Incorporating Biological Information as a Prior in an Empirical Bayes Approach to Analyzing Microarray Data

Wei Pan

1Division of Biostatistics, School of Public Health, University of Minnesota

December 2004; Revised April 2005

To appear in Statistical Applications in Genetics and Molecular Biology

Correspondence author: Wei Pan
Telephone: (612) 626-2705
Fax: (612) 626-0660
Email: weip@biostat.umn.edu
Address: Division of Biostatistics, MMC 303,
School of Public Health, University of Minnesota,
Minneapolis, Minnesota 55455-0392, U.S.A.
Incorporating Biological Information as a Prior in an Empirical Bayes Approach to Analyzing Microarray Data

ABSTRACT
Currently the practice of using existing biological knowledge in analyzing high-throughput genomic and proteomic data is mainly for the purpose of validations. Here we take a different approach of incorporating biological knowledge into statistical analysis to improve statistical power and efficiency. Specifically, we consider how to fuse biological information into a mixture model to analyze microarray data. In contrast to a standard mixture model where it is assumed that all the genes come from the same (marginal) distribution, including an equal prior probability of having an event, such as having differential expression or being bound by a transcription factor (TF), our proposed mixture model allows the genes in different groups to have different distributions while the grouping of the genes reflects biological information. Using a list of about 800 putative cell cycle-regulated genes as prior biological knowledge, we analyze a genome-wide location data to detect binding sites of TF Fkh1. We find that our proposal improves over the standard approach, resulting in reduced false discovery rates (FDR), and hence it is a useful alternative to the current practice.
INTRODUCTION

With the advent of high-throughput biotechnologies, large amounts of genomic and proteomic data have been accumulating, enabling one to test a null hypothesis against an alternative hypothesis for each of the thousands of genes or proteins. For example, with microarray gene expression data, it has become a common practice to test whether any gene has differential expression. Similarly, with genome-wide location data or ChIP-chip data, a goal is to detect which genes (or more precisely, their promoter regions) are bound by a transcription factor (TF) (Ren et al. 2000; Lee et al. 2000; Iyer et al. 2001; Simon et al. 2001). In these two cases, the null hypothesis $H_{0i}$ for gene $i$ is that gene $i$ either has no expression change or is not bound by the TF whereas the alternative hypothesis $H_{1i}$ is the opposite. A class of powerful statistical approaches to accomplishing these analysis goals are based on a mixture model, where it is assumed that each of the genes has an equal prior probability of $H_{0i}$’s being true (Allison et al, 2002; Broet et al, 2004; Efron et al, 2001; Ghosh 2004; Kendziorski et al, 2002; Liao et al, 2004; Newton et al, 2001, 2004; Pan 2002, 2003; Tusher et al, 2001). However, this mixture model ignores existing biological information and other sources of data. For example, based on the given experimental conditions and known gene functions in the form of the MIPS annotations (Mewes et al, 2002) or Gene Ontology (GO) (Ashburner et al, 2000), we may know a priori that some genes with similar functions or involved in related biological processes are more likely than others to have a correct $H_{0i}$ (or $H_{1i}$). Another well known example is to exploit the fact that gene expression is related to DNA-protein binding (Zhao et al, 2003), thus motivating borrowing information from one source of data to analyze the other. To take advantage of this kind of existing biological knowledge, we propose a more flexible mixture model that can assign different prior probabilities to different genes. For the purpose of illustration, we will analyze a yeast ChIP-chip dataset for TF Fkh1 (Simon et al, 2001). Fkh1 is known to regulate some genes whose expression is cell cycle dependent. Using expression data, Spellman et al (1998) have computationally identified about 800 putative cell cycle-regulated genes. Here we demonstrate that, in a nonparametric empirical Bayes (EB) approach (Efron et al, 2001), by incorporating the prior information about these 800 putative cell cycle-regulated
genes into our mixture model, we can obtain higher posterior probabilities to detect these genes’ being bound by Fkh1 and reduced false discovery rate (FDR) estimates (Benjamini and Hochberg, 1995) when compared to that of using the standard mixture model.

