

Degrees of differential gene expression

Detecting biologically significant expression differences and estimating their magnitudes

David R. Bickel

*Medical College of Georgia
Office of Biostatistics and Bioinformatics
Augusta, GA 30912-4900*

Abstract

Motivation: Many methods of identifying differential expression in genes depend on testing the null hypotheses of exactly equal means or distributions of expression levels for each gene across groups, even though a statistically significant difference in the expression level does not imply the occurrence of any difference of biological or clinical significance. This is because a mathematical definition of "differential expression" as any non-zero difference does not correspond to the differential expression biologists seek. Furthermore, while some current methods account for multiple comparisons in hypothesis tests, they do not accordingly adjust estimates of the *degrees* to which genes are differentially expressed. Both problems lead to overstating the relevance of findings.

Results: Testing whether genes have relevant differential expression can be accomplished with customized null hypotheses, thereby redefining "differential expression" in a way that is more biologically meaningful. When such tests control the false discovery rate, they effectively discover genes based on a desired quantile of differential gene expression. Estimation of the degree to which genes are differentially expressed has been corrected for multiple comparisons.

Availability: R code is freely available from <http://www.davidbickel.com> and may become available from www.r-project.org or www.bioconductor.org.

Contact: bickel@prueba.info

Supplementary information: Applications to cancer microarrays, an application in the absence of differential expression, pseudocode, and a guide to customizing the methods may be found at www.davidbickel.com and www.mathpreprints.com.

Key words and phrases: Statistical significance; biological significance; scientific significance; clinical significance; medical significance; effect size; multiple testing; microarray gene expression data.

Introduction

Microarray technology enables the simultaneous measurement of the expression levels of genes throughout the genome. The use of microarrays to discover which genes are differentially expressed between two or more groups of patients has many biomedical applications, including the identification of disease biomarkers that can potentially be used to better understand and diagnose diseases. While reliable statistical methods for discovering which genes are differentially expressed have been developed, the *degree* to which each gene is differentially expressed has been largely neglected from a statistical viewpoint. This paper addresses this deficiency within the statistical framework of multiple comparison procedures (MCPs).

MCPs, including p-value correction (Westfall and Young, 1993), false discovery rate control (Benjamini and Hochberg, 1995; Storey, 2002a; Efron and Tibshirani, 2002; Fernando *et al.* 2003), decision-theoretic optimization (Storey, 2003; Müller *et al.*, 2002; Bickel, 2003b), and achievement of a posterior probability (Efron *et al.*, 2001; Genovese and Wasserman, 2002) have been applied to microarray data to achieve the goal of discovering which genes are differentially expressed. This goal is distinct from that of classifying patients based on gene expression profiles, for which discrimination methods are better suited (e.g., Dudoit, Fridlyand, and Speed, 2002). It also differs from the goal of grouping genes or microarrays together based on similar expression patterns, as in cluster analysis (e.g., Bickel 2003a and references). The research herein builds on MCPs, with the first goal in mind for the analysis of data. However, the use of receiver operating characteristics in MCPs makes the discovered genes potentially useful as biomarkers for the classification goal. The use of such genes may improve the classification accuracy of discrimination methods.

Receiver operating characteristics and Wilcoxon statistics

While the degree of differential expression of a gene between two groups can be defined a number of ways, e.g., by a t-statistic, the area under the receiver operating characteristic curve (AUC) has several advantages. First, the AUC, by definition, takes into account all possible specificities and sensitivities; there is always a trade-off between sensitivity and specificity, but the AUC is an overall measure of classification accuracy that includes all possible decision thresholds, making the AUC particularly important for clinical settings. If X is a random variable representing the expression levels of a gene for one group of patients and Y is that for another group, then the AUC for the two groups is equal to the probability that a randomly selected patient from one group will have a greater expression value than a random patient from the other group: $AUC = \Pr(Y > X)$ (Green and Swets, 1966). This means that, given an abnormal individual and a normal individual, the AUC is the probability that the gene expression levels can correctly identify them, as Hanley and McNeil (1982) explained in terms of radiology. Thus, AUCs around 1/2 indicate a complete lack of discriminating power (50% chance of correct identification), whereas $AUC = 1$ indicates perfect discrimination (100% chance of correct identification; 100% sensitivity and specificity). Another advantage of the

AUC is that it is easily estimated by the familiar Wilcoxon rank-sum statistic (Hanley and McNeil, 1982), as Pepe *et al.* (2003) noted in the context of microarray data analysis. Since such AUC estimation is nonparametric, it does not require assumptions about the family of distributions of the levels of gene expression; even assumptions about equal shapes of the distribution are not needed. Although t-statistics do not require such assumptions when permutations are used to generate the null distribution, the permutations can be very costly in terms of computation time. Thus, this paper focuses on the AUC as the level of differential gene expression, as estimated by the rank-sum statistic (per Efron and Tibshirani, 2002), but the new methods apply to other measures of expression differences as well. Alternatives are discussed at the end of the paper.

