Advances and Applications in Bioinformatics and Chemistry

Dovepress

open access to scientific and medical research

Open Access Full Text Article

METHODOLOGY

FDR-FET: an optimizing gene set enrichment analysis method

This article was published in the following Dove Press journal:
Advances and Applications in Bioinformatics and Chemistry
14 March 2011
Number of times this article has been viewed

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

[1]Applied Genomics, Research and Development, Bristol-Myers Squibb, Pennington, NJ, USA; [2]Congenomics, Glastonbury, CT, USA

Abstract: Gene set enrichment analysis for analyzing large profiling and screening experiments can reveal unifying biological schemes based on previously accumulated knowledge represented as "gene sets". Most of the existing implementations use a fixed fold-change or *P* value cutoff to generate regulated gene lists. However, the threshold selection in most cases is arbitrary, and has a significant effect on the test outcome and interpretation of the experiment. We developed a new gene set enrichment analysis method, ie, FDR-FET, which dynamically optimizes the threshold choice and improves the sensitivity and selectivity of gene set enrichment analysis. The procedure translates experimental results into a series of regulated gene lists at multiple false discovery rate (FDR) cutoffs, and computes the *P* value of the 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. We also implemented improved methods to define a more relevant global reference set for the FET. We demonstrate the validity of the method using a published microarray study of three protease inhibitors of the human immunodeficiency virus and compare the results with those from other popular gene set enrichment analysis algorithms. Our results show that combining FDR with multiple cutoffs allows us to control the error while retaining genes that increase information content. We conclude that FDR-FET can selectively identify significant affected biological processes. Our method can be used for any user-generated gene list in the area of transcriptome, proteome, and other biological and scientific applications.

Keywords: gene set enrichment analysis, false discovery rate, Fisher's exact test, microarray profiling, protease inhibitors

Introduction

Expression profiling analysis usually begins with the generation of gene lists ranked by fold-changes or *P* values. Interpretation of the gene lists can be facilitated by analytical approaches such as gene set enrichment analysis,[1] which utilizes a priori constructed reference gene sets that groups genes by classifiers, such as biological function or chromosome location.[2] This type of analysis can help to identify the underlying biological mechanisms and increase the statistical power by reducing the dimensionality of the problem.

The general framework and methodology of gene set enrichment analysis approaches have been thoroughly analyzed and discussed.[2,3] 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 with a fixed standard, and is not dependent on genes outside of the set. These methods make use of the raw

Correspondence: Rui-Ru Ji
Mail Stop 3A0.06, 311 Pennington-Rocky Hill Road, Pennington, NJ 08534, USA
Tel +1 609 818 6036
Fax +1 609 818 3100
Email ruiruji@gmail.com

expression data, and some of them are based on logistic regression models while others utilize Hotelling's t^2-tests or the more general multivariate analysis of variance models.[4,5] By contrast, a competitive test compares the differential expression of a gene set with that of its complement. Most of these methods examine whether regulated genes are over-represented in a given gene set by a test of independence in a two-by-two contingency table, where the test statistic can be constructed based on $\chi 2$, hypergeometric, or binomial distribution.[6] A strict fold-change or *P* value cutoff is needed to obtain the regulated gene list, but the choice of the cutoff is often arbitrary and can have a significant influence on the test outcome and, subsequently, the interpretation of an experiment.[7,8] Alternatively, methods that utilize the whole vector of *P* values or fold-changes have been developed.[9,10] For example, parametric analysis of gene-set enrichment (PAGE) implements a computationally efficient solution based on the central limit theorem to define an enrichment probability.[10]

Implementation

We have implemented a new gene set enrichment analysis method, FDR-FET, which was first described by Ji et al[11] in a transcriptional profiling study of compound dose responses. The current implementation extends the original method and provides options to choose the reference set (ie, "gene universe").

