BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 22 2009, pages 3043–3044
doi:10.1093/bioinformatics/btp498

Databases and ontologies

Next generation software for functional trend analysis

Gabriel F. Berriz[1], John E. Beaver[1], Can Cenik[1], Murat Tasan[1] and Frederick P. Roth[1,2,*]

[1]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 250 Longwood Avenue and [2]Center for Cancer Systems Biology, Dana Farber Cancer Institute, 44 Binney Street, Boston, MA 02115, USA

Received on June 29, 2009; revised on July 31, 2009; accepted on August 11, 2009

Advance Access publication August 28, 2009

Associate Editor: Dmitrij Frishman

ABSTRACT

Summary: FuncAssociate is a web application that discovers properties enriched in lists of genes or proteins that emerge from large-scale experimentation. Here we describe an updated application with a new interface and several new features. For example, enrichment analysis can now be performed within multiple gene- and protein-naming systems. This feature avoids potentially serious translation artifacts to which other enrichment analysis strategies are subject.

Availability: The FuncAssociate web application is freely available to all users at http://llama.med.harvard.edu/funcassociate.

Contact: fritz_roth@hms.harvard.edu

1 INTRODUCTION

The task of characterizing large collections of genes and proteins has become routine in modern research. FuncAssociate is a popular web application (5000 hits/month; 150+ citations) designed to facilitate this task.

In its most basic mode of operation, FuncAssociate invites users to submit a set (or ranked list) of genes or proteins arising from their large-scale experiments. It then performs a Fisher's Exact Test (FET) analysis to identify Gene Ontology (GO Ashburner *et al.*, 2000) attributes that describe a fraction of the entries in this set (or toward the top of the list) that would be surprising by chance. The results are corrected for multiple hypotheses via empirical resampling, with significance thresholds that the user may provide. The application also allows the user to account for known ascertainment biases. For full details of the basic mode of computation performed by FuncAssociate, please consult Berriz *et al.* (2003). In this note we focus on features new to FuncAssociate 2 .

(Several computational tools have been developed in recent years to help researchers identify functional trends in the results of large-scale experiments. They have been reviewed in Huang *et al.* (2009), but note that this review erroneously describes FuncAssociate's method for multiple-hypothesis correction as 'Bonferroni; Holm', when in fact FuncAssociate uses a resampling-based method.)

2 NEW FEATURES

More species: FuncAssociate 2 now serves GO associations for 37 species: *Agrobacterium tumefaciens, Anaplasma phagocytophilum, Bacillus anthracis, Bos taurus, Caenorhabditis elegans, Campylobacter jejuni, Candida albicans, Carboxydothermus hydrogenoformans, Clostridium perfringens, Colwellia psychrerythraea, Coxiella burnetii, Danio rerio, Dehalococcoides ethenogenes, Dictyostelium discoideum, Drosophila melanogaster, Ehrlichia chaffeensis, Escherichia coli, Gallus gallus, Geobacter sulfurreducens, Homo sapiens, Hyphomonas neptunium, Leishmania major, Listeria monocytogenes, Magnaporthe grisea, Methylococcus capsulatus, Mus musculus, Neorickettsia sennetsu, Plasmodium falciparum, Pseudomonas aeruginosa, Pseudomonas fluorescens, Rattus norvegicus, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Shewanella oneidensis, Silicibacter pomeroyi, Trypanosoma brucei*, and *Vibrio cholerae*. These associations are obtained directly from the GO Consortium. See Section 3.

More namespaces: For 12 of the supported species, the availability of authoritative ID mappings enables FuncAssociate 2 to now accept queries in any of several different nomenclature schemes, or *namespaces.* For example, for *H.sapiens* FuncAssociate 2 currently accepts queries provided in any one of 23 namespaces, including UniProt accession (e.g. A5D905), RefSeq RNA (e.g. NM_007315), and Ensembl gene ID (e.g. ENSG00000115415). Handling these queries requires either (i) mapping the query gene set to the GO-annotated identifier system ('mapping the input') or (ii) mapping GO annotations to the users preferred namespace ('mapping the annotation'). However, complications may arise because mapping between namespaces is often not one-to-one. Under the mapping-the-input strategy, for example, an apparent enrichment might arise from a single query gene that has multiple synonyms within the GO-annotated namespace. FuncAssociate 2 has adopted the mapping-the-annotation strategy (and is unique among gene enrichment applications in this regard). Issues related to namespace mapping are discussed at length in the documentation (under the heading *'map-the-input' versus 'map-the-associations'*).