New methods

This section addresses two questions related to the degrees of differential expression:

1. **Which genes have not just statistical significance, but also biologically or clinically relevant differential expression?** Most determinations of statistical significance, including commercial software packages that rely on MCPs, consider a gene differentially expressed if it has *any* difference in expression, even scientifically negligible differences. Such approaches are misleading since a sufficiently large sample size could always find statistical significance for every gene. For example, the population ("true") AUC of two groups may be approximately 0.501, which is negligible from a clinical point of view since classification accuracy would only differ from chance by 0.1%, and yet such a difference would be detected if there are enough microarrays. This is not just a problem for large sample sizes: the following application to two cancer data sets shows that the detection of differential expression depends on the degree of difference that is considered biologically relevant. At best, standard methods give an upper bound on the number of genes that would be considered differentially expressed enough to be relevant. This question is answered by mathematically redefining what is meant by "differential gene expression."
2. **What is the extent to which each gene is differentially expressed?** Unbiased methods of estimating the degree of differential expression of one gene at a time become biased when applied to thousands of genes at once, as seen below. Even commercial software packages that implement p-value correction or other MCPs for hypothesis testing suffer from this problem. (Corrected p-values do not measure the extent of differential expression since a sufficiently large sample size will yield an arbitrarily small p-value for even the smallest nonzero difference in expression between populations.)

Method 1: Detecting biologically relevant differential expression

The method of detecting differential expression on the basis of observed microarray data will be presented in general mathematical terms that can make use of any reasonable test statistics, including statistics for both one-sided and two-sided tests. For the i th gene out of m genes represented on each microarray, let τ_i be the differential expression parameter for

which the test statistic t_i is an estimator, let ϵ_0 be the maximum amount of differential expression a gene can have without being relevantly differentially expressed from a clinical or biological perspective, and let ϵ_{\min} be the value of ϵ_i corresponding to no differential expression at all; $\epsilon_{\min} \geq 0$, $\epsilon_0 \geq \epsilon_{\min}$, $\epsilon_i \geq \epsilon_{\min}$, and $t_i \geq \epsilon_{\min}$ without loss of generality. ϵ_i will be referred to as the *expression difference* of the i th gene, so ϵ_0 is the maximum irrelevant expression difference. In statistical terms, if $\epsilon_i \leq \epsilon_0$, then the i th null hypothesis is true ($H_i = 0$), but if $\epsilon_i > \epsilon_0$, then it is false ($H_i = 1$). Thus, for the i th gene, $H_i = 0$ indicates that the expression level is zero or negligible, and $H_i = 1$ indicates that the expression level is biologically or clinically relevant. H_i is similar to R_i , which indicates whether the i th null hypothesis is rejected ($t_i < \tau$, $R_i = 1$) or not rejected ($t_i \geq \tau$, $R_i = 0$) for a positive threshold τ . That is, the i th gene is thought to have relevant differential expression if $R_i = 1$, but not if $R_i = 0$. In false discovery rate terminology, a rejection of the i th null hypothesis ($R_i = 1$) is called a *discovery* of differential expression in the i th gene, whereas the corresponding failure to reject ($R_i = 0$) is called a *nondiscovery*. A *true discovery* is the discovery of differential expression in a gene that is relevantly differentially expressed ($H_i = R_i = 1$), whereas a *false discovery* is the discovery of differential expression in a gene that is not really relevantly differentially expressed ($H_i = 0$, $R_i = 1$). True and false nondiscoveries are similarly defined. Most researchers select $\epsilon_0 = \epsilon_{\min}$ out of convenience, often unknowingly, but that is inadequate since such a selection has no regard for what would be considered a biologically or clinically relevant level of differential expression. It effectively tests whether there is any difference in expression at all, even when detected differences would be scientifically negligible. Furthermore, given $\epsilon_0 = \epsilon_{\min}$ and a large enough sample size, all genes would be discovered to be differentially expressed because $\forall_i \epsilon_i > \epsilon_{\min}$ for any real biological data set, even though $\exists_i \epsilon_i = \epsilon_{\min}$ would be possible for artificial theoretical models and their simulated data sets. The selection $\epsilon_0 = \epsilon_{\min}$ corresponds to defining "differential expression" as *any* expression difference, no matter how small. The results of the statistical analysis will have more meaning when genes are only defined to be differentially expressed if the expression difference must exceed some threshold level ϵ_0 that satisfies $\epsilon_0 > \epsilon_{\min}$. By the suggested definition of relevant differential expression, the i th gene is only relevantly differentially expressed if $\epsilon_i > \epsilon_0 > \epsilon_{\min}$, instead of the less stringent criterion $\epsilon_i > \epsilon_{\min}$. The proposed null hypotheses and definitions of differential expression are compatible with any MCP criterion for determining the value of τ . For example, the use of values of ϵ_0 that are greater than ϵ_{\min} can be applied to p-value correction (control of a family-wise error rate), control of a false discovery rate, satisfaction of a minimum posterior probability of differential expression, or decision-theoretic optimization. For illustrative purposes, a variant of false discovery rate control is used herein since such control is much more powerful than p-value adjustment and since it is more widely used than the decision-theoretic and probability-threshold approaches.

