An Integrated Approach for the Analysis of Biological Pathways using Mixed Models

Lily Wang[1]*, Bing Zhang[2], Russell D. Wolfinger[3], Xi Chen[4]

[1] Department of Biostatistics, Vanderbilt University, Nashville, Tennessee, United States of America, [2] Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, United States of America, [3] SAS Institute, Inc., Cary, North Carolina, United States of America, [4] Department of Quantitative Health Sciences, The Cleveland Clinic, Cleveland, Ohio, United States of America

Abstract

Gene class, ontology, or pathway testing analysis has become increasingly popular in microarray data analysis. Such approaches allow the integration of gene annotation databases, such as Gene Ontology and KEGG Pathway, to formally test for subtle but coordinated changes at a system level. Higher power in gene class testing is gained by combining weak signals from a number of individual genes in each pathway. We propose an alternative approach for gene-class testing based on mixed models, a class of statistical models that:

a) provides the ability to model and borrow strength across genes that are both up and down in a pathway,
b) operates within a well-established statistical framework amenable to direct control of false positive or false discovery rates,
c) exhibits improved power over widely used methods under normal location-based alternative hypotheses, and
d) handles complex experimental designs for which permutation resampling is difficult.

We compare the properties of this mixed models approach with nonparametric method GSEA and parametric method PAGE using a simulation study, and illustrate its application with a diabetes data set and a dose-response data set.

Citation: Wang L, Zhang B, Wolfinger RD, Chen X (2008) An Integrated Approach for the Analysis of Biological Pathways using Mixed Models. PLoS Genet 4(7): e1000115. doi:10.1371/journal.pgen.1000115

Editor: David B. Allison, University of Alabama at Birmingham, United States of America

Received August 30, 2007; **Accepted** June 4, 2008; **Published** July 4, 2008

Funding: LW was partially supported by NICHD Grant 5P30 HD015052-25 and NIH grant1P50 MH078028-01A1. BZ was partially supported by NIH P30 ES000267. XC was partially supported by NHLBI SCCOR Grant 1 P50 HL 077107-01.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: lily.wang@vanderbilt.edu

Introduction

To help increase power to detect microarray differential expression and to better interpret findings, gene-class testing or pathway analysis has become increasingly popular [1]. These approaches allow the integration of gene annotation databases such as Gene Ontology [2] and KEGG Pathway [3] to formally test for subtle but coordinated changes at the system level. Improved power of gene-class testing is gained by combining weak signals from a number of individual genes in each pathway. In addition, pathway analysis has been effectively used to examine common features between data sets [4].

The most commonly used approach for pathway analysis, the enrichment or overrepresentation analysis, uses Fisher's exact test. This method starts with a list of differentially expressed genes based on an arbitrary cutoff of nominal p-values, and compares the number of significant genes in the pathway to the rest of the genes to determine if any gene-set is overrepresented in the significant gene list. The Fisher's exact test is implemented in a number of software packages such as GOTM [5], WebGestalt [6], GENMAPP [7], ChipInfo [8], ONTO-TOOLS [9], GOstat [10], DAVID [11], and JMP Genomics (http://www.jmp.com/genomics). Although straightforward to implement and interpret, this method loses information by using only the significant genes resulted from arbitrarily dichotomizing p-values at some threshold.

More recent approaches such as Gene Set Enrichment Analysis (GSEA) [12,13] and its extensions use continuous distributions of evidence for differential expression and are based on a modified version of the Kolmogorov-Smirnov test that compares the distribution of test statistics in a pathway to the test statistics for the rest of the genes. However, as explained in [14], the specific alternative hypothesis for coordinated association between genes in a gene-set with phenotype is likely to be a location change from background distribution. The Kolmogorov-Smirnov test used by GSEA, which detects any changes in the distribution, is often not optimally powerful for detecting specific location changes. In addition, false positives may result when genes in a gene-set have different variances compared with genes outside the pathway. Methods that test for location changes include PAGE [15] and Functional Class Scoring [16]. PAGE uses normal distribution to approximate test statistics based on differences in means for gene-set genes and other genes; Functional Class Scoring method computes mean (-log(p-value)) from p-values for all genes in a gene-set, and compares this raw score to an empirically derived distribution of raw scores for randomly selected gene-sets of the same size using a statistical resampling approach.