BIOINFORMATICS APPLICATIONS NOTE

Vol. 24 no. 1 2008, pages 137–139
doi:10.1093/bioinformatics/btm541

Gene expression

Gene Set Expression Comparison kit for BRB-ArrayTools

Xiaojiang Xu, Yingdong Zhao and Richard Simon*

Biometric Research Branch, National Cancer Institute, 9000 Rockville Pike, Bethesda, MD 20892-7434, USA

Received on July 5, 2007; revised on August 27, 2007; accepted on October 22, 2007

Advance Access publication November 15, 2007

Associate Editor: David Rocke

ABSTRACT

Summary: A Gene Set Expression Comparison kit is developed as a module of BRB-ArrayTools for discovering biologically meaningful patterns in gene expression data. The kit consists of gene sets of transcription factor (TF) targets, gene sets containing genes whose protein products share the same protein domain and gene sets of microRNA targets. Using this module of BRB-ArrayTools, researchers can efficiently analyze pre-defined sets of gene whose expression is correlated with a categorical quantitative phenotype or patient survival.

Availability: Gene Set Expression Comparison kit is freely available as a module of BRB-ArrayTools for non-commercial users. BRB-ArrayTools is available at http://linus.nci.nih.gov/BRB-ArrayTools.html.

Contact: rsimon@mail.nih.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Discovering biologically meaningful gene patterns is very important in analyzing genome-wide transcription profiles. Instead of simply enumerating a list of genes that are differentially expressed between pre-specified classes of samples (e.g. tumor or normal), researchers are more interested in determining how those genes interact as parts of complexes, pathways and networks. Several approaches have been developed to utilize functional annotations of genes in interpreting microarray data (Curtis *et al.*, 2005; Draghici *et al.*, 2003; Khatri and Draghici, 2005; Khatri *et al.*, 2002; Manoli *et al.*, 2006; Pavlidis *et al.*, 2002). However, efficient and convenient approaches for utilization of functional information in the data analysis of data from gene expression arrays are still lacking. Here, we describe a Gene Set Expression Comparison kit that enables gene set enhancement types of analyses to be conducted based on transcription factor target gene sets, microRNA target gene sets and gene sets whose corresponding proteins contain a defined protein domain. We have incorporated this Gene Set Expression Comparison kit into BRB-ArrayTools, which is an integrated package for the visualization and statistical analysis of DNA microarray gene expression data (Simon *et al.*, 2006). BRB-ArrayTools contains multiple statistical methods for evaluating the significance of gene expression for gene sets in class comparison, correlation with a quantitative variable or correlation with a censored survival time. This new feature helps the users to identify biologically meaningful gene sets that account for the variation in gene expression in supervised analyses.

2 PREDEFINED GENE SETS

(1) Gene sets that contain genes whose protein products share a common domain. Pfam (Finn *et al.*, 2006) and SMART (Letunic *et al.*, 2006) protein domain links in Swiss-Prot/TrEMBL Protein knowledgebase (Boeckmann *et al.*, 2003) are used to group genes into sets. Proteins encoded by genes in each set contain the same domain. Pfam and SMART are high quality manually curated protein domain databases. Six hundred and thirty-seven human gene sets and 708 mouse gene sets are created based on Pfam annotations; 337 human gene sets and 349 mouse gene sets are created based on SMART annotations.

(2) Gene sets of TF targets. All genes in each gene set are either predicted or experimentally verified to be targets of the same TF. Predicted targets were obtained using the web-based software MATCH (Kel *et al.*, 2003) to search the upstream sequences of genes (~1500 bp) that we obtained from the EnsEMBL (Birney *et al.*, 2004) database. The search utilized TF binding weight matrices obtained from the TRANSFAC (Matys *et al.*, 2003) database and the MATCH cutoffs to minimize the number of false positive targets. With this approach, each set contains genes that are predicted to be potential targets of the same TF. Sixty-eight predicted gene sets for human and 49 gene sets for mouse are created. Moreover, separate sets of genes that have been experimentally verified as targets of the same TF are included. Curation information in the Transcriptional Regulatory Element Database (TRED) (Jiang *et al.*, 2007) is used to eliminate targets without any experimental verification. One hundred and thirty experimentally verified gene sets for human and 115 gene sets for mouse are collected.