Since H_i is a random variable in Bayesian approaches, ϵ_i is also a random variable for fixed ϵ_0 . (Bayesian approaches treat estimated parameters as random variables. The i th test statistic, t_i , is also a random variable, and has an observation denoted by T_i for a particular sample.) Storey (2003) showed that $\Pr(H_i = 0 | t_i \geq \tau) = \nabla(\tau)$, where $\nabla(\tau)$ is the proportion of false positives (FPF; Fernando *et al.* 2003), the ratio of the expected number of false rejections to the expected total number of rejections for the rejection region given by $\{t : t \geq \tau\}$. (The

rejection region is the space that contains only the test statistics corresponding to genes considered differentially expressed.) It follows that $\nabla(\tau) = \Pr(t_i \leq z_0 \mid t_i \geq \tau)$, i.e., $\nabla(\tau)$ is equal to a CDF of t_i for $t_i \geq \tau$ if $\Pr(t_i \geq \tau) \neq 0$. Although $\nabla(\tau)$ has several advantages over conventional false discovery rates (Fernando *et al.* 2003), it cannot be controlled in the sense that Benjamini and Hochberg (1995) control the false discovery rate since $\nabla(\tau)$ is undefined when $\forall_{i \in \{1, 2, \dots, m\}} \Pr(t_i \geq \tau) = 0$, which can occur when all null hypotheses are true. This can be addressed by introducing $\Delta(\tau)$, defined to be equal to $\nabla(\tau)$ if the expected number of rejections is positive, and to be equal to 0 otherwise. (Since $\Delta(\tau)$ has additional advantages in decision-theoretic analyses (Bickel, unpublished), it is called the *decisive false discovery rate* (dFDR).) $\Delta(\tau)$, when considered as a function of z_0 , is equal to a CDF of t_i for $t_i \geq \tau$ even when that CDF is defined, and is equal to 0 when that CDF is not applicable. It follows that z_0 is the $\Delta(\tau)$ quantile of $\{t_i : t_i \geq \tau\}$, i.e., of differential expression parameters that correspond to test statistics in the rejection region, when such a quantile is meaningful. This observation can provide solutions to the following problem. Given z_0 and Δ_{goal} , a fixed value of $\Delta(\tau)$, find the value of τ that defines the rejection region. For example, selecting $z_0 = z_{\text{min}}$ and $\Delta_{\text{goal}} = 5\%$ leads to standard control of the dFDR in the sense of rejecting as many null hypotheses as possible with the constraint that the dFDR does not exceed 5%. On the other hand, selecting $\Delta_{\text{goal}} = 50\%$ gives the rejection region for which some given z_0 is the median of t_i . If $\hat{\Delta}(\tau)$ is a conservative estimate of $\Delta(\tau)$, then finding the smallest value of τ for which $\hat{\Delta}(\tau) \leq \Delta_{\text{goal}}$ finds a conservative rejection region in the sense that the expected Δ_{goal} quantile of the differences in expression is at least z_0 . (If no null hypotheses are rejected, $\hat{\Delta}(\tau) = 0$, so there is always some τ for which $\hat{\Delta}(\tau) \leq \Delta_{\text{goal}}$.) This effect of conservative estimation was observed in the first simulation study described below. Storey (2002) proved that estimators of the form of those of Efron *et al.* (2001), Storey (2003), and Efron and Tibshirani (2002) are conservative under reasonable assumptions, so the quantile method of this paper applies to such estimators. They satisfy

$$\hat{\nabla}(\tau) \equiv \frac{\hat{\pi}_0(1 - \hat{F}_0(\tau))}{1 - \hat{F}(\tau)} \text{ for } \hat{F}(\tau) \neq 1, \quad (1)$$

where $\hat{\pi}_0$, $\hat{F}_0(\tau)$, and $\hat{F}(\tau)$ are estimators of the proportion of null hypotheses that are true (the proportion of genes that are not differentially expressed), of the distribution of null test statistics, and of the distribution of all test statistics, respectively. $\hat{\nabla}(\tau)$ can be used to estimate the *false discovery rate* of Benjamini and Hochberg (1995) or, with modification, the *positive false discovery rate* (Storey, 2003) under the independence or weak dependence of test statistics (Storey, 2003). Without assumptions about the dependence of test statistics, $\hat{\nabla}(\tau)$ also estimates the PFP (Fernando *et al.* 2003), which is equivalent for random H_i to the *Bayesian false discovery rate* (Efron and Tibshirani, 2002). The definition of $\Delta(\tau)$ suggests its estimation by

$$\hat{\Delta}(\tau) \equiv \begin{cases} \hat{\nabla}(\tau), & \hat{F}(\tau) \neq 1 \\ 0, & \hat{F}(\tau) = 1 \end{cases}. \quad (2)$$

\hat{F} is an empirical distribution of the test statistics, \hat{F}_0 can be an empirical distribution of test statistics after permutation, and there are various ways to compute $\hat{\pi}_0$. Here, we use the

conservative selection $\hat{\pi}_0 \equiv 1$ for simplicity, following Benjamini and Hochberg (1995).