Evidence codes: By default FuncAssociate 2 excludes from its analysis GO associations that have evidence codes NR ('Not Recorded') or ND ('No biological Data available'), but users may now search for trends among associations supported by any combination of the 18 evidence codes defined by the GO Consortium (The Gene Ontology, 2009), including NR and ND. This option enables users to, for example, restrict the analysis to GO annotation supported by direct experimentation. Alternatively, it allows the user to cast a broader net by including GO annotation based on reviewed computational analysis (RCA) or other indirect methods.

*To whom correspondence should be addressed.

© The Author 2009. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

User-supplied associations: FuncAssociate 2 may also be used to perform a multiple-hypothesis-adjusted trend analysis for any type of entity using any collection of descriptors. The user simply uploads a file describing the desired associations. FuncAssociate 2 then determines the set of entities and attributes to use for its analysis. The entities need not be genes or proteins, and the attributes need not be GO terms. In fact, the available data from any published large-scale experiment can be readily represented in the form of a set of associations between the entities being assayed (e.g. metabolites) and attributes that correspond to the experiment's measurements (e.g. 'increased abundance in response to IL-2').

It is also now possible to download the associations used by FuncAssociate 2 for any specific run. This, in combination with the above-described ability to upload associations, allows the results of queries (which may be the subject of publications) to be reproduced at a later date despite subsequent updates.

New interface: We have entirely revamped the web interface for FuncAssociate 2 to make it more responsive and easier to navigate, while remaining compatible with a broad range of browsers.

Web-service architecture: The web interface for FuncAssociate 2 is one of the potentially many specialized clients of the FuncAssociate 2 Service. This service understands the JSON-RPC protocol (JSON-RPC Working Group, 2008), and therefore can handle requests from any client that conforms to its API, as described in the FuncAssociate 2 documentation page. Two simple FuncAssociate 2 client modules (written in Perl and Python, respectively) are available for download (see http://llama.med.harvard.edu/software.html). They can be used directly or as templates for similar modules in other languages. Thus, developers of other bioinformatic applications may incorporate FuncAssociate 2 analysis within an application of their own.

3 METHODS

The associations used by FuncAssociate 2 reside in a PostgreSQL database called gofunc, built from several datasets. These are:

The GO DAG: The latest description of the directed acyclic graph (DAG) of GO terms is downloaded from http://archive.geneontology.org/latest-full whenever a new version becomes available.

The Synergizer database: The Synergizer database (Berriz and Roth, 2008) is a repository of mappings between different nomenclature schemes for biological entities such as genes and proteins. We refer to these nomenclature schemes as 'namespaces'. The data in the Synergizer database comes from multiple authorities. Currently, FuncAssociate uses the mappings from the following authorities:CGD (Arnaud *et al.*, 2007) for *C.albicans*, EcoCyc (Keseler *et al.*, 2009) for *E.coli*, Ensembl (Hubbard *et al.*, 2009) for *B.taurus, D.rerio, D.melanogaster, G.gallus, H.sapiens* and *M.musculus*, PomBase (GeneDB; Hertz-Fowler *et al.*, 2004) for *S.pombe*, NCBI (Sayers *et al.*, 2009) for *M.grisea*, RGD (Dwinell *et al.*, 2009) for *R.norvegicus*, SGD (Weng *et al.*, 2003) for *S.cerevisiae* and WormBase (Rogers *et al.*, 2008) for *C.elegans*.

The gene-association files: The latest versions of the GO gene-association files for the supported species are downloaded as they become available from ftp://ftp.geneontology.org/pub/go/gene-associations.