FDR-FET automatically optimizes the cutoff criterion for a gene list (L) under investigation using a false discovery rate (FDR) procedure that employs a series of linearly increasing critical values[12] and has been shown to control the FDR at prespecified levels for independent test statistics.[13] Rather than employing a single FDR criterion that would represent an arbitrary limitation of the analysis, we calculated a series of regulated gene lists (l_i, where $l_i \subset$ L, $1 \leq I \leq 35$), corresponding to FDR cutoff values of 1%–35% (default, or per user-specified) in 1% increments.

We denote the gene set collection as S. The overlap between l_i and a gene set *s* of interest ($s \subset$ S) is examined using a Fisher's exact test (FET). We utilize the right test that evaluates the significance of positive association between two lists, ie, an enrichment of elements of list A (eg, l_i) in list B (eg, *s*) or vice versa.[14] For each *s*, there are as many as 35 FETs to be performed by default, and the most significant *P* value is retained. This procedure is repeated for each gene set *s* in S.

We have implemented FDR-FET as a Perl module (Bio::FDR-FET) with C inline codes. The module expects that gene sets S consisting of gene identifiers and associated classifiers, and gene list L consisting of unique gene identifiers and associated *P* values from a study of interest. We also provide an executable program that uses this module and reads two input files containing these datasets. The Perl module will evaluate each gene set *s* and output detailed analysis information such as best *P* value, odds ratio, and the corresponding FDR cutoff, numbers in the contingency table, and genes in the overlap (between *s* and the l_i with the best *P* value). The C inline code of the Perl module is a slightly modified implementation of the FET code found in R[15] that is based on an elegant computation of binomial coefficients.[16] The test data in the module contains the Gene Ontology pathways and gene *P* values are used in the example in the next section.

Additional options are provided to deal more rigorously with the choice of reference set that has a major influence on the *P* value. We allow four options for the reference set, ie, genes in L ("genes"), union of genes in L and S ("union"), intersection of genes in L and S ("intersection"), and a user-specified arbitrary number ("user"). In particular, the third choice excludes genes with unknown classification from being counted as negative matches, which may be an issue with *P* value calculations. Details of how to use the Perl module can be found by searching for 'Bio::FdrFet' in the CPAN search website (http://search.cpan.org/).

Results and discussion

Here we demonstrate the performance of FDR-FET from three perspectives. First, we assessed the selectivity and sensitivity of the method. Second, we compared FDR-FET with other gene set enrichment analysis methods. Because FDR-FET takes *P* values as input and does not differentiate the directions of gene regulation, we chose two popular implementations of the same category, ie, a simple FET and PAGE. Third, we compared the results generated from different reference set options.

In general, the sensitivity of gene set enrichment analysis can be improved by removal of background noise, which can have a strong impact on the FDR result through removing the bottom n percentile of low intensity probes or probes flagged as "absent", or similar. Consolidation of probes onto the gene level is also recommended to improve independence of measures, which is one assumption of FET.[3] For example, Affymetrix probe sets can be consolidated by associating each gene with the most significant *P* value among all probe sets for the gene. Alternatively, one can utilize the updated probe set definitions, which have been

shown to improve the precision and accuracy of microarray data analysis.[17,18]

We utilized a microarray dataset from a published study on the cellular effects of three human immunodeficiency virus (HIV) protease inhibitors.[19] It is well known that patients taking protease inhibitor drugs to treat HIV-autoimmune deficiency syndrome often develop a lipodystrophy-like syndrome, including hyperlipidemia, peripheral lipoatrophy, and central fat accumulation.[20] Parker et al[19] 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 a basis for a molecular mechanism for the pathophysiology of protease inhibitor-induced lipodystrophy.

