BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 9 2004, pages 1464–1465
DOI: 10.1093/bioinformatics/bth088

GOstat: find statistically overrepresented Gene Ontologies within a group of genes

Tim Beißbarth and Terence P. Speed*

Walter and Eliza Hall Institute of medical Research, 1G Royal Parade, Parkville, Vic 3050, Australia

Received on July 14, 2004; revised on December 1, 2003; accepted on December 4, 2003
Advance Access publication February 12, 2004

ABSTRACT

Summary: Modern experimental techniques, as for example DNA microarrays, as a result usually produce a long list of genes, which are potentially interesting in the analyzed process. In order to gain biological understanding from this type of data, it is necessary to analyze the functional annotations of all genes in this list. The Gene-Ontology (GO) database provides a useful tool to annotate and analyze the functions of a large number of genes. Here, we introduce a tool that utilizes this information to obtain an understanding of which annotations are typical for the analyzed list of genes. This program automatically obtains the GO annotations from a database and generates statistics of which annotations are overrepresented in the analyzed list of genes. This results in a list of GO terms sorted by their specificity.

Availability: Our program GOstat is accessible via the Internet at http://gostat.wehi.edu.au

Contact: beissbarth@wehi.edu.au

Ontologies are a widely used concept to create a controlled vocabulary to communicate and annotate knowledge. The Gene Ontology Consortium defines GO as an international standard to annotate genes (Ashburner *et al*., 2000). GO has a hierarchical structure starting with top-levels ontologies for molecular functions, biological processes and cellular components. The GO database consists of two essential parts, the current ontologies, which define the vocabulary and structure, and the current annotations, which create a link between the known genes and the associated GOs that define their function. Currently, many groups are working on the development of the ontologies and annotations for different organisms. All the information can be downloaded from the web-site http://www.geneontology.org

Here, we would like to make use of the annotations and structure of the GOs in order to understand the biological processes present in a large dataset of genes. The usefulness of keyword hierarchies in interpreting large datasets has been demonstrated previously (Masys *et al*., 2001). Recently, a vast number of tools are evolving that make use of GOs (Doniger *et al*., 2003; Draghici *et al*., 2003; Al-Shahrour *et al*., 2004; Dennis *et al*., 2003). We consider GOstat an easy to use tool with a solid statistical foundation.

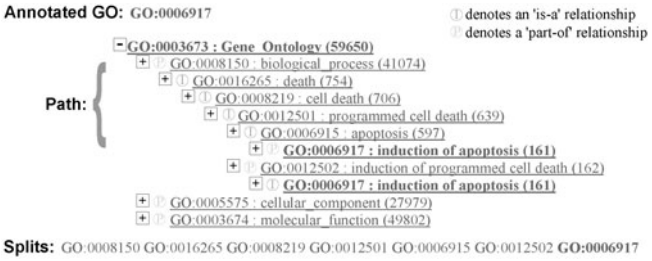

Fig. 1. Schema of GO annotation terms.

Each gene can have several associated GO terms. Further, due to the hierarchical structure of the GOs, each GO term can be connected to several other GO terms higher in the GO hierarchy and therefore associated with the gene as well (Fig. 1). We call the list of GO terms that are in between a top level and the annotated GO term its path. In fact, several such paths might lead to an individual GO term. Each GO term in the path we call a split. So in the end a list of 100 genes will usually have many hundreds of associated GO terms and several thousand associated splits.

GOstat requires a list of gene identifiers that specify the group of genes of interest. The program uses several synonyms, each of which is sufficient to identify a gene. These synonyms are derived from the release of the GO database as well as from Unigene (Boguski and Schuler, 1995). GO databases for several organisms (human, mouse, *Drosophila*, yeast, *Arabidopsis thaliana*, etc.) are provided. In order to find GO terms that are statistically significant within the group, a control set of genes needs to be used to obtain a total count of occurrences for each GO term. This can be the complete database of annotated genes, one of several subsets that are commonly used on widely available microarrays or a second list of gene identifiers that is passed to the program. In this case, the second list is used as a reference to search for GO

*To whom correspondence should be addressed.

