BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 6 2003, pages 788–789
DOI: 10.1093/bioinformatics/btg088

GoFish finds genes with combinations of Gene Ontology attributes

Gabriel F. Berriz[1], James V. White[2, 3], Oliver D. King[1] and Frederick P. Roth[1,*]

[1]Harvard Medical School, Department of Biological Chemistry and Molecular Pharmacology, 250 Longwood Avenue, Boston, MA 02115, USA, [2]JVWhite.Com, 5 Kelly Road, Cambridge, MA 02139, USA and [3]Boston University, Department of Biomedical Engineering, 44 Cummington Street, Boston, MA 02215, USA

Received on August 26, 2002; revised on November 20, 2002; accepted on December 13, 2002

ABSTRACT

Summary: GoFish is a Java application that allows users to search for gene products with particular gene ontology (GO) attributes, or combinations of attributes. GoFish ranks gene products by the degree to which they satisfy a Boolean query. Four organisms are currently supported: *Saccaromyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *M.musculus*.

Availability: GoFish can be used freely through the www as a Java applet for Windows, Mac OS X, or Unix: http://llama.med.harvard.edu/Software.html.

Contact: fritz_roth@hms.harvard.edu

As the volume of sequence data and the number of known genes mount, there is an urgent need for tools that help researchers navigate and retrieve this information. The Gene Ontology Consortium (GO; http://www.geneontology.org) has developed a standardized vocabulary to describe gene products (Ashburner *et al.*, 2000). This vocabulary has been designed to be species-independent, and has been widely adopted. The vocabulary contains over 11 000 attributes and continues to grow.

Efforts are well underway to scour the literature and assign GO attributes to gene products for at least eight model organisms: *Saccaromyces cerevisiae* (Cherry *et al.*, 1998), *Drosophila melanogaster* (The Flybase Consortium, 1999), *M.musculus* (Blake *et al.*, 2002), *Schizosaccharomyces pombe* (http://www.genedb.org/pombe), *A.thaliana* (Huala *et al.*, 2001), *O.sativa* (Ware *et al.*, 2002), *Caenorhabditis elegans* (Stein *et al.*, 2001), and *R.norvegicus* (Twigger, 2002). For these eight organisms, nearly 200 000 explicit annotations have already been made, describing more than 40 000 gene products. The total number of annotations is around 1.3 million if we also count annotations implied by the GO directed acyclic graph (DAG) of attribute relationships (Ashburner *et al.*, 2000).

It is difficult to navigate effectively through all this information without information retrieval software. Several GO browsers are currently available (http://www.geneontology.org/#tools), including some—e.g. the AmiGO Browser (http://www.godatabase.org) or the CGAP GO Browser (http://cgap.nci.nih.gov/Genes/)—that can list the gene products associated with a particular GO attribute. In addition, model organism databases (MODs) using GO terms list these attributes in the context of reports on specific genes. Notably, the Mouse Genome Informatics (MGI) GO Browser (http://www.informatics.jax.org/searches/GO_form.shtml), allows users to search for mouse gene products that have been associated with multiple user-specified GO attributes. But currently none of these tools can search for gene products that satisfy any user-defined Boolean expression containing multiple GO attributes. GoFish provides this functionality.

To use GoFish, the user must first select one of the four organisms currently supported: *S.cerevisiae*, *C.elegans*, *D.melanogaster*, and *M.musculus*. The user can then select one or more GO attributes, in one of two ways: either by using the tree-like attribute browser that is part of GoFish's interface, or by using the search utility to find attributes by name or attribute ID and selecting among the attributes returned by the search.

By default, GoFish performs the Boolean 'AND' query containing all the selected attributes. This means that, by default, GoFish will rank gene products according to how many of the selected attributes they have. This default query, however, can be edited by the user to be any Boolean expression. For example, the user may request a listing of all gene products that have either or both of two particular attributes and not a third attribute.

*To whom correspondence should be addressed.

Downloaded from https://academic.oup.com/bioinformatics/article/19/6/788/234623 by guest on 20 April 2025

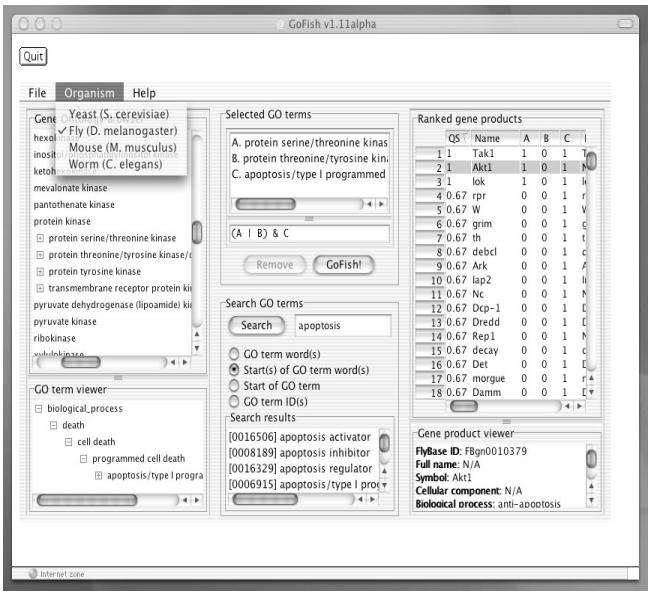

Fig. 1. Sample screen shot of the GoFish GUI.

