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

Alexander C. Zambon1, Stan Gaj2, Isaac Ho3, Kristina Hanspers3, Karen Vranizan3, Chris T. Evelo2, Bruce R. Conklin3,4, Alexander R. Pico3 and Nathan Salomonis1,2

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 or 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
Contact: nsalomonis@gladstone.ucsf.edu

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 exists for pathway over-representation, few support a wide array of 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 or 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
Contact: nsalomonis@gladstone.ucsf.edu

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 exists for pathway over-representation, few support a wide array of 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 or 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
Contact: nsalomonis@gladstone.ucsf.edu

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 exists for pathway over-representation, few support a wide array of 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 or 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
Contact: nsalomonis@gladstone.ucsf.edu

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 exists for pathway over-representation, few support a wide array of 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 or 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
Contact: nsalomonis@gladstone.ucsf.edu

Supplementary Information: Supplementary data are available at Bioinformatics online.

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

nodes, to report the most informative, highest scoring term for a network of nodes with the largest default filtering options and building all unique branch paths of these results (simple yet robust pruning method. Pruning occurs by importing these ORA 2003).

expression values can be clustered and visualized outside of GO-Elite stand-alone or GenMAPP-CS interface. Pathway or ontology summarized and metabolites can be immediately viewed on WikiPathways using the ORA upon pathways, ontologies or loaded gene sets. Regulated genes to a primary ID system (EntrezGene, Ensembl, HMD or custom) for (e.g. Affymetrix) and numerical values (optional). These IDs are mapped genes) and a denominator list (e.g. all genes examined), source ID type (e.g. Affy1(matrix) and numerical values (optional)). These IDs are mapped to a primary ID system (EntrezGene, Ensembl, HMD or custom) for ORA upon pathways, ontologies or loaded gene sets. Regulated genes and metabolites can be immediately viewed 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 approximation to the hypergeometric distribution along with a permutation or a Fisher’s exact test P-value. False-discovery rate adjusted P-values are calculated using a Benjamin–Hochberg correction (Reiner et al., 2003).

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.

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 (Supplementary Table).

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 (<a href="http://www.genome.msi.vt.edu/GoQuant">http://www.genome.msi.vt.edu/GoQuant</a>) allowing for downstream pathway-level expression cluster visualization (Supplementary 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 in the stand-alone interface or following GenMAPP-CS analysis. Relationships between all regulated IDs and ORA terms can also be easily visualized as networks in Cytoscape using produced output files (Supplementary Methods). This application should be of considerable interest to the genomics community, as it represents a highly customizable, simple to use and powerful framework for minimal ontology/pathway reporting. As GO-Elite is agnostic to the type of data input (e.g. gene, protein or metabolite), source ontology, pathway or gene set, we hope to rely further on community-contributed content to improve the utility of this tool in the years to come.

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


