BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 6 2004, pages 979–981
DOI: 10.1093/bioinformatics/bth040

Ontologizing gene-expression microarray data: characterizing clusters with Gene Ontology

Peter N. Robinson[1,*], Andreas Wollstein[1], Ulrike Böhme[1] and Brad Beattie[2]

[1]Institute of Medical Genetics, Charité University Hospital, Humboldt University, Augustenburger Platz 1, 13353 Berlin, Germany and [2]Department of Neurology, Memorial Sloan Kettering Cancer Center, New York, NY, USA

Received on September 26, 2003; accepted on November 18, 2003
Advance Access publication February 5, 2004

ABSTRACT

Summary: An XML-based Java application is described that provides a function-oriented overview of the results of cluster analysis of gene-expression microarray data based on Gene Ontology terms and associations. The application generates one HTML page with listings of the frequencies of explicit and implicit Gene Ontology annotations for each cluster, and separate, linked pages with listings of explicit annotations for each gene in a cluster.

Availability: http://www.charite.de/ch/medgen/ontologizer

Contact: peter.robinson@charite.de

INTRODUCTION

Microarray technologies for transcriptional profiling allow researchers to extract genome-wide expression data, and numerous methods have been developed to analyze this type of data, including clustering procedures that divide genes into disjoint or overlapping sets representing all or part of the genes being analyzed (Slonim, 2002).

A major challenge for researchers who analyze gene-expression microarray data is to make biological sense out of the large amounts of information proceeding from these experiments. The interpretation of these experiments can be facilitated by well-presented functional annotations, which provide an overview of the functions that predominate in clusters as well as functional annotations for each gene. In this report, we present the Ontologizer, a Java application that creates an XML tree representing the Gene Ontology terms to which the genes of individual clusters have been annotated and outputs the results as a set of interlinked HTML page. These pages allow users to visualize the functional pages of clusters and member genes (or gene products) at a glance.

DESCRIPTION

Gene Ontology: background

The Gene Ontology (GO) Consortium provides a structured, controlled vocabulary for describing the roles of genes and gene products in any organism. Three separate ontologies are provided: Biological process ('P') describes the biological objective to which the gene product contributes, molecular function ('F') describes the biochemical activity of a gene product and cellular component ('C') refers to the place in the cell in which a gene product exerts its activity (Ashburner *et al*., 2000). The terms in the GO database form a directed acyclic graph, in which terms are children of one or several more general terms. For instance, the term 'DNA replication' (GO:0006260) is a child of the term 'DNA replication and cell cycle' (GO:0000067) and also of the term 'S phase of mitotic cell cycle' (GO:0000084).

The GOs do not themselves describe specific genes or gene products. Rather, collaborating databases generate gene association files consisting of links between genes or gene products and GO terms. Genes and gene products are annotated at the most specific level possible, but are considered to share the attributes of all the parent nodes (Dwight *et al*., 2002). At the time of the writing of this paper, association files have been made available for about 20 organisms, including human, mouse, yeast and *Caenorhabditis elegans*.

Design and usage

The Ontologizer is a Java 1.4 application that generates a set of interlinked HTML pages that present three different 'views' of the functional annotations for the genes or gene products represented in the clusters being analyzed. The first view ('explicit annotations') presents a count of the number of explicitly annotated GO terms per cluster. As mentioned above, gene products are annotated at the most specific level possible but are considered to share the attributes of all their parent terms. The second view ('implicit annotations') therefore provides a count of the parent terms of the explicitly annotated terms of the first view. In cases where there is more than one path through the directed acyclic graph structure of GO from a term to a parent term, the parent term is counted only once. The main output page of the Ontologizer displays these views for each of the clusters being analyzed and is linked to additional HTML pages for each cluster with a third

*To whom correspondence should be addressed.

cluster04 (n = 213) Explicit Annotations Implicit Annotations Details

Explicit Annotations

Name	ID	Count	Aspect
cell motility	GO:0006928	42	P
structural molecule activity	GO:0005198	42	F
protein amino acid dephosphorylation	GO:0006470	7	P
ATP binding activity	GO:0005524	7	F
protein kinase activity	GO:0004672	5	F
protein amino acid phosphorylation	GO:0006468	5	P
intracellular signaling cascade	GO:0007242	5	P
binding activity	GO:0005488	4	F
transport	GO:0006810	4	P
mitochondrial inner membrane	GO:0005743	4	C

