A mixture model-based strategy for selecting sets of genes in multiclass response microarray experiments

Philippe Broët¹,³,*, Alex Lewin², Sylvia Richardson², Cyril Dalmasso¹,³ and Henri Magdelenat³

¹INSERM U472, 16 Avenue Paul Vaillant Couturier, 94807 Villejuif Cedex, France, ²Department of Epidemiology and Public Health, Imperial College, Norfolk Place, London W2 1PG, UK and ³Institut Curie, 26 rue d’Ulm, 75248 Paris Cedex, France

Received on February 3, 2004; revised on March 22, 2004; accepted on April 13, 2004
Advance Access publication April 29, 2004

ABSTRACT

Motivation: Multiclass response (MCR) experiments are those in which there are more than two classes to be compared. In these experiments, though the null hypothesis is simple, there are typically many patterns of gene expression changes across the different classes that led to complex alternatives. In this paper, we propose a new strategy for selecting genes in MCR that is based on a flexible mixture model for the marginal distribution of a modified F-statistic. Using this model, false positive and negative discovery rates can be estimated and combined to produce a rule for selecting a subset of genes. Moreover, the method proposed allows calculation of these rates for any predefined subset of genes.

Results: We illustrate the performance our approach using simulated datasets and a real breast cancer microarray dataset. In this latter study, we investigate predefined subset of genes and point out interesting differences between three distinct biological pathways.

Availability: http://www.bgx.org.uk/software.html
Contact: broet@vjf.inserm.fr

1 INTRODUCTION

Recent developments in transcriptome-oriented biotechnologies have made possible the comparative analysis of thousands of mRNA expression measurements in parallel. Typically, these data consist of measurements of gene expression under various experimental or biological conditions that can potentially provide information on the complex transcriptional activity for the biological system under study (Schena, 2000). In parallel to the rapid development of these technologies, research into ways of identifying gene expression changes in microarray experiments whilst taking into account false conclusions has become an active area. Up to now, statistical procedures have mostly relied on the multiple comparisons framework in order to control false positive conclusions (Hochberg and Tamhane, 1987; Westfall and Young, 1993). In this framework, two quantities have been considered: the familywise error rate (FWER) and the false discovery rate (FDR). The FWER, which is the oldest criterion considered in multiple comparisons, is defined as the probability of at least one false positive conclusion among all the true null hypotheses (a null hypothesis corresponds to the lack of relationship between gene expression measurement and a response variable). In contrast, the FDR introduced by Benjamini and Hochberg (1995) is defined as the expected proportion of erroneously rejected null hypotheses among the rejected ones. This latter criterion is now widely used for microarray analyses since it controls an error quantity that is relevant for exploratory analysis purpose and leads to more powerful procedures than those relying on the FWER. In this spirit, seminal work has been done for controlling (Benjamini and Hochberg, 1995) or estimating the FDR, or the pFDR as defined by Storey (2001), in a non-parametric way [for some key contributions, see Tusher et al. (2001); Efron et al. (2001); Storey and Tibshirani (2003)].

However, a drawback of these latter procedures is that they only focus on protecting against false positive conclusions. In the exploratory and screening context of most microarray data analysis, investigators may be seriously concerned that such methods do not take into account false negatives and lead to the discarding of too large a proportion of meaningful experimental information. Indeed, a large gene expression variation does not necessarily translate into a major role in the biological process studied and vice versa. This is especially true for microarray experiments in oncology where the ‘top genes’ (based on P-value or gene statistics) are not necessarily ‘key’ genes, whereas other interesting genes (related to biological pathway or target drug) may exhibit smaller transcriptional variations.

In this work, we pay particular attention to the modelling of the alternative hypothesis in order to obtain good estimates of the FDR and its dual quantity the FNR as defined

*To whom correspondence should be addressed.
by Genovese and Wasserman (2002). We consider multiclass response (MCR) experiments, where there are more than two experimental conditions to be compared. Although this situation is frequently encountered in biomedical microarray studies, it has received less attention than the classical two class comparison problem. In MCR microarray experiments, there are typically many patterns of gene expression changes across the different experimental conditions, leading to more complex alternatives than two class comparison situations.

