BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 18 2003, pages 2502–2504
DOI: 10.1093/bioinformatics/btg363

Characterizing gene sets with FuncAssociate

Gabriel F. Berriz[1], Oliver D. King[1], Barbara Bryant[2], Chris Sander[3] and Frederick P. Roth[1,*]

[1]Harvard Medical School, Department of Biological Chemistry and Molecular Pharmacology, 250 Longwood Avenue, Boston, MA 02115, USA, [2]Millennium Pharmaceuticals, One Kendall Square, Cambridge, MA 02139, USA and [3]Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10021, USA

Received on April 2, 2003; revised on June 26, 2003; accepted on July 8, 2003

ABSTRACT

Summary: FuncAssociate is a web-based tool to help researchers use Gene Ontology attributes to characterize large sets of genes derived from experiment. Distinguishing features of FuncAssociate include the ability to handle ranked input lists, and a Monte Carlo simulation approach that is more appropriate to determine significance than other methods, such as Bonferroni or Šidák *p*-value correction. FuncAssociate currently supports 10 organisms (*Vibrio cholerae*, *Shewanella oneidensis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Arabidopsis thaliana*, *Caenorhaebditis elegans*, *Drosophila melanogaster*, *Mus musculus*, *Rattus norvegicus* and *Homo sapiens*).

Availability: FuncAssociate is freely accessible at http://llama.med.harvard.edu/Software.html. Source code (in Perl and C) is freely available to academic users 'as is'.

Contact: fritz_roth@hms.harvard.edu

The recent proliferation of high-throughput functional genomics methods has brought with it a variety of ways to produce large lists of 'interesting genes' (e.g., a collection of genes with similar expression patterns). These lists of genes immediately prompt the question 'what else do these genes have in common?' The number of genes in such lists is often in the hundreds or thousands and the number of characteristics potentially held in common by them is in the thousands. Several solutions have been offered to address this seemingly simple question including: FunSpec (Robinson *et al.*, 2002), 6046 (http://www.esat.kuleuven.ac.be/~saerts/software/go4g.html), ProToGo (http://www.protogo.cs.huji.ac.il), GoSurfer (http://biosun1.harvard.edu/complab/gosurfer), and goTermFinder (http://genome-www4.stanford.edu/cgi-bin/SGD/GO/goTermFinder). However, each of these has limitations.

Chief among these limitations has been the lack of appropriate corrections for multiple hypothesis testing and the failure to consider which genes were actually assayed in the experiment that produced the list of interesting genes. Here we present FuncAssociate, a web-based application that addresses these problems. FuncAssociate has the added benefit that it can handle ranked lists of 'interesting genes' (e.g., a list of genes ranked by degree of differential expression). FuncAssociate currently supports 10 model organisms: *Vibrio cholerae*, *Shewanella oneidensis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Arabidopsis thaliana*, *Caenorhaebditis elegans*, *Drosophila melanogaster*, *Mus musculus*, *Rattus norvegicus* and *Homo sapiens*. This choice of organisms reflects the ones for which the Gene Ontology (GO) Consortium has published curated gene-attribute association lists. As lists for more organisms become available from the GO Consortium, the range of organisms supported by FuncAssociate will grow accordingly.

FuncAssociate accepts as its primary input a query list of genes Q. (Let us assume this list consists of q genes.) Then, for each GO attribute A, it first determines the number m of genes in Q that have been annotated with attribute A. In a typical run, FuncAssociate then computes the probability $p_+(A)$ (using Fisher's Exact Test) of finding at least m genes with attribute A in a query list length q if the null hypothesis (H_0) is true. (Here the null hypothesis H_0 is that belonging to the query list Q is independent of having attribute A.) If $p_+(A)$ is sufficiently small, then the number of genes in Q having attribute A is considered to be statistically significantly larger than what would be expected if H_0 were true; we say that A is *over-represented* in Q.

(The user may instead, or in addition, request FuncAssociate to rank attributes according to the probability $p_-(A)$ of having *at most* m genes with attribute A in a query of size q if H_0 were true. If $p_-(A)$ is sufficiently low, we say that A is *under-represented* in Q.)

FuncAssociate lets the user request that the list be treated as an ordered set. This option is useful when the user has some reason to rank the genes in the input list in order of interest. An example of this would be a list of genes ranked by significance or fold-change in mRNA abundance between two growth conditions. In this *ordered-input mode*, FuncAssociate will rank GO attributes according to which initial segment of the input list gives the most significant degree of over- or

*To whom correspondence should be addressed.

under-representation. (For each gene g in the original ordered set Q, the *initial segment* corresponding to g is the subset of Q consisting of g plus all the elements that precede g in the order.)

