Published in final edited form as:

Stat Biosci. 2014 May 1; 6(1): 38–54. doi:10.1007/s12561-012-9076-3.

Likelihood-Based Approach to Gene Set Enrichment Analysis with a Finite Mixture Model

Sang Mee Lee,

Division of Biostatistics, School of Public Health, University of Minnesota, A460 Mayo Building
MMC 303, 420 Delaware St SE, Minneapolis, MN 55455, USA

Baolin Wu, and

Division of Biostatistics, School of Public Health, University of Minnesota, A460 Mayo Building
MMC 303, 420 Delaware St SE, Minneapolis, MN 55455, USA

John H. Kersey

Masonic Cancer Center, University of Minnesota, Minneapolis, MN 55455, USA

Baolin Wu: baolin@umn.edu

Abstract

In this paper, we study a parametric modeling approach to gene set enrichment analysis. Existing methods have largely relied on nonparametric approaches employing, e.g., categorization, permutation or resampling-based significance analysis methods. These methods have proven useful yet might not be powerful. By formulating the enrichment analysis into a model comparison problem, we adopt the likelihood ratio-based testing approach to assess significance of enrichment. Through simulation studies and application to gene expression data, we will illustrate the competitive performance of the proposed method.

Keywords

Gene set enrichment analysis; Finite mixture model; EM

1 Introduction

Differential gene expression data analysis is a mainstream of microarray experiments. The classical statistical method is to test one gene at a time, compute a p-value for each gene and then adjust to a multiple comparison through controlling the familywise error rate or false discovery rate (FDR, [3]). Although single gene analysis gives many important insights, it has a few limitations [25]. A number of genes which contribute to subtle changes in expression may not be detected because cut-off is determined after a correction for multiple testing. On the other hand, statistical analysis results in a long list of significant genes, and it is not easy to interpret and figure out any genetic patterns. Often a set of genes jointly

influence a biological process or a critical function of metabolic pathways, and single-gene manner may ignore these. Recently many researchers have proposed methods to address challenges of gene set-based analyses. These approaches are often based on gene sets which have already been annotated by functional categories and yield more biologically interpretable result. One of the main research questions for gene set inference is called gene set enrichment analysis (GSEA): we want to evaluate whether the gene set is enriched in terms of certain characteristic of our interest (e.g., differential expression) relative to the other (random) gene sets.

A widely used approach starts from the list of differentially expressed genes derived from single gene analysis, and then evaluates over-representation of a gene set within a list of genes using Fisher's exact test, hypergeometric test, or other independent tests in a 2×2 contingency table. This approach has been modified by many authors (see, e.g., [13] for a review), but the results of significance could be highly dependent on the selected cutoff value and we may lose information as a result of discretizing continuous values. An alternative approach is based on distribution comparisons. Typically a gene score, known as the local statistic for each gene that measures the difference of that gene's expression across different experimental conditions, is computed. Then a gene set score (global statistic) associated with local statistics within a gene set is compared to those of its complement. Several different variations of testing methods have been developed (see, e.g., [2, 19, 22, 25] and [6]). Among the existing methods, the random set-based methods proposed by [9] and [20] have standardized test statistics, which are then compared to random gene sets with significance assessed by permutation and random sampling. These random set-based methods are state-of-the-art currently in the field. In this paper, we will approach the GSEA under a likelihood-based testing framework, and develop a parametric statistical method for enrichment analysis, which could offer very competitive performance by combining information across all genes.

The rest of the paper is organized as follows. Statistical methods are introduced in Sect. 2, and we develop efficient numerical algorithms for model estimation in Sect. 3. Section 4 is devoted to simulation studies and Sect. 5 discusses applications to a leukemia and p53 gene expression data. We end the paper with a discussion in Sect. 6. All technical details are delegated to the Appendix.

2 Statistical Methods

Gene pathway typically consists of a set of genes that jointly influence the system function. Genes are often divided into sets with similar functions based on their annotation information (e.g., the Gene Ontology, [1]). For the following discussion, we will summarize them as providing the gene set information.

Consider a two-class microarray data, and denote the normal transformed two-sample t -statistics for testing differential expressions as z_i for gene $i = 1, \dots, m$. We propose to model z_i with the following finite normal mixture model:

$$\sum_{k=0}^K \theta_k f_k(z), \quad f_k = N(\mu_k, \sigma_0^2), \quad \theta_k > 0, \quad \sum_{k=0}^K \theta_k = 1. \quad (1)$$

Here the first component, $N(\mu_0, \sigma_0^2)$, empirically models null genes, which is different from theoretical null (standard normal distribution) and could take into account the potential dependence among genes [7]. We can interpret θ_0 as the proportion of null genes, and θ_k the proportion of genes with μ_k magnitude of differential expression. In principle, the collection of all μ_k will capture the heterogeneity of differential expressions across all genes. We choose K based on BIC [23].

