BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 4 2004, pages 578–580
DOI: 10.1093/bioinformatics/btg455



FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes

*Fátima Al-Shahrour, Ramón Díaz-Uriarte and Joaquín Dopazo**

Bioinformatics Unit, Centro Nacional de Investigaciones Oncológicas (CNIO), Melchor Fernández Almagro 3, 28029 Madrid, Spain

Received on August 15, 2003; revised on September 30, 2003; accepted on October 1, 2003
Advance Access publication January 22, 2004

ABSTRACT

Summary: We present a simple but powerful procedure to extract Gene Ontology (GO) terms that are significantly over- or under-represented in sets of genes within the context of a genome-scale experiment (DNA microarray, proteomics, etc.). Said procedure has been implemented as a web application, FatiGO, allowing for easy and interactive querying. FatiGO, which takes the multiple-testing nature of statistical contrast into account, currently includes GO associations for diverse organisms (human, mouse, fly, worm and yeast) and the TrEMBL/Swissprot GOAnnotations@EBI correspondences from the European Bioinformatics Institute.
Availability: http://fatigo.bioinfo.cnio.es
Contact: jdopazo@cnio.es

Most resources available that collect information regarding gene or protein function, biological properties, etc., are based on the pre-genomic design in which the information is acceded and displayed in a one-gene-at-a-time format. Nevertheless, many problems related to functional genomics involve the detection of biological properties, functions, etc., shared by a set of genes, that sets them aside from the remaining ones. The practical application of methods of pre-genomic design to these problems is drastically limited when thousands of genes are involved in the comparative study. The use of techniques of automatic management of biological information, such as text mining, in studying the coherence of gene groups obtained from different methodologies has only recently been addressed (Oliveros *et al*., 2000; Raychaudhuri *et al*., 2002; Pavlidis *et al*., 2002), although its practical application still has many drawbacks (Blaschke *et al*., 2002). Furthermore, real implementations are often scarce and beyond the reach of many users.

An alternative to extracting information from scientific text sources is by using ontologies. In its most simple representation, ontologies provide a structured description of biological information that is extremely useful for computational management. One of the most widely accepted ontologies is Gene Ontology (GO; Ashburner *et al*., 2000), which organizes information for molecular function, biological processes and cellular components for a number of different organisms. The potential of GO terms as a structured source of information however, has yet to be fully exploited.

Here we present FatiGO, a web-based application (http://fatigo.bioinfo.cnio.es). Since the publication of FatiGO in the GO consortium web page (http://www.geneontology.org) less than a year ago, a number of tools have been implemented based on the same idea of mapping biological knowledge on sets of genes. Thus, Onto-Express (Khatri *et al*., 2002), which generates tables that correlate groups of genes to biochemical and molecular functions or MAPPFinder (Doniger *et al*., 2003), which, using a searchable web interface, identifies GO terms over-represented in the data. A similar tool, FunSpec (Robinson *et al*., 2002), evaluates groups of yeast genes in terms of their annotations in diverse databases. Many of these tools are stand-alone applications with user-friendly interfaces, but obviously suppose limitations in processing large amounts of data. Moreover, important issues such as the multiple-testing nature of the statistical contrasts are not well addressed.

FatiGO is used to extract relevant GO terms for a group of genes with respect to a set of genes of reference (typically the rest of genes). The terms are considered to be relevant by the application of a Fisher's exact test that considers the multiple-testing nature of the statistical contrast performed. Multiple testing is an important issue that is, nevertheless, scarcely addressed (Slonim, 2002). If the multiple-testing nature of the statistical contrast is not taken into account an increase in the rate of false positives (i.e. terms identified as over- or under-represented whose proportions, in reality, are not significantly different), occurs. FatiGO can deal with thousands of genes from different organisms (currently human, mouse, *Drosophila*, worm, yeast, as well as genes whose proteins are included in Swissprot database), and can be queried using

*To whom correspondence should be addressed.

Bioinformatics 20(4) © Oxford University Press 2004; all rights reserved.

different gene identifiers (GenBank ID, Unigene, ENSEMBL, systematic name, Swissprot/TrEMBL). FatiGO uses tables of correspondence between genes and their corresponding GO terms. The program can be used to list the proportions of GO terms in a set of genes. The output links the genes to the corresponding databases and the GO terms to AmiGO, a GO browser (http://godatabase.org/).