In most of the current approaches, existing biological information has been mainly used to validate analysis results. For example, it has become popular to associate gene annotations in GO to a group of supplied genes (e.g. Khatri et al 2002; Doniger et al 2003; Draghici et al 2003; Zeeberg et al 2003; Zhang et al 2003; Al-Shahrour et al 2004; Cui et al 2004; Robinson et al 2004; Zhong et al 2004). The group can be obtained from analyzing gene expression data, such as a set of differentially expressed genes or a cluster of genes with similar expression patterns. In these approaches, if there are statistically significant enrichments of the genes in one or more GO categories, the group of the supplied genes are regarded as biologically more meaningful. In contrast, our approach is to directly incorporate biological information into statistical analysis to improve statistical efficiency and power for new discoveries. However, once the biological information is incorporated into statistical analysis, it can no longer be used to validate analysis results; other sources of information or more direct validation procedures have to be used. Hence, for a given piece of biological information, the user has to decide at the beginning whether to use it directly in analysis or rather as a validation criterion. We regard our proposal as a complementary method to existing validation approaches.

METHODS

Statistical models

Suppose that $Z_i$ is a test statistic for gene $i$ to test a null hypothesis $H_{0i}$ against an alternative $H_{1i}$ for $i = 1, ..., n$. The test statistic can be as simple as a log-ratio of two-channel intensities, a t-type statistic or any appropriate statistic. In the standard mixture model (e.g. Efron et al 2001; Newton et al 2001; Kendziorski et al 2003; Pan et al 2003), the distribution of $Z_i$ is modeled as a mixture of two subpopulations:

$$f(Z_i) = \pi f_0(Z_i) + (1 - \pi) f_1(Z_i),$$

where $f_0$ corresponds to the distribution of $Z_i$ for genes for which $H_{0i}$ holds while $f_1$ is that
for genes whose $H_{1i}$ holds, and $\pi = Pr(H_{0i})$ is a prior probability of $H_{0i}$ for any gene. A key assumption in the above model is that every gene has an equal prior probability of $H_{0i}$; that is, the prior probability $\pi$ is independent of gene $i$. By the Bayes theorem, we have the posterior probability of $H_{1i}$ to be

$$Pr(H_{1i}|Z_i) = 1 - \frac{\pi f_0(Z_i)}{f(Z_i)},$$

which is to be used to draw inference.

With existing biological knowledge, we may identify a group of genes $G$ such that each gene in $G$ has a higher (or lower) a priori probability that its $H_{0i}$ holds. Hence, to take advantage of such existing information, we propose assigning a different prior probability to the genes in $G$ from that of the other genes not in $G$, leading to our new mixture model: i) for any gene $i$ in $G$,

$$f^1(Z_i) = \pi_1 f_0(Z_i) + (1 - \pi_1)f^1_1(Z_i);$$

and ii) for any gene $i$ not in $G$,

$$f^0(Z_i) = \pi_0 f_0(Z_i) + (1 - \pi_0)f^0_1(Z_i).$$

Again by the Bayes theorem, we have the posterior probability

$$Pr(H_{1i}|Z_i) = \begin{cases} 1 - \frac{\pi_1 f_0(Z_i)}{f^1(Z_i)} & \text{if } i \in G; \\ 1 - \frac{\pi_0 f_0(Z_i)}{f^0(Z_i)} & \text{otherwise}. \end{cases}$$

Note that the above mixture model allows the genes in $G$ and those not in $G$ to have different marginal distributions $f^1$ and $f^0$ even in the extreme case with an equal prior probability $\pi_1 = \pi_0$ because $f^1_1$ can be different from $f^0_1$. Therefore, rather than assuming that all the genes are from the same distribution as in the standard mixture model, our model assumes that only the genes in the same group come from the same distribution, and the grouping of the genes aims to incorporate prior biological information.

**Statistical inference using empirical Bayes**

Statistical inference can be accomplished by applying either parametric methods (e.g. Newton et al 2001; Kendziorski et al 2003), semi-parametric methods (Newton et al 2004)
or nonparametric methods (Tusher et al 2001; Pan et al 2003; Pan 2003). Here we consider a nonparametric EB method (Efron et al, 2001).