In enrichment analysis, we try to test whether a given gene set A is significantly different from any random gene set. Note that a random gene set can be treated as a random sampling from all genes. Thus comparing the given set A to a random set is equivalent to comparing A to all genes, which is again equivalent to comparing A to other genes (since A is a subset of all genes). Conceptually the (modified) two-sample t -statistics of genes in a given set can be modeled by a similar finite mixture model with different proportions of each component,

$$\sum_{k=0}^K \nu_k f_k(z), \quad \sum_{k=0}^K \nu_k = 1. \quad (2)$$

Under no enrichment, the gene set A and any random gene set have the same proportion of differentially expressed genes. Therefore gene set A and all the other genes (denoted as A^c) can be modeled, respectively, with

$$\nu_0 f_0(z) + \sum_{k=1}^K \nu_{jk} f_k(z), \quad \nu_0 + \sum_{k=1}^K \nu_{jk} = 1, \quad j=1, 2. \quad (3)$$

Under enrichment, the gene sets A and A^c have different proportions of differentially expressed genes, and hence can be modeled separately with

$$\sum_{k=0}^K \eta_{jk} f_k(z), \quad \sum_{k=0}^K \eta_{jk} = 1, \quad j=1, 2. \quad (4)$$

Enrichment analysis corresponds to evaluating $\eta_{10} = \eta_{20}$, which can be tested by likelihood ratio statistic, e_A , comparing models (3) and (4). The significance of e_A can be approximately assessed using chi-square distribution with one degree of freedom.

Enrichment analysis is an one-sided test: whether gene set A is enriched with more differentially expressed genes compared to a random set. Therefore we adjust the p-value calculation as $0.5 + F(e_A; 1)/2$ when $\hat{\eta}_{10} > \hat{\eta}_{20}$, and $0.5 - F(e_A; 1)/2$ otherwise, where $F(\cdot; df)$ is the χ_{df}^2 distribution function.

In the proposed model, we have assumed that the variance of individual gene test statistics z_i conditional on the mean expression is fixed and the different mixing proportions will capture the varying variation of different gene sets. Therefore it is important that we allow the individual mixing proportions to vary across different gene sets.

In the following we discuss estimation of the empirical null distribution, and EM algorithms [5] for solving the proposed models (1), (2), and (3).

3 Model Estimation

3.1 Empirical Null Distribution Estimation and Finite Mixture Model Fitting

Efron [8] proposed two methods for estimating $(\theta_0, \mu_0, \sigma_0)$: the geometric and analytical approaches. The geometric approach approximates the marginal log density with a quadratic curve near zero. The analytical approach is based on a truncated normal model by assuming non-null distribution has zero support in a pre-chosen small interval around zero. The geometric approach yields almost unbiased estimates if θ_0 exceeds 0.9, but it has large variation for estimating μ_0 . The analytical approach generally gives more stable estimates while it depends on the pre-chosen interval. Both methods have been implemented in the R package, *locfdr*. In our simulation studies, we have observed that the analytical approach gives satisfactory results.

Given K and estimated empirical null distribution parameters $(\hat{\theta}_0, \hat{\mu}_0, \hat{\sigma}_0^2)$, we can estimate (μ_k, θ_k) for model (1) iteratively based on the EM algorithm as follows (see the Appendix for technical details)

$$\theta_k \leftarrow (1 - \hat{\theta}_0) \frac{\sum_{i=1}^m T_{k,i}}{\sum_{j=1}^K \sum_{i=1}^m T_{j,i}}, \quad \mu_k \leftarrow \frac{\sum_{i=1}^m T_{k,i} z_i}{\sum_{i=1}^m T_{k,i}}, \quad k=1, \dots, K,$$

where

$$T_{k,i} = \frac{\theta_k f_k(z_i; \mu_k, \hat{\sigma}_0^2)}{\hat{\theta}_0 f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2) + \sum_{j=1}^K \theta_j f_j(z_i; \mu_j, \hat{\sigma}_0^2)}, \quad k > 0,$$

$$f_j(z; \mu, \sigma^2) = \frac{1}{\sigma} \phi\left(\frac{z - \mu}{\sigma}\right).$$

Here $\phi(\cdot)$ is the standard normal distribution density function.

Occasionally, analytical approach implemented in *locfdr* gives abnormal estimate of $\hat{\theta}_0$ which is larger than 1. We then estimate θ_0 in the EM algorithm together with other parameters as follows:

$$\theta_k \leftarrow \frac{1}{m} \sum_{i=1}^m T_{k,i}, \quad k \geq 0, \quad \mu_k \leftarrow \frac{\sum_{i=1}^m T_{k,i} z_i}{\sum_{i=1}^m T_{k,i}}, \quad k > 0,$$

where

$$T_{0,i} = \frac{\theta_0 f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2)}{\theta_0 f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2) + \sum_{j=1}^K \theta_j f_j(z_i; \mu_j, \hat{\sigma}_0^2)}.$$

3.2 Gene Set Model Fitting

Given estimated $(\hat{\mu}_0, \hat{\sigma}_0^2, \hat{\mu}_k)$ based on all genes, we can estimate the individual model (2) for a given set A iteratively as follows (see the Appendix for technical details)

$$\nu_k \leftarrow \frac{\sum_{i \in A} T_{k,i}}{m_A},$$

where m_A is the size of set A and for gene i in set A

$$T_{k,i} = \frac{\nu_k f_k(z_i; \hat{\mu}_k, \hat{\sigma}_0^2)}{\sum_{j=0}^K \nu_j f_j(z_i; \hat{\mu}_j, \hat{\sigma}_0^2)}.$$

