Databases and Ontologies

GO-Elite: a flexible solution for pathway and ontology over-representation


1Departments of Pharmacology and Medicine, University of California at San Diego, La Jolla, CA 92093, USA,
2Department of Bioinformatics—BiGCaT, Maastricht University, Maastricht, The Netherlands, 3Gladstone Institute of Cardiovascular Disease and 4Departments of Medicine and Molecular and Cellular Pharmacology, University of California, San Francisco, CA 94158, USA

ABSTRACT

Summary: We introduce GO-Elite, a flexible and powerful pathway analysis tool for a wide array of species, identifiers (IDs), pathways, ontologies and gene sets. In addition to the Gene Ontology (GO), GO-Elite allows the user to perform over-representation analysis on any structured ontology annotations, pathway database or biological IDs (e.g. gene, protein or metabolite). GO-Elite exploits the structured nature of biological ontologies to report a minimal set of non-overlapping terms. The results can be visualized on WikiPathways or as networks. Built-in support is provided for over 60 species and 50 ID systems, covering gene, disease and phenotype ontologies, multiple pathway databases, biomarkers, and transcription factor and microRNA targets. GO-Elite is available as a web interface, GenMAPP-CS plugin and as a cross-platform application.

Availability: http://www.genmapp.org/go_elite

Supplementary Information: Supplementary data are available at Bioinformatics online.

Received on December 1, 2011; revised on May 25, 2012; accepted on June 22, 2012

1 INTRODUCTION

The analysis of pathways, ontologies and other gene sets has become the preferred method for biologists looking to identify global trends from genomic datasets. Although a myriad of tools exist for pathway over-representation, few consider the structured nature of associated ontology data, alternative ontologies and diverse gene sets; few support a wide array of genomes or biological measurements, and they are often limited in scope [Huang da et al. 2009].

Unlike ontologies, pathways provide valuable qualitative contexts (interactions, reactions, metabolites and cellular compartments) that highlight biological relevance. Although various pathway resources now exist [Soh et al. 2010], most over-representation analysis (ORA) tools are limited to one resource that is often outdated. To address these deficiencies, GO-Elite was developed to provide an interchangeable and updatable model of pathway, ontology, species and gene ID system relationships. Using these relationships, GO-Elite performs ontology pruning to report a minimally non-redundant set of results (Fig. 1). Multiple options for naming GO-Elite exist: source-code, cross-platform binaries, OpenAP web service [Ren et al. 2010], online interface or as extensions to the programs GenMAPP-CS (http://www.genmapp.org/) and AltAnalyze (http://www.altanalyze.org). The stand-alone versions of GO-Elite provide an intuitive user interface and command-line control. As previously shown, GO-Elite can be applied to a broad range of biological applications and data types [Hochstenbach et al. 2010; Lemay et al. 2009].

2 METHODS AND IMPLEMENTATION

2.1 Database architecture

Users working with GO-Elite can create their own databases (species, ID systems, relationships) or download official GO-Elite species databases available for each release of Ensembl. The official databases are created primarily from the Ensembl database, which include all external ID systems related to Ensembl (e.g. EnitreGene, UniProt, EMBL) as well as supported microarray platforms (e.g. Affymetrix, Agilent, CodeLink, Illumina). The database is augmented with relationships directly from NCBI EntrezGene and Affymetrix. Currently, relationships to multiple biological Ontology [Gene Ontology (GO), Disease and Phenotype], pathway (WikiPathways, PathwayCommons, KEGG) and gene set resources (e.g. PAZAR, Amadeus, miRanda, RNAhybrid, InterProt and Lineage Biomarkers) are supported [Supplementary Methods]. In addition to gene relationships, metabolomics analyses are available for WikiPathways and KEGG. Although only a select few (ID) systems link directly to pathway and ontology annotations [Ensembl, EntrezGene and HMDB (http://www.hmdb.ca)] by default, all secondary ID systems (e.g. Affymetrix, RefSeq, MGI and Symbol) connect to these through relationship tables. Thus, users can import and analyze ID lists for dozens of supported or user added ID systems.

