BIOINFORMATICS

Vol. 19 no. 10 2003, pages 1236–1242
DOI: 10.1093/bioinformatics/btg148

Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values

Stan Pounds [1,*] and Stephan W. Morris [2]

[1]Department of Biostatistics and [2]Departments of Pathology and Hematology/Oncology, St. Jude Children's Research Hospital, 332 N. Lauderdale St., Memphis, TN 38105-2794, USA

Received on October 18, 2002; revised on January 10, 2003; accepted on January 22, 2003

ABSTRACT

Motivation: The occurrence of false positives and false negatives in a microarray analysis could be easily estimated if the distribution of p-values were approximated and then expressed as a mixture of null and alternative densities. Essentially any distribution of p-values can be expressed as such a mixture by extracting a uniform density from it.

Results: A model is introduced that frequently describes very accurately the distribution of a set of p-values arising from an array analysis. The model is used to obtain an estimated distribution that is easily expressed as a mixture of null and alternative densities. Given a threshold of significance, the estimated distribution is partitioned into regions corresponding to the occurrences of false positives, false negatives, true positives, and true negatives.

Availability: An S-plus function library is available from http://www.stjuderesearch.org/statistics.

Contact: stanley.pounds@stjude.org

INTRODUCTION

Microarray technology allows investigators to simultaneously measure the expression of thousands of genes. Use of microarray technology allows investigators to ask questions such as "Which genes' expressions are affected by selected treatments?" or "Which genes' expressions are correlated with another relevant variable?" These types of questions comprise a series of questions that are applied to each gene. For each gene, one could ask 'Is the expression of this gene affected by selected treatments?' or 'Is the expression of this gene correlated with another relevant variable?' For each gene, the question is termed as a statistical hypothesis test placing a null hypothesis ('The gene's expression is unaffected by the selected treatments' or 'The gene's expression is not correlated with another relevant variable') versus an alternative hypothesis ('The gene's expression is affected by the selected treatments' or 'The gene's expression is correlated with another relevant variable'). Thousands of hypotheses are tested, one hypothesis per gene, resulting in a very complex multiple-testing problem. The multiple-testing problem concerns the occurrence of erroneous conclusions among such a large set of hypothesis tests. Conclusions should be based on a statistical method that adequately addresses the multiple-testing problem by appropriately controlling the probability of making errorenous conclusions.

Many methods that address the multiple-testing issue compare the expressions of genes across two treatments (Pan, 2002). As mentioned above, questions of interest are not limited to the comparison of two treatments. A more unified approach to the analysis of microarray data is needed. In classical statistics, p-values have unified the determination of significance for a wide variety of experimental designs and hypotheses. In microarray studies, a technique based upon the analysis of the set of p-values could unify the determination of significance across a wide variety of experimental designs and adequately address the multiple-testing issue.

TWO IMPORTANT PROPERTIES

Two fundamental properties provide the basis of the analysis of a large set of p-values. First, p-values arising from the null hypothesis are distributed uniformly on the interval $(0, 1)$ (Casella and Berger, 1990). Second, the distribution of the set of p-values can be expressed as a mixture consisting of a uniform$(0, 1)$ component and another component.

THEOREM 1. *Let X be any continuous random variable with a probability density function (pdf) $f(x)$ such*

*To whom correspondence should be addressed.

that $f(x) > 0$ *only if* $0 \leq x \leq 1$. *Then* $f(x) = \pi + (1 - \pi) f_1(x)$ *where* $0 \leq \pi \leq \min(f(x)) \leq 1$ *and* $f_1(x)$ *is a well-defined pdf.*

PROOF. Clearly, $\int_0^1 f(x)dx = 1$ because $f(x)$ is a pdf such that $f(x) > 0$ only if $0 \leq x \leq 1$. By the mean-value theorem (Anton, 1992), $\min(f(x)) \leq 1$. Now let $0 \leq \pi \leq \min(f(x))$. Note that

$$f(x) = \pi + f(x) - \pi$$
$$= \pi + (1 - \pi)\frac{f(x) - \pi}{1 - \pi}. \quad (1)$$

Let $f_1(x) = \frac{f(x)-\pi}{1-\pi}$. Obviously $f_1(x) \geq 0$ and $\int_0^1 f_1(x)dx = 1$, implying that $f_1(x)$ is a well-defined pdf. □

A p-value is a random variable that satisfies the conditions of Theorem 1 because p-values represent a null hypothesis based probability and, therefore, must fall in the interval $[0, 1]$. The two properties imply that the distribution of a set of p-values can be expressed as a mixture of two components: one arising from the null hypothesis and one arising from the alternative hypothesis. The null component is the uniform density extracted as π in (1). The alternative component is the remainder of the overall distribution expressed as $(1 - \pi) f_1(x)$ in Theorem 1. However, the distribution of p-values must be approximated before it can be expressed as a mixture of null and alternative components.