Equations (1) and (2) are appropriate when the test statistics are defined such that a sufficiently high test statistic implies a discovery, the rejection of a null hypothesis that a gene is not differentially expressed, e.g., T_i could be the absolute value of a t statistic for a two-sided test. (The notation can be slightly modified for more general rejection regions, for example, for asymmetric regions.) If p-values $\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_m$ are used in the place of test statistics, so that $R_i = 1$ if $\mathcal{P}_i \leq \Pi$ or $R_i = 0$ if $\mathcal{P}_i > \Pi$ for some positive threshold Π , equations (1) and (2) become

$$\hat{V}_\Pi \equiv \frac{\hat{\pi}_0 \Pi}{\hat{F}(\Pi)} \text{ for } \hat{F}(\Pi) \neq 0, \quad (3)$$

and

$$\hat{\Delta}_\Pi \equiv \begin{cases} \hat{V}_\Pi, & \hat{F}(\Pi) \neq 0 \\ 0, & \hat{F}(\Pi) = 0 \end{cases}, \quad (4)$$

since the distribution of p-values is uniform under the null hypothesis (Storey, 2002a; Genovese and Wasserman, 2002). In this case, $\hat{F}(\Pi)$ is the proportion of observed p-values that are less than or equal to Π .

The dFDR is controlled by rejecting as many null hypotheses as possible, subject to the constraint $\hat{\Delta}(\tau) \leq \Delta_{\text{goal}}$ or $\hat{\Delta}_\Pi \leq \Delta_{\text{goal}}$. Since the AUC is important in clinical settings, and since the Wilcoxon rank-sum statistic estimates the AUC, the p-values of the two-sided Wilcoxon rank-sum test were used as $\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_m$ to compute $\hat{\Delta}_\Pi$. Then $\hat{\Delta}(\tau)$ was defined to satisfy $\hat{\Delta}(\tau) = \hat{\Delta}_\Pi$ for the purpose of the above quantile interpretation. (Wilcoxon statistics also have the advantages that statistical tests can be performed without distributional assumptions, and that they are invariant to monotonic transforms of the data such as the log transform. Furthermore, a normal approximation can be used with Wilcoxon statistics to speed calculations by obviating permutations.)

Method 2: Correcting the bias in estimates of expression differences

Let ρ_i be the rank of T_i , the observed test statistic of the i th gene, such that T_i is the ρ_i th smallest observed test statistic among all m genes. Similarly, let τ_i be the rank of t_i . The problem to be corrected is that, for a large rank r , $E\{t_i | \rho_i = r\} > E\{t_i | \tau_i = r\}$ if t_i is an unbiased estimator of τ_i , i.e., if $E\{t_i\} = E\{\tau_i\}$. What is desired is an estimator \hat{t}_i that satisfies $E\{\hat{t}_i | \hat{\rho}_i = r\} \approx E\{t_i | \tau_i = r\}$, where $\hat{\rho}_i$ denotes the rank of \hat{t}_i . If $\forall_i \tau_i = i$, then $E\{t_i | \tau_i = r\} = E\{t_r\}$, so that \hat{t}_i should satisfy $E\{\hat{t}_i | \hat{\rho}_i = r\} \approx E\{t_r\}$. (T_i is an observation of the variable t_i , but for simplicity of notation, other observations are not distinguished from their random variables.) Since the bias, $E\{t_i | \rho_i = r\} - E\{t_i | \tau_i = r\}$, is due to using the same data to rank the genes that is used to estimate the effect sizes (expression differences), it can be reduced by randomly designating half of the microarrays as "microarrays for ranks" to determine the ranks, and the other half as "microarrays for expression" to estimate the expression differences. This is accomplished by estimating the expression difference of the gene with the r th smallest expression difference in $\{T_i\}_{i=1}^m$ by the mean expression difference of the microarrays for expression

for the genes whose ranks among the microarrays for ranks are equal to r . If the new expression difference is greater than the original one, the original one is used in its place since a positive bias is being corrected. This procedure may be implemented by the first 12 steps of the algorithm of Appendix A (online supplement). In mathematical terms, the expression difference of the i th gene is estimated by $\hat{\epsilon}_i = \min\left(T_i, \frac{1}{J} \sum_{j=1}^J t_{j,(\rho_i)}^{***}\right)$, where J is the number of times microarrays are chosen without replacement for ranks or for expression, and where, for the j th iteration, $t_{j,(\rho_i)}^{***}$ is the test statistic of the microarrays for expression for the gene with the ρ_i th lowest test statistic of the microarrays for ranks. It can be seen that the ranks and the expression differences are determined from different microarrays since the ranks of $\{T_i\}_{i=1}^m$ correspond to the ranks of $\{t_{j,i}^{**}\}_{i=1}^m$, whereas $\{t_{j,i}^{**}\}_{i=1}^m$, $\{t_{j,i}^*\}_{i=1}^m$, and $\{\hat{\epsilon}_i\}_{i=1}^m$ correspond in genes. Thus, unlike $\{T_i\}_{i=1}^m$, $\{\hat{\epsilon}_i\}_{i=1}^m$ can be used to make inferences about the expression differences of genes with certain ranks given by $\{T_i\}_{i=1}^m$. For example, by using $\{\hat{\epsilon}_i\}_{i=1}^m$, the investigator will have a better idea of what the expression values are of the genes that were identified as differentially expressed by controlling the false discovery rate at the 5% level.