3.3 Model Fitting for a Gene Set and All the Other Genes Under no Enrichment

Under no enrichment, we can similarly estimate the mixture model (3) using the EM algorithm. Denote by A^c the complement of set A . Let

$$T_{0,i} = \frac{\nu_0 f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2)}{\nu_0 f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2) + \sum_{l=1}^K \nu_{1l} f_l(z_i; \hat{\mu}_l, \hat{\sigma}_0^2)}, \quad i \in A,$$

$$T_{k,i} = \frac{\nu_{1k} f_k(z_i; \hat{\mu}_k, \hat{\sigma}_0^2)}{\nu_0 f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2) + \sum_{l=1}^K \nu_{1l} f_l(z_i; \hat{\mu}_l, \hat{\sigma}_0^2)}, \quad i \in A, \quad k > 0,$$

$$T_{0,j} = \frac{\nu_0 f_0(z_j; \hat{\mu}_0, \hat{\sigma}_0^2)}{\nu_0 f_0(z_j; \hat{\mu}_0, \hat{\sigma}_0^2) + \sum_{l=1}^K \nu_{2l} f_l(z_j; \hat{\mu}_l, \hat{\sigma}_0^2)}, \quad j \in A^c,$$

$$T_{k,j} = \frac{\nu_{2k} f_k(z_j; \hat{\mu}_k, \hat{\sigma}_0^2)}{\nu_0 f_0(z_j; \hat{\mu}_0, \hat{\sigma}_0^2) + \sum_{l=1}^K \nu_{2l} f_l(z_j; \hat{\mu}_l, \hat{\sigma}_0^2)}, \quad j \in A^c, \quad k > 0.$$

We then iteratively solve parameters as follows (see the Appendix for technical details)

$$\nu_0 \Leftarrow \frac{1}{m} \left(\sum_{i \in A} T_{0,i} + \sum_{j \in A^c} T_{0,j} \right),$$

$$\nu_{1k} \Leftarrow (1 - \nu_0) \frac{\sum_{i \in A} T_{k,i}}{\sum_{l=1}^K \sum_{i \in A} T_{l,i}},$$

$$\nu_{2k} \Leftarrow (1 - \nu_0) \frac{\sum_{j \in A^c} T_{k,j}}{\sum_{l=1}^K \sum_{j \in A^c} T_{l,j}}.$$

Next we conduct a simulation study to compare the proposed likelihood-based method (denoted as Lrt) to the GSA approach (using the maxmean test statistic) studied at [9].

4 Simulation Study

For 2×10^4 genes from two groups each with n samples, we simulate their expressions based on the conditional normal distribution. Expression variance σ_j^2 is simulated individually for each gene from a χ^2 distribution with 10 degrees of freedom. This mimics the commonly observed large variation of gene variances in microarray data. We simulate the dependence by dividing genes into 200 blocks each with $m_g = 100$ genes and within-block pairwise gene correlation being ρ_g . Gene block correlation parameter ρ_g is randomly simulated from a Beta distribution, Beta(2, 2). We randomly set $m_g \theta_0$ genes in each block as null. The standardized differences of non-null genes, $(\mu_{1j} - \mu_{2j})/\sigma_j$ are randomly simulated from a mixture of two scaled Beta distributions, $0.5 + \text{Beta}(2, 2)$ and $-0.5 - \text{Beta}(2, 2)$, with equal probabilities.

We consider three types of gene set, each with m_e genes and different dependence structures. The first type has similar dependence structure as all genes and is sampled from all $G = 200$ blocks. The other two types of gene set exhibit relatively stronger dependence and are sampled from the first $G = 30$ and 50 blocks respectively. This mimics the commonly observed gene pathways with genes highly interacting with each other. For every type of gene set, we consider two enrichment scenarios. Firstly, the non-null genes in the gene set are randomly sampled from all differentially expressed genes. Secondly, the non-null genes in the gene set are all up-regulated (i.e., the gene set is enriched with different differential expression categories compared to all the other genes).

For size evaluation, we randomly sample $m_e \theta_0$ null and $m_e(1 - \theta_0)$ non-null genes, and compute the enrichment p-values based on Lrt and GSA in each simulation. For power comparison, we consider gene set with randomly sampled $m_e \theta_e$ null and $m_e(1 - \theta_e)$ non-null genes.

In the simulation, we set $n = 15$, $\theta_0 = 0.9$, and consider two sets of scenarios: (1) $\theta_e = (0.86, 0.82, 0.78)$ and $m_e = (100, 200, 300)$, and (2) $\theta_e = (0.8, 0.7, 0.6)$, and $m_e = (10, 20, 50)$, which will investigate the performance under different gene set sizes. In the second scenario with relatively small gene set, θ_e is selected to define a meaningful number of differentially expressed genes.

The proposed Lrt performs better than GSA under all simulation settings, and we have observed similar patterns. Here we report the results for $m_e = (100, 200, 300)$ and non-null genes are sampled from all differentially expressed genes. The complete results are provided at the supplementary materials.

Table 1 summarizes the estimated sizes for true Type I error $\alpha = (0.01, 0.05, 0.10)$ over 1000 simulations. We can see that both methods have approximately the right size. The proposed Lrt in general is more conservative compared to GSA, which could over estimate the Type I error under relatively large significance level.

