Application of Mixture Models to Detect Differentially Expressed Genes

Liat Ben-Tovim Jones[1], Richard Bean[1], Geoff McLachlan[1,2,3], and Justin Zhu[1]

[1] ARC Centre in Bioinformatics, Institute for Molecular Bioscience, UQ
[2] Department of Mathematics, University of Queensland (UQ)
[3] ARC Special Research Centre for Functional and Applied Genomics, UQ

Abstract. An important and common problem in microarray experiments is the detection of genes that are differentially expressed in a given number of classes. As this problem concerns the selection of significant genes from a large pool of candidate genes, it needs to be carried out within the framework of multiple hypothesis testing. In this paper, we focus on the use of mixture models to handle the multiplicity issue. With this approach, a measure of the local FDR (false discovery rate) is provided for each gene. An attractive feature of the mixture model approach is that it provides a framework for the estimation of the prior probability that a gene is not differentially expressed, and this probability can subsequently be used in forming a decision rule. The rule can also be formed to take the false negative rate into account. We apply this approach to a well-known publicly available data set on breast cancer, and discuss our findings with reference to other approaches.

1 Introduction

DNA microarrays allow the simultaneous measurement of the expression levels of tens of thousands of genes for a single biological sample; see, for example, McLachlan et al. (2004). A major objective in these experiments is to find genes that are differentially expressed in a given number of classes. In cancer studies, the classes may correspond to normal versus tumour tissues, or to different subtypes of a particular cancer. Comparing gene expression profiles across these classes gives insight into the roles of these genes, and is important in making new biological discoveries. Yet now a real goal for microarrays is to establish their use as tools in medicine. This requires the identification of subsets of genes (marker genes) potentially useful in cancer diagnosis and prognosis.

In the early days of microarray technology, a simple fold change test with an arbitrary cut-off value was used to determine differentially expressed genes. This method is now known to be unreliable as it does not take into account the statistical variability. In order to determine statistical significance, a test such as the t-test, can be performed for each gene. However, when many hypotheses are tested the probability of a type I error (false positive) occurring increases sharply with the number of hypotheses. This multiplicity poses a considerable

M. Gallagher, J. Hogan, and F. Maire (Eds.): IDEAL 2005, LNCS 3578, pp. 422–431, 2005.

problem in microarray data, where there are many thousands of gene expression values.

Recently, a number of sophisticated statistical methods have been proposed, including several nonparametric methods. Tusher et al. (2001), in their significance analysis method (SAM), proposed a refinement on the standard Student's t-statistic. Because of the large number of genes in microarray experiments, there will always be some genes with a very small sum of squares across replicates, so that their (absolute) t-values will be very large whether or not their averages are large. The modified t-statistic of Tusher et al. (2001) avoids this problem. Pan et al. (2003) also considered a nonparametric approach in their mixture model method (MMM). These methods are reviewed in Pan (2002).

In this paper, we initially present the statistical problem and show how a prediction rule based on a two-component mixture model can be applied. In particular, we show how the mixture model approach can handle the multiplicity issue. It provides a measure of the local FDR (false discovery rate), but can be used in the spirit of the q-value. In the latter case, an upper bound, c_o, can be obtained on the posterior probability of nondifferential expression, to ensure that the FDR is bounded at some desired level α.

We finally apply this method to real data, in the well-known breast cancer study of Hedenfalk et al. (2001), with the aim of identifying new genes which are differentially expressed between BRCA1 and BRCA2 tumours. We compare our findings with those of Storey and Tibshirani (2003), and of Broët et al. (2004), who also analysed this data set using different approaches.

2 Two-Component Mixture Model Framework

2.1 Definition of Model

We focus on a decision-theoretic approach to the problem of finding genes that are differentially expressed. We use a prediction rule approach based on a two-component mixture model as formulated in Lee et al. (2000) and Efron et al. (2001). We let G denote the population of genes under consideration. It can be decomposed into G_0 and G_1, where G_0 is the population of genes that are not differentially expressed, and G_1 is the complement of G_0; that is, G_1 contains the genes that are differentially expressed.

We let the random variable Z_{ij} be defined to be one or zero according as the jth gene belongs to G_i or not $(i = 0, 1; j = 1, \ldots, N)$. We define H_j to be zero or one according as to whether the null hypothesis of no differential expression does or does not hold for the jth gene. Thus Z_{1j} is zero or one according as to whether H_j is zero or one.

