FDR-FET – ~~yet another~~An optimizing gene set enrichment analysis method

Rui-Ru Ji[1*], Karl-Heinz Ott[1], Roumyana Yordanova[1], and Robert E. Bruccoleri[2*]

[1]Applied Genomics, Research and Development, Bristol-Myers Squibb, 311 Pennington-Rocky Hill Road, Pennington, NJ 08534 and

[2]Congenomics, 60 Gates Farm Road, Glastonbury, CT 06033

[*]To whom correspondence should be addressed.

Running Title: FDR-based gene set enrichment analysis

ABSTRACT

Summary:

Gene set enrichment analysis (GSEA) is an important approach in microarray data analysis since it can reveal unifying biological schemes based on previously accumulated knowledge. We describe a new GSEA method, FDR-FET, which employs a False Discovery Rate (FDR) procedure to select a series of differentially expressed gene lists at multiple FDR cutoff values and computes the *P* value of overrepresentation of a gene set using a Fisher's exact test (FET) in each of these gene lists. The lowest P value is retained to represent the significance of the gene set of interest, thereby dynamically setting the most sensitive FDR cutoff. We demonstrate the validity of the method using a published microarray dataset.

Availability: In CPAN with GPL or Perl Artistic License

Contacts: ruiru.ji@bms.com; bruc@acm.org

1 INTRODUCTION

Microarray data analysis usually begins with the generation of a gene list sorted by their fold changes between treatment groups or differential expression *P* values from either a t-test or analysis of variance (ANOVA). Interpretation of the list is often a daunting task, but can be greatly assisted by a group of analytical approaches generally referred to as gene set enrichment analysis (GSEA) (Allison et al 2005). Many varieties of GSEA have been proposed, all of which utilize a priori constructed gene sets that contain related genes with the same annotation such as biological function or chromosome location (Ackermann and Strimmer, 2009). Focusing on gene sets instead of individual genes has obvious advantages. First, it can make use of previously accumulated biological knowledge and thus allow for a more biology-driven analysis. Second, from a statistical point of view it increases power and reduces the dimensionality of the problem.

The general framework and methodology of GSEA approaches have been thoroughly analyzed and discussed recently (Goeman and Buhlmann, 2007; Ackermann and Strimmer, 2009). These methods can be classified as either self-contained or competitive based on the definition of the null hypothesis. A self-contained test compares a gene set to a fixed standard and is not dependent on genes outside of the set. These methods make use of the raw expression data, some of them are based on logistic regression models while others utilize Hotelling's T^2-tests or the more general MANOVA (multivariate analysis of variance) models. By contrast, a competitive test compares the differential expression of a gene set to that of its complement. Majority of the proposed GSEA methods belong to this category. They start with a differentially expressed gene list and test whether a gene set is overrepresented in the list. This is usually achieved by a test of independence in a two by two contingency table, where the test statistic can be constructed based on χ2, hypergeometric, or binomial distribution (Khatri and Draghici, 2005). Since a strict cutoff is needed to obtain the differentially expressed gene list, what criterion constitutes a 'good' cutoff is often debated. Alternative methods have been proposed that make use of the whole vector of *P* values or fold changes. For example, the PAGE (Parametric Analysis of Gene Set Enrichment) method is based on well grounded statistical theory (i.e. the Central Limit Theorem), fully parametric, and computationally efficient as no permutation is needed to derive the gene set *P* value (Kim and Volsky, 2005).

2 DESCRIPTION OF METHOD AND IMPLEMENTATION

We have devised and implemented a new GSEA method, FDR-FET, which belongs to the competitive test category. The key difference between FDR-FET and other competitive methods is that it employs a False Discovery Rate (FDR) procedure to select the differentially expressed gene list. This FDR correction uses the Simes procedure, which employs a series of linearly increasing critical values (Simes, 1986) and has been shown to control the FDR at pre-specified levels for independent test statistics (Benjamini and Hochberg, 1995). Since a single FDR criterion also represents an arbitrary limitation of analysis, we calculate a series of differentially expressed gene lists corresponding to FDRs from 1% to 35% (default; or per user specified). The employment of the FDR procedure and multiple cutoffs provides statistical rigor with additional flexibility to the method.

The overrepresentation of a gene set in a differentially expressed gene list is examined using a Fisher's exact test (FET). We utilize the right test that evaluates the significance of the intersection between two lists for positive association, i.e. an enrichment of elements of list A in list B or *vice versa* (Agresti, 1992). By default there are as many as 35 distinct differentially expressed gene lists and thus up to 35 FETs may be performed for every gene set. The most significant *P* value (i.e. lowest) from these tests is retained as the significance value for the gene set.

Like all other tests based on the two by two contingency table, FDR-FET also relies on the assumption that the observations for every gene are independent and identically distributed (Goeman and Buhlmann, 2007). While it is unrealistic to assume complete independence among genes from the same gene set, it is at minimal desirable to consolidate values for probes representing the same gene. For example, in the case of Affymetrix microarrays, probesets need to be consolidated to the locus level where every locus can be associated with the most significant *P* value among all probesets mapped to the single locus. This consolidation should be performed prior to the FDR-FET analysis.

