BIOINFORMATICS APPLICATIONS NOTE
Vol. 25 no. 4 2009, pages 541–542
doi:10.1093/bioinformatics/btn655

Gene expression

SAFEGUI: resampling-based tests of categorical significance in gene expression data made easy

Daniel M. Gatti[1,*], Myroslav Sypa[1], Ivan Rusyn[1], Fred A. Wright[2] and William T. Barry[3]

[1]Department of Environmental Sciences & Engineering, [2]Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599 and [3]Department of Biostatistics & Bioinformatics, Duke University, Durham, NC 27705, USA

Received and revised on November 19, 2008; accepted on December 16, 2008

Advance Access publication December 19, 2008

Associate Editor: David Rocke

ABSTRACT

Summary: A large number of websites and applications perform significance testing for gene categories/pathways in microarray data. Many of these packages fail to account for expression correlation between transcripts, with a resultant inflation in Type I error. Array permutation and other resampling-based approaches have been proposed as solutions to this problem. SAFEGUI provides a user-friendly graphical interface for the assessment of categorical significance in microarray studies, while properly accounting for the effects of correlations among genes. SAFEGUI incorporates both permutation and more recently proposed bootstrap algorithms that are demonstrated to be more powerful in detecting differential expression across categories of genes.

Availability: http://cebc.unc.edu/software/

Contact: fwright@bios.unc.edu; dmgatti@email.unc.edu

1 INTRODUCTION

The computational pipeline for gene expression studies has improved in recent years with a degree of consensus surrounding pre-processing and normalization, analysis of differential expression and biological interpretation of the findings (Allison *et al*., 2006). A typical experiment involves a set of microarrays from one or more treatment groups, where series of gene-specific statistical tests are carried out to produce a list of differentially expressed genes of interest to the investigator. This list is then passed to software which tests for enrichment of biological categories among significant genes, aiding in biological interpretation and generating new leads for followup. Several software packages and websites exist to carry out this type of analysis (Khatri and Draghici, 2005).

As recently noted, software of this type has one property in common; they employ an enrichment score that assumes independence between genes, a property which does not apply in microarray studies (Goeman and Buhlmann, 2007; Rhee *et al*., 2008). The violation of this assumption leads to inflated *P*-values (Barry *et al*., 2008), even for categories that are moderately correlated, and may lead to seriously biased conclusions.

Array permutation, in which the arrays are randomly assigned to treatment groups, produces an empirical null distribution for enrichment statistics, and has been shown to reduce the Type I error associated with overrepresentation analysis (ORA) methods (Barry *et al*., 2005; Goeman and Buhlmann, 2007). Several packages exist to perform permutation-based significance testing (Breslin *et al*., 2004; Subramanian *et al*., 2005). More recently, bootstrapping has been implemented as an alternative and more powerful resampling scheme to test for functional category enrichment (Barry *et al*., 2008). One stumbling block for biologists is that some of these packages require scripting or the use of a command-line interface. SAFEGUI introduces a graphical interface to R code, for loading gene expression datasets and performing resampling-based significance testing of user-specified gene sets, Gene Ontology (GO) categories, Kyoto Encyclopedia of Gene and Genomes (KEGG) pathways or PFAM families.

2 SAFEGUI

SAFEGUI is written in Java in order to provide cross-platform compatibility, and relies upon the Significance Analysis of Function and Expression (SAFE) package (Barry *et al*., 2005) written in R (R Development Core Team, 2006). The release of SAFE 2.0 coincides with the release of SAFEGUI, and adds several new features, including new statistics for differential expression and pathway enrichment, as well as new procedures for error control and resampling. SAFE provides a highly generalized environment for category testing, with a greater variety of options than other resampling category enrichment procedures.

When SAFEGUI starts, the user is presented with the main window. The user selects a data file which consists of a row of sample group labels, followed by gene expression measurements. The user selects the appropriate microarray platform and SAFEGUI automatically retrieves the corresponding annotation data from Bioconductor (Gentleman *et al*., 2004). The user selects a gene-specific 'local' statistic to test for differential expression (e.g. an *F*-statistic in a one-way ANOVA for a multi-class experimental design). Finally, the user can select from several options to test category enrichment correct for multiple testing of categories/pathways. These 'global' enrichment statistics can be categorical, such as Fisher's Exact Test of category membership versus statistical significance of the genes, thus providing a resampling version of common genelist enrichment tests (Khatri and Draghici, 2005). Other global statistics can directly use the full continuous range of the local statistics. Examples include the Wilcoxon rank sum or Kolmogorov–Smirnov statistics.

*To whom correspondence should be addressed.