Figures 12 and 3 summarize the power averaged over 1000 simulations for $m_e = (300, 200, 100)$, respectively. The red solid/dashed/dotted lines are estimated power for Lrt under $\theta_e = (0.86, 0.82, 0.78)$, and black lines are the corresponding power for GSA. Overall we can see that the proposed Lrt has very competitive performance compared to GSA under all settings. In general both methods have reduced power with increasing gene interactions within a given set and decreasing gene set size m_e . With increasing gene set size m_e , we observe relatively larger performance difference between the two methods.

Next we analyze a leukemia and p53 gene expression microarray data to illustrate the relative performance of the proposed likelihood-based method and GSA.

5 Application to Leukemia and p53 Gene Expression Data

The leukemia gene expression data reported at [15] measured the expressions of 45101 genes from five paired controls and Meis1-knockdown cases. We identified 522 gene pathways from C2 functional collection in the *Molecular Signature Database* [25]. Pathway sizes range from 2 to 365 genes. We analyze in total 357 pathways that have more than 10 genes.

To improve the accuracy of the normal distribution approximation, we apply the empirical Bayes modeling approach of [24], which computed a moderated t -statistic, t_i , for gene i by pooling information across all genes for an improved sample variance estimate (implemented in the R package, *limma*). We then apply the normal distribution transformation to the moderated t -statistic, $z_i = \varphi^{-1}(T_d(t_i))$, where $\varphi(\cdot)$ is the standard normal distribution function and $T_d(\cdot)$ is the t -distribution function with d degrees of freedom. Here, the degree of freedom d is estimated from all genes using the empirical Bayes modeling approach.

When applied to the leukemia microarray data, controlling FDR at 0.05/0.1, the proposed Lrt detected 29/51 significant gene sets, while no gene pathway is identified as significant

with GSA. Figure 4 shows the number of significant pathways versus the estimated FDR for Lrt and GSA.

Table 2 lists the top 29 significant pathways identified by the proposed method. Many of them are closely related to cancer development. For example, several identified pathways are related to cell cycle, which is known to play an important role in cancer development: cell cycle machinery controls cell proliferation, and cancer is a disease of inappropriate cell proliferation [4]. The *atrbcrPathway* is also closely related to cell cycle and cancer. Specifically the ATR gene serves as a checkpoint kinase that halts cell cycle progression and induces DNA repair when DNA is damaged. Loss of ATR results in a loss of checkpoint control in response to DNA damage, leading to cell death (see http://www.biocarta.com/pathfiles/h_ATRBRCAPATHWAY.asp). Liu et al. [18] have shown the important role of ATR in cell cycle control in MLL/Meis1 leukemia. The DNA damage signaling pathway is linked to DNA repair, cell-cycle control, growth arrest, and plays an important role in cancer development.

The p53 expression data are available at <http://www.broadinstitute.org/gsea/datasets.jsp>, and consists of 12625 genes from 33 p53 mutant and 17 p53+ cancer cell lines. We analyze in total 453 pathways that have more than 10 genes from the C2 functional collection.

Controlling FDR at 0.01/0.05, the proposed Lrt detected 26/50 significant gene sets, and GSA detected 3/8 significant gene sets. Figure 5 shows the number of significant pathways versus the estimated FDR for Lrt and GSA. Table 3 listed the top ranked pathways by Lrt and GSA (controlling FDR at 0.05).

We can see that the eight significant pathways identified by GSA are all detected by Lrt. Many of the identified pathways by Lrt are Biocarta pathways, e.g., *ceramidePathway*, *fm1ppathway*, *hivnefPathway*, *hsp27Pathway*, *insulinPathway*, *badPathway*, *integrinPathway*, *igf1Pathway*, *g2Pathway*, and *atmPathway* etc. Most of them have been studied and shown related to the p53 gene (see, e.g., [10–12, 14, 16, 17, 21]). For example, the ATM gene interacts with p53 gene to cause the disease ataxia telangiectasia which involves an inherited predisposition to some cancers (http://www.biocarta.com/pathfiles/h_atmPathway.asp). The hsp27 gene modulated the p53 signaling [21]. The igf1Pathway highly interacts with the p53 signaling pathway and they regulate cell growth, proliferation, and death [16]. The g2Pathway consists of genes involved in the cell cycle G2/M checkpoint event, and the p53 gene plays an important role (http://www.biocarta.com/pathfiles/h_g2Pathway.asp).

6 Discussion

The GSEA approach firstly proposed and studied at [19] and [25] provides a very novel way to interpret the large-scale gene expression data. Compared to individual gene oriented analysis, gene set-based inference can often produce meaningful and easy to interpret results and provide additional insights into the underlying biological processes. Many simple and ad hoc statistical methods based on categorization are becoming routinely used in practice (e.g., the widely used hypergeometric testing approach) for gene set significance assessment.

Nonparametric methods based on permutation and random sampling have been proposed and proven to be more powerful but might be quite computing intensive. We approach the GSEA from a likelihood framework and transform it into a model comparison problem, which can be addressed using the powerful likelihood ratio test approach. Through applications and simulation studies we have demonstrated the competitive performance of the proposed method. An interesting extension is to develop similar method for multi-group comparison problems, which can be approached using a finite chi-square distribution mixture model. We will report the results elsewhere in the future.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported in part by a Biomedical Informatics and Computational Biology research grant from the University of Minnesota-Rochester, and National Institute of Health grant CA134848 and GM083345. We are grateful to the University of Minnesota Supercomputing Institute for assistance with the computations. We would like to thank the associate editor and two anonymous referees for their constructive comments, which have dramatically improved the presentation of the paper.