Parameter Estimation

All the parameters in the above models may not be known and need to be estimated. In general, $f$ (or $f^1$ or $f^0$) can be empirically estimated based on observed $Z_i$'s of all genes (or genes in or not in $G$), and permutation methods can be used to estimate $f_0$ (Efron et al 2001; Tusher et al, 2001; Pan et al, 2003). Here we follow the approach of Pan et al (2003) to use finite Normal mixture models (McLachlan and Peel 2002) to estimate $f$, $f^1$, $f^0$ and possibly $f_0$. Specifically, we assume

$$f(Z) = \sum_{k=1}^{K} p_k \phi(Z; \mu_k, \sigma_k^2),$$

(1)

where $\phi(Z; \mu_k, \sigma_k^2)$ is the density function for a Normal distribution with mean $\mu_k$ and variance $\sigma_k^2$, $p_k$’s are mixing proportions with constraints i) $0 \leq p_k \leq 1$ for any $1 \leq k \leq K$ and ii) $\sum_{k=1}^{K} p_k = 1$, and the number of components $K$, along with parameters $\Phi_K = \{ (\mu_k, \sigma_k^2, p_k) : i = 1, ..., K \}$, needs to be estimated.

It is well known that, for any given $K \geq 1$, based on the observed $Z_i$’s, the Normal mixture model (1) can be fitted using the EM algorithm (McLachlan and Peel 2002), resulting in the maximum likelihood estimate (MLE) $\hat{\Phi}_K$. To estimate $K$, one can try a series of the mixture models with $K = 1, 2, ..., K_1$ and choose the one that fits the data best. The Bayesian Information Criterion (BIC) is one of the most widely used model selection criteria (Schwarz 1978):

$$BIC_K = -2 \log L(\hat{\Phi}_K) + (3K - 1) \log(n),$$

where $\log L(\hat{\Phi}_K)$ is the maximized log-likelihood with $K$ components. We choose $K_0$ such that $BIC_{K_0} = \max_{1 \leq K \leq K_1} BIC_K$. Substituting $K$ and $\Phi_K$ with $K_0$ and $\hat{\Phi}_{K_0}$ in $f$, we obtain an estimate $\hat{f}$.

We can similarly estimate $f^1$ and $f^0$ using the observed $Z_i$’s of the genes in $G$ and not in $G$ respectively.

In general, we use permutation to construct so-called null statistics $z_i$’s, which are then used to estimate $f_0$. The motivation of constructing $z_i$ is that, under $H_0$, that the mean of $Z_i$
is zero (and with the implicit assumption that $f_0$ is symmetric about its mean), both $Z_i$ and $-Z_i$ have the same distribution $f_0$. Hence we define $z_i = S_iZ_i$ where $S_i = 1$ or $-1$ with an equal probability $1/2$, $S_i$'s are independent of each other and of $Z_i$'s. That is, we randomly keep or flip the sign of the test statistic $Z_i$ to obtain a null statistic $z_i$. Similar permutation methods have been widely used in microarray data analysis (Efron et al, 2001; Pan et al, 2003; Tusher et al; 2001; Xu et al, 2002). As before, we model $f_0$ as a finite Normal mixture and use $z_i$'s to estimate its parameters. In principle, multiple sets of null statistics (based on multiple runs of permutations across all the genes) can be obtained and thus used; in this report, mainly to reduce computational demand in fitting Normal mixture models, we only used one random set of the null statistics. In practice, as expected, because some $H_{0i}$'s do not hold, the above permutation method may lead to conservative inference (Pan 2003).

It is challenging to estimate $\pi$ (or $\pi_1$ or $\pi_0$) in a nonparametric mixture model: as pointed out by Efron et al (2001), $\pi$ is not identifiable. However, several sensible estimators of $\pi$ have appeared. Here we apply a method proposed by Efron et al (2001); other approaches will be discussed later. The basic idea of the approach is quite simple. Based on the mixture model, we have $(1 - \pi)f_1 = f - \pi f_0 \geq 0$ because $f_1$ is a density function. Hence we have $\pi \leq f/f_0$, which motivates the estimator

$$\hat{\pi} = \frac{\int_A \hat{f}(z)dz}{\int_A \hat{f}_0(z)dz},$$

(2)

where $\hat{f}$ and $\hat{f}_0$ are estimated $f$ and $f_0$ respectively, and $A$ is a small interval around 0. We estimate $\pi_1$ and $\pi_0$ similarly by using only the genes in $G$ or not in $G$ respectively.