In this paper, we propose to summarize the profiles by a modified $F$-statistic and then to use a flexible mixture of normal distributions to model the marginal distribution of this modified $F$-statistic. This new approach is based on a fully Bayesian mixture model, presented in Section 2, which extends previous work (Broët et al., 2002) on two class comparison in microarray experiments. We illustrate the performance in estimating FDR and FNR using simulated microarray datasets in Section 3 and compare our estimates with those obtained using the $q$-value approach proposed by Storey (Storey, 2001; Storey and Tibshirani, 2003). The usefulness of this new approach is illustrated on real data investigating breast cancer (Hedenfalk et al., 2001) in Section 4. We conclude with a discussion.

2 BAYESIAN MIXTURE MODELLING

2.1 PROCEDURE FOR MCR EXPERIMENTS

In this subsection, we define a gene-based statistic for MCR experiments. In the following, let $X_{ijk}$ denote the measurement from the $i$-th gene ($1, \ldots, I$), in the $j$-th sample ($1, \ldots, J$) belonging to the $k$-th class ($1, \ldots, K$). We recall that MCR corresponds to a situation where there are more than two classes (or groups) to be compared ($K \geq 3$). We first define a classical $F$-statistic for each gene:

$$F_i = \left[ \frac{(N - K) \sum_k J_k (m_{ik} - m_{i..})^2}{(K - 1) \sum_k \sum_j (x_{ijk} - m_{ik})^2} \right]^{1/2}$$

where $m_{i..}$ and $m_{ik}$ denote, respectively, the general mean and mean expression for class $k$ relative to gene $i$ and $N = \sum_k J_k$. In the classical testing hypothesis framework, the null hypothesis $H_0$ is that the mean level for a gene is the same across the different classes. The general alternative is that at least one class mean is different from the others. Under the null hypothesis and with the usual weak assumptions for each gene (normality and equality of the variances for gene expression $X_{ijk}$ across replicates and conditions), the statistic $F_i$ follows a Fisher distribution, denoted $F_{N-k, (K-1)}$, and $(N-K)$ degrees of freedom.

The gene-based statistic $D_i$ used in our model-based approach is a transformation of the gene statistic $F_i$:

$$D_i = \left[ (1 - \frac{2}{9(N - K)}) \frac{F_i^{1/3}}{1 - (1 - \frac{2}{9(N - K)})} \right] \times \left[ \frac{2}{9(N - K)} F_i^{2/3} + \frac{2}{9(K - 1)} \right]^{-1/2}$$

This transformation normalizes the distribution of the $F_i$ and is remarkably accurate for $N - K \geq 10$ (Johnson and Kotz, 1970). Under $H_0$ (corresponding to truly unmodified gene expression), $D_i$ is approximately distributed as a standard normal distribution, while $D_i$ has a more complex decentered distribution otherwise. Note that the one-dimensional (1D) decentered $D_i$ may correspond to several different patterns of expression across the conditions. Thus, the marginal distribution of $D_i$ is a mixture of different distributions related to modified and unmodified gene expression patterns over the different classes.

2.2 Proposed model

Our purpose is to model flexibly the mixture distribution of $D_i$ and thus to estimate for each gene the posterior probability of belonging to the null component, representing no difference of expression over the classes, conditional on the observed data. We stress that we are using a mixture model to separate efficiently the unmodified genes (corresponding to the null component of the mixture) from the modified ones (corresponding to the other components); we are not interpreting the other components to discover classes of gene profiles since different profiles can lead to similar values of the $F$-distribution. This is in contrast to our previous work on two class comparison (Broët et al., 2002) in which a mixture was used to model differential expression and the components interpreted to form clusters corresponding to different levels of expression.