Appendix

EM Algorithm for Estimating the Finite Mixture Model

We begin with the finite mixture model in (1) given $(\hat{\theta}_0, \hat{\mu}_0, \hat{\sigma}_0^2)$ and K . Define indicators $w_{ik} \in \{0, 1\}$ following a multinomial distribution, $\Pr(w_{ik} = 1) = \theta_k, \sum_{k=0}^K w_{ik} = 1$, and conditionally we assume $z_i | w_{ik} = 1 \sim f_k$. The complete data likelihood function for (z_i, w_{ik}) can be written as

$$\prod_{i=1}^m \{\hat{\theta}_0 f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2)\}^{w_{i0}} \prod_{k=1}^K \{\theta_k f_k(z_i; \mu_k, \hat{\sigma}_0^2)\}^{w_{ik}}.$$

In the E-step, the conditional probabilities can be checked to be

$$T_{0,i} = \frac{\hat{\theta}_0 f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2)}{\hat{\theta}_0 f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2) + \sum_{k=1}^K \theta_k f_k(z_i; \mu_k, \hat{\sigma}_0^2)},$$

$$T_{k,i} = \frac{\theta_k f_k(z_i; \mu_k, \hat{\sigma}_0^2)}{\hat{\theta}_0 f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2) + \sum_{j=1}^K \theta_j f_j(z_i; \mu_j, \hat{\sigma}_0^2)}.$$

In the M-step, the conditional expected log likelihood can be checked to be proportional to

$$\sum_{i=1}^m \left\{ T_{0,i} \left(\log \hat{\theta}_0 - \log \hat{\sigma}_0 - \frac{(z_i - \hat{\mu}_0)^2}{2\hat{\sigma}_0^2} \right) + \sum_{k=1}^K T_{k,i} \left(\log \theta_k - \log \hat{\sigma}_0 - \frac{(z_i - \mu_k)^2}{2\hat{\sigma}_0^2} \right) \right\},$$

which can be easily verified to be maximized by

$$\hat{\theta}_k = (1 - \hat{\theta}_0) \frac{\sum_{i=1}^m T_{k,i}}{\sum_{j=1}^K \sum_{i=1}^m T_{j,i}}, \quad \hat{\mu}_k = \frac{\sum_{i=1}^m T_{k,i} z_i}{\sum_{i=1}^m T_{k,i}}, \quad k \geq 1.$$

Given only $(\hat{\mu}_0, \hat{\sigma}_0^2)$ with θ_0 also being a parameter, we have

$$(\hat{\theta}_0, \dots, \hat{\theta}_K) = \arg \max_{\theta_k} \sum_{k=0}^K \sum_{i=1}^m T_{k,i} \log \theta_k,$$

$$\hat{\mu}_k = \arg \min_{\mu_k} \sum_{i=1}^m \sum_{k=1}^K T_{k,i} (z_i - \mu_k)^2, \quad k > 0.$$

We can easily check that

$$\hat{\theta}_k = \frac{1}{m} \sum_{i=1}^m T_{k,i}, \quad k \geq 0, \quad \hat{\mu}_k = \frac{\sum_{i=1}^m T_{k,i} z_i}{\sum_{i=1}^m T_{k,i}}, \quad k > 0.$$

EM Algorithm for Estimating the Gene Set Model

The complete data likelihood function for a gene set A given $(\hat{\theta}_0, \hat{\mu}_0, \hat{\sigma}_0^2, \hat{\mu}_k)$ is

$$\prod_{i \in A} \prod_{k=0}^K \{ \nu_k f_k(z_i; \hat{\mu}_k, \hat{\sigma}_0^2) \}^{w_{ik}}.$$

The conditional expected log likelihood can easily be checked to be

$$\sum_{i \in A} \sum_{k=0}^K T_{k,i} \log \nu_k, \quad T_{k,i} = \frac{\nu_k f_0(z_i; \hat{\mu}_k, \hat{\sigma}_0^2)}{\sum_{j=0}^K \nu_j f_j(z_i; \hat{\mu}_j, \hat{\sigma}_0^2)}, \quad k \geq 0.$$

We can easily verify that

$$\hat{\nu}_k = \frac{\sum_{i \in A} T_{k,i}}{m_A}, \quad k \geq 0.$$

EM Algorithm for Estimating the Model Under no Enrichment

The complete data likelihood can be written as

$$\prod_{i \in A} \{\nu_0 f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2)\}^{w_{i0}} \prod_{k=1}^K \{\nu_{1k} f_k(z_i; \hat{\mu}_k, \hat{\sigma}_0^2)\}^{w_{ik}} \prod_{j \in A^c} \{\nu_0 f_0(z_j; \hat{\mu}_0, \hat{\sigma}_0^2)\}^{w_{j0}} \times \prod_{k=1}^K \{\nu_{2k} f_k(z_j; \hat{\mu}_k, \hat{\sigma}_0^2)\}^{w_{jk}},$$

where $\nu_0 + \sum_{k=1}^K \nu_{lk} = 1, l=1, 2$. The conditional expected log likelihood can be easily checked to be

$$\sum_{i \in A} \left\{ T_{0,i} \log \nu_0 + \sum_{k=1}^K T_{k,i} \log \nu_{1k} \right\} + \sum_{j \in A^c} \left\{ T_{0,j} \log \nu_0 + \sum_{k=1}^K T_{k,j} \log \nu_{2k} \right\},$$

