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

Xinan Yang1, Jianrong Li1, Younghee Lee1 and Yves A. Lussier1,2,*

1Department of Medicine, Section of Genetic Medicine and Center for Biomedical Informatics and 2Comprehensive 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

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 GOBrowser dynamically calculates the most informative subset of GO terms from a prioritized gene list by disregarding other statistically significant GO terms. These approaches are designed to eliminate a massive number of GO terms deemed as significant by statistics, thus severely limiting the expressiveness of the remaining 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
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’ descendents. 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), with theoretical median of 100%). Specifically, we report that on average 28% of input terms were found to be false positives (F), with theoretical median of 100%).

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
<th>Journal (author year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>#GO</td>
<td>#K</td>
<td>#T</td>
</tr>
<tr>
<td>1</td>
<td>15</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>162</td>
<td>46</td>
</tr>
<tr>
<td>4</td>
<td>657</td>
<td>37</td>
</tr>
<tr>
<td>5</td>
<td>109</td>
<td>76</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>20</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>23</td>
<td>21</td>
</tr>
<tr>
<td>9</td>
<td>35</td>
<td>30</td>
</tr>
<tr>
<td>10</td>
<td>21</td>
<td>19</td>
</tr>
<tr>
<td>11</td>
<td>30</td>
<td>21</td>
</tr>
<tr>
<td>12</td>
<td>13</td>
<td>6</td>
</tr>
<tr>
<td>13</td>
<td>20</td>
<td>8</td>
</tr>
<tr>
<td>14</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>15</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>16</td>
<td>10</td>
<td>5</td>
</tr>
</tbody>
</table>

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

**Funding:** National Institute of Health [1U54CA121852 (MAGNET), CTSA ULI RRO24999-03, K22 LM008308-04], and the Cancer Research Foundation (T-AML).

**ACKNOWLEDGEMENTS**

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

**Conflict of Interest:** none declared.
REFERENCES