Our modelling approach assumes that the marginal density of $D_i$ can be written as a mixture of Gaussian distributions, i.e. in the form $f(D_i) = \sum_{g=0}^{G} w_g f(\cdot | \mu_g, \sigma^2_g)$, where $f(\cdot | \mu_g, \sigma^2_g)$ are Gaussian densities, with unknown parameters $(\mu_g, \sigma^2_g)$ for the $g$-th component density in the mixture. The quantities $w_g$ are the mixing proportions or weights with $0 \leq w_g \leq 1$ and $\sum_{g=0}^{G} w_g = 1$.

For convenience, we define $g = 0$ to be the unmodified component corresponding to no expression change over the different conditions. This component has a centered normal distribution, i.e. we fix $\mu_0 = 0$. The number of modified components $G$ in the mixture is treated as unknown since the alternative is expected to have a complex distribution summarizing various patterns of gene expression. The prior distribution for $G$ is a Poisson distribution with parameter $m$, with $m$ chosen small so as to encourage a parsimonious number of components to be fitted. Note that this allows $G = 0$, i.e. all observations can be assigned to the unmodified component. The mean parameter for the
unmodified component $\mu_0$ is set to 0 and we impose that $\mu_G > \mu_{G-1} \cdots > \mu_0 = 0$. This choice is motivated by the remark that the distribution of $F_i$ are Fisher distributions $F_{N-k}^{-1}$ under $H_0$ and non-central Fisher distributions $F_{N-k}^{-1}(\eta)$ where $\eta > 0$ is the non-centrality parameter under the alternative. These constraints led to a mixture model, which is different from the one developed previously (Broët et al., 2002) for differential expression and is not fitted by standard implementations for estimating mixtures in a maximum-likelihood set-up using the EM algorithm, as provided e.g. in EMMIX (McLachlan et al., 1999, http://www.maths.uq.edu.au/~gjm/emmix/emmix.html) and used by Pan et al. (2002). Being able to fix $\mu_0 = 0$ and making that the other component means positive is a key element for the fitting and interpretability of our method. The prior distributions specify that $\mu_G \not= 0$ it will indicate that the gene $i$ does not belong to the null component. A joint posterior distribution for all unknowns is formed. Inference is then undertaken by simulating realizations from the resulting posterior distribution using a reversible-jump Metropolis–Hastings algorithm similar to the one used in Broët et al. (2002) and Richardson and Green (1997). The reversible-jump algorithm allows the fitting of models of variable dimension simultaneously, e.g. a mixture with variable number $G$ of components. In this particular case, it proceeds by including in the steps of the algorithm, besides the usual moves that aim to re-estimate the weights, means and variances of the current mixture (with $G$ component), new types of moves that propose to fit mixtures with $G + 1$ or $G - 1$ components that are created by splitting or merging some of the components of the current mixture. Of course, all the moves and their associated acceptance rates are designed so that a sample from the joint posterior distribution of all unknowns is formed.

As usual for mixture models, we introduce $L_i$, an unobserved (latent) categorial variable taking the values 0, . . . , $G$ with probability $w_0, \ldots, w_G$, respectively (McLachlan and Peel, 2000). Thus, when $L_i \neq 0$ it indicates that the gene $i$ does not belong to the null component. A joint posterior distribution for all unknowns is formed. Inference is then undertaken by simulating realizations from the resulting posterior distribution using a reversible-jump Metropolis–Hastings algorithm similar to the one used in Broët et al. (2002) and Richardson and Green (1997). The reversible-jump algorithm allows the fitting of models of variable dimension simultaneously, e.g. a mixture with variable number $G$ of components. In this particular case, it proceeds by including in the steps of the algorithm, besides the usual moves that aim to re-estimate the weights, means and variances of the current mixture (with $G$ component), new types of moves that propose to fit mixtures with $G + 1$ or $G - 1$ components that are created by splitting or merging some of the components of the current mixture. Of course, all the moves and their associated acceptance rates are designed so that a sample from the joint posterior distribution of all unknowns is formed.

The full output of the Bayesian analysis includes information on the posterior distribution of $G$ as well as our main quantities of interest, the posterior probabilities $p_{0i} = p(L_i = 0 | \text{data})$ for each gene. The $p_{0i}$ are estimated within the algorithm by the number of times when $L_i = 0$ divided by the length of the simulation run. Note that these probabilities are integrated over the range of normal mixtures (with different $G$), which are used to fit the marginal density of $D_i$ in a semi-parametric fashion, a unique feature of our model that avoids the need to test several values for $G$ and the consequent uncertainties.