THE STATISTICAL MODEL

The beta-uniform mixture (BUM) distribution

The pdf

$$f(x|a, \lambda) = \lambda + (1 - \lambda)ax^{a-1} \quad (2)$$

for $0 < x \leq 1$, $0 < \lambda < 1$, and $0 < a < 1$ provides a reasonable model for the distribution of p-values arising from a microarray experiment. The pdf $f(x|a, \lambda)$ is a curve that asymptotes at $x = 0$ and monotonically decreases to its minimum of $\lambda + (1 - \lambda)a$ at $x = 1$. This curve approximates the anticipated distribution of the p-values arising from a microarray experiment. Under the null hypothesis, the p-values will have a uniform density corresponding to a flat horizontal line. Under the alternative hypothesis, the p-values will have a distribution that has high density for small p-values and the density will decrease as the p-values increase. The overall distribution will be a mixture of p-values arising from the two hypotheses and will have a shape similar to the pdf defined in (2). As shown below, the distribution obtained by maximum likelihood estimation (MLE) of the parameters provides an excellent approximation to the observed distribution of the p-values arising from microarray experiments in practice. The distribution in (2) will be referred to as the BUM distribution or the BUM density because the distribution is a mixture of a special case of the beta distribution ($b = 1$) and the uniform$(0, 1)$ distribution.

Parameter estimation

Given a set of p-values $\boldsymbol{x} = x_1, \ldots x_n$, one can calculate MLEs for the parameters of the BUM distribution. First, $f(x|a, \lambda)$ should be expressed in terms of the new parameters $\psi \equiv \text{logit}(a)$, and $\phi \equiv \text{logit}(\lambda)$. Second, use numerical optimization techniques to find $\hat{\psi}$ and $\hat{\phi}$, the values of ψ and ϕ that maximize the log of the likelihood $l(\psi, \phi|\boldsymbol{x}) = \sum \log(f(x|a, \lambda)$. Finally, let $\hat{a} = \frac{\exp(\hat{\psi})}{1+\exp(\hat{\psi})}$, and $\hat{\lambda} = \frac{\exp(\hat{\phi})}{1+\exp(\hat{\phi})}$. The invariance property of the MLE ensures that the estimates $\hat{a}$ and $\hat{\lambda}$ are the MLEs for a and λ (Casella and Berger, 1990).

Expressing the distribution as a mixture

The estimated density $\hat{f}(x) = f(x|\hat{a}, \hat{\lambda})$ can be expressed as a mixture, as in (1). Note that in Theorem 1, π must be less than or equal to the minimum of $\hat{f}(x)$. Unfortunately, the proportion π that actually arises from the null hypothesis cannot be estimated. However, the logical upper bound of π can be estimated by using

$$\hat{\pi}_{ub} = \hat{\lambda} + (1 - \hat{\lambda})\hat{a}. \quad (3)$$

The maximum proportion of the set of the p-values that could arise from the null hypothesis is $\hat{\pi}_{ub}$. The remaining $1 - \hat{\pi}_{ub}$ portion of the set of p-values cannot arise from the null hypothesis; thus it must arise from the alternative hypothesis. The alternative component is described by $\hat{f}(x) = (\hat{f}(x) - \hat{\pi}_{ub})/(1 - \hat{\pi}_{ub})$. Finally, $\hat{f}(x)$ is expressed as a mixture of a null component (the uniform density) and an alternative component ($\hat{f}_1(x)$). When the distribution of p-values is expressed in this manner, the occurrence of errors in hypothesis testing can be estimated.