Background noise is known to have strong impact on the FDR result and needs to be minimized before applying the FDR-FET method. There are many ways to achieve this depending on sample, microarray type and normalization method. One possibility is to sort probesets based on their maximal expression values across all samples, then remove probesets from the bottom using a pre-specified percentile (e.g. 15%). Typically for expression data generated from the MAS 5.0 algorithm, up to 50% of the probesets on the HG-U133A array are flagged as ‘absent’ and may be removed.

We implemented FDR-FET as a P~~p~~erl module (FdrFet) with C inline codes. The module expects two sets of data: a set of genes consisting of gene names and their associated P values from a study of interest and a set of pathways each of which references a subset of the genes. We have also provided an executable program that uses this module which ~~The program~~ expects two input files containing these sets. ~~One of them contains genes with corresponding *P* values from a study of interest; the other file is the gene set file.~~ It is expected that the gene ID is unique for every gene and matches in the two input files. The Perl module will calculate the gene sets ~~Two output files will be generated by FDR-FET. One of them contains the gene sets~~ and their respective *P* values, ~~while the other contains~~ with detailed information of the analysis such as best P value, odd ratio and the corresponding FDR cutoff, numbers in the two by two contingency table, and genes in the overlapping set (between the regulated gene list and the gene set being tested), etc. The Perl module has options controlling the size of “gene universe”, i.e. how many total genes are used for “N” in the Fisher Exact Test, as well as control over whether pathway genes whose expression is unknown are counted as being part of the pathway. The Fisher Exact Test implementation in R (R Development Core Team 2009) was used, and this implementation is based on an elegant computation of binomial coefficients (Loader, 2000). The test data in the module contains the GO pathways and gene P values used in the example in the next section.

~~(Bob, please fill in any performance details and platform issues)~~

3 RESULTS AND DISCUSSION

We tested the FDR-FET method using microarray data from a published study on the cellular effects of three HIV protease inhibitors (Parker et al., 2005). It is well known that patients taking protease inhibitor drugs to treat HIV-AIDS often develop a lipodystrophy-like syndrome such as hyperlipidermia, peripheral lipoatrophy and central fat accumulation (Calza et al., 2004). Parker et al. have shown that protease inhibitors could induce gene expression changes indicative of dysregulation of lipid metabolism, endoplasmic reticulum stress, and metabolic disturbance. These results are consistent with clinical observations and provide basis for a molecular mechanism for the pathphysiology of protease inhibitor-induced lipodystrophy.

The probeset level expression data was generated using the MAS 5.0 algorithm with quantile normalization (Bolstad et al., 2003) and 20% lowest expressed probesets were removed as described above. A one-way ANOVA with "drug treatment" as the factor was performed to generate the input file that contains genes and their associated *P* values. We utilized gene sets from both the Gene Ontology (Ashburner et al., 2000) and KEGG (Kanehisa et al., 2008) and the maximal FDR cutoff was set to 35%. The top 10 gene sets by *P* value are shown in Table 1. It is immediately apparent that this list includes all the major targets of the HIV protease inhibitors including lipid metabolism, amino acid metabolism, gluconeogenesis, and endoplasmic reticulum. By contrast, when a single arbitrary *P* value cutoff (0.01 or 0.05) is used, the compound effect on gluconeogenesis is missed.

Table 1. Top 10 gene sets for the HIV protease inhibitor dataset.

Gene set	Description	*P* value
KEGG:hsa00970	Aminoacyl-tRNA biosynthesis	11.12
GO:0005783	endoplasmic reticulum	11.02
GO:0004812	aminoacyl-tRNA ligase activity	10.27
GO:0006418	tRNA aminoacylation for protein translation	9.96
KEGG:hsa00100	Biosynthesis of steroids	9.63
GO:0008610	lipid biosynthetic process	9.03
GO:0005789	endoplasmic reticulum membrane	8.87
KEGG:hsa00010	Glycolysis / Gluconeogenesis	8.52
GO:0016126	sterol biosynthetic process	8.4
GO:0006695	cholesterol biosynthetic process	6.76

REFERENCES

Ackermann, M and Strimmer, K. (2009) A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10, 47.

Agresti, A. (1992) A survey of exact inference for contingency tables. *Statist. Sci*., 7, 131-153.

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*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*., 25, 25-9.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser*., B 57, 289-300.

Bolstad, B.M. *et al*. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics,* 19, 185-93.

Calza, L. *et al*. (2004) Dyslipidaemia associated with antiretroviral therapy in HIV-infected patients. *J. Antimicrob. Chemother*., 53, :10-4.

Goeman J.J. and Bühlmann, P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23, 980-7.

Kanehisa, M. *et al*. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, 36, D480-D484.

Khatri, P. and Drăghici, S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21, 3587-95.

Kim S.-Y. and Volsky, D.J. (2005) PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*, 6, 144.

Loader, C (2000) Fast and Accurate Computation of Bionomial Probabilities http://projects.scipy.org/scipy/raw-attachment/ticket/620/loader2000Fast.pdf

Parker, R.A. *et al*. (2005) Endoplasmic reticulum stress links dyslipidemia to inhibition of proteasome activity and glucose transport by HIV protease inhibitors. Mol Pharmacol., 67, 1909-19.

R Development Core Team (2009) R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna Austria, http://www.r-project.org.

Simes, R.J. (1986) An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73, 751-754.