The probe set level expression data was generated using the MAS 5.0 algorithm with quantile normalization,[21] and the 20% lowest expressed probe sets were removed. A one-way analysis of variance with respect to the "drug treatment" factor was performed to generate the sorted gene list by *P* values. We utilized gene sets from both the Gene Ontology[22] project and the *Kyoto Encyclopedia of Genes and Genomes* (KEGG).[23]

Validation of FDR-FET

To demonstrate the sensitivity and selectivity of FDR-FET, we generated 1000 randomized gene lists while retaining the same set of *P* values from the analysis of variance. We ran FDR-FET on each of these gene lists using reference set option 1 (ie, "genes") and maximal FDR at 35% for every gene set in KEGG. The 95th and 99th percentiles of the negative log of *P* values were calculated for every gene set, and these values are found to center around 1.9 and 2.6, respectively (Figure 1). As expected, no gene set shows any large deviation from the others. By contrast, the *P* values generated from the real dataset exhibit a nonuniform distribution with only a few highly significant gene sets. Importantly, the top three gene sets with the largest separations from the 99th percentiles are the targets of HIV protease inhibitors, ie, aminoacyl-tRNA biosynthesis (KEGG:hsa00970), biosynthesis of steroids (KEGG:hsa00100), and glycolysis/gluconeogenesis (KEGG:hsa00010).

Comparison of FDR-FET with a simple FET test

Many of the existing gene set enrichment analysis implementations are based on FET with a fixed *P* value or fold-change cutoff. To compare the performance of FDR-FET, which employs a flexible cutoff criterion, with that of a typical

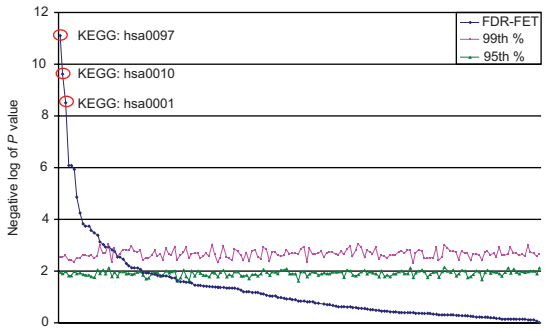

Figure 1 Performance assessment of FDR-FET using simulated datasets. *P* values are calculated for gene sets from the KEGG for each of the 1000 randomized gene lists using FDR-FET (with the option "genes" and maximal FDR 35%). The 95th (red, squares) and 99th (green, triangles) percentiles of the *P* values are calculated for each of the gene sets. Gene sets are ordered by their *P* values calculated from the real dataset (blue, diamonds). The top three gene sets (highlighted in red circles) with the largest separations from the 99th percentiles are the targets of human immunodeficiency virus protease inhibitors, ie, aminoacyl-tRNA biosynthesis (KEGG:hsa00970), steroid biosynthesis (KEGG:hsa00100), and glycolysis/gluconeogenesis (KEGG:hsa00010).
Abbreviations: FDR, multiple false discovery rate; FET, Fisher's exact test; KEGG, *Kyoto Encyclopedia of Genes and Genomes.*

gene set enrichment analysis, we analyzed the regulated gene list generated with an arbitrary FDR cutoff (35%). Table 1 contains the10 most significant gene set hits calculated by FDR-FET using reference set option 1 (ie, "genes") and maximal FDR at 35%. This list includes all the established major targets of the HIV protease inhibitors (lipid metabolism, amino acid metabolism, gluconeogenesis, and endoplasmic reticulum). By contrast, when a single arbitrary FDR cutoff (35%) is used, the effect on gluconeogenesis associated with the pathophysiology of protease inhibitors is missed. Moreover, as depicted in Figure 2, the *P* values for three representative gene sets reach the maximal significance at different FDR cutoffs, demonstrating that the utilization of a flexible cutoff criterion indeed maximizes the signal to noise ratio of a gene list for individual gene sets.