Confidence regions and intervals

Clearly the quantities $\hat{a}$, $\hat{\lambda}$, and $\hat{\pi}_{ub}$ are sample based estimates that are subject to variation. Therefore, it is important to characterize the uncertainty in the estimation of a and λ with a joint confidence region. In microarray studies, $\hat{a}$ and $\hat{\lambda}$ will be based upon thousands of p-values. In such situations, a $1 - \alpha$ confidence region for a and λ is given by all values of $a^\star$ and $\lambda^\star$ such that

$$2(l(\hat{a}, \hat{\lambda}|\boldsymbol{x}) - l(a^\star, \lambda^\star|\boldsymbol{x})) \leq \chi^2_{2,1-\alpha} \quad (4)$$

where the function l represents the log likelihood and $\chi^2_{2,1-\alpha}$ is the $1 - \alpha$ quantile of the chi-square distribution

Table 1. Outcomes of a hypothesis test

	Declare significance	Fail to declare significance
False null hypothesis	True Positive (A)	False Negative (B)
True null hypothesis	False Positive (C)	True Negative (D)

with two degrees of freedom (Casella and Berger, 1990). The confidence region for a and λ can be transformed into a confidence interval for π_{ub}. The $1-\alpha$ confidence interval for π_{ub} is given by the range of $\pi^{\star}_{ub} = \lambda^{\star} + (1 - \lambda^{\star})a^{\star}$ for all $a^{\star}$ and $\lambda^{\star}$ within the confidence region defined by (4). Confidence intervals for other quantities (such as the error quantities mentioned in what follows) based on the estimates of a and λ can be constructed by finding the range of the quantity for values of a and λ within the confidence region.

ESTIMATING THE OCCURRENCE OF ERRORS

Four hypothesis testing outcomes

A hypothesis test is an attempt to use available information to infer whether the null hypothesis is false. A hypothesis test can reach one of two decisions: to declare significance (i.e. to conclude that there is sufficient evidence to infer that the null hypothesis is false) or to fail to declare significance (i.e. to conclude that there is insufficient evidence to infer that the null hypothesis is false). Because the null hypothesis is either true or false, a hypothesis test can have four possible outcomes: (A) declaring significance when the null hypothesis is false, also known as a true positive; (B) failing to declare significance when the null hypothesis is false, also known as a false negative or a Type II error; (C) declaring significance when the null hypothesis is true, also known as a false positive or a Type I error; and (D) failing to declare significance when the null hypothesis is true, also known as a true negative. The four possible outcomes are illustrated in Table 1.

Partitioning the estimated density

When p-values are used in hypothesis testing, significance is determined on the basis of the comparison of the p-value to a threshold τ. Significance is declared when the p-value is less than τ. Failure to declare significance occurs when the p-value is greater than τ. Once τ is selected, the estimated density $\hat{f}(x)$ can be partitioned into four regions, and each region corresponds to a unique hypothesis testing outcome (Fig. 1). There are two straight lines that form the partition: the vertical line at $x = \tau$ and the horizontal line at $y = \pi$. The region to the left of the vertical line corresponds to the p-values declared significant and the region to the right corresponds to p-values declared insignificant. The region above the horizontal line $y = \pi$ is the alternative component of the distribution: the portion of the distribution of p-values arising from the alternative hypothesis. The region below the horizontal line is the null component of the distribution: the portion corresponding to the distribution of p-values arising from the null hypothesis. The area of each region is an estimate of the proportion of hypothesis tests resulting in the corresponding outcome. The areas of the regions A, B, C and D in Figure 1 are computed by

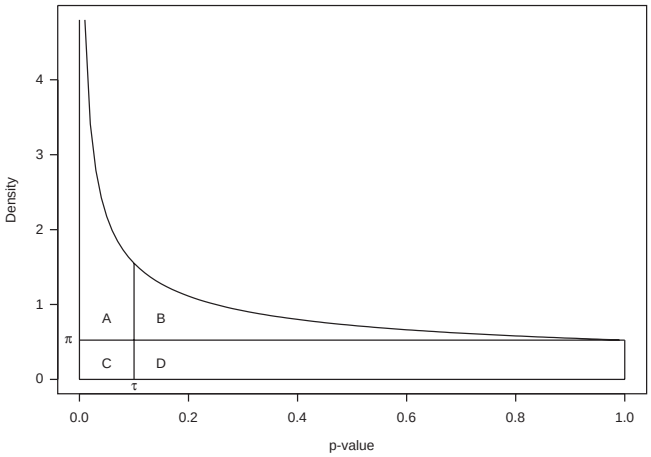

Fig. 1. Graphical illustration of error-control quantities. Region A corresponds to the occurrence of true positives because it lies above the horizontal line (the alternative component) and to the left of the vertical line (declared significant). Region B corresponds to the occurrence of false negatives because it lies above the horizontal line (the null component) and to the right of the vertical line (not declared significant). Region C corresponds to the occurrence of false positives because it lies below the horizontal line (the null component) and to the left of the vertical line (declared significant). Region D corresponds to the occurrence of true negatives because it lies below the horizontal line (the null component) and to the right of the vertical line (declared significant).

$$\begin{aligned} \hat{p}_A(\tau) &= \hat{F}(\tau) - \hat{\pi}_{ub}\tau, \\ \hat{p}_B(\tau) &= 1 - \hat{F}(\tau) - (1-\tau)\hat{\pi}_{ub}, \\ \hat{p}_C(\tau) &= \hat{\pi}_{ub}\tau, \quad \text{and} \\ \hat{p}_D(\tau) &= (1-\tau)\hat{\pi}_{ub}, \end{aligned} \tag{5}$$

respectively, where $\hat{F}(\tau) = \hat{\lambda}\tau + (1-\hat{\lambda})\tau^{\hat{a}}$.