where

$$T_{0,i} = \frac{\nu_0 f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2)}{\nu_0 f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2) + \sum_{l=1}^K \nu_{1l} f_l(z_i; \hat{\mu}_l, \hat{\sigma}_0^2)}, \quad i \in A,$$

$$T_{k,i} = \frac{\nu_{1k} f_k(z_i; \hat{\mu}_k, \hat{\sigma}_0^2)}{\nu_0 f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2) + \sum_{l=1}^K \nu_{1l} f_l(z_i; \hat{\mu}_l, \hat{\sigma}_0^2)}, \quad i \in A, \quad k > 0,$$

$$T_{0,j} = \frac{\nu_0 f_0(z_j; \hat{\mu}_0, \hat{\sigma}_0^2)}{\nu_0 f_0(z_j; \hat{\mu}_0, \hat{\sigma}_0^2) + \sum_{l=1}^K \nu_{2l} f_l(z_j; \hat{\mu}_l, \hat{\sigma}_0^2)}, \quad j \in A^c,$$

$$T_{k,j} = \frac{\nu_{2k} f_k(z_j; \hat{\mu}_k, \hat{\sigma}_0^2)}{\nu_0 f_0(z_j; \hat{\mu}_0, \hat{\sigma}_0^2) + \sum_{l=1}^K \nu_{2l} f_l(z_j; \hat{\mu}_l, \hat{\sigma}_0^2)}, \quad j \in A^c, \quad k > 0.$$

To maximize the conditional log likelihood, we use the Lagrange multiplier method

$$Q = \sum_{i \in A} \left\{ T_{0,i} \log \nu_0 + \sum_{k=1}^K T_{k,i} \log \nu_{1k} \right\} + \sum_{j \in A^c} \left\{ T_{0,j} \log \nu_0 + \sum_{k=1}^K T_{k,j} \log \nu_{2k} \right\} - \lambda_1 \left(\nu_0 + \sum_{k=1}^K \nu_{1k} - 1 \right) - \lambda_2 \left(\nu_0 + \sum_{k=1}^K \nu_{2k} - 1 \right).$$

Setting the gradient vector $\nabla Q = 0$ yields the following equations:

$$\frac{\partial Q}{\partial \nu_0} = \frac{\sum_{i \in A} T_{0,i}}{\nu_0} + \frac{\sum_{j \in A^c} T_{0,j}}{\nu_0} - \lambda_1 - \lambda_2 = 0,$$

$$\frac{\partial Q}{\partial \nu_{1k}} = \frac{\sum_{i \in A} T_{k,i}}{\nu_{1k}} - \lambda_1 = 0, \quad k > 0,$$

$$\frac{\partial Q}{\partial \nu_{2k}} = \frac{\sum_{j \in A^c} T_{k,j}}{\nu_{2k}} - \lambda_2 = 0, \quad k > 0,$$

$$\frac{\partial Q}{\partial \lambda_1} = \nu_0 + \sum_{k=1}^K \nu_{1k} - 1 = 0,$$

$$\frac{\partial Q}{\partial \lambda_2} = \nu_0 + \sum_{k=1}^K \nu_{2k} - 1 = 0.$$

From the first three equations we can obtain

$$\nu_0 = \frac{\sum_{i \in A} T_{0,i} + \sum_{j \in A^c} T_{0,j}}{\lambda_1 + \lambda_2}, \quad \nu_{1k} = \frac{\sum_{i \in A} T_{k,i}}{\lambda_1},$$

$$\nu_{2k} = \frac{\sum_{j \in A^c} T_{k,j}}{\lambda_2}, \quad k > 0.$$

When plugging these into the last two equations, we obtain

$$\hat{\nu}_0 = \frac{\sum_{i \in A} T_{0,i} + \sum_{j \in A^c} T_{0,j}}{m},$$

and

$$\hat{\nu}_{1k} = (1 - \hat{\nu}_0) \frac{\sum_{i \in A} T_{k,i}}{\sum_{l=1}^K \sum_{i \in A} T_{l,i}}, \quad \hat{\nu}_{2k} = (1 - \hat{\nu}_0) \frac{\sum_{j \in A^c} T_{k,j}}{\sum_{l=1}^K \sum_{j \in A^c} T_{l,j}}.$$

References

1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet. 2000; 25(1):25–29. [PubMed: 10802651]