As discussed in a recent paper of Newton et al. (2003), from these posterior probabilities we can obtain model-based estimates of the observed FDR and FNR. Note that our estimates are conditional upon the data, whereas in the seminal paper of Benjamini and Hochberg (1995) the definition of the FDR is more general. To be precise, for a selected subset $(S)$ with cardinality $\#S$ selected from the total set of $I$ genes, the expected FDR can be estimated from the $\{p_{0i}\}$ by using the relation

$$E(\text{FDR} | \text{data}) = \sum_{i \in S} p_{0i} / \#S.$$  

Similarly, the false non-discovery rate can be estimated by averaging the posterior probability of not belonging to the unmodified component for the complementary set of genes:

$$E(\text{FNR} | \text{data}) = \sum_{i \not\in S} [1 - p_{0i}] / [I - \#S].$$

Selection rules can be then constructed from appropriate linear combinations of the FNR and FDR assuming different penalties for misallocation.

### 3 SIMULATION STUDY

#### 3.1 Simulation procedure

Data were generated to mimic a MCR array study with normalized log-ratio measurements for 500 genes ($i = 1, \ldots, 500$) obtained from 24 experiments corresponding to 3 different experimental conditions (classes) $k = 1, 2, 3$ (e.g. 3 different types of tumours), each with 8 replicated samples ($j = 1, \ldots, J = 8$). All values were sampled from normal distributions, $X_{ijk} \sim N(\alpha_{ik}, 0.2^2)$ for all $j$. In each dataset, we simulated three different profiles for the set of genes as follows. The first profile (unmodified) comprises 300 genes with $\alpha_{ik} = 0$ for $i = 1, \ldots, 300$ and for all $k$. The second profile comprising 100 genes is characterized by moderate overexpression for the first condition compared with the other two ($\alpha_{i1} = 0.4/a, \alpha_{i2} = 0, \alpha_{i3} = 0$ for $i = 301, \ldots, 400$). The third profile comprising 100 genes is characterized by a small overexpression for the first and second conditions in comparison with the third ($\alpha_{i1} = 0.2/a, \alpha_{i2} = 0.2/a, \alpha_{i3} = 0$ for $i = 401, \ldots, 500$).

Three cases were considered ($a = 1, 1.5$ and 2 which we call Cases 1, 2, and 3, respectively) to investigate the sensitivity of the method to the blurring of the distinction between the profiles. For each case, 50 replications of the basic set-up were simulated. Figure 1 displays the histogram of the gene statistics $D_i$ for the real dataset (presented in the next subsection) and for one simulated dataset from each case. Note that Case 1, which has the most heterogeneous profiles, corresponds to the heavily right skewed distribution of $D_i$ whilst that of Case 3 is more homogeneous and will clearly be more challenging for teasing out the mixture of modified and unmodified components. The gene statistics in Cases 2 and 3 are quite close to those for the real data.
Mixture model in multiclass microarray experiments

For each set, we allowed 100,000 iterations of the MCMC algorithm to converge (the so-called burn-in period) and computed the relevant quantities over the next 100,000 iterations. Visual checks show that this burn-in period was sufficient. The parameter for the Dirichlet distribution was set to 1, corresponding to uniform weights. The prior distribution for the means was a continuous uniform distribution on the interval [0, b], where b equals the maximum observed gene-based statistics. For the variances $\sigma_g^2$, we adopt the hierarchical choice of inverse gamma prior as stated in Broët et al. (2002). The parameter for the Poisson distribution was set to 2. To interpret the posterior distribution of $G$ independently of the choice of prior for $G$, we report the Bayes factor between different values of $G$ rather than the posterior probabilities:

$$B_{k_1,k_2} = \frac{\Pr(G = k_1 | \text{Data}) / \Pr(G = k_2 | \text{Data})}{\Pr(G = k_1) / \Pr(G = k_2)}.$$ 

3.2 Results

For all three cases, the Bayesian mixture fit includes at least one modified component. Case 1 requires the highest number of components, with half the simulations finding support for $G$ up to 4 or more. For Case 3, the Bayes factors are more weighted towards $G = 1$ but there is still support for $G$ up to 3.