When possible, curated associations of gene IDs to GO terms have been used. Associations for different organisms have been included in FatiGO (at present human, mouse, *Drosophila*, yeast, *Caenorhabditis elegans* and, in general, genes whose proteins have been included in Swissprot). In the GO Consortium web page (http://www.geneontology.org) tables relating GO terms with gene IDs can be found.

A common problem when dealing with thousands of genes annotated in origin in different places is the lack of standardization in the annotations. Several manufacturers of microarrays use unigene codes, but often GenBank IDs and even gene names are also found embedded in the annotation. **Xref** EBI codes are used to relate GenBank, ENSEMBL and unigene to Swissprot/TREMBL, which is associated to GO by means of the GO annotations@EBI tables. For the other databases specific gene ID to GO associations have been used. Locuslink contains information on official nomenclature genes, aliases and unigene clusters and provides a link between different gene identifiers. Unigene IDs present an additional problem: sequences can change unigene cluster from release to release, being necessary to maintain a historical record to keep track of the last version's ID for a gene that was annotated several releases ago. An additional table of correspondence between old and new unigene versions (downloaded from the unigene web site) has been used. Using this table we can obtain the updated version of the unigene ID for any ID belonging to an old version, if any. It is possible that (a) there is a different ID because it has been merged to another cluster and has taken its ID, (b) there is more than one ID because the cluster has split into several clusters and (c) the unigene cluster has been withdrawn from the database. On the other hand, unigene IDs are easy to parse (they are 'Hs.' or 'Mm.' followed by a number) which is convenient for processing files of DNA microarray results in which the manufacturers tend to include the unigene code among other additional information. Gene associations include all evidence codes. Part of them, the Inferred from Electronic Annotation, have been assigned automatically without the intervention of human curators, so they are less reliable (see the GO evidence code documentation at http://www.geneontology.org/doc/GO.evidence.html).

The FatiGO tool can be used to find GO terms that are over- and under-represented in a set of genes with respect to a reference group. Once both sets of genes have been uploaded (as lists of gene IDs) the GO level at which the statistical contrast is going to be performed must be chosen. Deeper terms in the GO hierarchy are more precise. Obviously, the number of genes with annotations decreases at deeper GO levels. GO level 3 constitutes a good compromise between information quality and number of genes annotated at this level (Mateos *et al*., 2002). For genes annotated at deeper levels than the selected level, FatiGO climbs up the GO hierarchy until the terms for said level are reached. The use of the parent terms increases the sizes of the classes (genes annotated with a given GO term), making it easier to find relevant differences in distributions of GO terms among clusters of genes. The information is not lost and can be recovered later. In the instance of repeated genes (something common in microarray data) only one is used. Once the collections of GO terms corresponding to the two datasets of genes are prepared, a Fisher's exact test for 2×2 contingency tables is applied.

FatiGO returns adjusted p-values based on three different ways of accounting for multiple testing. One of them is the step-down minP method of Westfall and Young (1993). This method provides control of the family wise error rate (i.e. the probability of making a Type I error rate over the family of tests). Adjusted p-values were also calculated using the false discovery rate (FDR), that is, the expected number of false rejections among the rejected hypothesis can be controlled. The FDR method of Benjamini and Hochberg (1995), which offers control of the FDR only under independence and some specific types of positive dependence of the tests statistics, and the FDR method of Benjamini and Yekutieli (2001), which offers strong control under arbitrary dependency of test statistics were also implemented in FatiGO (see also Dudoit *et al*., 2002; Reiner *et al*., 2003).

A permutation test that preserves the pattern of among GO co-variation in the calculation of the adjusted p-values was used. For each random permutation and for each GO term the exact p-value from Fisher's exact test for each contingency table is calculated. The FatiGO program returns four columns: the unadjusted p-value, which is the p-value from Fisher's exact test without adjusting for multiple comparisons, and the adjusted p-values based on the three methods described above. The results are ordered by decreasing value of the adjusted p-value, thus facilitating the selection of GO terms with the most significant differences.