The false discovery rate (FDR)

Various error control quantities of interest can be estimated by using (5). For example, an estimated upper bound of the FDR (the proportion of tests declared significant that are false positives) introduced by Benjamini and Hochberg (1995) is given by

$$\hat{\mathrm{FDR}}_{ub}(\tau) = \frac{\hat{p}_C(\tau)}{\hat{p}_A(\tau) + \hat{p}_C(\tau)} \tag{6}$$

because $\hat{p}_C(\tau)$ estimates the proportion of all tests resulting in false positives, and $\hat{p}_C(\tau) + \hat{p}_A(\tau)$ estimates the proportion of all tests declared significant. The threshold τ could then be selected to ensure that $\hat{\text{FDR}}_{ub}(\tau) \leq \tilde{\alpha}$ for some prespecified $\tilde{\alpha}$. Selecting the threshold in this manner yields

$$\hat{\tau}(\tilde{\alpha}) = \left(\frac{\hat{\pi} - \tilde{\alpha}\hat{\lambda}}{\tilde{\alpha}(1-\hat{\lambda})} \right)^{1/(\hat{a}-1)}. \quad (7)$$

Empirical Bayes' probability (EBP)

Expressing the estimated density in terms of its two components leads immediately to the empirical Bayes' interpretation introduced by Efron *et al.* (2001). For a given p-value x, the estimated lower bound for the EBP $\hat{P}_{lb}(x)$ that x arises from the alternative hypothesis is given by

$$\hat{P}_{lb}(x) = \frac{\hat{f}(x) - \hat{\pi}_{ub}}{\hat{f}(x)}. \quad (8)$$

The right-hand side of (8) is simply the proportion of the distribution at x that lies above the horizontal line separating the null and alternative components. Significance can be determined by comparing the empirical probability in (8) with a preselected threshold γ. In such a comparison, the null hypothesis would be rejected when $\hat{P}_{lb}(x) \geq \gamma$. In this case, the estimated threshold $\hat{\tau}$ is

$$\hat{\tau}(\gamma) = \left(\frac{\gamma\hat{\lambda} + \hat{a}(1-\hat{\lambda})}{\hat{a}(1-\gamma)(1-\hat{\lambda})} \right)^{1/(\hat{a}-1)}. \quad (9)$$

AN EXAMPLE

Description of the experiment

The B-cell lymphoma/leukemia-10 (BCL10) protein, which is aberrantly overexpressed in so-called mucosa-associated lymphoid tissue (MALT) lymphomas, is believed to contribute to the genesis of these hematopoietic cancers by enhancing the proliferation and survival of their normal cellular counterpart, the marginal zone (MZ) B lymphocytes. The normal function of BCL10 is essential for the activation of the transcription factor NF-κB upon the initiation of signaling in lymphocytes through the B-cell receptor (BCR). The tumorigenic effects of BCL10 may be mediated in part by the ability of the overexpressed protein to inappropriately activate the function of NF-κB, which in turn regulates the expression of a number of genes that positively affect the growth of MZ and other B cells (Zhang *et al.*, 1999).

Microarray studies were performed to compare the global gene expression pattern in MZ B lymphocytes purified to homogeneity from transgenic FVB strain mice engineered to overexpress BCL10 in their B cells and purified MZ B cells from wild-type FVB mice. Additional array analyses were performed using RNAs prepared from these two MZ B-cell populations following a brief incubation of the cells with an anti-IgM antibody to cross-link and, therefore, activate BCR signaling and NF-κB activation.

A total of 29 expression measurements were made in these microarray expression profiling studies: eight were performed using the wild-type, non-anti-IgM-activated MZ B-cells; 10 from the BCL10-overexpressing, non-anti-IgM-activated MZ B cells; seven from the wild-type, anti-IgM-activated MZ B cells; and four from the BCL10-overexpressing, anti-IgM-activated MZ B cells. A primary focus of the experiments was to determine the identity of genes that were differentially expressed across the four treatments by comparing all four cellular populations simultaneously.