Figure 1 displays the estimated mixture densities integrated over $G$ superimposed on the histograms. It demonstrates that the mixture model adapts well to the various shapes of the marginal distribution of the $D_i$.

Figure 2 shows the Bayesian estimates of FDR and FNR conditional upon the data [as given by Equations (3) and (4)] as a function of the number of genes selected from a cutoff rule based on $p_0$. Also plotted are the realized FDR and FNR (the true proportions of false positive and negative discoveries for each list of selected genes). For both estimated and realized rates, the curves shown are the averages over the 50 simulations for each number of selected genes. In all three cases the FDR is well estimated by the Bayesian mixture, with realized and estimated Bayes curves following each other closely. The estimate of FNR is also good, except in Case 1 for large numbers of selected genes, where the FNR is somewhat overestimated (the mixture gives all genes a non-zero probability of being modified). As a point of comparison with EM-like procedures that would use mixtures with a fixed number of components, we have also computed the FDR and FNR estimates for a Bayesian mixture fitted with a fixed number of components, this number being chosen on the basis of the highest Bayes factor. In all cases, we obtain results that are inferior to those which use the full mixture model integrating over $G$, highlighting that the additional flexibility of using

![Fig. 1. Histograms of the gene statistics $D_i$ for the breast cancer data and for three simulated datasets.](image-url)
parsimonious mixtures with a variable number of components pays off.

As a further point of comparison with non-Bayesian methods we have computed the recently proposed estimation of FDR as implemented by Storey in the QVALUE program (QVALUE: The manual, http://faculty.washington.edu/jstorey/qvalue/manual.pdf; Storey, 2003) since the qvalue can be viewed as an empirical Bayes quantity (Storey, 2003). In this case, the estimated FDR for a list of genes corresponds to the maximum of the qvalues of the list. For the sake of comparability,
we have also estimated the corresponding FNR by
\[
(1 - \hat{\pi}_0) - [1 - \text{FDR}] \times \left\lceil \frac{\#\{\text{gene } i \notin S\}}{\#\{\text{gene } i \in S\}} \right\rceil
\]
where \(\hat{\pi}_0\) is the estimator for the proportion of unmodified genes given by the QVALUE program (Broberg, 2003). Storey’s method is based on the assumption that under the null hypothesis the distribution of \(P\)-values is uniform on \([0,1]\) while under the alternative these will tend to cluster around 0, and uses a spline approach to identify the fraction of unmodified genes. Modelling the distribution of \(P\)-values as a finite mixture of beta distributions has been proposed in earlier work by Allison et al. (2002), but to the best of our knowledge, the associated algorithm and performance has not been compared with that of Storey which is now in wider use. Thus in our performance evaluation, we limit ourselves to a comparison with Storey’s QVALUE.

Averages over the 50 simulations of the Storey estimates of FDR and FNR are also shown in Figure 2. The Storey method performs well for the clearly separated Case 1, but becomes less good as \(a\) increases (i.e. as there is less differentiation between the profiles). In general, the FDR is overestimated and the FNR is correspondingly underestimated. This is not surprising since the Storey (2001) point estimate of \(\pi_0\) is conservative.

We have also investigated the sensitivity of our approach to the number of replicates and to unbalanced designs due for example to missing data. From standard formula for the non-central \(F\)-distribution, it can be seen that we get a small shift in the mean but a 2-fold increase in the variance when the number of replicates is decreased from eight to four. Consequently, there is more overlap between the null and alternative \(F\)-distributions. For Case 1 with \(J_k = 4\), we find similar results as for \(J_k = 8\). For Case 3, we find an overestimation of the FDR which is nevertheless smaller than that of Storey (data not shown). In some experimental set-ups, unbalanced numbers of replicates can occur due to missing data. Our method does not seem particularly sensitive to this; even with 30% of genes having only four replicates instead of eight we get almost identical results to those displayed in Figure 2.