However, this method has the problem that a gene thought to be more differentially expressed than another can have a lower estimated expression difference, i.e., $\rho_i \leq \rho_j$ does not imply that $\hat{\epsilon}_i \leq \hat{\epsilon}_j$. That can be solved by enforcing monotonicity as in step-down methods of controlling a family-wise Type I error rate (Westfall and Young, 1993), using a bias-corrected estimator $\tilde{\epsilon}_i$ that satisfies $\forall_{i,j \in \{1,2,\dots,m\}} \rho_i \leq \rho_j \Rightarrow \tilde{\epsilon}_i \leq \tilde{\epsilon}_j$. It is computed as follows. Let $l(r)$ be the function of a rank such that $l(m)$ is the index of the gene with the highest value of the uncorrected statistic ($\rho_{l(m)} = m$), $l(m-1)$ is that of the next highest ($\rho_{l(m-1)} = m-1$), and so on, so that $\forall_{r \in \{1,\dots,m\}} \rho_{l(r)} = r$ and the index of the gene with the lowest value of the uncorrected statistic is $l(1)$. Then $\tilde{\epsilon}_{l(m)} = \hat{\epsilon}_{l(m)}$ and $\forall_{r \in \{1,\dots,m-1\}} \tilde{\epsilon}_{l(r)} = \min(\hat{\epsilon}_{l(r)}, \tilde{\epsilon}_{l(r+1)})$. Thus, except for ties, the ranks of the corrected estimates $\{\tilde{\epsilon}_i\}_{i=1}^m$ are the same as those of the uncorrected estimates $\{T_i\}_{i=1}^m$, as desired. This enforcement of monotonicity adds three steps to the algorithm of Appendix A.

Results

Simulation

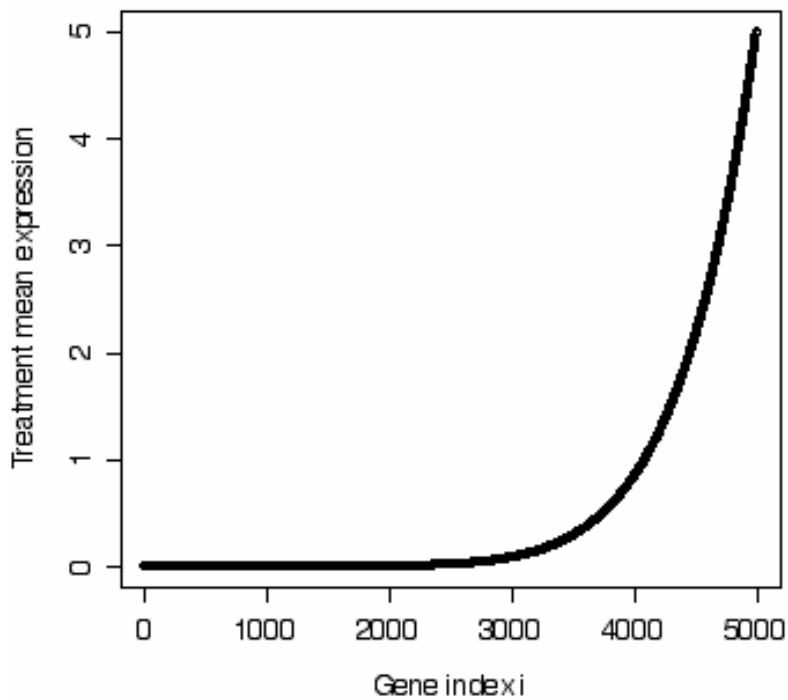
Determining which simulated genes have relevant differential expression

How closely ϵ_0 matches the actual Δ_{goal} quantiles for actual rejection regions was determined by simulation. Each simulated set of microarrays consists of two samples of 20 microarrays each, with 5000 genes per microarray ($m = 5000$). For the first sample, called the *control sample*, each expression value is drawn from $N(0,1)$, where $N(\mu, \sigma^2)$ is the normal distribution with mean μ and standard deviation σ . For the second sample, called the *treatment sample*, each expression value is drawn from $N(\mu_i, 1)$, where i , the gene index, satisfies $i \in \{1, 2, 3, \dots, 5000\}$ and $\mu_{i,\text{treatment}} = 5(i/5000)^8$. As with real expression data, even the

irrelevant, negligible expression differences are not exactly equal to zero, though $\mu_{i,\text{treatment}} - \mu_{i,\text{control}} \approx 0$ for $i \lesssim 2000$ (Fig. 1). The probability that, for the gene with index i , a randomly selected observation in the treatment sample is greater than one in the control sample is equal to the AUC for that gene, as noted under "Significance and Background":