Table 2. Confidence intervals

Quantity	Estimate	95% CI	99.9% CI
a	0.367	(0.352, 0.382)	(0.345, 0.389)
λ	0.173	(0.130, 0.214)	(0.107, 0.235)
π_{ub}	0.476	(0.457, 0.496)	(0.447, 0.506)
$\tau(\tilde{\alpha} = 0.05)$	0.022	(0.016, 0.029)	(0.014, 0.034)
$\text{FDR}_{ub}(\tau = 0.022)$	0.050	(0.046, 0.054)	(0.045, 0.056)
$P_{lb}(\tau = 0.022)$	0.868	(0.850, 0.885)	(0.840, 0.893)
$p_A(\tau = 0.022)$	0.196	(0.189, 0.204)	(0.185, 0.208)
$p_B(\tau = 0.022)$	0.327	(0.311, 0.344)	(0.303, 0.353)
$p_C(\tau = 0.022)$	0.010	(0.009, 0.011)	(0.009, 0.011)
$p_D(\tau = 0.022)$	0.466	(0.447, 0.485)	(0.437, 0.495)

Analysis of the experiment

The genes were not filtered before conducting the following analysis. A Kruskal–Wallis (1952) test comparing all four groups was applied to the expression values for each of 12 488 probes. For each test, a p-value was generated by using the χ^2 approximation, because the sample sizes were large enough. For the resulting set of p-values, the BUM distribution MLEs are $\hat{a} = 0.367$ and $\hat{\lambda} = 0.173$, which implies that up to $\hat{\pi}_{ub} = 0.476$ of the probes' expressions are not affected by the treatments. Figure 2 shows 90, 95, 99 and 99.9% confidence regions for a and λ. Corresponding confidence intervals for a, λ, π_{ub} and other estimated quantities discussed in this section are given in Table 2.

The threshold $\hat{\tau} = 0.022$ was selected by using (6) to ensure that the estimated false discovery rate $\tilde{\alpha}$ is less than 0.05. The threshold $\hat{\tau} = 0.022$ corresponds to declaring significance for an EBP greater than 0.868. By declaring significance for all p-values less than $\hat{\tau} = 0.022$, approximately $\hat{p}_A(\hat{\tau}) = 0.196$ of the tests result

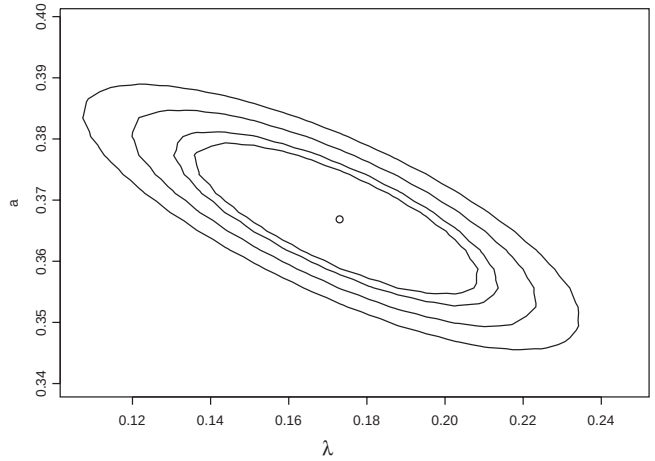

Fig. 2. Confidence regions for λ and a. The contours above represent confidence regions for a and λ. Beginning with the innermost contour and moving outward, the contours represent 90, 95, 99, and 99.9% confidence regions respectively. The point in the center is the maximum likelihood estimate $(\hat{\lambda}, \hat{a})$.

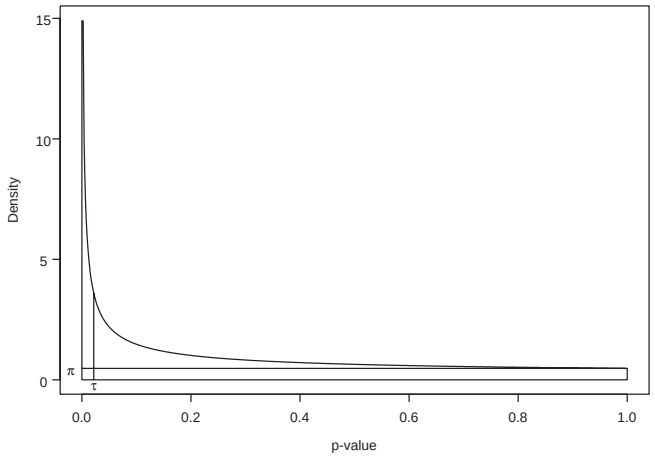

Fig. 3. Error regions. Significance was determined by comparing p-values to $\tau = 0.022$. The bottom left region corresponds to the occurrence of false positives; the top left region corresponds to true positives; the top right region corresponds to false negatives; and the bottom right region corresponds to true negatives.