**FDR and its estimation**

It is critical to have a means to control multiple test errors in genomic- or proteomic discoveries as encountered here. The false discovery rate (FDR) has become increasingly popular for such a purpose (Benjamini and Hochberg 1995; Storey and Tibshirani 2003). A nice property of the posterior probability in an empirical Bayes approach is its connection with FDR. In fact, the posterior probability $Pr(H_{0i}|Z_i) = 1 - Pr(H_{1i}|Z_i)$ can be interpreted as a local FDR (Efron et al 2001). Furthermore, as pointed out by Newton et al (2004), a direct estimator of FDR can be constructed using $Pr(H_{0i}|Z_i)$. Specifically, for any given
cut-off value $c$, we claim that gene $i$ is statistically significant (i.e. rejecting $H_{0i}$) if and only if $Pr(H_{ii}|Z_i) > 1 - c$. Then the corresponding FDR is

$$FDR(c) = \frac{\sum_{i=1}^{n} \beta_i 1(\beta_i \leq c)}{\sum_{i=1}^{n} 1(\beta_i \leq c)}$$

with $\beta_i = Pr(H_{0i}|Z_i)$. Plugging-in parameter estimates, we obtain an estimated FDR.

**Which estimates to use**

For any gene $i$ in the list $G$, applying the standard mixture model to all the genes results in a posterior estimate, say $\hat{Pr}(H_{ii}|Z_i)$. Similarly, in our new method with a subgroup analysis of only the genes in $G$, we obtain another posterior estimate, say $\tilde{Pr}(H_{ii}|Z_i)$. Because a posterior estimate is going to be used to draw inference on $H_{0i}$, and the above two estimates can be quite different, the natural question is which one to be used. Our recommendation is to use the one resulting in a smaller FDR estimate.

**RESULTS AND DISCUSSION**

**Data**

A genome-wide location study was conducted by Simon et al (2001) to detect binding sites of nine transcription factors (TFs) in the yeast *S. cerevisiae*. On each microarray $n = 6270$ DNA subsequences of genes or intergenic regions were spotted. These TFs all have been previously linked to regulating a small set of genes whose expression is cell-cycle dependent. For each TF, three microarray experiments were conducted. The original authors used a statistical model to obtain a p-value for each gene with any given TF, based on which and a selected cut-off value, they identified a set of binding sites of any given TF. The publicly available data contain only a p-value and an averaged ratio of two-channel intensities for each gene and each TF.

In a separate study, Spellman et al (1998) computationally identified about 800 putative cell cycle-regulated genes using microarray gene expression data. In the original analysis by Simon et al (2001), 794 of these cell cycle-regulated genes were present on each array. However, the prior information of the identities of these cell cycle-regulated genes was not used to identify binding sites; rather, after the analysis was done with a list of genes as
candidate binding sites, they were used to (partially) validate the binding sites detected. In this study, we formulate these 794 cell cycle-regulated genes as a source of prior information to help identify binding sites of a TF; that is, the 794 cell cycle-regulated genes consist of a gene list G, which will be treated differently from the other remaining genes.

![Distributions of P-values](image)

**Figure 1:** Distributions of P-values.

We only consider the TF Fkh1; applications to other TFs are exactly the same. We downloaded the ChIP-chip data containing the intensity ratios and p-values $p_i$ from the authors’ web page. The null hypothesis $H_{0i}$ holds if Fkh1 does not bind to gene $i$ (or more precisely its promoter region); otherwise, the alternative hypothesis $H_{1i}$ holds. Figure 1 shows that, as expected, there are more significant p-values for (putative) cell cycle-regulated genes than that for the other genes, and hence it is more likely that $H_{1i}$ holds for a cell cycle-regulated gene.
Results based on ratios

We first used the downloaded ratios to do analysis. The test statistic $Z_i$ is the log-ratio of the two-channel intensities for gene $i$. As discussed earlier, we randomly maintain or flip the sign of $Z_i$ to obtain a null statistic $z_i$.

Figure 2: Distributions of $Z_i$'s (i.e. observed log intensity ratios). The solid lines represent fitted finite Normal mixture models for $Z_i$'s, and the dashed lines for $z_i$'s.

