

Systems biology

BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks

Steven Maere, Karel Heymans and Martin Kuiper*

Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052, Ghent, Belgium

Received on May 18, 2005; revised on June 13, 2005; accepted on June 17, 2005

Advance Access publication June 21, 2005

ABSTRACT

Summary: The Biological Networks Gene Ontology tool (BiNGO) is an open-source Java tool to determine which Gene Ontology (GO) terms are significantly overrepresented in a set of genes. BiNGO can be used either on a list of genes, pasted as text, or interactively on subgraphs of biological networks visualized in Cytoscape. BiNGO maps the predominant functional themes of the tested gene set on the GO hierarchy, and takes advantage of Cytoscape's versatile visualization environment to produce an intuitive and customizable visual representation of the results.

Availability: <http://www.psb.ugent.be/cbd/papers/BiNGO/>

Contact: martin.kuiper@psb.ugent.be

1 INTRODUCTION

Over the last decade, the development of high-throughput technologies, such as microarray-based transcript profiling, has led to an exponential increase in the volume of functional genomics data. Interpretation of these data is greatly facilitated by a structured description of known biological information at different levels of granularity. The Gene Ontology (GO) project (Ashburner *et al.*, 2000), initiated in the late 1990's, aims at capturing the increasing knowledge on gene function in a controlled vocabulary applicable to all organisms. GO consists of three hierarchically structured vocabularies that describe gene products in terms of their associated biological processes, molecular functions and cellular components. Gene products may be annotated to one or several nodes in each hierarchy.

The increasing complexity of functional genomics data also drives the development of methods and tools for data integration and visualization. Cytoscape (Shannon *et al.*, 2003) is an open-source software platform for visualizing molecular interaction networks and integrating these interactions with gene expression profiles and other functional genomics data. The Cytoscape platform actively supports the development of plugin tools that extend the core functionality. We developed the Biological Networks Gene Ontology tool (BiNGO) as a plugin for Cytoscape. BiNGO assesses the overrepresentation of GO categories in a subgraph of a biological network, or any other set of genes. Several tools exist that analyze GO term enrichment in a given gene set (Berriz *et al.*, 2003; Hosack *et al.*, 2003; Zeeberg *et al.*, 2003; Al-Shahrour *et al.*, 2004; Beißbarth and Speed, 2004; Boyle *et al.*, 2004; Masseroli *et al.*,

2004; Zhang *et al.*, 2004). A comprehensive list can be found at <http://www.geneontology.org/GO.tools.shtml>. The main advantage of BiNGO over these tools is its interactive use on molecular interaction networks, e.g. protein interaction networks or transcriptional coregulation networks, visualized in Cytoscape. Furthermore, BiNGO offers great flexibility in the use of ontologies and annotations. Besides the traditional GO ontologies, BiNGO also supports the use of GOSlim ontologies, as well as custom ontologies and annotations. Finally, the Cytoscape graphs produced by BiNGO can be viewed, laid out, modified and saved in various manners.

2 METHODS AND IMPLEMENTATION

There are two modes for selecting the set of genes to be functionally profiled. In the default mode, a set of nodes can be selected from a Cytoscape network, either manually or using other plugins such as MCODE (Bader and Hogue, 2003). Alternatively, a test set can be compiled from other sources, for instance a set of genes that are up-regulated in a microarray experiment, and pasted in a text input box. BiNGO retrieves the relevant GO annotations and propagates them upwards through the GO hierarchy, i.e. any gene annotated to a certain GO category is also explicitly included in all parental categories. BiNGO currently provides two statistical tests for assessing the enrichment of a GO term in the test set. The basic question answered by these tests is as follows: when sampling X genes (test set) out of N genes (reference set, either a graph or an annotation), what is the probability that x or more of these genes belong to a functional category C shared by n of the N genes in the reference set? The hypergeometric test, in which sampling occurs without replacement, answers this question in the form of a P -value. Its counterpart with replacement, the binomial test, provides only an approximate P -value, but requires less calculation time.

Because BiNGO tests the significance of all GO labels present in the test set, the number of statistical tests performed in a single analysis may amount to several hundreds. When testing multiple hypotheses, the obtained P -values have to be corrected in order to control the type I error (false positive) rate (Ge *et al.*, 2003). One of the most basic multiple testing corrections is the Bonferroni correction, which strongly controls the family-wise error rate (FWER), i.e. the probability of making at least one type I error. The Bonferroni correction is generally assumed to be conservative, although it might actually be rather liberal (Boyle *et al.*, 2004), at least for FWER control, when used for correcting tests that are not mutually independent, as is the case when testing GO categories (see below). An alternative to FWER corrections is to control the false discovery rate (FDR), i.e. the expected proportion of false positives among the positively identified tests. Generally, this type of correction is more appropriate for our purposes, because we would typically prefer to have more power (less false negatives) at the cost of a few more false positives. One of the most popular FDR corrections is the Benjamini and Hochberg correction, which provides strong control over the FDR under positive regression

*To whom correspondence should be addressed.