Comparison of FDR-FET with PAGE

PAGE analysis was performed using the whole vector of *P* values from the one way analysis of variance as input. Because PAGE is based on the central limit theorem that requires gene sets to be sufficiently large, we only examined those gene sets with sizes $\geq$10. The negative log of *P* values for three gene sets (ie, GO:0006418, GO:0004812, and KEGG:hsa00970) are set to 20 because they all have a *P* value of zero by PAGE analysis. Again, we could identify all the major targets of HIV protease inhibitors in the top 10 gene set hits from PAGE output (Appendix 1). Interestingly, the results from FDR-FET and PAGE show high concordance, despite the fundamental difference in their underlining methodologies (Figure 3). Using a gene set negative log *P* value cutoff of 3,

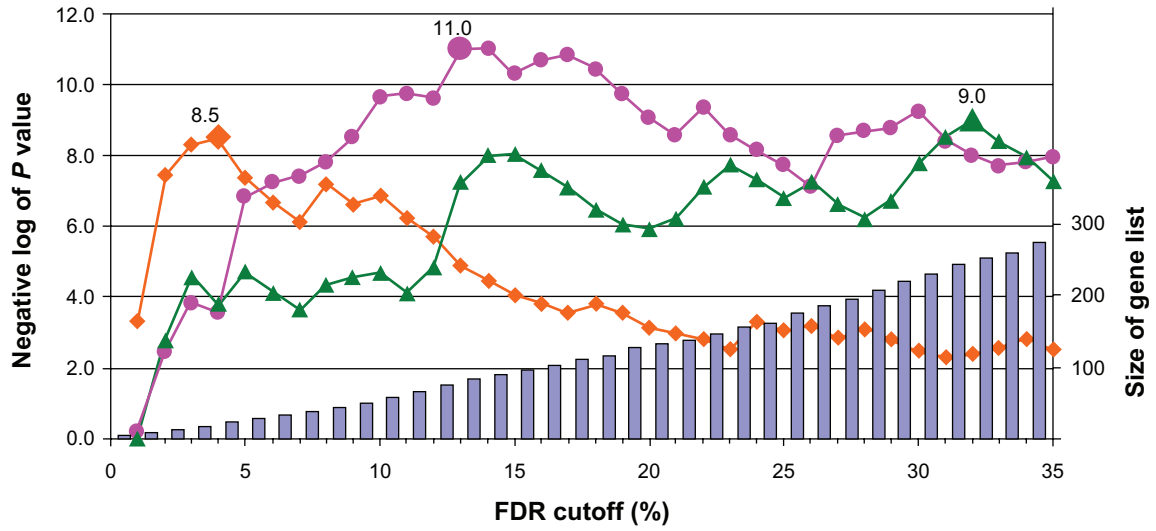

Figure 2 The impact of cutoff criterion on gene set analysis result. The influence of the FDR cutoff on the size of regulated gene list (bars, right axis) and on the significance of selected gene sets (calculated with the option "genes") for the human immunodeficiency virus protease inhibitor experiment, ie, endoplasmic reticulum (GO:0005783; red, circles), lipid biosynthetic process (GO:0008610; green, triangles), and glycolysis/gluconeogenesis (KEGG:hsa00010; orange, diamonds). The highlighted data points indicate the maximal *P* values (labeled) for the respective hits in the gene sets.
Abbreviations: FDR, multiple false discovery rate; KEGG, *Kyoto Encyclopedia of Genes and Genomes*; GO, gene ontology project.

PAGE identified 76 significant affected gene sets, whereas FDR-FET identified 79, among which 63 are shared between the two methods. In particular, the two top 10 hit lists have eight gene sets in common.

Because PAGE is a parametric test, it is generally more liable to gene outliers. In other words, a gene (or a few genes) with a sufficiently large fold-change may lead to significant testing results for the gene set of which the gene is a member. For instance, GO:0008652 and GO:0000049 have highly significant *P* values by PAGE, but only modest *P* values by FDR-FET (Figure 3). A close examination of the genes annotated to these two gene sets reveals that both contain a couple of genes with extremely low *P* values from the analysis of variance test (Appendix 2). By contrast, genes in FET-based methods have equal weight, and the *P* value reflects the gene set enrichment in the regulated gene list, true to the name of gene set enrichment analysis. There are areas where FDR-FET and PAGE can complement each other. For example, FDR-FET is more robust when the gene set size is small and when PAGE cannot produce a reliable *P* value. On the other hand, incomplete gene annotation may affect FET-based methods more than PAGE because lack of knowledge is counted as a "true negative" in the contingency table.