An automated script checks nightly whether any of the component sources listed above has changed, in which case the gofunc database is rebuilt, as follows. First, the associations in the gene-association files are minimally processed and loaded into gofunc. In the process we filter out all associations with a non-empty qualifier field (e.g. NOT, COLOCALIZES_WITH, CONTRIBUTES_TO, and NOT|CONTRIBUTES_TO), a small fraction of the total (0.3%). Next, the associations are 'up-propagated' according to the ancestor–descendant relationships given in the GO DAG. Specifically, for every entity X and associated GO attribute Y, we also associate X with each of the ancestors of Y in the GO DAG. We also up-propagate the supporting evidence codes for each association. Then, for many supported species, associations are mapped to a number of different 'namespaces'. For example, the set of namespaces for human includes HGNC symbols, UniProt accession numbers, Entrez gene identifiers, and Ensembl gene identifiers. This mapping is done using the frequently updated Synergizer database (Berriz and Roth, 2008). Lastly, a complete set of up-propagated, and possibly mapped, associations is stored in gofunc for each available species/namespace combination. The associations are stored in a format optimized to be rapidly read (and possibly filtered) by the FuncAssociate engine.

ACKNOWLEDGEMENTS

We thank Syed Haider and Arek Kasprzyk for help with the BioMart Service, and John Reid for sharing with us his Synergizer service client. For helpful advice and comments on id mappings and the GO associations data, we also thank Tomer Altman, Siddhartha Basu, Ewan Birney, Judith Blake, J. Michael Cherry, Emily Dimmer, Syed Haider, Todd Harris, Pallavi Kaipa, Peter Karp, Donna Maglott, Fiona McCarthy, Quaid Morris, Chris Mungall, Victoria Petri, Monica Romiti, Prachi Shah, David Swarbreck, Majda Valjavec-Gratian, Valerie Wood and Kimberly van Auken. We also thank members of the Roth lab for helpful feedback, and the West Quad Computing Group at Harvard Medical School for computational support.

Funding: US National Institutes of Health (grants NS054052, NS035611, HL081341, HG0017115, HG004233 and HG003224, in part); Canadian Institute for Advanced Research Fellowship (to F.P.R.).

Conflict of Interest: none declared.

REFERENCES

Arnaud,M.B. *et al.* (2007) Sequence resources at the Candida Genome Database. *Nucleic Acids Res.*, **35**, D452–D456.

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

Berriz,G.F. and Roth,F.P. (2008) The Synergizer service for translating gene, protein and other biological identifiers. *Bioinformatics*, **24**, 2272–2273.

Berriz,G.F. *et al.* (2003) Characterizing gene sets with FuncAssociate. *Bioinformatics*, **19**, 2502–2504.

Dwinell,M.R. *et al.* (2009) The Rat Genome Database 2009: variation, ontologies and pathways. *Nucleic Acids Res.*, **37**, D744–D749.

Hertz-Fowler,C. *et al.* (2004) GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Res.*, **32**, D339–D343.

Huang,D.W. *et al.* (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.

Hubbard,T.J.P. *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.

JSON-RPC Working Group (2008) JSON-RPC 2.0 specification proposal. Available at http://groups.google.com/group/json-rpc/web/json-rpc-1-2-proposal?pli=1 (last accessed date June 29, 2009).

Keseler,I.M. *et al.* (2009) EcoCyc: a comprehensive view of Escherichia coli biology. *Nucleic Acids Res.*, **37**, D464–D470.

Rogers,A. *et al.* (2008) WormBase 2007. *Nucleic Acids Res.*, **36**, D612–D617.

Sayers,E.W. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–D15.

The Gene Ontology (2009) Guide to go evidence codes. Available at http://www.geneontology.org/GO.evidence.shtml (last accessed date June 29, 2009).

Weng,S. *et al.* (2003) Saccharomyces Genome Database (SGD) provides biochemical and structural information for budding yeast proteins. *Nucleic Acids Res.*, **31**, 216–218.