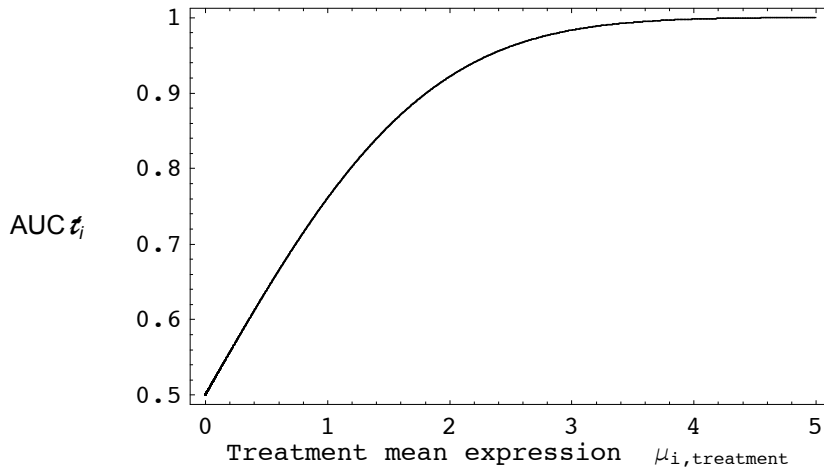
$$\text{AUC}(X_{i,\text{treatment}}, X_{i,\text{control}}) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \exp\left(-\frac{x^2 + (y - \mu_{i,\text{treatment}})^2}{2}\right) dy \right) dx. \quad (5)$$

A numerical integration shows that small differences in the mean expression level of a gene correspond to values of AUC near 0.5, making them clinically irrelevant, whereas large differences ($\mu_{i,\text{treatment}} - 0 \gtrsim 4$) correspond to values of AUC approaching 1, as seen in Fig. 2.



$\mu_{i,\text{treatment}}$ as a function of i . As with real expression data, even the irrelevant, negligible expression differences are not exactly equal to zero, though they appear to be zero for low i .

Figure 1



True area under the ROC curve ($\Pr(X_{i,treatment} > X_{i,control})$), as a function of $\mu_{i,treatment}$ for the simulated data. As expected, small differences in the mean expression level of a gene correspond to values of AUC near 0.5, whereas large differences correspond to values of AUC approaching 1.

Figure 2

The best possible AUC is taken to be the expression difference of interest for gene i :

$$\begin{aligned}
 \mathcal{t}_i &= \max \{ \Pr(X_{i,treatment} > X_{i,control}), \Pr(X_{i,control} > X_{i,treatment}) \} \\
 &= | \Pr(X_{i,treatment} > X_{i,control}) - 1/2 | + 1/2; \\
 \mathcal{t}_{\min} &= 1/2.
 \end{aligned} \tag{6}$$

This equation is used, instead of simply setting \mathcal{t}_i to $\Pr(X_{i,treatment} > X_{i,control})$, in order to ensure that the analysis would also apply to the more general case in which the treatment causes genes to tend to be less expressed than in the control group. In other words, the absolute value enables two-sided tests of the AUC. In the case of the simulated data, the equation reduces to $\mathcal{t}_i = \Pr(X_{i,treatment} > X_{i,control})$. Then, since W_i , the Wilcoxon rank-sum statistic for the i th gene, is an estimator of the AUC, an estimator of \mathcal{t}_i is

$$t_i = \max \{ W_i, 1 - W_i \} = | W_i - 1/2 | + 1/2. \tag{7}$$

For W_i to estimate the AUC, it must be normalized to lie between 0 and 1, as follows. For all possible treatment-control comparisons of the i th gene, W_i is the sum of the number of comparisons for which the expression value of the treatment group is greater than that of the control group and half the number of comparisons for which the expression value of the treatment group is equal to that of the control group, divided by the total number of comparisons. In other words,

$$W_i = \frac{1}{n_{\text{treatment}} n_{\text{control}}} \left(\left| \{(j, k) : j \in \{1, 2, \dots, n_{\text{treatment}}\}, k \in \{1, 2, \dots, n_{\text{control}}\}, x_{i,j,\text{treatment}} > x_{i,k,\text{control}}\}_{i,j=1}^n \right| + \left| \{(j, k) : j \in \{1, 2, \dots, n_{\text{treatment}}\}, k \in \{1, 2, \dots, n_{\text{control}}\}, x_{i,j,\text{treatment}} = x_{i,k,\text{control}}\}_{i,j=1}^n \right| / 2 \right) \quad (8)$$

where $x_{i,j,\text{treatment}}$ is the observed expression value of the i th gene of the j th treatment microarray, and $x_{i,j,\text{control}}$ is the observed expression value of the i th gene of the j th control microarray.