We used R (Ihaka and Gentleman, 1996) package mclust to fit a finite Normal mixture model (Fraley and Raftery 1998, 2002, 2003). By default, it first fitted nine models with the number of components ranging from 1 to 9, then it used BIC to select the best $K_0$. There were 3, 2 and 3 components for estimated $f$, $f^1$ and $f^0$ respectively. As discussed in Pan (2003), it is not optimal to use $z_i$'s of the genes whose $H_{0i}$ do not hold to estimate $f_0$ because the resulting estimate tends to have heavier tails than the true $f_0$. In this example, because, presumably, there are fewer genes whose $H_{0i}$ do not hold among the non-cell-cycle-
Table 1: Estimates of the prior probabilities $\pi$, $\pi_1$ and $\pi_0$ when various intervals $A$ were used.

<table>
<thead>
<tr>
<th></th>
<th>$A = [-0.1, 0.1]$</th>
<th>$A = [-0.3, 0.3]$</th>
<th>$A = [-0.5, 0.5]$</th>
<th>$A = [-0.7, 0.7]$</th>
<th>$A = [-1, 1]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\pi}$</td>
<td>0.978</td>
<td>0.980</td>
<td>0.985</td>
<td>0.991</td>
<td>0.998</td>
</tr>
<tr>
<td>$\hat{\pi}_1$</td>
<td>0.790</td>
<td>0.842</td>
<td>0.897</td>
<td>0.935</td>
<td>0.974</td>
</tr>
<tr>
<td>$\hat{\pi}_0$</td>
<td>1.005</td>
<td>0.999</td>
<td>0.998</td>
<td>1.001</td>
<td>1.001</td>
</tr>
</tbody>
</table>

regulated genes, we used $z_i$’s of only those non-cell-cycle-regulated genes to estimate $f_0$. The resulting estimate of $f_0$ contained 3 components. Comparing the histograms of observed $Z_i$’s (and $z_i$’s) and the estimates $f$, $f^1$, $f^0$ (and $f_0$), we found the mixture models all fitted well. Figure 2 shows the histograms of $Z_i$’s of all the genes, the cell-cycle-regulated genes and other genes, with the estimated $f_0$ imposed. It can be seen that, for each gene group, because the distribution of $Z_i$’s is shifted to the right of the estimated $f_0$, there are genes with their log-ratios $Z_i$ larger than what would be expected under the null hypothesis $H_0$, lending support to the existence of the binding sites of the TF. In particular, there is a larger discrepancy between the estimated $f_0$ and $f^1$, implying that there is a higher proportion of the genes’ being bound by the TF among the cell-cycle-regulated genes.

To estimate $\pi$, $\pi_1$ and $\pi_0$, we took various intervals $A$ in equation (2) and obtained various estimates (Table 1). It can be seen that some estimates (e.g. $\hat{\pi}_1$) are sensitive to the choice of the interval $A$, while others (e.g. $\hat{\pi}$) are more robust. Nonetheless, in any case, the results agree with what we observed earlier based on the distributions of the log-ratios (or p-values) in Figure 2 (or Figure 1) for the three groups of the genes: it is more likely that the TF binds to cell-cycle-regulated genes; that is, we would expect $\pi_1 < \pi < \pi_0$. For example, when taking $A = [-0.5, 0.5]$, we obtained the estimates $\hat{\pi} = 0.985$, $\hat{\pi}_1 = 0.897$ and $\hat{\pi}_0 = 0.998$.

In addition to their possible instability, because, strictly speaking, each of the above estimates $\hat{\pi}$, $\hat{\pi}_1$ and $\hat{\pi}_0$ provides only an estimated upper bound of its true value, one can argue against their use; see Dalmaso et al (2005) for some new development to sharpen the upper bound. As an alternative, in the following, we consider the most conservative approach
Figure 3: Estimated posterior probabilities of a gene’s being bound by the TF, given that its binding log-ratio is $Z$. The standard mixture model was applied to all the genes, while the new method was applied to only cell cycle-regulated genes and other genes respectively.

of estimating all of them at 1, their common upper bound, and show that nevertheless our proposal still gives better results. Note that, because both $\hat{\pi}$ and $\hat{\pi}_0$ in the earlier approach are very close to 1, the earlier approach and the current approach yield similar posterior probability estimates. However, replacing $\hat{\pi}_1 = 0.897$ by 1 leads to lower estimates of the posterior probabilities $Pr(H_{1i}|Z_i)$ for cell cycle-regulated genes.