One pitfall in the interpretation of this type of analysis is the multiple testing problem (Westfall and Young, 1993). FuncAssociate provides what we believe is the most accurate criterion of significance to account for multiple hypothesis testing. The number of hypotheses tested in the case of an unordered input list is the number of GO attributes for which there is any evidence for the organism in question (call this number z). In the case of an ordered input list, the number of hypotheses tested is given by the product qz, where q is the size of the input list, since q is also the number of initial segments in the query list. Some tools (such as FunSpec and GO4G) use Bonferroni or Holm adjustments to the computed p-values to correct for multiple hypotheses. While such corrections may be appropriate when the multiple hypotheses are independent, they can be misleading when there is substantial dependence among the hypotheses, as is the case here given the dependence between GO attributes (Gibbons and Roth, 2002).

FuncAssociate estimates an adjusted p-value (p_{adj}) from the results of 1000 simulated null hypothesis queries. (See mathematical details for FuncAssociate at http://llama.med.harvard.edu/FuncAssociate_Methods.html) From these simulations (in which, by design, all null hypotheses hold) we directly estimate the probability of obtaining at least one false positive for any desired cutoff in the *hypothesis-wise* p-value. This probability corresponds exactly to the definition of an *experiment-wise* p-value (Westfall and Young, 1993; Ewens and Grant, 2001). By contrast, methods such as Bonferroni's and Holm's compute upper bounds on this probability, and only under the assumption that all the hypotheses are independent. Šidák's correction computes the probability exactly, but also under the assumption of indepedent hypotheses (Ewens and Grant, 2001).

To illustrate that this theoretical improvement does in fact alter the results, we applied FuncAssociate to the 30 clusters of *S.cerevisiae* genes reported by Tavazoie *et al*. (1999), and compared FuncAssociate's experiment-wise p-values against p-values corrected using Holm's procedure (Holm, 1979). Specifically, we compared the number of 'hits' (rejected null hypotheses) as a function of significance level. As expected, in all cases, and at all significance levels, we found that the number of hits according to FuncAssociate's p-value was equal or greater, sometimes markedly greater, than those obtained with the Holm-adjusted p-value. Out of 30 clusters, we found 17 that showed significant enrichment for some GO attributes. Among these, our method for adjusting p-values resulted in an average of four more hits at the usual 0.05 significance level than when using Holm's correction. The most extreme examples were clusters 2 and 14 where the differences in the number of hits were 19 and 11, respectively. (To count the hits at the 0.05 significance level given by our simulation approach, we ran the simulation protocol 20 times for each cluster, and compared the lowest number of hits to the number of hits obtained by using the Holm-corrected p-values.) We obtained results nearly identical to these when we used a Šidák sequentially-rejective method (Ewens and Grant, 2001; Holland and Copenhaver, 1987) instead of Holm's. (Both Šidák and Holm corrections are less conservative than the original Bonferroni correction under the assumption of independent hypotheses.)

Table 1 illustrates FuncAssociate's output. The query set used for this example is cluster 3 as reported in Tavazoie

Table 1. FuncAssociate sample output (truncated)

No.	m	n	LOD	p_+	p_{adj}	GO attribute
1	22	174	1.081	4.7×10^{-15}	<0.001	Ribosome biogenesis and assembly
2	20	138	1.141	6.5×10^{-15}	<0.001	Ribosome biogenesis
3	22	203	1.003	1.2×10^{-13}	<0.001	Nucleolus
4	18	143	1.061	2.1×10^{-12}	<0.001	Transcription from PolI promoter
5	15	113	1.076	7.8×10^{-11}	<0.001	rRNA processing
6	25	398	0.748	4.3×10^{-10}	<0.001	RNA metabolism
7	43	1124	0.580	5.4×10^{-10}	<0.001	Nucleus
8	23	361	0.746	1.9×10^{-9}	<0.001	RNA processing
9	44	1235	0.547	3.1×10^{-9}	<0.001	Nucleobase, nucleoside, nucleotide and nucleic acid metabolism
10	8	30	1.428	8.3×10^{-9}	<0.001	snoRNA binding
11	8	32	1.391	1.5×10^{-8}	<0.001	Processing of 20S pre-rRNA
12	25	494	0.642	3.7×10^{-8}	<0.001	Transcription, DNA-dependent
13	25	516	0.620	8.7×10^{-8}	<0.001	Transcription
14	7	31	1.333	2.6×10^{-7}	<0.001	Small nucleolar ribonucleoprotein complex
15	35	1005	0.493	5.1×10^{-7}	<0.001	Cell organization and biogenesis
16	28	698	0.538	6.6×10^{-7}	<0.001	Cytoplasm organization and biogenesis
17	17	335	0.615	7.7×10^{-6}	0.001	RNA binding
18	27	809	0.442	3.5×10^{-5}	0.007	Nucleic acid binding