All resources and annotations provided by GO-Elite can be easily updated or further customized using built-in importers. These importers connect online to the various resources (e.g. WikiPathways, GO and Ensembl) or import local relationships from multiple file formats (e.g. GPML, BioPax and GMT). Alternative ontologies can also be added in GO-Elite, by specifying the URL for any OBO ontology file and importing a species-specific ontology ID relationship file through the user interface.

2.2 Optimized pathway over-representation

For ORA, ontologies, pathways and gene sets are analyzed by a method similar to the program MAPFinder [Bonner et al. 2003]. GO-Elite ranks each analyzed term according to a Z-score, calculated with a normal
The compared scores can be optionally weighted based on the number of IDs associated with each term. This adjustment can result in more or less reported results, by favoring higher level parent nodes with more associated genes, resulting in up to an 80% reduction in the number of reported terms. 

Since several alternative ORA methods exist, such as GSEA (Huang da et al., 2009), users wishing to load results from such algorithms can restrict their analysis to this pruning step.

### 2.3 Data representation

From these analysis steps, multiple results files are produced. The most informative of these is the pruned summary report, which includes all summary term statistics and associated gene or metabolite symbols for both ontology and non-ontology terms. Gene content redundancy between reported terms is also provided, to highlight unrelated terms with similar or identical gene content. When numerical values, such as fold changes, are included with each input ID, GO-Elite will also report mean and standard deviation ontology/pathway-level values in this summary file, analogous to GO-Quant. Xu et al. (2009) allow for downstream pathway-level expression clustering: quantitative integrative methods.

In addition, a full list of ontology and pathway statistics, associated IDs (e.g. gene symbol and associated Ensembl), comparison of reported ontology/pathway statistics between input files (where applicable) and additional gene redundancy focused files are provided. Regulated genes and metabolites can also be immediately visualized on WikiPathways using the stand-alone or GenMAPP-CS interface. Pathway or ontology summarized expression values can be clustered and visualized outside of GO-Elite.

**ACKNOWLEDGEMENTS**

We thank Dr Conrad C. Huang and Dr John H. Morris for their assistance in establishing the GO-Elite web service. This work was supported by grants from the National Institutes of Health (GM080223, GM080223-06S1, HG003053) and the BioRange 1.2.4 research program of the Netherlands Bioinformatics Centre.

Conflict of Interest: none declared.

**REFERENCES**


**ID Systems**

<table>
<thead>
<tr>
<th>Available</th>
<th>ID Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>EnsDB</td>
<td>Entrez</td>
</tr>
<tr>
<td>Affy</td>
<td>HMDB</td>
</tr>
<tr>
<td>Custom</td>
<td></td>
</tr>
</tbody>
</table>

**Ontology Pruning**

**Pathway Filtering**

**Gene Redundancy**

**Results Comparison**

**Pathway-Gene Summarization**

**Gene Associations**

**Pruned Results**

**Gene Rankings**

**Pathway Clustering**

**Pathway Visualization**

**Input IDs**

<table>
<thead>
<tr>
<th>ID Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affy1</td>
</tr>
<tr>
<td>Affy2</td>
</tr>
<tr>
<td>Entrez1</td>
</tr>
</tbody>
</table>

**denominator IDs**

<table>
<thead>
<tr>
<th>Mapping &amp; ORA</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:1</td>
</tr>
<tr>
<td>GO:2</td>
</tr>
<tr>
<td>GO:3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Available Annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td>WP1</td>
</tr>
<tr>
<td>WP2</td>
</tr>
</tbody>
</table>

**CONCLUSIONS**

The ontology ORA results from this step are further evaluated by a simple yet robust pruning method. Pruning occurs by importing these ORA statistics (Z-score, P-values and gene counts), matching user-defined or default filtering options and building all unique branch paths of these results based on the ontology tree structure. Branch paths are pruned to obtain the nodes with the largest Z-score relative to all corresponding child and parent nodes, to report the most informative, highest scoring term for a network of related terms.