Fig. 1. Sample output of the Ontologizer. Explicit annotations for a cluster of *C.elegans* genes. The first 10 annotations are shown here. Forty-two genes in this cluster have been annotated with GO:0006928, or 'cell motility'; this term has aspect 'P', meaning that it belongs to the Biological Process ontology. Forty-two genes possess the annotation GO:005198, or 'structural molecule activity' ('F' or Molecular Function), seven have been annotated as GO:0006470, or 'protein amino acid dephosphorylation', and so on. The view 'Implicit Annotations' (data not shown) displays a similar listing of the parent terms of the explicitly annotated GO terms. The view 'Details' (data not shown) presents annotations in each of the three GO categories for each gene in the cluster. Full examples of the Ontologizer's output are available on the accompanying website, http://www.charite.de/ch/medgen/ontologizer

view in which all genes or gene products are listed together with their annotations in each of the three GO categories (Fig. 1).

The user is required to download the GO termdb.xml file (which is updated monthly) from the GO website, as well as the association file specific for the species being analyzed (a listing of these files can be found at the main GO website: http://www.geneontology.org). The user must then provide a list of files containing lists of genes for each cluster. Currently supported file types are FASTA and plain text (one gene to a line).

The Ontologizer first parses the termdb.xml file to obtain a list of GO terms with corresponding accession numbers and parent terms, then it parses the cluster files to obtain a list of genes (and optional descriptions) for each file. Finally, it parses the association file to obtain correspondences between GO terms and the specific genes or gene products belonging to the cluster files supplied by the user. An XML tree is produced with data for GO associations of each gene in each cluster. XML allows for a flexible but well-structured intermediate representation of the parse or it can be the output to a file itself.

The source code for the Ontologizer, as well as an executable Java archive (jar) file, and detailed instructions for using the program are provided on the accompanying website.

Explicit and implicit annotations

The program produces listings of explicit annotations (the count reflects the number of genes annotated as having the GO term in question) and implicit annotations derived from traversing the directed acyclic graph from each explicit annotation up to the root node, and counting each visited node once for each explicit term. The implicit annotations may be particularly useful for the recognition of the general functional roles of a cluster of genes.

(Lack of) uniformity in gene nomenclature

The Ontologizer, and indeed any program or database used for the analysis of genes and their annotations, depends on the user entering the gene names according to defined nomenclature. Unfortunately, a multitude of different naming schemes exist for genes and gene sequences (including National Center for Biotechnology Information (NCBI) accession numbers for mRNA and expressed sequence tag (EST) sequences, RefSeqs, gene names and their abbreviations, ENSEMBL Gene IDs, Affymetrix identifiers . . .), and users may wish to analyze data in which genes are denoted according to any one of these schemes. The Ontologizer uses the appropriate *gene_association* file for the species being analyzed. These files contain database-specific identifiers (such as the Mouse Genome Informatics accession number), identifiers that have meaning to biologists (such as Swiss-Prot names) and optional synonyms, and the Ontologizer will recognize all these.

Although many datasets will conform to these norms, there are others that will not. For instance, in order to analyze clusters of mouse genes denoted as accession numbers for mRNA and EST sequences, it will be necessary to transform

the data according to a nomenclature that conforms to that of the gene association file. On the accompanying website, we present several Perl scripts that show how to do this with the help of freely available files from the NCBI ftp site.

DISCUSSION

Many types of cluster analysis of gene expression microarray data produce lists of up to hundreds of genes in groups or clusters of putatively related genes. We developed the Ontologizer as a tool to help in the biological interpretation of such clusters by means of functional annotations based on GO (Ashburner *et al*., 2000; Dwight *et al*., 2002).

One limitation of this method is that the GO annotations often refer to gene products, and not to genes, whose expression is the quantity being measured in gene expression microarray experiments. For species whose gene annotation files actually refer to proteins (e.g. *Homo sapiens*), gene names must first be converted to corresponding protein names. However, we feel that the Ontologizer is a useful tool for the initial, exploratory analysis of gene expression microarray data. It is not limited to the analysis of cluster data, and may be useful for the analysis of any set of genes or proteins for which GO annotations are available.

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.

Dwight,S.S., Harris,M.A. and Dolinski,K. (2002) *Saccharomyces* Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.*, **30**, 69–72.

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