in true positives; $\hat{p}_B(\hat{\tau}) = 0.327$ of the tests result in false negatives; $\hat{p}_C(\hat{\tau}) = 0.010$ of the tests result in false positives; and $\hat{p}_D(\hat{\tau}) = 0.466$ of the tests result in true negatives. The regions corresponding to the four outcomes are illustrated in Figure 3.

Obviously, other values of τ could have been selected. Figure 4 illustrates how the selection of τ would affect the FDR, EBP, $\hat{p}(B)$ and $\hat{p}(C)$. Figure 5 illustrates the same quantities for values of τ less than 0.10. The analysis has a very rich interpretation. For each probe, the analysis can provide the p-value, the FDR (if τ set to that probe's p-value), the EBP, the quantities $\hat{p}_A$, $\hat{p}_B$, $\hat{p}_C$, and $\hat{p}_D$, or any other error-control quantity that can be derived from the partition of the estimated density. In addition, confidence intervals for the quantities are available.

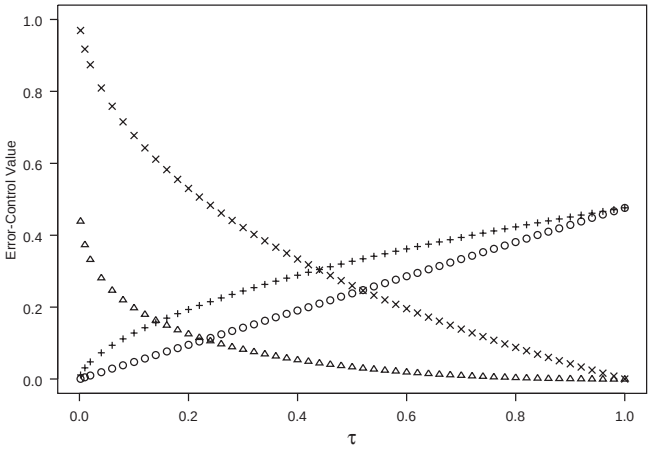

Fig. 4. Various error control quantities. The triangles show $p_B(\tau)$; the circles show $p_C(\tau)$; the crosses show $\hat{\text{FDR}}_{ub}(\tau)$; and the $\times$'s show $\hat{P}_{lb}(\tau)$.

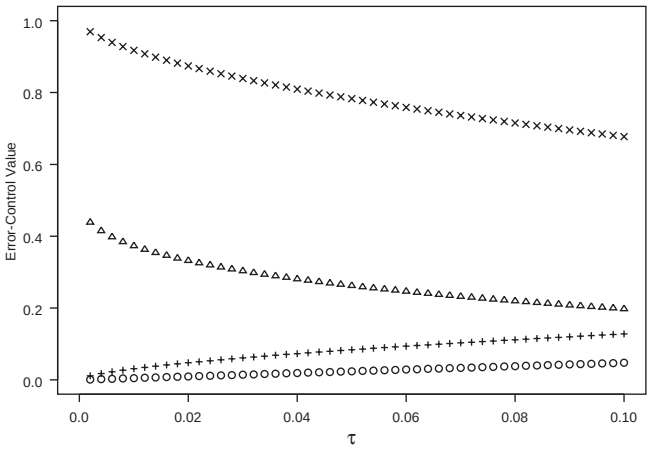

Fig. 5. Various error control quantities. The same quantities as in Figure 4 for small τ.

The estimated BUM density $\hat{f}(x)$ is an excellent approximation of the distribution of the p-values, as shown by a quantile–quantile plot (Fig. 6) and by comparing the estimated density to a histogram (Fig. 7). This indicates that the assumption regarding the form of the distribution of the p-values is a reasonable approximation.