Besides being less conservative, the Bayesian model gives us different information from the Storey method, as it includes for each gene the probability \(p_{0i}\) of being classified in the null conditional on its observed value. Instead, the Storey \(q\)-value, which can be interpreted in Bayesian setting, gives the probability that a gene is unmodified given its statistic is in a rejection region. Having this gene-specific information in the Bayesian model enables us to plot the number of selected genes for different cut-off values for \(p_{0i}\), as suggested by Newton et al. (2003). These plots are presented in Figure 3 (for one simulated dataset from each of the three cases) together with the expected numbers of false discoveries. Note that the well-contrasted Case 1 has a steep increase followed by a much smaller gradient. For Case 3, the increase is linear whereas Case 2 is intermediate. For Cases 1 and 2, these graphs could be used to maximize the size of the list subject to small FDR, by taking the cut-off probability where the curve first starts to level out.

Choosing different cut-off probabilities for declaring genes as having a modified profile over the classes corresponds to minimizing the Bayes error
\[
\lambda \text{FNR} \left( \frac{\#\{\text{gene } i \notin S\}}{I} \right) + (1 - \lambda) \text{FDR} \left( \frac{\#\{\text{gene } i \in S\}}{I} \right).
\]
for different misclassification penalties \(\lambda \in [0,1]\). This can be seen by considering Equations (3) and (4). The penalty \(\lambda\) acts as the cut-off probability, hence a graph of the number of genes selected by minimizing the Bayes error versus \(\lambda\) is identical to that shown in Figure 3. Increasing \(\lambda\) increases the importance of the FNR relative to the FDR. When \(\lambda = 0.5\), the list of genes is the same as that found by the Bayes rule, where genes are assigned to the null or alternative components based on the mode of their classification (in this case genes are assigned to the null if \(p_{0i} > 0.5\)). Using the Bayes rule, the numbers of genes declared positive are (averaged over the 50 simulations) 181, 163 and 133 for Cases 1, 2 and 3, respectively, which correspond to estimates of the proportion of unmodified genes of 36, 33 and 27%. These numbers are close to the true fraction of positive genes (40%) though there is an expected underestimation when the profiles are less separated. This trend to conservatism is more severe when using methods based on fixing FDR at the typical 10%, with no regard for FNR. In our example, at FDR of 10% we would declare on average 34, 21 and 7% of genes for Cases 1, 2 and 3, respectively.

3.3 The analysis of the Hedenfalk dataset

3.3.1 Dataset We analyzed the cDNA microarray dataset publicly available from the breast cancer study conducted by Hedenfalk et al. (2001). The aim of the study was to examine breast-cancer tissues from patients with BCRA1-, BCRA2-related cancer and sporadic cases of breast cancer for determining global gene expression patterns in these three classes of tumours. Details of this study are given in the paper of Hedenfalk et al. (2001). The initial dataset consists of 3226 gene expression ratios derived from the fluorescent intensities from a tumour sample divided by those from a common reference sample. For each gene, a log-expression ratio was available. In this paper, we focus on the subset of 2471 genes having a nominal denomination [expressed sequenced tags (ESTs) and unknown genes were excluded] and log-ratio values >0.1 and <10.

Here, we do not elaborate on data preprocessing since it is not our main purpose [for a detailed discussion on data preprocessing strategies see Speed (2003)]. We consider each log-ratio measurement to be an additive sum of four terms: (i) a gene effect, (ii) a differential effect between the tumour...
Fig. 3. Number of selected genes for different cut-off values for $p_{0i}$ (solid lines) and corresponding expected numbers of false discoveries for three simulated samples (high-density dashed lines). Low-density dashed lines show the number on the list for the Bayes Rule.