The prior probability that the jth gene belongs to G_0 is assumed to be π_0 for all j. That is, $\pi_0 = \text{pr}\{H_j = 0\}$ and $\pi_1 = \text{pr}\{H_j = 1\}$. Assuming that the test statistics W_j all have the same distribution in G_i, we let $f_i(w_j)$ denote the density of W_j in $G_i\,(i = 1, 2)$. The unconditional density $f(w_j)$ of W_j is given by the two-component mixture model

$$f(w_j) = \pi_0\, f_0(w_j) + \pi_1\, f_1(w_j). \tag{1}$$

Using Bayes Theorem, the posterior probability that the jth gene is not differentially expressed (that is, belongs to G_0) is given by

$$\tau_0(w_j) = \pi_0 f_0(w_j)/f(w_j) \quad (j = 1, \ldots, N). \tag{2}$$

In this framework, the gene-specific posterior probabilities $\tau_0(w_j)$ provide the basis for optimal statistical inference about differential expression.

2.2 Bayes Decision Rule

Let e_{01} and e_{10} denote the two errors when a rule is used to assign a gene to either G_0 or G_1, where e_{ij} is the probability that a gene from G_i is assigned to G_j $(i, j = 0, 1)$. That is, e_{01} is the probability of a false positive and e_{10} is the probability of a false negative. Then the risk is given by

$$\text{Risk} = (1 - c)\pi_0 e_{01} + c\pi_1 e_{10}, \tag{3}$$

where $(1 - c)$ is the cost of a false positive. As the risk depends only on the ratio of the costs of misallocation, they have been scaled to add to one without loss of generality.

The Bayes rule, which is the rule that minimizes the risk (3), assigns a gene to G_1 if

$$\tau_0(w_j) \leq c; \tag{4}$$

otherwise, the jth gene is assigned to G_0. In the case of equal costs of misallocation ($c = 0.5$), the cutoff point for the posterior probability $\tau_0(w_j)$ in (4) reduces to 0.5.

2.3 The FDR and FNR

When many hypotheses are tested, the probability that a type I error (false positive) is made increases rapidly with the number of hypotheses. The Bonferroni method is perhaps the best known method for dealing with this problem. It controls the family-wise error rate (FWER), which is the probability that at least one false positive error will be made. Control of the FWER is useful for situations where the aim is to identify a small number of genes that are truly differentially expressed. However, in the case of exploratory type microarray analyses, approaches to control the FWER are too strict and will lead to missed findings. Here it is more appropriate to emphasize the proportion of false positives among the identified differentially expressed genes. The false discovery rate (FDR), introduced by Benjamini and Hochberg (1995), is essentially the expectation of this proportion and is widely used for microarray analyses. Similarly, the false nondiscovery rate (FNR) can be defined as the expected proportion of false negatives among the genes identified as not differentially expressed (Genovese and Wasserman 2002).

2.4 Estimated FDR

In practice, we do not know π_0 nor the density $f(w_j)$, and perhaps not $f_0(w_j)$. In some instances, the latter may be known as we may have chosen our test statistic so that its null distribution is known (or known to a good approximation). For example, we shall work with the oneway analysis of variance F-statistic, which can be so transformed that its null distribution is approximately the standard normal.

Alternatively, null replications of the test statistic might be created, for example, by the bootstrap or permutation methods. We shall estimate the population density $f(w)$ by maximum likelihood after its formulation using a mixture model. But it can be estimated also nonparametrically by its empirical distribution based on the observed test statistics w_j.

If $\hat{\pi}_0, \hat{f}_0(w_j)$, and $\hat{f}(w_j)$ denote estimates of $\pi_0, f_0(w_j)$, and $f(w_j)$, respectively, the gene-specific summaries of differential expression can be expressed in terms of the estimated posterior probabilities $\hat{\tau}_0(w_j)$, where

$$\hat{\tau}_0(w_j) = \hat{\pi}_0 \hat{f}_0(w_j)/\hat{f}(w_j) \quad (j = 1, \ldots, N) \tag{5}$$

is the estimated posterior probability that the jth gene is not differentially expressed. An optimal ranking of the genes can therefore be obtained by ranking the genes according to the $\hat{\tau}_0(w_j)$ ranked from smallest to largest. A short list of genes can be obtained by including all genes with $\hat{\tau}_0(w_j)$ less than some threshold c_o or by taking the top N_o genes in the ranked list.