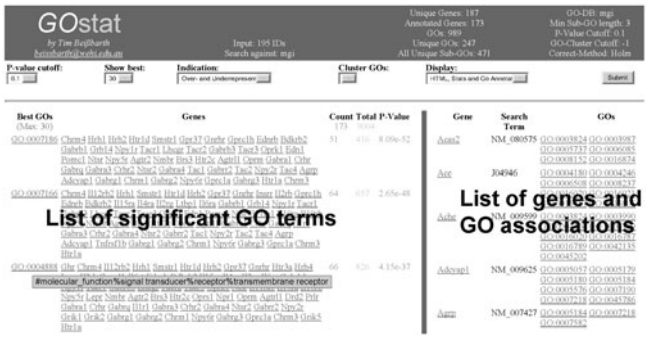

Fig. 2. GOstat Output.

terms, which are significantly more represented in the first list compared with the second.

For all of the genes analyzed, GOstat will determine the annotated GO terms and all splits. The program will then count the number of appearances of each GO term for the genes in the group as well as in the reference group. For each GO term, a *p*-value is calculated representing the probability that the observed numbers of counts could have resulted from randomly distributing this GO term between the tested group and the reference group. A χ^2 test is used in order to approximate this *p*-value. If the expected value for any count is below 5, the χ^2 approximation is inaccurate. Therefore, we use Fisher's Exact Test in these cases. The resulting list of *p*-values is sorted. The GO terms that are most specific for the analyzed list of genes will have the lowest *p*-values.

As the number of GO terms for which we test significance is large, the computed *p*-values have to be corrected in order to control the rate of errors we expect with multiple testing (Shaffer, 1995; Dudoit *et al*., 2002). Two methods for correcting the *p*-value are offered in GOstat. The Holm correction controls the familywise error rate, e.g. selecting genes with a *p*-value below 0.1 we expect a 10% chance that any of the selected GO terms are not specific. The Benjamini and Hochberg correction controls the false discovery rate, e.g. selecting genes with a *p*-value below 0.1, we expect that 10% of the selected GO terms are not specific.

However, there are dependences between various GO terms in the resulting list. Frequently, genes share more or less the same set of annotations, as several GO terms are indicative of the same process. Also, GO terms that are within one path have strongly correlated results. In order to make the resulting list of GO terms more interpretable, GOstat has the option to cluster the GO terms. In this process, GO terms that are annotated in the same set of genes or where one set of genes is a subset of the other are grouped.

GOstat will result in a list of *p*-values that state how specific certain GO terms are for a given list of genes (Fig. 2). The output is sorted by the *p*-value and can be limited by various cutoff values. It is possible to display the over or underrepresented terms only. *p*-values of GO terms that are overrepresented in the dataset are typeset in green, *p*-values of underrepresented GO terms are colored red. GO terms that are annotated in more or less the same subsets of genes can be grouped together. GOstat will also output the complete list of the associations for the supplied genes to the annotated GO terms. The GO IDs in the output are linked to AmiGO, a visualization tool for the hierarchy in the GO database (http://www.godatabase.org). It is possible to format the output in HTML or as a tabular text.

GOstat provides a useful tool in order to find biological processes or annotations characteristic of a group of genes. This is greatly helpful in analyzing lists of genes resulting from high-throughput screening experiments, such as microarrays, for their biological meaning.

ACKNOWLEDGEMENTS

Thanks to Joelle Michaud, Lavinia Hyde, Gordon Smyth and Hamish Scott for helpful suggestions and testing of the program. This work was funded by the Deutsche Forschungsgemeinschaft.

REFERENCES

Al-Shahrour,F., Diaz-Uriarte,R. and Dopazo,J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.

Ashburner,M., Ball,C., Blake,J., Botstein,D., Butler,H., Cherry,J., Davis,A., Dolinski,K., Dwight,S., Eppig,J. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*., **25**, 25–29.

Boguski,M. and Schuler,G. (1995) Establishing a human transcript map. *Nat. Genet*., **10**, 369–371.

Dennis,G.,Jr, Sherman,B., Hosack,D., Yang,J., Gao,W., Lane,H. and Lempicki,R. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol*., **4**, P3.

Doniger,S., Salomonis,N., Dahlquist,K., Vranizan,K., Lawlor,S. and Conklin,B. (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol*., **4**, R7.

Draghici,S., Khatri,P., Bhavsar,P., Shah,A., Krawetz,S. and Tainsky,M. (2003) Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res*., **31**, 3775–3781.

Dudoit,S., Shaffer,J. and Boldrick,J. (2002) Multiple hypothesis testing in microarray experiments. *Technical Report 110*, Division of Biostatistics, UC Berkeley.

Masys,D., Welsh,J., Fink,J.L., Gribskov,M., Klacansky,I. and Corbeil,J. (2001) Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics*, **17**, 319–326.

Shaffer,J. (1995) Multiple hypothesis testing. *Annu. Rev. Psychol*., **46**, 561–584.