The input is the list of genes in cluster 3 reported by Tavazoie *et al*. (1999) the row's GO attribute (out of 102 in the query set); n: total number of genes having the row's GO attribute; LOD: 10-base logarithm of the 'odds ratio' (Agresti, 2002)

$$\frac{(m+\epsilon)/(q-m+\epsilon)}{(n-m+\epsilon)/(N-q-n+m+\epsilon)},$$

where q is the number of genes in the query set, N is the total number of genes in the organism and ϵ is a pseudo-count with value 0.5; p_+ is the raw (hypothesis-wise) p-value; p_{adj} is the experiment-wise p-value, as described in the text. A few attributes (such as 'obsolete' and 'Gene_Ontology') are excluded from FuncAssociate's analysis; see http://llama.med.harvard.edu/FuncAssociate_Methods.html. Of the remaining GO attributes, FuncAssociate tests the 3151 attributes for which there is at least one positive association with a gene of *S.cerevisiae*. For this many hypotheses, the too-conservative Holm's-corrected p-value for the last entry of this table is 0.11 and, therefore, this hit would have been missed under a typical 0.05 p-value cutoff.

et al. (1999). (This cluster was one of those that remained uncharacterized in the original paper.)

One further FuncAssociate feature not found in other similar tools is the ability to restrict the universe of genes used in the analysis. By default, FuncAssociate uses all genes in the organism for the analysis, but a researcher may only be interested in characterizing a list of genes relative to a more restricted superset, such as, for example, all the genes tested in a particular microarray experiment. This is important since biases in the set of genes examined in a high-throughput experiment will necessarily bias the set of genes returned as interesting.

FuncAssociate handles web-submitted queries consisting of 953 yeast genes in about 8 s (unordered query) and 4 min (ordered query). (The gene space for these queries consists of 6905 genes; 3151 GO attributes are tested per query.)

FuncAssociate's engine is written in C, and the CGI script that handles requests over the web is written in Perl, using standard modules (notably CGI.pm). Source code is freely available to academic users 'as is'.

We plan several improvements for coming versions of FuncAssociate, including giving the users the ability to supply their own attributes (and ontologies), and to provide the option to control for false discovery rate instead of the family-wise error (Yekutieli and Benjamini, 1999).

ACKNOWLEDGEMENTS

We thank C. Reich and M. Kauffman for helpful discussions and M. Keogh and S. Buratowski for helpful feedback on FuncAssociate. We also thank one of our referees for making us aware of false discovery rate control as an alternative to family-wise error control. This work was in part supported by the HHMI Biomedical Research Support Program for Medical Schools. O.D.K. was supported by an NRSA Fellowship from the NIH/NHGRI.

REFERENCES

Agresti,A. (2002) *Categorical Data Analysis*. John Wiley & Sons, Inc., New York.

Ewens,W.J. and Grant,G.R. (2001) *Statistical Methods in Bioinformatics*. Springer-Verlag, New York.

Gibbons,F.D. and Roth,F.P. (2002) Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res.*, **12**, 1574–1581.

Holland,B.S. and Copenhaver,M.D. (1987) An improved sequentially rejective Bonferroni test procedure. *Biometrics*, **43**, 417–423.

Holm,S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, **6**, 65–70.

Robinson,M.D., Grigull,J., Mohammad,N. and Hughes,T.R. (2002) Funspec: a web-based cluster interpreter for yeast. *BMC Bioinformatics*, **3**, 35.

Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, 281–285.

Westfall,P.H. and Young,S.S. (1993) *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. John Wiley & Sons, Inc., New York.

Yekutieli,D. and Benjamini,Y. (1999) Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Stat. Plan. Infer.*, **82**, 171–196.