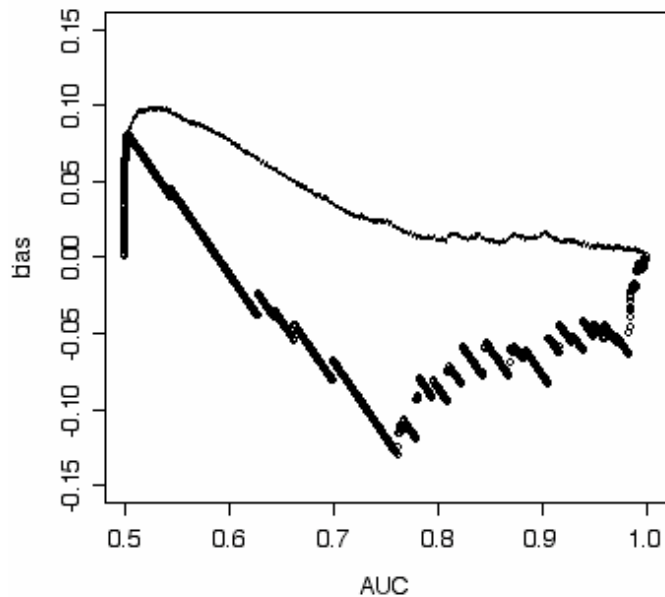
The simulations quantify how close the Δ_{goal} quantile of the expression differences of the genes with test statistics in the rejection region is to τ_0 , the desired quantile. This was accomplished by using the known values of τ_i and finding the value of Π for which $\hat{\Delta}_{\Pi} \leq \Delta_{\text{goal}}$ and $\hat{\Delta}_{\Pi} \approx \Delta_{\text{goal}}$. For example, the settings $\Delta_{\text{goal}} = 50\%$ and $\tau_0 = 0.60$ were used to find the value of τ for which the median value of the τ_i s with p-values in the rejection region is at least 0.60. (An AUC of 0.60 means that 60% of the time, a random observation from the treatment group will be greater than a random observation from the control group for a given gene.) Likewise, setting $\Delta_{\text{goal}} = 5\%$ and $\tau_0 = 0.60$ determined which genes to consider differentially expressed such that the 5th percentile of the τ_i s with p-values in the rejection region is at least 0.60. The results for the simulated data set are displayed in Table 1, which gives the simulation quantiles of genes in the rejection region for $\Delta_{\text{goal}} = 5\%$ (5th percentile) and $\Delta_{\text{goal}} = 50\%$ (50th percentile or median). For real data, the true quantiles, those of $\{\tau_i : T_i \geq \tau\}$, are unknown, but Table 1 suggests using τ_0 and the quantile of observed test statistics in the rejection region as lower and upper bounds on each true quantile, respectively. The desired quantile, τ_0 , is a good lower bound on the true quantile because $\hat{\Delta}_{\Pi}$ is a conservative estimate of Δ_{Π} , i.e., $E(\hat{\Delta}(\tau)) \geq E(\Delta(\tau))$ under the independence of test statistics (Storey, 2002). The sample quantile is a good upper bound because the observed test statistics, $\{T_i : T_i \geq \tau\}$, have an upward bias due to the multiple comparisons problem: they were chosen because they are high enough to be in the rejection region. Nonetheless, lower and upper bounds on the quantiles do not estimate the degree to which each individual gene is differentially expressed.

Goal 5 th percentile, t_0	5 th percentile of $\{t_i : T_i \geq \tau\}$	5 th percentile of $\{T_i : T_i \geq \tau\}$	Goal median, t_0	Median of $\{t_i : T_i \geq \tau\}$	Median of $\{T_i : T_i \geq \tau\}$
0.50	0.59	0.74	0.50	0.70	0.74
0.60	0.81	0.86	0.60	0.94	0.94
0.70	0.93	0.96	0.70	0.98	0.98

Table 1

Estimating the expression differences of the simulated genes

The estimator \tilde{t}_i was tested by applying it to the simulated data set of Table 1 with $J = 50$. The AUC bias is plotted as a function of the true AUC in Fig. 3. The bias for each of 5000 genes is estimated as the estimate of expression difference minus the true expression difference t_i , as a function of t_i . (Here, expression differences are AUCs.) The top curve is the estimated bias of t_i , and the circles are the estimated biases of \tilde{t}_i , for $i \in \{1, 2, \dots, 5000\}$. As expected, the bias of \tilde{t}_i is lower than that of t_i . For many values of the AUC, the negative bias of \tilde{t}_i is greater than the positive bias of t_i , but negative biases are less of a problem since they are conservative. (A negative bias means that the estimate is probably less than the true AUC.)



The estimated bias of two estimators of expression difference.

The bias for each of 5000 genes is estimated as the estimate of expression difference minus the true expression difference t_i , as a function of t_i . (Here, expression differences are AUCs). The top curve is the estimated bias of t_i , and the circles are the estimated biases of \tilde{t}_i , for $i \in \{1, 2, \dots, 5000\}$.

Figure 3

Applications

Applications of the new methods to two cancer data sets and to a case of no differential expression are described in the online supplement. They illustrate the methods and indicate how much the methods can improve data analysis.

Discussion

While either of the two methods proposed could be applied without the other, they naturally go together since both are concerned with effect sizes (expression differences), since both rely on a set of test statistics $\{t_i\}_{i=1}^m$, and since an investigator interested in the results of one method would often be interested in the results of the other: the first method tells the investigator which genes can be considered differentially expressed without saying anything about the levels of differential expression, and the second method yields estimates of expression differences without saying which genes have differential expression between two larger populations represented by the data at hand.