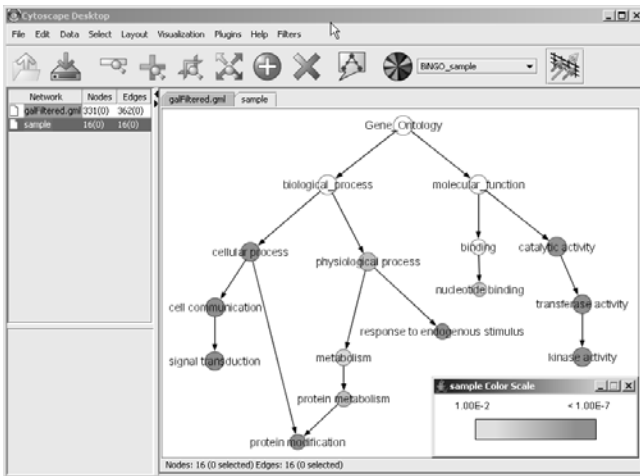


Fig. 1. A sample BiNGO result for a set of *Arabidopsis* protein kinases, as visualized in Cytoscape. Dark grey categories are most significantly overrepresented. White nodes are not significantly overrepresented, they are included to show the grey nodes in the context of the GO hierarchy. The area of a node is proportional to the number of genes in the test set annotated to the corresponding GO category.

dependency of the test statistics (Benjamini and Yekutieli, 2001). In fact, it is unclear whether the GO hierarchy fulfils the requirement of positive regression dependency. Nevertheless, the Benjamini and Hochberg correction is used widely. BiNGO is an open-source Java package, distributed under the GNU General Public Licence (<http://www.gnu.org/>). Extra statistical tests or more refined multiple testing corrections can be added to BiNGO through the implementation of interfaces provided for this purpose. However, the safest way to minimize the impact of multiple testing issues is to test fewer categories. To this end, we provide several GOSlim ontologies in BiNGO that are organism-specific slimmed-down versions of the full GO hierarchy. When using these GOSlims in combination with a standard or custom annotation, the organism's full GO annotation is remapped onto the chosen GOSlim, with the full GO ontology as a remapping guide, thereby eliminating the need for separate GOSlim annotation files.

BiNGO provides annotations for a wide range of organisms. These default annotations are parsed from the GO information available from NCBI (<http://www.ncbi.nlm.nih.gov/Ftp/>). Several gene identifiers are supported. Probably the most stable one is the Entrez GeneID, which is the unique identifier for a gene in NCBI's Entrez Gene (formerly LocusLink) database, and to a lesser extent the LocusTag identifier, which is unique to a particular locus (e.g. ORF names for baker's yeast and AGI codes for *Arabidopsis*). Official Gene Symbols and Unigene IDs are generally less stable, or at least more prone to misinterpretation. We deliberately chose not to support the use of synonyms or other commonly used names. These alternative names are frequently non-unique and may lead to confusion, so we do not want to encourage their use.

Although BiNGO is primarily designed for use with GO ontologies, other classification systems [e.g. the MIPS Functional Catalogue (Ruepp *et al.*, 2004)] can be used as well, provided that the classification information is parsed into the right format. More information about the required formats is available at our website (<http://www.psb.ugent.be/cbd/papers/BiNGO/>).

BiNGO assesses the functional themes that are present in a set of genes. Eventhough a *P*-value gives a good indication about the prominence of a certain functional category, it is risky to draw conclusions solely based on *P*-values. The *P*-values returned by BiNGO should be regarded as suggestions, and interpreted in the light of other evidence. Due to the interdependency of functional categories in the GO hierarchy, it is very likely that

not one category, but a whole branch of the GO hierarchy is highlighted as being significantly overrepresented (Fig. 1). In such cases, interpretation can be more difficult. The most intensely colored nodes that are farthest down the hierarchy are probably the most relevant ones. For example, let us suppose that a branch of 'catalytic activity' subcategories is highlighted (Fig. 1), then we cannot conclude that genes involved in 'catalytic activity' as a whole are significantly overrepresented in the test set. In fact, if 'kinase activity' is the relevant category, the overrepresentation of the 'transferase activity' and 'catalytic activity' categories merely results from the presence of 'kinase activity' genes. Would there be a substantial contribution of genes in the 'catalytic activity' category other than 'kinase activity' genes, then the 'catalytic activity' node would be bigger in size, which is not the case. Next to the visual representation, BiNGO produces a tab-delimited text file containing more detailed results. Apart from a listing of the analysis options, the results file contains the (adjusted) *P*-value for each significantly overrepresented GO class, the number of genes in the test set annotated to that class and their identity, and the number of genes annotated to that class in the reference set.

BiNGO is a flexible, extendable tool used to analyze GO term overrepresentation in biological networks. We believe that embedding BiNGO in Cytoscape will further contribute to the establishment of Cytoscape as an integrated suite of tools for the analysis of biological networks. As Cytoscape continues to evolve, BiNGO will evolve alongside it. Comments and feature requests will be considered thoroughly.

ACKNOWLEDGEMENTS

The authors would like to thank Benno Schwikowski, Iliana Avila, Gary Bader, Rowan Christmas, Andrew Markiel and the whole Cytoscape development team for their technical support and useful comments. S.M. is a research fellow at the Fund for Scientific Research (Flanders, Belgium).

Conflict of Interest: none declared.

REFERENCES

Al-Shahrour,F. *et al.* (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.
 Ashburner,M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
 Bader,G.D. and Hogue,C.W.V. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.
 Beißbarth,T. and Speed,T.P. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.
 Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, **29**, 1165–1188.
 Berriz,G.F. *et al.* (2003) Characterizing gene sets with FuncAssociate. *Bioinformatics*, **19**, 2502–2504.
 Boyle,E.I. *et al.* (2004) GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.
 Ge,Y., Dudoit,S. and Speed,T.P. (2003) Resampling-based multiple testing for microarray data analysis. *Technical Report #633*, Dept. of Statistics, UC Berkeley.
 Hosack,D.A. *et al.* (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol.*, **4**, R70.
 Masseroli,M. *et al.* (2004) GFINDER: Genome Function INtegrated Discoverer through dynamic annotation, statistical analysis, and mining. *Nucleic Acids Res.*, **32**, W293–W300.
 Ruepp,A. *et al.* (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res.*, **32**, 5539–5545.
 Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
 Zeeberg,B.R. *et al.* (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.*, **4**, R28.
 Zhang,B. *et al.* (2004) GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics*, **5**, 16.