The estimated posterior probability of a gene’s being bound by TF Fkh1 as a function of its observed binding log-ratio $Z$ using the standard method or our new method is shown in Figure 3. Obviously, for cell cycle-regulated genes, our method yields higher posterior
Figure 4: *Estimated FDR vs the cut-off value c. The standard mixture model was applied to all the genes, while the new method was applied to only cell cycle-regulated genes and other genes respectively.*

probabilities of their being bound by the TF than that of the standard mixture model method, which is expected based on the existing biological knowledge. Unsurprisingly, for cell cycle-regulated genes, with any given cut-off value our method results in a lower FDR than that of the standard approach (Figure 4). From another perspective (Figure 5), if we look at only cell cycle-regulated genes, for any given number of significant genes, our new method gives a smaller FDR than that given by the standard method. Hence, to detect the binding sites of Fkh1 among cell cycle-regulated genes, we will use the posterior probability estimates based on the new method; on the other hand, for non-cell-cycle-regulated genes,
because of the closeness of the FDR estimates between the two methods, either can be used.

Figure 5: Estimated FDR vs the number of significant genes among only cell cycle-regulated genes identified by the standard mixture model method and by the new method respectively.

Results based on p-values

The above use of the intensity ratios may not be optimal because, like the use of the fold change to detect differential expression, it essentially treats all the ratios as having an equal variability, which is questionable. A possibly better way is to use the p-values \( p_i \) provided by the original authors, which should have taken account of varying variabilities across the genes. As we will show later, there may be some problems with the use of these p-values. Unfortunately, due to the unavailability of the individual intensity ratios from the three replicates for each gene, other rigorous statistical approaches are impossible.
We took a transformation $Z_i = \psi^{-1}(p_i)$ to facilitate fitting a finite Normal mixture model for the distribution of $Z_i$’s, where $\psi$ is the cumulative distribution function of the standard Normal distribution (with mean 0 and variance 1). As before, $f_1$, $f_2$ and $f_0$ were estimated using finite Normal mixture models and observed $Z_i$’s. Under $H_0$, $p_i$ has a uniform distribution, and hence $f_0$ is the standard Normal distribution. The results and conclusions based on using the p-values are essentially the same as that reached based on using the intensity ratios. In the following, we present some interesting findings on estimating the prior probability in a mixture model, which may have implications to other applications, such as calculating FDR and q-values (Storey and Tibshirani 2003).

**Issues in Estimating the Prior Probability**

It was problematic to estimate the prior probability in a nonparametric mixture model using the supplied p-values. Trying various intervals $A$ around 0 in equation (2) led to the estimates of $\pi$, $\pi_1$ and $\pi_0$ all close to 1. Among others, because we expected a smaller $\pi_1$ based on the previous studies, we investigated whether the estimation could be improved using three other methods.

First, we tried a simpler estimator proposed by Efron et al (2001): $\hat{\pi} = \min_x \hat{f}(x)/\hat{f}_0(x)$, and similarly for other two parameters $\pi_1$ and $\pi_0$. It turned out that such estimates for $\pi$, $\pi_1$ and $\pi_0$ were 0.081, 3.15e-10 and 4.34e-8 respectively, which are unreasonably small.

Second, we tried a more elaborate method proposed by Storey and Tibshirani (2003). The method works in the following steps:

1. For $\lambda = 0, 0.01, \ldots, 0.95$, calculate
   $$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda\}}{n(1 - \lambda)}$$

2. Fit a cubic smoothing spline with 3 df $\hat{g}$ of $\hat{\pi}_0(\lambda)$ on $\lambda$.

3. $\hat{\pi} = \hat{g}(1)$.