(3) Gene sets of predicted microRNA targets. The predicted microRNA target gene information in the miRBase Targets database (Griffiths-Jones *et al.*, 2006) is used to group genes into sets. Genes in each set are predicted to be potential targets of the same microRNA.

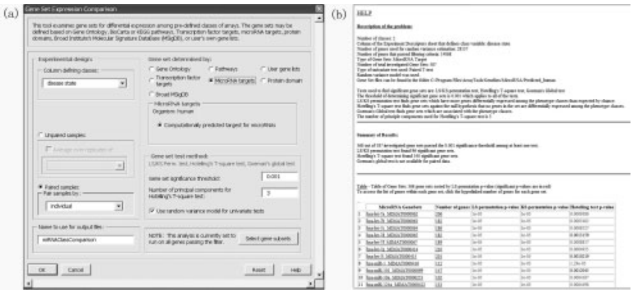
*To whom correspondence should be addressed.

The prediction steps include first detecting potential binding sites with a large degree of complementary to the microRNA, followed by filtering out those sites that do not appear to be conserved in multiple species. Five hundred and eighty-seven predicted gene sets for human and 576 predicted gene sets for mouse are included.

3 IMPLEMENATION

The Gene Set Expression Comparison kit is developed as a module of BRB-ArrayTools. This system uses Excel as front end, powerful R statistical system for analysis and Java applications environment for visualization. Background FORTRAN functions are used for computationally intensive calculations. The tool analyzes pre-defined gene sets for differential expression among phenotype classes using several statistical approaches. It identifies gene sets that contain more differentially expressed genes among the phenotype classes than would be expected by chance. Users can select one class of gene sets as input for analysis (Fig. 1a), e.g. computationally predicted targets for microRNAs. Several statistical methods are used for identifying differentially expressed gene sets (Pavlidis *et al*., 2002). With one of the statistical methods incorporated, first a P-value is computed for each gene in a gene set using a random variance model for univariate tests (Wright and Simon, 2003). Then, the set of P-values for a gene set is summarized by the LS and Kolmogorov–Smirnov (KS) summary statistics. For a set of N genes, the LS statistic is defined as the mean negative natural logarithm of the P-values of the appropriate single gene univariate tests. The KS statistic is defined as the maximum difference between i/N and P_i, where P_i is the ith smallest P-value of the univariate tests. Finally, the statistical significance (P-value) of a gene set containing N genes is evaluated based on computing the empirical distribution of these summary statistics in random samples of N genes. The tests are applied separately to each gene set. A gene set is selected if its corresponding LS or KS summary P-value is below the threshold specified by the user (default is 0.005). The multivariate Hoteling's T^2 analysis is also used to evaluate the statistical significance of a gene set (Kong *et al*., 2006). This approach is based on analysis of the largest principal components of the expression levels of the genes in the set. The differential expression of these principal components among the classes is compared to its null distribution using Hoteling's T^2 test. If the P-value is below the threshold specified by the user (default is 0.005), this gene set is selected. A default significance threshold of 0.005 is employed but can be changed by the user by typing in the input box provided. A default of 0.005 results in an expected number of five false positive gene sets per 1000 gene sets examined. Using a small significance threshold is an easy way to provide some control on the multiplicity of testing for multiple gene sets.

The output is presented to users in HTML files (Fig. 1b). A table of selected significant gene sets provides the P-values for LS and KS tests and Hoteling's T^2 analysis. The selected gene sets are incrementally ordered by the P-value for the LS test with links to the websites containing the detailed information of the functional pattern. For each gene set, the table lists the unique gene sets name, the number of genes represented on the array that belongs to the set and the P-values. In addition, users can obtain annotations for all genes in the pre-defined functional pattern by clicking the link to the gene sets file. Supplementary tables list all significant genes found in the selected functional patterns with numerous annotations for these genes and links to websites containing additional information. For each class, the geometric means of gene expression value are also provided.



This tool examines gene sets for differential expression among pre-defined classes of arrays. The gene sets may be defined based on Gene Ontology, BioCarta or KEGG pathways, Transcription factor targets, microRNA targets, protein domains, Broad Institute's Molecular Signature DataBase (MSigDB), or user's own gene lists.