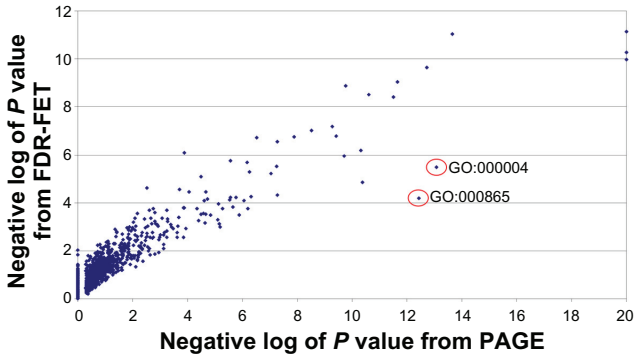

Figure 3. Comparison of the analysis result of FDR-FET with that of PAGE. *P* values are calculated for gene sets from the Gene Ontology and KEGG for the human immunodeficiency virus protease inhibitor experiment using FDR-FET (with the option "genes" and maximal FDR 35%) and PAGE (using the whole vector of gene *P* values as input). Gene sets of size ≥10 are included in the plot.
Abbreviations: FDR, multiple false discovery rate; FET, Fisher's exact test; KEGG, *Kyoto Encyclopedia of Genes and Genomes*; PAGE, parametric analysis of gene-set enrichment.

Comparison of different reference set options

When the biological experiment is performed using a focused gene array (ie, a subset of genes from a genome), the whole genome is used as the reference set, and the number of "true negative" is inflated, leading to unrealistic small *P* values in gene set enrichment analysis outputs. Therefore, one must evaluate what is (close to) the true "universe" for an enrichment analysis. We have introduced new options to address this issue:

- "Genes" whereby all genes tested are counted in the gene set enrichment analysis calculation, assuming that the gene sets are universally representing the genome universe
- "Intersection" can be used when the gene sets are selected to represent a restricted universe, eg, signaling pathways; in this case, only genes that are present in at least one of the signaling pathways are counted

- "Union" represents the general case by which any genes are counted once they are present in either the regulated gene list or the gene sets ("genome as reference set").

In options "genes" and "union", annotated and unannotated genes are both counted in the reference set, while in option "intersection", genes are only counted when they are annotated in at least one of the gene sets. Table 2 contains the 10 most significant gene set hits by the option "genes" and the corresponding *P* values, and ranks by options "union" and "intersection" calculated using maximal FDR at 35%. All three options identified the main HIV protease inhibitor targets, present in the top 10 s, except for gluconeogenesis, which is ranked 12th in results generated from the "union" option. Using a gene set negative log *P* value cutoff of 3, the options "genes" and "intersection" identified similar numbers of affected gene sets, 79 and 73, respectively, among which 71 gene sets are shared between the two hit lists. By contrast, the "union" option identified 96 gene sets, of which 21 are unique to this option and appear to be nonspecific and unrelated to the drug effects upon close examination, suggesting a possible loss of selectivity with this option (Appendix 1). The effect of "intersection" becomes more apparent when smaller gene sets are used. The *P* values and the order of the hits are altered when considering smaller reference sets (Appendix 1 and Appendix 3). By selecting an appropriate reference set, we can enhance the sensitivity and selectivity and reduce the number of spurious hits.

Conclusion

In summary, the employment of FDR and multiple cutoffs provides statistical rigor with additional flexibility. The gene list size is dynamically adjusted so that genes that increase information content are retained, but the addition of noise is limited. This methodology can be applied to results from divergent experiments (eg, hit lists from expression profiling and proteomics studies) as often found in chemogenomics and systems biology approaches.

Authors' contributions

RRJ: method conception and development of Fdr-Fet, data analysis, writing of manuscript; KHO: expert advice of method, conception of altered reference sets, data analysis, writing of manuscript; RY: expert advice of method, data analysis, revision of manuscript; REB: implementation and testing of Fdr-Fet, revision of manuscript. All authors read and approved the final manuscript.