Using this method, we obtained the estimates 0.424, 0.216, 0.454 for $\pi$, $\pi_1$ and $\pi_0$ respectively. Although the ordering of these estimates is reasonable, but they seem to be too small. Figure 6 offers an explanation: $\pi_0(\lambda)$ keeps dropping as $\lambda$ tends to 1, in contrary to
our expectation that $\hat{\pi}_0(\lambda)$ should stabilize as $\lambda \to 1$, which is the key heuristics motivating the above algorithm.

Finally, we applied Pounds and Cheng’s (2004) more sophisticated method. We used their Splosh software and obtained 0.215, 0.259 and 0.239 as the estimates of $\pi$, $\pi_1$ and $\pi_0$ respectively. They are again unreasonably too small.

![Figure 6: Estimating prior probabilities using Storey and Tibshirani’s method.](image)

This puzzling issue may be related to how the p-values were calculated by the original authors (Simon et al 2001) and the commonly used independence assumption across the genes. In theory, under $H_{0i}$, the p-value $p_i$ should have a uniform distribution $U(0,1)$. For the given data, because some $H_{0i}$’s do not hold, we will have more p-values close to 0; however, this should not influence the right tail of the distribution of the p-values if the p-values are independent with each other. Figure 1 shows that i) as expected, there are more p-values close to 0; ii) surprisingly, there are also fewer p-values close to 1. We would expect a flat
right-tail (near 1) from a Uniform distribution of $U(0,1)$, which does not show up in any of the three panels in Figure 1. We suspect that either i) the statistical model used to calculate p-values by the original authors may not fit the data well or ii) the common assumption that the genes’ p-values are independent with each other is not valid, or both, resulting in fewer p-values close to 1 (i.e. highly non-significant p-values) than expected. This issue may thus influence estimating the prior probability in a mixture model; in particular, both Storey and Tibshirani’s method and Pounds and Cheng’s method of estimating $\pi$ critically depends on the right-tail of the p-value distribution (i.e. when p-values are close to 1). Hence some cautions should be taken for any analysis based on using these p-values. It also suggests that more reliable methods of estimating the prior probability that are robust to the deviation from the $U(0,1)$ distribution of the p-value (under the null hypothesis) be worth further investigations.

Possible extensions

Our idea can be easily extended to cover more general situations where the genes can be classified into more than two groups based on existing biological knowledge or data. For concreteness, we consider a specific scenario. Suppose that there are two gene functional categories that are likely to be related to the experimental condition. We partition the genes into three groups $G_1$, $G_2$ and $G_3$ according to whether they are in either or neither of the two functional categories; a gene with both functions is assigned to both $G_1$ and $G_2$. Then we will have three mixture distributions: if gene $i$ is in $G_j$ for $j = 1, 2, 3$, the distribution of its test statistic $Z_i$ is $f_j(Z_i) = (1 - \pi_j)f_0(Z_i) + \pi_j f_1(Z_i)$. A further simplified model can impose that $f_1 = f_1$ for any $j$. In any case, we can proceed to draw inference under the frameworks of nonparametric or parametric EB as discussed before.

Alternatively, rather than using existing and incomplete gene annotations in MIPS or GO, we can use other existing data, such as gene expression profiles, to group genes into various related categories, such as clusters of the genes with similar expression patterns. Then we can proceed as proposed above.

Although our proposal can be applied to multiple groups of genes, there is a trade-off in practice: partitioning the genes into more groups does give a more flexible model that
can account for possibly different characteristics of the genes in different groups, but at the same time it introduces more unknown parameters $\pi_j$'s and $f_j$'s to be estimated. In practice, it may be known a priori that many partitioning (e.g. gene functional) groups may not be related to the experimental condition under investigation and hence can be collapsed together. With the total number of genes around 10000 in a typical array, we envision an upper bound of the number of groups around 10 when nonparametric methods are adopted. In spite of its simplicity and flexibility, there is a limitation with our proposed method when we have a large number of genes groups. For example, we may want to incorporate the whole GO annotations or a hierarchical clustering result of all the genes into analysis. We are currently investigating how to extend our method to such situations; in particular, we would like to take account and take advantage of the hierarchical structure of GO annotations and hierarchical clustering results. In addition to its potential use in practice, our method proposed here also serves as a proof-of-concept and motivates for developing more flexible and powerful approaches.

Acknowledgments

This research was supported by NIH grant HL65462 and a UM AHC Development grant.

References


533-538.