sample and the reference sample co-hybridized on a defined array, (iii) an interaction gene × cell line effect that reflects differential gene expression among the three tumour classes specific to each gene and (iv) an error term. The interaction term quantifies the extent to which the mean expression for a specific gene in a class differs from its global mean (over the three classes). As the term of interest is the interaction term, we estimate this term through a classical analysis of variance model. In practice, row and column effects are subtracted, and we let $X_{ijk}$ be the resulting interaction term for the $j$-th gene ($i = 1, \ldots, n = 2471$), $j$-th array ($j = 1, \ldots, J = 22$) and $k$-th tumour class ($k = 1, 2, 3$). From these values, we calculate the gene-based statistics $D_i$ as presented in Section 2.1.
3.3.2 Results We proceeded to fit the mixture model on the \( D_i \) with the prior for \( G \) as specified in Section 3.1. For these data, the mixture is weighted towards finding only one extra component as the Bayes factor for two (respectively three or four) additional component(s) versus one is 0.67 (0.31 and 0.14, respectively). The predicted mixture density (integrated over \( G \)) is superimposed on the histogram of the data in Figure 1. The mixture integrated over different numbers of components provides a good semi-parametric fit to the gene-based statistics. We also investigated on this dataset the sensitivity of our procedure to the choice of prior distribution of components provides a good semi-parametric fit to the data as a function of the number of selected genes. This dataset appears to have a large number of differentiatied genes [the Bayes estimate for the proportion of truly modified genes is 48% and the Storey estimate (1 - \( \hat{p}_0 \)) is 53%]. The Bayes rule with the mixture model would give us a list of 995 genes, which is too many for practical purposes. Typically, other practically and biologically motivated considerations would come into play to define sets of genes of interest for which FDR and FNR are evaluated.

Figure 4a displays estimated FDR and FNR conditional upon the data as a function of the number of selected genes from a cut-off rule based on the posterior probability of belonging to the null component. Figure 4b shows the number of genes selected for different cut-off values. This dataset appears to have a large number of differentiatied genes [the Bayes estimate for the proportion of truly modified genes is 48% and the Storey estimate (1 - \( \hat{p}_0 \)) is 53%]. The Bayes rule with the mixture model would give us a list of 995 genes, which is too many for practical purposes. Typically, other practically and biologically motivated considerations would come into play to define sets of genes of interest for which FDR and FNR are evaluated.

Here, we wish to discuss cases where either monotonic or non-monotonic sets of genes are of interest. The monotonic case corresponds to the common situation where the investigator is interested in obtaining a list of ‘top genes’ for a defined FDR and FNR based on the ordered posterior probabilities \( p_{0i} \). From classical FDR-based strategies, such a monotonic set of genes forms a natural filtration based on the ordered \( P \)-values (e.g. Benjamini and Hochberg, 1995) or \( q \)-values (e.g. Storey and Tibshirani, 2003). The non-monotonic case corresponds to the situation where the investigator is keen to consider a set of genes built from separate and additional biological information. Using the posterior probability estimates for each gene, we can easily estimate FDR and FNR for such non-monotonic sets of genes, a feature that cannot be obtained from the Storey or other approaches discussed in Section 1.

The monotonic case is common in the context of selecting a fixed size list of ‘top genes’ for further automated reverse transcription–polymerase chain reaction (RT–PCR) analyses. Using ordered \( p_{0i} \), our method will provide FDR estimates for a list of 96 or 384 genes (corresponding to the typical 96 or 384 well plates) of 1.6 and 6.1%, whereas FNR estimates are 39 and 31.6%, respectively. Note that clones selected in the original paper (Hedenfalk et al., 2001) based on a modified \( F \)-test with a \( P \)-value cut-off rule of 0.001 (‘top genes’ strategy) all had a \( p_{0i} \) ranging from 0.28 to 7.06%, which shows a good match between the two approaches.