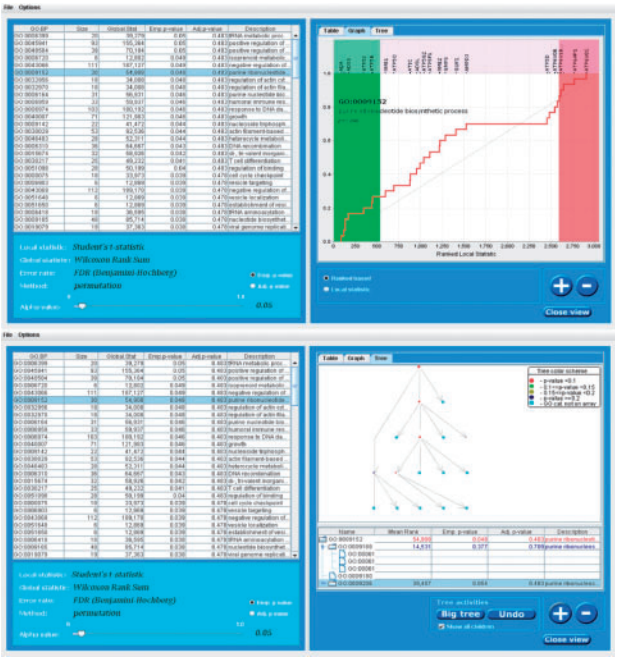

Fig. 1. Results window showing significant GO categories on the left and (**a**) a safe plot for the selected GO category and (**b**) the GO tree with significant categories highlighted.

SAFEGUI enables permutation and bootstrap resampling of arrays, and both methods preserve the correlation structure of the expression measurements. Using existing Bioconductor annotation, SAFEGUI enables testing of GO categories (Ashburner *et al*., 2000), KEGG pathways (Ogata *et al*., 1999) and PFAM motifs (Bateman *et al*., 2000).

Once the options are selected, SAFEGUI performs the analysis and presents the output in results window (Fig. 1a). The table on the left shows the significant categories/pathways sorted by nominal or corrected *P*-value. Upon selecting a category, several tabs appear on the right: the first shows lists of up- and downregulated genes (for two-condition experiments). The second tab shows the *safeplot* (Barry *et al*., 2005), a cumulative distribution of ranked local statistics. The third tab shows a GO graph (if applicable) in which each category is colored according to its statistical significance (Fig. 1b). All tabular and graphical results can be written to output for further analysis.

SAFE 2.0 and SAFEGUI also allow the user to select the univariate Cox proportional hazards model to relate gene expression to censored survival data for category testing. The use of a Cox model on the expression of gene categories has been shown to be more predictive of survival than the use of individual genes, and may provide greater biological insight (Barry *et al*., 2005).

2.1 Bootstrap resampling

The release of the Bioconductor package SAFE 2.0 adds several bootstrap resampling methods for testing gene categories, which have been shown to control Type I error while providing more power than permutation methods (Barry *et al*., 2008). Empirical category *P*-values are calculated using two standard approaches. (i) The argument method = 'bootstrap.t' invokes pivot tests for the exclusion of a null value from Gaussian confidence interval, using on resampled estimates of the mean and variance of the global statistic; (ii) method = 'bootstrap.q' invokes tests based on the exclusion of a null value from the alpha-quantile interval of the resampled global statistic. For both the bootstrap-based tests, a null value for the global statistic must be specified that is insensitive to expression correlation, and bootstrapping is thus limited to the use of the Pearson, Wilcoxon and *T*-tests.

3 CONCLUSIONS

SAFEGUI introduces a user-friendly graphical user interface to a powerful statistical package for the analysis of category or pathway significance from microarray data.

ACKNOWLEDGEMENTS

The research described in this article has not been subjected to the Agency's peer review and policy review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred.

Funding: United States Environmental Protection Agency (RD832720, RD833825 and F08D20579, in part). However, the research described in this article has not been subjected to the Agency's peer review and policy review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred. UNC Environmental Sciences & Engineering Interdisciplinary Fellowship (to D.M.G.).

Conflict of Interest: none declared.

REFERENCES

Allison,D.B. *et al*. (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.*, **7**, 55–65.

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

Barry,W.T. *et al*. (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, **21**, 1943–1949.

Barry,W.T. *et al*. (2008) A statistical framework for testing functional categories in microarray data. *Ann. Appl. Stat.*, **2**, 286–315.

Bateman,A. *et al*. (2000) The Pfam protein families database. *Nucleic Acids Res*., **28**, 263–266.

Breslin,T. *et al*. (2004) Comparing functional annotation analyses with Catmap. *BMC Bioinformatics*, **5**, 193.

Gentleman,R.C. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Goeman,J.J. and Buhlmann,P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.

Khatri,P. and Draghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.

Ogata,H. *et al*. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*., **27**, 29–34.

R Development Core Team (2006) R: a Language and Environment for Statistical Computing, Vienna, Austria, available at http://www.R-project.org.

Rhee *et al*. (2008) Use and misuse of the gene ontology annotations. *Nat. Rev. Genet.*, **7**, 509–515.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.