Suppose that we select all genes with

$$\hat{\tau}_0(w_j) \leq c_o. \tag{6}$$

Then an estimate of the FDR rate is given by

$$\widehat{\mathrm{FDR}} = \sum_{j=1}^{N} \hat{\tau}_0(w_j)\, I_{[0,c_o]}(\hat{\tau}_0(w_j))/N_r, \tag{7}$$

where

$$N_r = \sum_{j=1}^{N} I_{[0,c_o]}(\hat{\tau}_0(w_j)) \tag{8}$$

is the number of the selected genes in the list. Here $I_A(w))$ is the indicator function that is one if w belongs to the interval A and is zero otherwise.

Thus we can find a data-dependent $c_o \leq 1$ as large as possible such that $\widehat{\mathrm{FDR}} \leq \alpha$. This assumes that there will be some genes with $\hat{\tau}_0(w_j) \leq \alpha$, which will be true in the typical situation in practice. This bound is approximate due to the use of estimates in forming the posterior probabilities of nondifferential expression and so it depends on the fit of the densities $f_0(w_j)$ and $f(w_j)$.

2.5 Bayes Risk in Terms of Estimated FDR and FNR

The Bayes prediction rule minimizes the risk of an allocation defined by (3). We can estimate the error of a false positive e_{01} and the error of a false negative e_{10} by

$$\hat{e}_{01} = \sum_{j=1}^{N} \hat{\tau}_0(w_j)\hat{z}_{1j} / \sum_{j=1}^{N} \hat{\tau}_0(w_j) \tag{9}$$

and

$$\hat{e}_{10} = \sum_{j=1}^{N} \hat{\tau}_1(w_j)\hat{z}_{0j} / \sum_{j=1}^{N} \hat{\tau}_1(w_j) \tag{10}$$

respectively, where $\hat{z}_{0j}$ is taken to be zero or one according as to whether $\hat{\tau}_0(w_j)$ is less than or greater than c in (4), and $\hat{z}_{1j} = 1 - \hat{z}_{0j}$. Also, we can estimate the prior probability π_0 as

$$\hat{\pi}_0 = \sum_{j=1}^{N} \hat{\tau}_0(w_j)/N. \tag{11}$$

On substituting these estimates (9) to (11) into the right-hand side of (3), the estimated risk can be written as

$$\widehat{\text{Risk}} = (1-c)\hat{\omega}\widehat{\text{FDR}} + c(1-\hat{\omega})\widehat{\text{FNR}}, \tag{12}$$

where

$$\widehat{\text{FDR}} = \sum_{j=1}^{N} \hat{\tau}_0(w_j)\hat{z}_{1j} / \sum_{j=1}^{N} \hat{z}_{1j} \tag{13}$$

and

$$\widehat{\text{FNR}} = \sum_{j=1}^{N} \hat{\tau}_1(w_j)\hat{z}_{0j} / \sum_{j=1}^{N} \hat{z}_{0j} \tag{14}$$

are estimates of the FDR and FNR respectively, and where

$$\begin{aligned} \hat{\omega} &= \sum_{j=1}^{N} \hat{z}_{1j}/N \\ &= N_r/N \end{aligned} \tag{15}$$

is an estimate of the probability that a gene is selected.

Thus unlike the tests or rules that are designed to control just the FDR, the Bayes rule approach in its selection of the genes can be viewed as controlling a linear combination of the FDR and FNR. The balance between the FDR and the FNR is controlled by the threshold c.

3 Estimation of Posterior Probabilities

3.1 Mixture Model Approach

We choose our test statistic W_j so that it has a normal distribution under the null hypothesis H_j that the jth gene is not differentially expressed. For example, if F_j denotes the usual test statistic in a one-way analysis of variance of M observations from g classes, then we follow Broët et al. (2002) and transform the F_j statistic as

$$W_j = \frac{\left(1 - \frac{2}{9(M-g)}\right) F_j^{\frac{1}{3}} - \left(1 - \frac{2}{9(g-1)}\right)}{\sqrt{\frac{2}{9(M-g)} F_j^{\frac{2}{3}} + \frac{2}{9(g-1)}}} \tag{16}$$

The distribution of the transformed statistic W_j is approximately a standard normal under the null hypothesis that the jth gene is not differentially expressed (that is, given its membership of population G_0). As noted in Broët et al. (2002), it is remarkably accurate for $(M - g) \geq 10$.