Comments on the analysis

The estimate of π implies that at least 0.524 of the 12 488 genes' expressions are affected by the treatments. This may seem to be excessive, but this fraction includes *any* gene that is even *slightly* differentially expressed

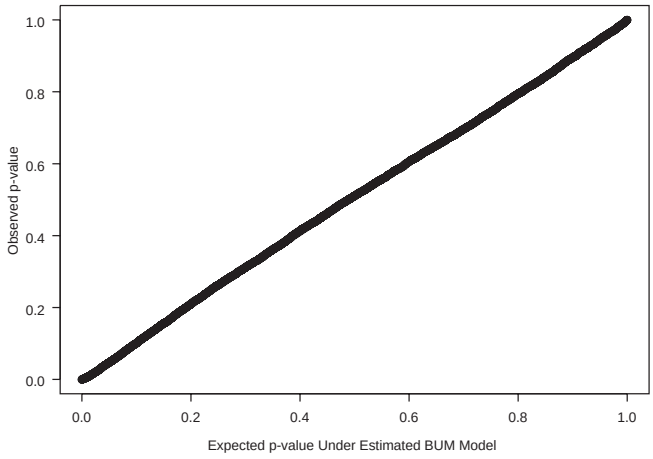

Fig. 6. A quantile–quantile plot comparing the estimated BUM distribution with the observed distribution.

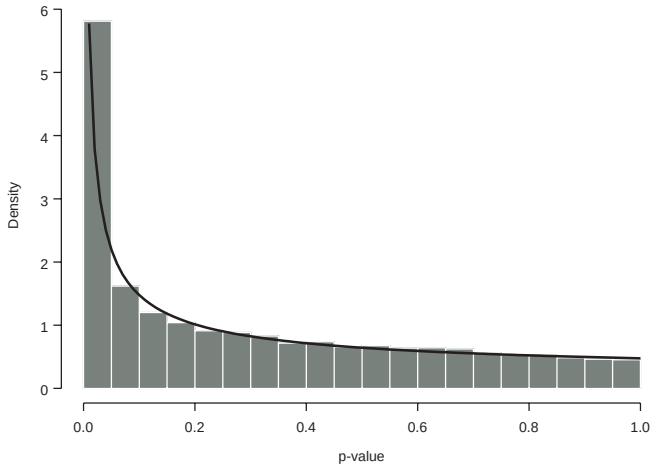

Fig. 7. A comparison of the fitted BUM model to the histogram.

in any one of the four treatments. The experimental design is anticipated to strongly affect the expression of NF-κB, which is known to have over 150 targets (Pahl, 1999). Included among these NF-κB target genes are cytokines/chemokines, immunoreceptors, cell adhesion molecules, acute phase proteins, stress response genes, cell-surface receptors, regulators of apoptosis (programmed cell death), a number of growth factors, early response genes, enzymes, cell-cycle control proteins, as well as other transcription factors that possess a number of their own unique target genes. The up- or down-regulation of these genes by NF-κB leads in turn to alterations in a huge number of additional genes that are involved in multiple cell signaling pathways.

DISCUSSION

To use the proposed technique, first determine the best statistical method to address the question, as if only one gene or probe was being studied. Second, apply the chosen method to the set of all genes (or probes) to obtain a set of p-values. Third, use maximum likelihood to estimate the distribution of p-values with a BUM distribution. Fourth, express the estimated distribution as a mixture of a null and uniform component. Finally, partition the estimated distribution for the selected value of τ, and compute the error control quantities for each p-value.

Because p-values can be produced by essentially all existing statistical methods, the technique can be applied to any experiment for which an appropriate statistical method exists for testing the same question for a single gene or probe. The technique has been successfully applied to p-values obtained from rank correlation analysis, the Kruskal-Wallis test, and two-way ANOVA. In all analyses conducted to date, the agreement between the fitted model and the empirical distribution has been excellent, very similar to that shown in the example.

The technique addresses the multiple-testing issues by producing estimates of the occurrence of each of the four hypothesis-testing outcomes. Unlike existing methods, which focus almost solely on the control of false positives, the proposed technique allows one to focus on the control of false negatives as well. Most techniques have focused on controlling false positives, because the processes to study a particular gene in greater detail are costly and time consuming. Therefore, it is currently very undesirable to study a false positive in greater detail. As the processes for studying genes in greater detail improve, the cost of a false positive relative to a false negative will decrease. If and when these processes improve, the control of false negatives may become as important as the control of false positives.

One needs to carefully consider how to compute p-values in an appropriate fashion. In one analysis, the technique was applied to a set of p-values computed from the F-distribution for two-way ANOVA. The assumptions of two-way ANOVA were not met. As a consequence, the resulting p-values were not well approximated by the estimated BUM distribution. However, when permutation was used to compute the p-values, the proposed technique worked very well. To date, the technique has worked very well for all appropriately computed sets of p-values to which it has been applied. However, it is conceivable that there will be a set of appropriately computed p-values that will not be well represented by the BUM distribution. In such a case, one could examine methods to estimate the density nonparametrically. Any nonparametric technique to estimate the distribution should appropriately address

the endpoints at $x = 0$, which may have an asymptote, and $x = 1$.