2. Barry WT, Nobel AB, Wright FA. A statistical framework for testing functional categories in microarray data. *Ann Appl Stat.* 2008; 2(1):286–315.
3. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B.* 1995; 57:289–300.
4. Collins K, Jacks T, Pavletich NP. The cell cycle and cancer. *Proc Natl Acad Sci USA.* 1997; 94(7): 2776–2778. [PubMed: 9096291]
5. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B.* 1977; 39(1):1–38.
6. Dørum G, Snipen L, Solheim M, Saebø S. Rotation testing in gene set enrichment analysis for small direct comparison experiments. *Stat Appl Genet Mol Biol.* 2009; 8:34.
7. Efron B. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J Am Stat Assoc.* 2004; 99:96–104.
8. Efron B. Correlation and large-scale simultaneous significance testing. *J Am Stat Assoc.* 2007; 102:93–103.
9. Efron B, Tibshirani R. On testing the significance of sets of genes. *Ann Appl Stat.* 2007; 1(1):107–129.
10. Ferbeyre G, Stanchina ED, Lin AW, Querido E, McCurrach ME, Hannon GJ, Lowe SW. Oncogenic ras and p53 cooperate to induce cellular senescence. *Mol Cell Biol.* 2002; 22(10): 3497–3508. [PubMed: 11971980]
11. Greenway AL, McPhee DA, Allen K, Johnstone R, Holloway G, Mills J, Azad A, Sankovich S, Lambert P. Human immunodeficiency virus type 1 nef binds to tumor suppressor p53 and protects cells against p53-mediated apoptosis. *J Virol.* 2002; 76(6):2692–2702. [PubMed: 11861836]
12. Jiang P, Du W, Wu M. p53 and bad: remote strangers become close friends. *Cell Res.* 2000; 17(4): 283–285. [PubMed: 17404594]
13. Khatri P, Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics.* 2005; 21(18):3587–3595. [PubMed: 15994189]
14. Kim SS, Chae HS, Bach JH, Lee MW, Kim KY, Lee WB, Jung YM, Bonventre JV, Suh YH. p53 mediates ceramide-induced apoptosis in SKN-SH cells. *Oncogene.* 2002; 21(13):2020–2028. [PubMed: 11960374]
15. Kumar AR, Li Q, Hudson WA, Chen W, Sam T, Yao Q, Lund EA, Wu B, Kowal BJ, Kersey JH. A role for MEIS1 in MLL-fusion gene leukemia. *Blood.* 2009; 113(8):1756–1758. [PubMed: 19109563]
16. Levine AJ, Feng Z, Mak TW, You H, Jin S. Coordination and communication between the p53 and IGF-1-AKT-TOR signal transduction pathways. *Genes Dev.* 2006; 20(3):267–275. [PubMed: 16452501]
17. Lewis JM, Truong TN, Schwartz MA. Integrins regulate the apoptotic response to DNA damage through modulation of p53. *Proc Natl Acad Sci USA.* 2002; 99(6):3627–3632. [PubMed: 11904424]
18. Liu H, Takeda S, Kumar R, Westergard TD, Brown EJ, Pandita TK, Cheng EH, Hsieh JJ. Phosphorylation of MLL by ATR is required for execution of mammalian s-phase checkpoint. *Nature.* 2010; 467:343–346. [PubMed: 20818375]
19. Mootha V, Lindgren C, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, Houstis N, Daly M, Patterson N, Mesirov J, Golub T, Tamayo P, Spiegelman B, Lander E, Hirschhorn J, Altshuler D, Groop L. PGC-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet.* 2003; 34:267–273. [PubMed: 12808457]
20. Newton M, Quintana F, den Boon J, Sengupta S, Ahlquist P. Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Ann Appl Stat.* 2007; 1(1):85–106.
21. O’Callaghan-Sunol C, Gabai VL, Sherman MY. Hsp27 modulates p53 signaling and suppresses cellular senescence. *Cancer Res.* 2007; 67(24):11779–11788. [PubMed: 18089808]
22. Pavlidis P, Qin J, Arango V, Mann JJ, Sibille E. Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochem Res.* 2004; 29(6):1213–1222. [PubMed: 15176478]
23. Schwarz G. Estimating the dimension of a model. *Ann Stat.* 1978; 6:461–464.

24. Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 2004; 3:1.
25. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. From the Cover: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*. 2005; 102:15545–15550. [PubMed: 16199517]

Fig. 1. Power of Lrt and GSA averaged over 1000 simulations for $m_e = 300$. The horizontal axis corresponds to type I error

Fig. 2. Power of Lrt and GSA averaged over 1000 simulations for $m_e = 200$. The horizontal axis corresponds to type I error

Fig. 3. Power of Lrt and GSA averaged over 1000 simulations for $m_e = 100$. The horizontal axis corresponds to type I error

Fig. 4.
The number of significant pathways versus FDR for the leukemia data

Fig. 5.

The number of significant pathways versus FDR for the p53 data

Table 1

Estimated type I error of Lrt and GSA over 1000 simulations (listed within parentheses are the standard errors). Non-null genes are randomly sampled from all differentially expressed genes

α		$\hat{\alpha}$			
		0.01	0.05	0.1	
$m_e = 100$	$G = 200$	Lrt	0.002 (5e-5)	0.014 (4e-4)	0.038 (1e-3)
		GSA	0.009 (3e-4)	0.076 (2e-3)	0.176 (5e-3)
	$G = 50$	Lrt	0.005 (2e-4)	0.019 (6e-4)	0.046 (1e-3)
		GSA	0.013 (4e-4)	0.085 (2e-3)	0.172 (5e-3)
	$G = 30$	Lrt	0.005 (2e-4)	0.026 (8e-4)	0.068 (2e-3)
		GSA	0.012 (4e-4)	0.071 (2e-3)	0.156 (4e-3)
$m_e = 200$	$G=200$	Lrt	0.001 (3e-5)	0.011 (3e-4)	0.035 (1e-3)
		GSA	0.012 (4e-4)	0.077 (2e-3)	0.166 (4e-3)
	$G = 50$	Lrt	0.002 (6e-5)	0.034 (1e-3)	0.071 (2e-3)
		GSA	0.012 (4e-4)	0.066 (2e-3)	0.173 (5e-3)
	$G = 30$	Lrt	0.012 (4e-4)	0.046 (1e-3)	0.089 (3e-3)
		GSA	0.009 (3e-4)	0.067 (2e-3)	0.140 (4e-3)
$m_e = 300$	$G = 200$	Lrt	0.001 (3e-5)	0.009 (3e-4)	0.030 (9e-4)
		GSA	0.007 (2e-4)	0.072 (2e-3)	0.158 (4e-3)
	$G = 50$	Lrt	0.007 (2e-4)	0.028 (9e-4)	0.072 (2e-3)
		GSA	0.006 (2e-4)	0.060 (2e-3)	0.149 (4e-3)
	$G = 30$	Lrt	0.020 (6e-4)	0.051 (2e-3)	0.105 (3e-3)
		GSA	0.008 (3e-4)	0.055 (2e-3)	0.126 (3e-3)

