BIOINFORMATICS APPLICATIONS NOTE

Vol. 20 no. 7 2004, pages 1196–1197
DOI: 10.1093/bioinformatics/bth056

CLENCH: a program for calculating Cluster ENriCHment using the Gene Ontology

N.H. Shah* and N.V. Fedoroff

The Huck Institute of Life Sciences, Pennsylvania State University, University Park, PA 16802, USA

Received on May 29, 2003; revised on September 29, 2003; accepted on November 5, 2003
Advance Access publication February 5, 2004

ABSTRACT

Summary: Analysis of microarray data most often produces lists of genes with similar expression patterns, which are then subdivided into functional categories for biological interpretation. Such functional categorization is most commonly accomplished using Gene Ontology (GO) categories. Although there are several programs that identify and analyze functional categories for human, mouse and yeast genes, none of them accept *Arabidopsis thaliana* data. In order to address this need for *A.thaliana* community, we have developed a program that retrieves GO annotations for *A.thaliana* genes and performs functional category analysis for lists of genes selected by the user.

Availability: http://www.personal.psu.edu/nhs109/Clench

Contact: nigam@psu.edu

The results of DNA and oligonucleotide microarray studies pose a significant analytical challenge. After the data are subjected to statistical testing for significance, the experimental biologist is often confronted with long lists of genes (Zeeberg *et al*., 2003). Extracting biological meaning from the activation (or repression) of the listed genes is an important but time-consuming task, which is exacerbated by the lack or incompleteness of gene annotations. In such a situation, the controlled vocabulary and the hierarchical classification of the vocabulary terms developed by the Gene Ontology (GO) consortium (Ashburner *et al*., 2000) are a very useful resource for the initial interpretation of gene lists. GO's hierarchical classification of terms gives a biologically meaningful basis for grouping genes in an unambiguous manner. By determining whether GO terms associated with particular biological process, molecular function or cellular component are either over- and under-represented among the genes in each subgroup generated by a cluster algorithm, it may be possible to gain insight into the biological significance of alterations in the gene expression levels. Although there are several programs that carry out such category assignments and analysis for human, mouse and yeast genes, none of them is useful for *Arabidopsis* genes because they do not accept AGI identifiers (Khatri *et al*., 2002; Doniger *et al*., 2003; FatiGO, 2003, http://fatigo.bioinfo.cnio.es/; GoSurfer, 2002, http://biosun1.harvard.edu/complab/gosurfer). We have developed such a tool for the *Arabidopsis* community.

Here we describe CLENCH, a program for calculating CLuster ENriCHment using the GO. CLENCH calculates functional enrichment in gene groups derived from the analysis of gene expression data. CLENCH uses TAIR (Rhee *et al*., 2003) as the source of annotations because it is the most reliable public source of curated information for *Arabidopsis thaliana*. The program accepts two lists of genes: (1) Total-Genes (which is the set of genes that each 'cluster' is to be compared with and may be all the genes in the genome, all the genes on the microarray or even the complete list of genes identified as differentially expressed) and (2) Changed-Genes (a subset of genes, e.g. a cluster of genes that is to be analyzed for enrichment of functional categories). It retrieves GO annotations for both gene lists to calculate the number of genes (n and m) belonging to a particular GO category in both lists and then calculates the hypergeometric probability (p-value) for finding at least n genes belonging to that category in the Changed-Genes list given that m genes were annotated to that category in the Total-Genes list. This p-value tells us how likely it is to find at least n genes of a particular category in the Changed-Genes list by chance alone, given the number of genes in that category in the total set. A category is called enriched if the p-value is less than 0.05.

CLENCH results are presented as an HTML page containing a table for each primary GO category (molecular function, biological process and cellular component). There is a row for each category with fields for the name of the category, the GO id of that category, the p-value for enrichment of the category and the genes annotated to the category. Each GO id in the result is hyper-linked to the AmiGO browser allowing easy visualization of its place in the GO hierarchy. The gene identifiers are linked to their TAIR annotations.