With this transformation, we can take the null density $f_0(w_j)$ to be the standard normal density (which has mean zero and unit variance). In order to estimate the mixing proportion π_0 and the mixture density $f(w_j)$, we postulate it to have the h-component normal mixture form

$$f(w_j) = \sum_{i=0}^{h-1} \pi_i \phi(w_j;\, \mu_i,\, \sigma_i 2), \tag{17}$$

where we specify $\mu_0 = 0$ and $\sigma_i 2 = 1$. In (17), $\phi(w_j; \mu_i, \sigma_i 2)$ denotes the normal density with mean μ_i and unit variance $\sigma_i 2$. We suggest starting with $h = 2$, adding more components if considered necessary as judged using the Bayesian Information Criterion (BIC).

3.2 Use of *P*-Values

An an alternative to working with the test statistic W_j, we could follow the approach of Allison et al. (2002) and use the associated P-value p_j. We can find these P-values using permutation methods whereby we permute the class labels. Using just the B permutations of the class labels for the gene-specific statistic W_j, the P-value for $W_j = w_j$ is assessed as

$$p_j = \frac{\#\{b : w_{0j}^{(b)} \geq w_j\}}{B}, \tag{18}$$

where $w_{0j}^{(b)}$ is the null version of w_j after the bth permutation of the class labels.

3.3 Link with FDR

Suppose that $\tau_0(w)$ is monotonic (decreasing in w). Then the rule (6) for declaring the jth gene to be differentially expressed is equivalent to

$$w \geq w_o, \tag{19}$$

where w_o is the value of w such that $\tau_0(w_o) = c_o$. The associated FDR, actually the positive FDR (Storey 2004), is given by

$$\pi_0 \frac{1 - F_0(w_o)}{1 - F(w_o)}. \tag{20}$$

Using (17), the positive FDR can be approximated using the fully parametric estimate for $F(w_o)$,

$$\hat{F}(w_o) = \pi_0 \Phi(w_o) + \sum_{i=1}^{h-1} \hat{\pi}_i \Phi(\frac{w_o - \hat{\mu}_i}{\hat{\sigma}_i}) \tag{21}$$

in the right-hand side of (21).

Alternatively, we could choose w_o, and hence c_o, so that (20) is equal to α. It thus also has an interpretation in terms of the q-value of Storey (2004). For if all genes with $\tau_0(w) \leq c_o$ are declared to be differentially expressed, then the FDR will be bounded above by α; see Efron et al. (2001).

Concerning the link of this approach with the tail-area methodology of Benjamini and Hochberg (1995), suppose that the right-hand side of (20) is monotonic (decreasing) in w_0. Then as shown explicitly in Wit and McClure (2004), if we set π_0 equal to one and estimate $F(w_0)$ by its empirical distribution in the right-hand side of (20), the consequent rule is equivalent to the Benjamini-Hochberg procedure.

4 Application to Hedenfalk Breast Cancer Data

We analyze the publicly available cDNA microarray data set of Hedenfalk et al. (2001). They studied the gene expression profiles of tumours from women with hereditary BRCA1- ($n_1 = 7$) and BRCA2-mutation positive cancer ($n_2 = 8$), here referred to as BRCA1 and BRCA2, as well as sporadic cases of breast cancer.

Hedenfalk et al. initially considered genes which could differentiate between the three types of breast cancer (BRCA1, BRCA2 and sporadic). They computed a modified F-statistic and used it to assign a P-value to each gene. A threshold of $\alpha = 0.001$ was selected to find 51 genes from a total of $N = 3,226$ that show differential gene expression. One of the main goals of the study was to identify the genes differentially expressed between the BRCA1 and BRCA2 cancers. They used a combination of three methods (modified t-test, weighted gene analysis and mutual-information scoring), and identified 176 significant genes.

Here we consider the gene expression data from the BRCA1 and BRCA2 tumours only. We use a subset of 3,170 genes, having eliminated genes with one or more measurements greater than 20, which was several interquartile ranges away from the interquartile range of all the data (as in Storey and Tibshirani 2003). We applied our decision-theoretic approach to this data set. In Table 1, we report the estimated values of the FDR, calculated using (13), for various levels of the threshold c_o.

Table 1. Estimated FDR for various levels of c_o

c_o	N_r	$\widehat{\text{FDR}}$
0.5	1702	0.29
0.4	1235	0.23
0.3	850	0.18
0.2	483	0.12
0.1	175	0.06

It can be seen that if we were to declare the jth gene to be differentially expressed if $\tau_0(w_j) \leq 0.1$, then 175 genes would be selected as being significant, with an estimated FDR equal to 0.06. The prior probability of a gene not being differentially expressed (π_0) was estimated to be 0.465. We found that the above estimates, based on the semi-parametric version (13), were the same (to the second decimal place) as those calculated using the fully parametric estimate given in (20).