As previously mentioned, the growing interest by functional genomics makes more evident than ever the necessity of methods for studying properties shared by groups of genes that have a common behaviour. Direct use of biological information extracted from biomedical literature (Oliveros *et al*., 2000; Jenssen *et al*., 2000; Raychaudhuri *et al*., 2002) for studying such properties still supposes serious drawbacks (Blaschke *et al*., 2002). One of them is that unless preprocessing is carried out, the volume of information to deal with is excessive for common on-line, interactive applications. The advantage of using GO terms is that interactivity is feasible, as our implementation, FatiGO, demonstrates. In addition, GO terms have a clear biological meaning,

something that is not guaranteed with the approaches based on processing of free text. Besides, the number of terms is not as large as to expect a high number of artifactual associations of terms to clusters of genes just by chance. Despite a number of genes still lacking GO annotations the results found in the examples analysed (see GEPAS examples page http://gepas.bioinfo.cnio.es/data) are informative enough to characterize the biological processes. In addition, as the number of genes with GO annotations increases in the next few years, functional genomics will benefit from the use of applications, such as FatiGO, based on GO terms.

FatiGO addresses another common problem: the multiple ways in which genes are annotated. Different manufacturers of genomic platforms use distinct gene IDs. The most common gene IDs can be used as input for the application.

Ontologies can be used as a quick and efficient information mining tool for the identification and validation of clusters of co-expressing genes studied (Pavlidis *et al*., 2002; Mateos *et al*., 2002). To facilitate this use, FatiGO is coupled with the clustering programs of the GEPAS (http://gepas.bioinfo.cnio.es), a suite of programs for microarray gene expression data analysis (Herrero *et al*., 2003).

We are currently implementing different types of biological information which include Interpro functional motifs, OMIM terms for diseases, protein interactions, pathways, etc. Finally, more organisms will be included as soon as their corresponding genomic projects produce high quality annotation results.

ACKNOWLEDGEMENTS

We are indebted to Amanda Wren for correcting the English of this manuscript. F.A. is supported by a grant BIO2001-0068 from MCYT, R.D.U. is supported by a Ramón y Cajal research contract from the MCYT.

REFERENCES

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

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165–1188.

Blaschke,C., Hirschman,L. and Valencia,A. (2002) Information extraction in molecular biology. *Brief. Bioinform.*, **3**, 154–165.

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

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

Herrero,J., Al-Shahrour,F., Díaz-Uriarte,R., Mateos,Á., Vaquerizas,J.M., Santoyo,J. and Dopazo,J. (2003) GEPAS, a web-based resource for microarray gene expression data analysis. *Nucleic Acids Res.*, **31**, 3461–3467.

Jenssen,T.-K., Laegreid,A., Komorowski,J. and Hovig,E. (2000) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.*, **28**, 21–28.

Khatri,P., Draghici,S., Ostermeier,G.C. and Krawetz,S.A. (2002) Profiling gene expression using onto-express. *Genomics*, **79**, 1–5.

Mateos,A., Herrero,J., Tamames,J. and Dopazo,J. (2002) Supervised neural networks for clustering conditions in DNA array data after reducing noise by clustering gene expression profiles. In Lin,S. and Johnson,K. (eds), *Methods of Microarray Data Analysis II*. Kluwer Academic Publishers, Boston. pp. 91–103.

Oliveros,J.C., Blaschke,C., Herrero,J., Dopazo,J. and Valencia,A. (2000) Expression profiles and biological function. *Genome Inform.*, **10**, 106–117.

Pavlidis,P., Lewis,D.P. and Noble,W.S. (2002) Exploring gene expression data with class scores. *Pac. Symp. Biocomput.*, **7**, 474–485.

Raychaudhuri,S., Schutze,H. and Altman,R.B. (2002) Using text analysis to identify functionally coherent gene groups. *Genome Res.*, **12**, 1582–1590.

Reiner,A., Yekutieli,D. and Benjamini,Y. (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**, 368–375.

Robinson,M.D., Grigull,J., Mohammad,N. and Hughes,T.R. (2002) FunSpect: a web-based cluster interpreter for yeast. *BMC Bioinformatics*, **3**, 1–5.

Slonim,D.K. (2002) From patterns to pathways: gene expression data analysis comes of age. *Nat. Genet.*, **32**, (suppl. The Chipping Forecast), 502–508.

Westfall,P.H. and Young,S.S. (1993) *Resampling-Based Multiple Testing*. John Wiley & Sons, New York.