Investigators can choose statistics and parameters to tailor the methods to their specific goals in examining the data. These are application-specific decisions that cannot be made on the basis of statistics alone. The definition of expression difference as the AUC (6) and the choice of the Wilcoxon statistics (7) as estimators have many advantages, especially for clinical settings, as mentioned above. Since the AUC is the area under a plot of the sensitivity versus 1 minus the specificity, a biomedical researcher can choose τ_0 on the basis of the minimum sensitivity that is clinically acceptable at each specificity. Alternately, to avoid such plots, τ_0 may be chosen as the lowest clinically useful probability of correctly classifying two individuals, with one individual randomly selected from each of two groups. However, a definition of the expression difference that is a function of a fold change, with an appropriate estimator similar to a t-statistic, may be better for some studies. For example, an investigator may consider average fold changes of at least 1.5 to be biologically relevant, in which case $\tau_0 = 1.5$ for an appropriately defined τ_i . (This differs from the statistically unsound procedures that consider genes differentially expressed that have *estimates* of fold changes of at least a given threshold.)

While it may be helpful for comparison purposes to make the conventional choices $\tau_0 = \tau_{\min}$ and $\Delta_{\text{goal}} = 5\%$, the choice $\tau_0 = \tau_{\min}$ is too low for a thorough study of biological data since it would imply that all genes are differentially expressed and would be discovered as such, given a large enough sample size. The exact choices of τ_0 and Δ_{goal} may not always be suggested by the biology, in which case investigators can benefit from seeing results for different choices presented as in Tables 2 and 3. Alternately, investigators may decide on values of τ_0 and Δ_{goal} on the basis of the bias-corrected estimates $\{\tilde{\tau}_i\}_{i=1}^m$. Some of the choices available to investigators are presented in Table 4 of Appendix B of the online supplement.

Acknowledgements

I thank an anonymous reviewer for suggestions to write the estimation algorithm as pseudocode and to apply it to a random subsample of the TEL-AML1 data (online supplementary material). I also thank Carey Priebe and Dan Nettleton for helpful discussions.

References

- 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
- Bickel, D. R. (2003a) "Robust cluster analysis of microarray gene expression data with the number of clusters determined biologically," *Bioinformatics* **19**, 818-824
- Bickel, D. R. (2003b) "Selecting an optimal rejection region for multiple testing: A decision-theoretic alternative to FDR control, with an application to microarrays," submitted; available from www.arxiv.org and www.mathpreprints.com
- Dudoit, S., Fridlyand, J., and Speed, T. P. (2002) "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association* **97**, 77-87
- Efron, B. and Tibshirani, R. (2002) "Empirical Bayes methods and false discovery rates for microarrays," *Genetic Epidemiology* **23**, 70-86
- 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
- Fernando, R. L., Nettleton, D., Southey, B. R., Dekkers, J. C. M., Rothschild, M. F., and Soller, M. (2003) "Controlling the proportion of false positives (PFP) in multiple dependent tests," submitted
- Genovese, C. and Wasserman, L. (2002) "Bayesian and frequentist multiple testing," unpublished draft of 4/12/02
- Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999) "Molecular classification of cancer: Class discovery and class prediction by gene expression modeling," *Science* **286**, 531-537
- Green, D. M. and Swets, J. A. (1966) *Signal Detection Theory and Psychophysics*, John Wiley and Sons, Inc.: New York
- Hanley, J. A. and McNeil, B. J. (1982) "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology* **143**, 29-36
- Müller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2002) "Optimal sample size for multiple testing: the case of gene expression microarrays," Technical Report, M. D. Anderson Cancer Center; available from <http://www.ceremade.dauphine.fr/~xian/publications.html> or http://www.ceremade.dauphine.fr/CMD_2002.php

-
- Pepe, M. S., Longton, G., Anderson, G. L., and Schummer, M. (2003) "Selecting differentially expressed genes from microarray experiments," *Biometrics* **59**, 133-142
- Storey, J. D. (2002) "A direct approach to false discovery rates," *Journal of the Royal Statistical Society, Series B* **64**, 479-498
- Storey, J. D. (2003) "The positive false discovery rate: A Bayesian interpretation and the Q-value," *Annals of Statistics* **31** (6)
- Westfall, P. H. and Young, S. S. (1993) *Resampling-Based Multiple Testing: Examples and Methods for p-value Adjustment*, John Wiley & Sons: New York
- Yeoh, E.-J., Mary E. Ross, Sheila A. Shurtleff, W. Kent Williams, Divyen Patel, Rami Mahfouz, Fred G. Behm, Susana C. Raimondi, Mary V. Relling, Anami Patel, Cheng Cheng, Dario Campana, Dawn Wilkins, Xiaodong Zhou, Jinyan Li, Huiqing Liu, Ching-Hon Pui, William E. Evans, Clayton Naeve, Limsoon Wong, and James R. Downing (2002) "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling," *Cancer Cell* **1**, 133-143