Of these 175 significant genes, 137 are over-expressed in BRCA1 tumours relative to BRCA2. Hedenfalk et al. (2001), and also Storey and Tibshirani (2003) in their further analysis of this data set, found too that a large block of genes are over-expressed in BRCA1. In particular, these included genes involved in DNA repair and cell death, such as MSH2 (DNA repair) and PDCD5 (induction of apoptosis), also identified by us. In their study, Storey and Tibshirani (2003) identified 160 genes to be significant for differential expression between BRCA1 and BRCA2 by thresholding genes with q-values less than or equal to $\alpha = 0.05$ (an arbitrary cut-off value). Here the q-value of a particular gene is the expected proportion of false positives incurred when calling that gene significant, so that 8 of their 160 genes were expected to be false positives.

On comparing our 175 genes with the 160 identified by Storey and Tibshirani (2003), we found that there were 140 genes in common. Of the 35 excluded genes, 12 were included in the Hedenfalk set of 176. The functional classes (where known) of the remaining 23 genes are shown in Table 2, and interestingly include several genes involved in cell death as well as cell cycle control.

Broët et al. (2004) recently also applied a mixture model appproach to identify differentially expressed genes in this data set. However, they implemented a Bayesian approach, in contrast to the frequentist approach as applied here. They obtained a slightly different estimate for π_0 of 0.52, hence rejecting 52 %

Table 2. Functional classes for uniquely identified genes

Functional Class	Gene Identifier
Cell death	ITPK1, NALP1, GADD34
Cell cycle	MAPK6
Transcription	GATA3, TLE1, HDAC2, GTF2B
Cell-to-cell signalling	ANXA1
Cell growth/adhesion/motility	COL5A1, ACTB1
Protein synthesis	EIF2S2
Protein modification	PRKACA, CSTB
Metabolism	OXCT1, POX1

of the genes as not differentially expressed, as opposed to our value of 46.5 %. In their approach, they did not constrain the variance of the first component to be one because it presents computational problems implementing the Bayesian solution via MCMC methods. However, using the frequentist approach, we were able to fix the variance to be one.

In conclusion, we feel that a mixture model-based approach towards finding differentially expressed genes in microarray data can provide useful information beyond that of other methods. In particular, genes which score as most significant using standard methods for multiple hypothesis testing may not necessarily be of most biological relevance (see Broët et al. 2004). Genes with more subtle changes in their expression levels, indicating that they are more tightly regulated, may be of more importance in the biology of tumour formation.

References

Allison, D.B., Gadbury, G.L., Heo, M., Fernandez, J.R., Lee, C.-K., Prolla, T.A., and Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis* **39**, 1–20.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society* **B** 57, 289–300

Broët, P., Richardson, S., and Radvanyi, F. (2002). Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. *Journal of Computational Biology* **9**, 671–683.

Broët, P., Lewin, A. Richardson, S., Dalmasso, C. and Magdelenat, H. (2004). A mixture model-based strategy for selecting sets of genes in multiclass response microarray experiments. *Bioinformatics* **20**, 2562–2571.

Efron, B., Tibshirani, R., Storey, J.D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**, 1151–1160.

Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology* **23**, 70-86.

Genovese, C.R. and Wasserman, L. (2002). Operating Characteristics and Extensions of the False Discovery Rate Procedure. *Journal of the Royal Statistical Society B* **64**, 499–517

Hedenfalk, I., Duggan, D., Chen, Y.D., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O.P., et al. (2001) Gene-expression profiles in hereditary breast cancer. *The New England Journal of Medicine* **344**, 539–548.

Lee, M.-L.T., Kuo, F.C., Whitmore, G.A., and Sklar, J. (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences USA* **97**, 9834–9838.

McLachlan, G.J., Do, KA, and Ambroise C. (2004). *Analyzing Microarray Gene Expression Data.* New York: Wiley.

Pan, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* **18**, 546–554.

Pan, W., Lin, J. and Le, C.T. (2003). A mixture model approach to detecting differentially expressed genes with microarray data. *Functional and Integrative Genomics* **3**, 117–124.

Storey, J.D. and Tibshirani, R. (2003). Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences USA* **100**, 9440–9445.

Storey, J. (2004). The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics* **31**, 2013–2035.

Tusher, V.G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences USA* **98**, 5116–5121.

Wit, E. and McClure, J. (2004). *Statistics for Microarrays: Design, Analysis and Inference*. Chichester: Wiley.