to be false positives (F), and among the remaining K and T terms, 33% are linked together in *GO-Module* of two or more terms (data not shown). Relatively general terms were rejected in several of the independently published GO lists because their *P*-values were larger than that of their children terms. In two cases in Table 1, 'regulation of tumor necrosis factor biosynthetic process' (GO:0042534) was rejected, while its child term 'positive regulation of tumor necrosis factor biosynthetic process' (GO:0042535) was retained (Heinig *et al.*, 2010 in Fig. 1; Marcucci *et al.*, 2008), demonstrating that this feature reduction eliminated redundant GO terms, while retaining useful informative ones. Self-evident uninformative terms such as 'positive regulation of biological process' (GO:0048518) are also annotated as false positive by GO-Module in Figure 1.

Conclusively, by synthesizing significant GO outputs and constructing biomodules, *GO-Module* better facilitates the biological understanding and interpretation of genetic analyses. Additionally, *GO-Module* can refine GO signatures found using conventional methods without requiring that the user revisit any gene expression or GO databases. By requiring prioritized GO terms rather than gene lists as its input, *GO-Module* prunes and modularizes results of conventional enrichment studies (Table 1) while remaining remarkably simple for end-users such as biologists and translational scientists.

Table 1. *GO-Module* applied to lists of prioritized GO terms[a]

Retained features (#K/#GO) (%)	Input	Output			Journal (author year)
	#GO	**#K**	#T	*#F*	
53	15[b]	**8**	2	*5*	*Oncogene* (Yamaguchi *et al*., 2010)
63	24	**15**	4	*5*	*Nature* (Burke *et al.*, 2010)
28	162	**46**	63	*53*	*Nature* (Heinig *et al.*, 2010)
57	65[b]	**37**	8	*20*	*Cell* (Smukalla *et al.*, 2008)
41	189[b]	**76**	47	*64*	*Cell* (Smukalla *et al.*, 2008)
50	16	**8**	2	*6*	*N. Engl. J. Med.* (Marcucci *et al.*, 2008)
45	20	**9**	0	*11*	*Genome Res.* (Swanson-Wagner *et al*. 2010)
91	23	**21**	2	*0*	*Genome Res*. (Mortazavi *et al*. 2010)
86	35[b]	**30**	2	*3*	*Genome Res*. (Somel *et al*. 2010)
90	21[b]	**19**	1	*1*	*Genome Res*. (Vinuela *et al*. 2010)
70	30	**21**	2	*7*	*Genome Res*. (Atanur *et al*. 2010)
46	13	**6**	2	*5*	*Genome Res*. (Hoffman *et al*. 2010)
40	20[b]	**8**	6	*2*	*Genome Res*. (Hoffman *et al*. 2010)
44	9	**4**	3	*2*	*Genome Res*. (Hoffman *et al*. 2010)
50	10	**5**	1	*4*	*Genome Res*. (Hoffman *et al*. 2010)
50	10	**5**	0	*5*	*Genome Res*. (Liu *et al*. 2010)

[a]Twelve scientific papers (16 GO lists) were selected in the following way: the first five were prioritized by a Google search for 'significant GO terms' and each of four renowned journal names and the remaining seven were sequentially found in the journal *Genome Research* by manually searching from 12/2010 back to 5/2010 for GO lists containing more than eight prioritized terms.
[b]Only the reported terms with trackable official GO IDs using AmiGO (Ashburner *et al.*, 2000) version 1.7 were used as the input for *GO-Module*.

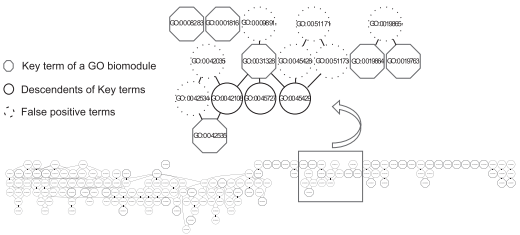

Fig. 1. Visualization of *GO-Module* output from the online portal. The key terms are seen as octagons, descendants of identified key terms as circles and false positive terms as dash lined circles. The contiguous hierarchical terms are linked together by lines. GO list from Heinig *et al*., 2010.

ACKNOWLEDGEMENTS

We thank Ellen Rebman and Yang Liu for their advice and review.

Funding: National Institute of Health [1U54CA121852 (MAGNET), CTSA UL1 RR024999-03, K22 LM008308-04], and the Cancer Research Foundation (T-AML).

Conflict of Interest: none declared.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

REFERENCES

Alterovitz,G. *et al.* (2007) GO PaD: the Gene Ontology partition database. *Nucleic Acids Res.*, **35**, D322–D327.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Atanur,S.S. *et al*. (2010) The genome sequence of the spontaneously hypertensive rat: analysis and functional significance. *Genome Res.*, **20**, 791–803.

Burke,M.K. *et al.* (2010) Genome-wide analysis of a long-term evolution experiment with Drosophila. *Nature*, **467**, 587–590.

Cho,Y.R. *et al.* (2007) Semantic integration to identify overlapping functional modules in protein interaction networks. *BMC Bioinformatics*, **8**, 265.

Dennis,G. Jr *et al.* (2003) DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol.*, **4**, P3.

Draghici,S. *et al.* (2003a) Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res.*, **31**, 3775–3781.

Draghici,S. *et al.* (2003b) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.

Falcon,S. and Gentleman,R. (2007) Using GOstats to test gene lists for GO term association. *Bioinformatics*, **23**, 257–258.

Grossmann,S. *et al.* (2007) Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. *Bioinformatics*, **23**, 3024–3031.

Harris,M.A. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.

Heinig,M. *et al.* (2010) A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. *Nature*, **467**, 460–464.

Hoffman,M.M. and Birney,E. (2010) An effective model for natural selection in promoters. *Genome Res.*, **20**, 685–692.

Lee,Y. *et al.* (2010) Network modeling identifies molecular functions targeted by mir-204 to suppress head and neck tumor metastasis. *PLoS Comput. Biol.*, **6**, e1000730.

Liu,G.E. *et al*. (2010) Analysis of copy number variations among diverse cattle breeds. *Genome Res.*, **20**, 693–703.

Marcucci,G. *et al.* (2008) MicroRNA expression in cytogenetically normal acute myeloid leukemia. *N. Engl. J. Med.*, **358**, 1919–1928.

Mortazavi,A. *et al*. (2010) Scaffolding a Caenorhabditis nematode genome with RNA-seq. *Genome Res.*, **20**, 1740–1747.

Rhee,S.Y. *et al.* (2008) Use and misuse of the gene ontology annotations. *Nat. Rev. Genet.*, **9**, 509–515.

Smukalla,S. *et al.* (2008) FLO1 is a variable green beard gene that drives biofilm-like cooperation in budding yeast. *Cell*, **135**, 726–737.

Somel,M. *et al*. (2010) MicroRNA, mRNA, and protein expression link development and aging in human and macaque brain. *Genome Res.*, **20**, 1207–1218.

Swanson-Wagner,R.A. *et al*. (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res.*, **20**, 1689–1699.

Vinuela,A. *et al*. (2010) Genome-wide gene expression regulation as a function of genotype and age in *C.elegans*. *Genome Res.*, **20**, 929–937.

Wolfe,C.J. *et al.* (2005) Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics*, **6**, 227.

Yamaguchi,K. *et al.* (2010) Identification of an ovarian clear cell carcinoma gene signature that reflects inherent disease biology and the carcinogenic processes. *Oncogene*, **29**, 1741–1752.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016