Experimental design:

Column defining classes: disease state

○ Unpaired samples

● Paired samples

Pair samples by: individual

Gene set determined by:

○ Gene Ontology ○ Pathways ○ User gene lists

○ Transcription factor targets ● MicroRNA targets ○ Protein domain

○ Broad MSigDB

MicroRNA targets

Organism: human

● Computationally predicted targets for microRNAs

Gene set test methods

LS/KS Perm. test, Hotelling's T-square test, Goeman's global test

Gene set significance threshold: 0.001

Number of principal components for Hotelling's T-square test: 3

☑ Use random variance model for univariate tests

Name to use for output files: miRNAClassComparison

NOTE: This analysis is currently set to run on all genes passing the filter.

HELP

Description of the problem:

[illegible]

Summary of Results:

[illegible]

	MicroRNA Genebank	Number of genes	LS permutation p-value	KS permutation p-value	Hotelling test p-value
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Fig. 1. (**a**) Screen shot of input window for Gene Set Expression Comparison kit. Three gene sets are provided with two subsets for each selection. Statistic significance threshold of LS/KS test and Hotelling's T^2 test can be specified by users. In the case shown here, the gene set of MicroRNAs target and the subset of computationally predicted targets are selected; the cutoff of P-values is set to 0.001. (**b**) Screen shot of outputs html file.

4 EXAMPLE

We examined differential gene expression between the NCI-60 cell lines containing mutations in the p53 gene and those not containing such mutations (Subramanian *et al*., 2005). We applied the Gene Set Expression Comparison kit to identify verified TF functional gene sets that were differentially expressed. Fourteen functional gene sets were identified as significant at the 0.005 significance level among the 107 tested. One set was that of p53 itself. Among the others the target gene set of E2F-1, JUN, NFIC, SP1 and CEBPB were identified. These TFs are known to be related with p53. For example, E2F-1 interacts with p53 to regulate transcription of some genes including Apaf-1 (Moroni *et al*., 2001). NFIC acts as a cofactor to regulate the transcription of p53. Transcription of AP1 and SP1 are regulated by p53. CEBPB works together with p53 to regulate transcription of some genes including IL-6 (Margulies and Sehgal, 1993).

ACKNOWLEDGEMENTS

We would like to acknowledge Deyun Pan, Yong Li, Ming-chung Li and Supriya Menezes who have contributed to the development of Gene Set Expression Comparison kit.

Conflict of Interest: none declared.

REFERENCES

Birney,E. *et al*. (2004) Ensembl 2004. *Nucleic Acids Res*., **32**, D468–D470.

Boeckmann,B. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*., **31**, 365–370.

Curtis,R.K. *et al*. (2005) Pathways to the analysis of microarray data. *Trends Biotechnol*., **23**, 429–435.

Draghici,S. *et al*. (2003) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.

Finn,R.D. *et al*. (2006) Pfam: clans, web tools and services. *Nucleic Acids Res*., **34**, D247–D251.

Griffiths-Jones,S. *et al*. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*., **34**, D140–D144.

Jiang,C. *et al*. (2007) TRED: a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Res*., **35**, D137–D140.

Kel,A.E. *et al*. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res*., **31**, 3576–3579.

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

Khatri,P. *et al*. (2002) Profiling gene expression using onto-express. *Genomics*, **79**, 266–270.

Kong,S.W. *et al*. (2006) A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics*, **22**, 2373–2380.

Letunic,I. *et al*. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res*., **34**, D257–D260.

Manoli,T. *et al*. (2006) Group testing for pathway analysis improves comparability of different microarray datasets. *Bioinformatics*, **22**, 2500–2506.

Margulies,L. and Sehgal,P.B. (1993) Modulation of the human interleukin-6 promoter (IL-6) and transcription factor C/EBP beta (NF-IL6) activity by p53 species. *J. Biol. Chem*., **268**, 15096–15100.

Matys,V. *et al*. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*., **31**, 374–378.

Moroni,M.C. *et al*. (2001) Apaf-1 is a transcriptional target for E2F and p53. *Nat. Cell Biol*., **3**, 552–558.

Pavlidis,P. *et al*. (2002) Exploring gene expression data with class scores. *Pac. Symp. Biocomput*., 474–485.

Simon,R.M. *et al*. (2006) Analysis of gene expression data using BRB-Array Tools. *Cancer Inform*., **2**, 1–7.

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.

Wright,G.W. and Simon,R.M. (2003) A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics*, **19**, 2448–2455.