Table 2

Top 29 most significant pathways identified with the proposed likelihood-based method

Pathway	# genes	p-value
Cell_Cycle	73	2E-13
CR_CELL_CYCLE	74	5E-11
atrbcaPathway	18	7E-07
CR_REPAIR	35	4E-06
GLUT_DOWN	230	5E-06
cell_cycle_checkpoint	22	1E-05
DNA_DAMAGE_SIGNALING	85	1E-05
HTERT_UP	94	2E-05
CR_DNA_MET_AND_MOD	20	3E-05
LEU_DOWN	130	3E-05
cell_cycle_regulator	20	6E-05
rbPathway	11	0.0001
cell_cycle_arrest	27	0.0005
hdacPathway	28	0.0008
RAP_DOWN	169	0.0010
SA_REG_CASCADE_OF_CYCLIN_EXPR	12	0.0011
il7Pathway	16	0.0015
mRNA_processing	40	0.0018
shh_lisa	15	0.0019
GLUCOSE_DOWN	122	0.0020
MAP00020_Citrate_cycle_TCA_cycle	16	0.0022
celcyclePathway	22	0.0022
mRNA_splicing	45	0.0023
SIG_IL4RECEPTOR_IN_B_LYPHOCYTES	26	0.0025
caspasePathway	21	0.0028
crebPathway	25	0.0030
eif4Pathway	24	0.0034
MAP00240_Pyrimidine_metabolism	38	0.0035
nfatPathway	49	0.0040

Table 3

Significantly enriched pathways for the p53 data identified by Lrt and GSA (FDR = 0.05)

Lrt		
Pathway	# genes	p-value
P53_UP	49	6.3E-09
p53Pathway	43	1.2E-08
rasPathway	41	4.3E-08
GLUT_UP	294	4.0E-07
SA_PROGRAMMED_CELL_DEATH	24	5.5E-07
mitochondriaPathway	32	6.4E-07
HTERT_UP	135	6.7E-07
p53hypoxiaPathway	36	7.3E-07
SA_G1_AND_S_PHASES	26	2.3E-06
ceramidePathway	48	1.2E-05
radiation_sensitivity	61	1.3E-05
fmlppathway	65	1.3E-05
hivnefPathway	100	3.0E-05
DNA_DAMAGE_SIGNALING	154	3.0E-05
hsp27Pathway	34	3.9E-05
XINACT_MERGED	26	3.9E-05
insulinPathway	44	1.1E-04
badPathway	43	1.3E-04
integrinPathway	60	1.6E-04
igf1Pathway	47	2.9E-04
g2Pathway	41	3.4E-04
atmPathway	43	3.6E-04
tcrPathway	85	3.6E-04
Glycogen_Metabolism	50	4.6E-04
tsp1Pathway	17	4.6E-04
tall1Pathway	23	5.0E-04
cdmacPathway	32	8.2E-04
metPathway	71	1.0E-03
at1rPathway	64	1.2E-03
ngfPathway	36	1.3E-03
cxcr4Pathway	41	1.3E-03
bcl2family_and_reg_network	50	1.4E-03
eif2Pathway	12	1.5E-03
mef2dPathway	27	1.5E-03
spryPathway	27	1.8E-03
eea1Pathway	12	1.8E-03
CR_DEATH	114	2.1E-03
pgc1aPathway	35	2.6E-03

Lrt		
Pathway	# genes	p-value
relaPathway	26	2.7E-03
rnaPathway	17	3.2E-03
ecmPathway	36	3.3E-03
INSULIN_2F_UP	200	3.5E-03
pyk2Pathway	57	3.7E-03
SA_FAS_SIGNALING	14	4.0E-03
chemicalPathway	46	4.0E-03
deathPathway	56	4.5E-03
Cell_Cycle	115	4.6E-03
breast_cancer_estrogen_signaling	162	4.7E-03
tollPathway	45	5.1E-03
SA_B_CELL_RECEPTOR_COMPLEXES	46	5.3E-03

GSA			
Pathway	# genes	p-value	Lrt rank
P53_UP	49	8.0E-6	1
p53Pathway	43	9.0E-6	2
p53hypoxiaPathway	36	1.0E-5	8
badPathway	43	1.1E-4	18
radiation_sensitivity	61	3.3E-4	11
SA_PROGRAMMED_CELL_DEATH	24	3.5E-4	5
rasPathway	41	7.2E-4	3
SA_G1_AND_S_PHASES	26	7.4E-4	9