The p-values reported by CLENCH should be interpreted with caution because the Total-Genes list to which the Changed-Genes list is being compared affects the p-value.

*To whom correspondence should be addressed.

For microarrays that comprise a random subset of all genes, using the list of all genes on the microarray as the Total-Genes list is equivalent to using the complete list of genes in the genome. However, for microarrays containing a selected subset of genes associated with a biological process, the choice of the gene set to use as the total set is not obvious. Comparing with all genes in the genome is very lenient because it biases the results towards categories that were already enriched during the pre-selection process. Comparing with the set of arrayed genes is too stringent because it biases the results against categories that were enriched during the pre-selection process. Thus, depending on the choice of the total set, the p-values should be interpreted in light of other available information, keeping in mind what is known about involvement of a particular biological process in the process under study, the number of genes in the category reported to be significant and the relative size of the category in the two lists. To aid in this process, CLENCH reports the number of genes assigned to each category in the Changed-Genes list, a ratio of the percentage of genes in each GO category in the two lists analyzed in a column titled ‘relative enrichment’ or RE and p-values calculated using the χ^2 and binomial distributions.

Another caveat of deciding significance using the calculated p-value (and a cutoff of 0.05) is the multiple testing that occurs. Multiple testing happens because we do not pre-select which category to test for enrichment, but rather test each existing category. This allows multiple opportunities (equal to the number of categories tested) to obtain a statistically significant p-value, by chance alone, from a given gene list. However, correcting for this (e.g. using a Bonferroni correction in which the critical p-value cutoff is divided by the number of tests made) is too restrictive and is not advised because the correction assumes independence of categories and even truly enriched categories are not detected (Zeeberg *et al.*, 2003).

Because annotation of genes and proteins by TAIR is an ongoing effort, there may be categorization errors and some categories may not yet be completely covered. One approach to managing such inconsistencies is to use GO Slim, a list of high level GO terms covering all three GO categories (GO Slim, 2003). The GO Slim terms convey biological meaning at a coarser level of resolution and each fine-level annotation can be mapped to a GO Slim term before performing the functional analysis. To support such an approach, CLENCH can also accept a list of GO categories at a coarser resolution level and automatically map the annotation returned by TAIR to the coarse level terms. In order to map the fine-level gene annotation to arbitrary coarse levels, CLENCH uses a local installation of the GO database (Gene Ontology database, 2003, http://www.godatabase.org/dev/database) and searches the path, made by the parent–child relationships between terms, from the fine level annotation towards the root of each GO category. The first term from the coarse terms list found in the path is assigned as the annotation of the gene. In cases of multiple parents for a fine level term, the coarse term nearest to the fine level annotation or farthest from the root will be assigned as the annotation of the gene. Such mapping to coarse terms is particularly useful when the annotation is very sparse and direct analysis results in a long list of categories found ‘enriched’, but each containing just one gene.

CLENCH is written in Perl 5.6, runs on Windows NT/2000 and XP platforms and is available free to academic users. It uses a local MySQL installation of the GO database for mapping TAIR annotations to user-defined coarser levels. These installation files are available from the GO consortium and MySQL is available from www.mysql.com (both free). CLENCH does not use local annotation files; instead it fetches annotations from TAIR during execution and hence updates made by TAIR are immediately passed on to users. We believe that CLENCH will aid *Arabidopsis* researchers in performing functional category analysis of microarray data using GO annotations.

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. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

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

GO Slim (2003) European Bioinformatics Institute, Saccharomyces Genome Database.

Khatri,P., Draghici,S., Ostermeier,G.C. and Krawetz,S.A. (2002) Profiling gene expression using onto-express. *Genomics*, **79**, 266–270.

Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G., Montoya,M. *et al.* (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res.*, **31**, 224–228.

Zeeberg,B.R., Feng,W., Wang,G., Wang,M.D., Fojo,A.T., Sunshine,M., Narasimhan,S., Kane,D.W., Reinhold,W.C., Lababidi,S. *et al.* (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.*, **4**, R28.