To understand how GoFish ranks gene products in response to a Boolean query, it is useful to think of each gene product g as having a Boolean value $a(g)$ for each attribute a in the query. $a(g) = 1$ (true) whenever there is an annotation that associates g with a; otherwise $a(g) = 0$ (false). Let Q denote a query, and, for each gene product g, let $Q(g)$ represent the Boolean expression resulting from replacing each attribute a appearing in Q by the corresponding Boolean value $a(g)$. If $Q(g)$ evaluates to 1 (true), then g receives a query satisfaction (QS) score of 1. If $Q(g)$ evaluates to 0 (false), GoFish determines the smallest number n of the query's attributes a for which $a(g)$ needs to be toggled to make $Q(g)$ evaluate to 1. It then assigns to g a QS score of $1 - (n/N)$, where N is the total number of attributes in the query. For example, suppose that Q is $(a_1 \ \& \ a_2) \mid !(a_3 \ \& \ a_4)$ (where '&' means AND, '|' means OR, and '!' means NOT), and suppose that for some gene product g, $a_1(g) = 1$, $a_2(g) = 0$, $a_3(g) = 1$, and $a_4(g) = 1$. Replacing the attributes in the query with these values results in the expression $(1 \ \& \ 0) \mid !(1 \ \& \ 1) = 0$. But reversing the value corresponding to either a_2, a_3, or a_4 makes the expression evaluate to 1. Therefore, for this g, QS = $1 - 1/4 = 0.75$.

Note that for the default 'AND' queries, this QS score corresponds to the fraction of the selected attributes that have been associated with g.

The gene product and attribute data that GoFish uses are obtained from the ontology and gene association files that are periodically released by the Gene Ontology Consortium. These data are pre-processed and bundled with GoFish. The versions of the associations used by GoFish may be found (once an organism is selected) by choosing the 'About GoFish' option under the Help menu.

Explicit associations are made by curators only to the most specific, or 'leafiest', GO attributes for which there is supporting evidence (Ashburner *et al.*, 2000). GoFish expands the explicit associations by propagating them upwards along the GO DAG (Ashburner *et al.*, 2000). This means that if there is an association between a gene product P and a GO attribute A, GoFish associates P also with all of A's ancestors in the GO DAG. For example, searching for gene products associated with the nucleus will retrieve, among other products, those that are associated with the nucleolus.

Although GoFish uses Gene Ontology attributes, the software can easily be extended to use gene product annotations based on additional attribute ontologies (e.g. phenotypes or post-translational modifications). As with the Gene Ontology, these additional ontologies may have a directed acyclic graph structure.

Software allowing a search for genes with any user-defined combination of gene function annotations (for example, the query shown in Fig. 1) has not previously been available. Therefore, we expect that GoFish will be a useful addition to the biologist's toolkit.

ACKNOWLEDGEMENTS

We thank F. Gibbons, D. Miller, C. Brockel, M. Heuer, M. Ashburner, S. Lewis, M. Cherry, J. Blake, and N. Perrimon for encouragement and helpful discussions. This work was sponsored by a grant from Aventis Pharmaceuticals, and by an institutional grant from HHMI. O.D.K. was supported by a fellowship from the NIH/NHGRI.

REFERENCES

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Blake,J.A. *et al.* (2002) The Mouse Genome Database (MGD): the model organism database for the laboratory mouse. *Nucleic Acids Res.*, **30**, 113–115.

Cherry,J.M. *et al.* (1998) SGD: Saccharomyces Genome Database. *Nucleic Acids Res.*, **26**, 73–79.

Huala,E. *et al.* (2001) The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res.*, **29**, 102–105.

Stein,L. *et al.* (2001) Worm Base: network access to the genome and biology of *Caenorhabditis elegans*. *Nucleic Acids Res.*, **29**, 82–86.

The Flybase Consortium (1999) The FlyBase database of the Drosophila Genome Projects and community literature. *Nucleic Acids Res.*, **27**, 85–88.

Twigger,S. *et al.* (2002) Rat Genome Database (RGD): mapping disease onto the genome. *Nucleic Acids Res.*, **30**, 125–128.

Ware,D. *et al.* (2002) Gramene: a resource for comparative grass genomics. *Nucleic Acids Res.*, **30**, 103–105.

Downloaded from https://academic.oup.com/bioinformatics/article/19/6/788/234623 by guest on 20 April 2025