Disclosure

This work was supported by Bristol–Myers Squibb, the past and current employer of the authors.

References

1. Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: From disarray to consolidation and consensus. *Nat Rev Genet*. 2006;7:55–65.
2. Ackermann M, Strimmer K. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*. 2009;10:47.
3. Goeman JJ, Bühlmann P. Analyzing gene expression data in terms of gene sets: Methodological issues. *Bioinformatics*. 2007;23:980–987.
4. Sartor MA, Leikauf GD, Medvedovic M. LR path: A logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics*. 2009;25:211–217.
5. Ucar D, Neuhaus I, Ross-MacDonald P, et al. Construction of a reference gene association network from multiple profiling data: Application to data analysis. *Bioinformatics*. 2007;23:2716–2724.
6. Khatri P, Drăghici S. Ontological analysis of gene expression data: Current tools, limitations, and open problems. *Bioinformatics*. 2005;21:3587–3595.
7. Breitling R, Amtmann A, Herzyk P. Iterative group analysis (iGA): A simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC Bioinformatics*. 2004;5:34.
8. Al-Shahrour F, Díaz-Uriarte R, Dopazo J. Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. *Bioinformatics*. 2005;21:2988–2993.
9. Luo W, Friedman MS, Shedden K, Hankenson KD, et al. GAGE: Generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*. 2009;10:161.
10. Kim SY, Volsky DJ. PAGE: Parametric analysis of gene set enrichment. *BMC Bioinformatics*. 2005;6:144.
11. Ji RR, de Silva H, Jin Y, et al. Transcriptional profiling of the dose response: A more powerful approach for characterizing drug activities. *PLoS Comput Biol*. 2009;5:e1000512.
12. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*. 1986;73:751–754.
13. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J Roy Statist Soc Ser*. 1995;B57:289–300.
14. Agresti A. A survey of exact inference for contingency tables. *Stat Sci*. 1992;7:131–153.
15. R Development Core Team. R: A language and environment for statistical computing. 2009. The R Foundation for Statistical Computing, Vienna, Austria. Available from: http://www.r-project.org. Accessed February 4, 2011.
16. Loader C. Fast and accurate computation of bionomial probabilities. 2000. Available from: http://projects.scipy.org/scipy/raw-attachment/ticket/620/loader2000Fast.pdf. Accessed February 4, 2011.
17. Dai M, Wang P, Boyd AD, et al. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res*. 2005;33:e175.
18. Sandberg R, Larsson O. Improved precision and accuracy for microarrays using updated probe set definitions. *BMC Bioinformatics*. 2007;8:48.
19. Parker RA, Flint OP, Mulvey R, et al. Endoplasmic reticulum stress links dyslipidemia to inhibition of proteasome activity and glucose transport by HIV protease inhibitors. *Mol Pharmacol*. 2005;67:1909–1919.
20. Calza L, Manfredi R, Chiodo F. Dyslipidaemia associated with antiretroviral therapy in HIV-infected patients. *J Antimicrob Chemother*. 2004;53:10–14.
21. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*. 2003;19:185–193.
22. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: Tool for the unification of biology. The gene ontology consortium. *Nat Genet*. 2000;25:25–29.
23. Kanehisa M, Araki M, Goto S, et al. KEGG for linking genomes to life and the environment. *Nucleic Acids Res*. 2008;36:D480–D484.

Appendices

Appendix 1

http://expertbioinfo.com/Papers/FDR_FET/FDR_FET_Appendix1.xls

Appendix 2

http://expertbioinfo.com/Papers/FDR_FET/FDR_FET_Appendix2.xls

Appendix 3

Comparison of the negative log of the *P* values calculated with the option "Genes" and "Intersection". Only pathways associated with KEGG metabolism were chosen for this example. With the "Genes" option, ~11,000 genes were considered for the FDR-FET analysis. With the "Intersection" option only ~1000 genes that are represented at least once in 46 metabolism related pathways were considered.