The non-monotonic case is relevant when investigators are interested in different biological pathways (e.g. apoptosis, cell cycle, DNA damage, and so on). As an example, we consider three subsets of genes based on their known biological functions (obtained from publicly available databases): apoptosis, cell cycle regulation and cytoskeleton. This gives us list sizes of 26, 21 and 25 genes, respectively. Estimates for the FDR were 70% for apoptosis, 26% for cell cycle regulation and 66% for cytoskeleton. These results suggest that genes involved in cell cycle regulation show differences over the three tumour classes, in contrast to those associated with other biological functions considered. Among the genes belonging to the cell cycle regulation pathway, Cyclin D1 \( (p_{0i} = 2\%) \) and Cyclin D2 \( (p_{0i} = 8\%) \) are shown.
to be up-regulated in BRCA2 and down-regulated in BRCA1 with respect to sporadic cases. The opposite profile is seen for Cyclin-dependent kinase 4 (\(p_0 = 1\%\)) and Cyclin-dependent kinase inhibitor 2C (\(p_0 = 3\%\)). Retinoblastoma 1 (\(p_0 = 16\%\)) exhibits a down-regulation for and BRCA2 and BRCA1 compared with sporadic cases. The opposite is seen for cyclin-dependent kinase inhibitor 1A (\(p_21\)) (\(p_0 = 8\%\)). More detailed exploration of these modified gene expression patterns is however beyond the scope of this paper.

4 DISCUSSION

Although MCR experiments are frequently encountered in biomedical microarray studies, they have received less attention than the two class comparison problem. Of course, generic approaches based on modelling the distribution of \(P\)-values are applicable to this case, but here we have shown that by using specific 1D gene statistics based on a transformed \(F\)-statistic followed by a model-based mixture estimation of their marginal distribution, we gain a better evaluation of the FDR and FNR and an extended applicability that covers monotonic and non-monotonic sets of genes. The normal mixture model has been successfully used in recent microarray studies in a variety of contexts (Lee et al., 2000; Pan et al., 2002; McLachlan et al., 2002; Broët et al., 2002). The complete Bayesian approach that we have implemented to model our derived gene statistic has been specifically adapted to the case of a right skewed distribution containing one component of known mean although preserving the additional flexibility gained by treating the number of components as unknown. This model-based approach can be considered as a convenient parsimonious representation of a complex mixture density in a semi-parametric way. In the same spirit, discrete mixtures have been used by Newton et al. (2003) for providing non-parametric density estimates in differential gene expression studies.

As seen from the simulation study, the Bayes estimates obtained using our mixture model for the FDR and FNR are accurate over a range of cases. When there is a substantial overlap between truly modified and unmodified gene profiles, the estimates outperform those obtained from the non-parametric qvalue approach. This is not surprising since the conservative bias for the FDR increases when the proportion of modified gene increases and when modified and unmodified gene distributions are close. The converse is true for the FNR.

Moreover, our approach gives an estimate for each gene of the posterior probability of belonging to the null component integrated over all the possible mixture models. This allows us to estimate the FDR and FNR for any subset of genes, a feature that cannot be obtained from any approach such as the qvalue that only considers monotone rejection regions. Allowing computation for any prespecified list of genes derived from external sources is particularly appealing since the investigator may be more interested in focusing on subset of genes based upon defined biological pathways (e.g. apoptosis, cell cycle, DNA damage and so on.)

We have applied the model to a cDNA microarray dataset from the breast cancer study conducted by Hedenfalk et al. (2001). In our analysis, we have focused on FDR and FNR error criteria for exploratory purposes, whereas in the original analysis the authors focused on the FWER criterion. When comparing three subsets of genes defined by their biological functions, our results would suggest that transcriptional expression for genes involved in the cell cycle regulation pathway differ among BRCA1, BRCA2 and sporadic tumours.

In conclusion, we think that the application of such a modelling approach to the distribution of gene statistics gives an efficient way to analyze this type of experiment that may bring new insights in to the biological problem investigated.

ACKNOWLEDGEMENTS

The authors wish to thank Hedenfalk et al. for providing free access to their data on the http://www.nhgri.nih.gov/DIR/Microarray website. We are grateful to Peter Green for stimulating discussion on implementation of Bayesian mixture models with a variable number of components. We also want to thank the two referees for useful comments. This work was supported by Alliance: a Franco–British partnership programme (grant PN03060) and by the BBSRC ‘Exploiting Genomics’ grant 28EGM16093.

REFERENCES


