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 Salomonis3,∗

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

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 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 over-representation analyses are available for WikiPathways and KEGG only for a select few (ID) systems link directly to pathway and ontology annotations (Ensembl, EntrezGene and Affymetrix). Currently, relationships to multiple biological Ontology (Gene Ontology (GO), Disease and Phenotype), pathway (WikiPathways, PathwayCommons, KEGG) and gene sets (e.g. PAZAR, Amadeus, miRanda, RNAShybrid, InterPro and Lineage Biomarkers) are supported. In addition to gene relationships, metabolomics analyses are available for WikiPathways and KEGG.

Unlike ontologies, pathways provide valuable qualitative contexts (interactions, reactions, metabolites and cellular compartments) that highlight biological relevance. Although various pathway resources now exist, 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 running GO-Elite exist: source code, cross-platform binaries, Oup web service online interface or as extensions to the programs GenMAPP-CS and AltAnalyze. As previously shown, GO-Elite can be applied to a broad range of biological applications and data types.

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. EntrezGene, 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 sets (e.g. PAZAR, Amadeus, miRanda, RNAShybrid, InterPro and Lineage Biomarkers) are supported. 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) 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) and 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. 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 (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 (Yu et al. 2009), allowing for downstream pathway-level expression clustering (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 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

Dongen,S.W.et al. (2003) MAPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biol. 4, R7.

Hochstein, K. et al. (2010) Transcriptomic profile indicative of immunotoxic exposure in vitro studies in peripheral blood mononuclear cells. Toxicol Sci. 118, 19–30.

Huang da,W. et al. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37, 1–13.

Lemay,D.G.et al. (2009) The bovine lactation genome: insights into the evolution of mammalian milk. Genome Biol. 10, R43.

Ren,J.et al. (2010) Opal web services for biomedical applications. Nucleic Acids Res. 38, W724–W731.

Soh,D. et al. (2010) Consistency, comprehensiveness, and compatibility of pathway databases. BMC Bioinformatics. 11, 449.

Yu,X. et al. (2000) A system-based approach to interpret dose- and time-dependent microarray data: quantitative integration of gene ontology analysis for risk assessment. Toxicol Sci. 92, 560–577.