One-sided tests are likely to result in distributions of p-values that do not follow the BUM distribution, because the p-values for the 'untested' alternative will concentrate around $x = 1$. This will result in a distribution that is U-shaped, and will not be well approximated by a BUM density. To perform one-sided tests, first obtain p-values using the corresponding two-sided test. After fitting the BUM model to the p-values, addressing multiplicity, and declaring significance, select only those probes with test statistics corresponding to the alternative of interest. Alternatively, one could consider using a more general form of the beta distribution in the mixture to capture the U-shape. In this case, the extracted uniform component would correspond to the strict null hypothesis, the region above the uniform component on the left side would correspond to the alternative hypothesis of interest, and the region above the uniform component on the right side would correspond to the other alternative hypothesis.

This technique is similar to the one proposed by Efron *et al.* (2001). However, the proposed method will require substantially less computing time when a method is available that can estimate p-values without resorting to permutation techniques, as was the case in the example. Even when permutation is required, the proposed technique will require much less memory, because it requires less memory to store p-values than to store the entire permutation distribution of the test statistics. Additionally, the local logistic regression step in Efron's technique is not required.

The proposed technique implicitly assumes that the p-values for the gene are independently and identically distributed according to the BUM model. This assumption is made primarily for computational ease, but also could imply that genes behave independently, which is obviously untrue. Nevertheless, the model appears to be a very useful and reasonable approximation for the example data set, as shown in Figure 6. Future statistical advances may allow the assumption of independence to be relaxed. In the meantime, the validity of the assumptions can be assessed by utilizing the quantile–quantile plot and by examining the correlations of genes to one another, keeping in mind the inherent multiplicity in examining those correlations. If correlation analysis or the quantile–quantile plot suggests radical violation of these assumptions, it is recommended that one explore other techniques for analysis. The validity of all estimates will depend upon how well the assumptions approximate reality. No attempts have been made to date to experimentally verify the obtained estimated error quantities obtained by the proposed method. The proposed technique is intended primarily to provide a reasonable estimate of the occurrence of errors for use in screening genes to find some candidates for further study. Genes that are strongly correlated and affected by the treatment will still likely show up as reasonable candidates for further study. Additionally, absolutely no filtering should be applied before utilizing the technique. If filtering is applied, the error estimates will apply only to the set of genes included when the technique is applied. The technique will not provide any insight as to how many genes were erroneously filtered prior to its application. The proposed technique could itself serve as an informative method of filtering.

The actual algorithm for finding confidence intervals works in the following manner. For quantities that are monotone in both a and λ, intervals are found by determining the extrema of the quantity along the boundary of the confidence region given by (4). For other quantities, conservative intervals are found by finding the extrema of the quantity within the smallest rectangle that bounds the confidence region described by (4).

It should be noted that Efron *et al.* (2001) define the FDR differently from Benjamini and Hochberg (1995). Efron *et al.* claim that the FDR is equal to $1 - \hat{P}(x)$. Efron's quantity should be considered a localized version of the FDR introduced by Benjamini and Hochberg. The estimate of the FDR proposed in (6) corresponds more closely to the definition of Benjamini and Hochberg.

ACKNOWLEDGEMENTS

The authors wish to thank Xiaoli Cui, Cheng Cheng, Shesh Rai, and the reviewers for the constructive comments. This research was supported in part by NCI grant R01 CA-87064, NIH Cancer Center Support Core Grant CA-21765, and the American Lebanese Syrian Associated Charities (ALSAC).

REFERENCES

Anton,H. (1992) *Calculus*, 4th edn, Wiley, New York.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. B*, **57**, 289–300.

Casella,G. and Berger,R. (1990) *Statistical Inference*. Wadsworth & Brooks/Cole, Pacific Grove, CA.

Efron,B. *et al.* (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.*, **96**, 1151–1160.

Kruskal,W. and Wallis,W. (1952) Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.*, **53**, 814–861.

Pahl,H.L. (1999) Activators and target genes of Rel/NF-κB transcription factors. *Oncogene*, **18**, 6853–6866.

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

Zhang,Q. *et al.* (1999) Inactivating mutations and overexpression of *BCL10*, a caspase recruitment domain-containing gene, in MALT lymphoma with t(1;14)(p22;q32). *Nat. Genet.*, **22**, 63–68.