BIOINFORMATICS

Vol. 18 no. 2 2002
Pages 251–260

Adjustments and measures of differential expression for microarray data

A. Tsodikov, A. Szabo and D. Jones*

Huntsman Cancer Institute and Department of Oncological Sciences, University of Utah, 2000 Circle of Hope, Salt Lake City, UT 84112-5550, USA

Received on February 2, 2001; revised on August 17, 2001; accepted on September 5, 2001

ABSTRACT

Motivation: Existing analyses of microarray data often incorporate an obscure data normalization procedure applied prior to data analysis. For example, ratios of microarray channels intensities are normalized to have common mean over the set of genes. We made an attempt to understand the meaning of such procedures from the modeling point of view, and to formulate the model assumptions that underlie them. Given a considerable diversity of data adjustment procedures, the question of their performance, comparison and ranking for various microarray experiments was of interest.

Results: A two-step statistical procedure is proposed: data transformation (adjustment for slide-specific effect) followed by a statistical test applied to transformed data. Various methods of analysis for differential expression are compared using simulations and real data on colon cancer cell lines. We found that robust categorical adjustments outperform the ones based on a precisely defined stochastic model, including some commonly used procedures.

Availability: A program implementing the proposed adjustment and test procedures is available at http://www.hci.utah.edu/groups/biostat/szabo.html.

Contact: alexander.tsodikov@hci.utah.edu

1 INTRODUCTION

The global composition of biologically active molecules provides the molecular basis for cell behavior. By regulating gene expression the cell fulfills its intended biological function. This direct correlation between biological activity and expression profile provides an avenue for characterizing and comparing cells.

The microarray technology offers an exciting opportunity to simultaneously screen the expression pattern of thousands of distinct genes. Researchers can track the effect of interventions or natural processes on gene expression levels thus identifying the functions of genes and the biochemical pathways they participate in.

*To whom correspondence should be addressed.

The central application of microarray analysis is to quantitatively characterize the difference between expression profiles from two distinct cell samples. Making gene expression information useful is a challenge for a number of reasons. A very large number of genes, unknown dependency structure and a relatively small number of measurement replicates complicate a rigorous mathematical treatment of the problem. Gene expressions represent a vector of huge dimension. Complex biological processes regulated by gene expressions make the components of gene expression vector a system of dependent random variables. Additional artifact dependency is induced by the measurement process. Changes in the multivariate dependency structure of the gene expression vector as well as changes in the absolute values of gene expression could be responsible for the difference between normal and pathological cells.

Statistical approaches to the analysis of microarray data can be divided into two main groups: methods for analysis of gene dependency and methods for analysis of differential expression. The primary tool for the first group of methods is based on cluster analysis (Alon *et al.*, 1999; Ben-Dor *et al.*, 1999; Heyer *et al.*, 1999; Herwig *et al.*, 1999; Carr *et al.*, 1997; Michaels *et al.*, 1998; Eisen *et al.*, 1998; Basett *et al.*, 1999). This approach holds much promise for determining groups of genes with a similar function. Usually a measure of correlation between expressions of different genes is used to cluster together genes with dependent expressions. Although these methods address important questions of mutual dependency within groups of genes, there are a number of interesting issues that remain beyond the scope of clustering methods. These include identification of genes differentially expressed in the target (pathological) cells as compared to the reference cells (normal). Such candidate genes may provide a direct recipe for design of new therapeutic strategies. This paper will focus on methods for identification of differential expression.

A sample of measurement replicates is the starting point for any statistical inference. Reproducibility of measurements and between-slide variation has been a major prob-

lem of microarray technology. A popular concept of using ratios is believed to provide an adjustment for slide-specific effects (Khan *et al.*, 1998; Chen *et al.*, 1997).

Each method of analysis of differential expression provides unbiased statistical inference under specific assumptions; and it may be misleading or inefficient if these assumptions do not hold. However, little has been done to formulate statistical models of the experiments and to characterize situations where particular methods provide unbiased results, and where they should be avoided. Newton *et al.* (2000) propose a Bayesian model-based approach to the statistical analysis of differential expression. In this paper we develop methods from the frequentist prospective and compare them with a number of existing practices of testing for differential expression using simulations and real data.

Most experimental designs associate reference cells with one microarray channel (green or red), and the target cells with the other channel. A paired (matched) sample structure can be used to reduce the effect of between-slide variability. In the microarray experiments conducted at the Huntsman Cancer Institute each slide consists of two half-slides. We treat the observations from different half-slides as independent experiments and refer to systematic changes of expression signals between half-slides as the slide effect. As a refinement of the analysis, a nested error structure allowing for effects shared by two half-slides belonging to the same slide is worth further exploration. Let m be the number of half-slides (even), and n be the number of genes. Associated with each half-slide, indexed by $j = 1, \ldots, m$, is a pair of dependent random variables X_{ij}, Y_{ij} representing paired (two channels) measurements of gene expression for each gene indexed by $i = 1, \ldots, n$, where X refers to one particular channel, and Y to the other channel. Subject to between-slide variability, the pairs X_{ij}, Y_{ij} are not identically distributed for different values of j. Let A_i, B_i be a pair of random variables representing the outcome in an ideal reproducible experiment and let x_{ij}, y_{ij} be the observed realization of X_{ij}, Y_{ij}. An adjustment is a method to restore the ideal sample a_{ij}, b_{ij} (drawn from A_i, B_i) or its surrogate by transforming the observed sample x_{ij}, y_{ij}. Under the null hypothesis of no differential expression, A_i and B_i are, by definition, independent and identically distributed (i.i.d.) for any given i. Informally, the model behind a microarray experiment can be written as

$$\varphi_1(A_i, (\text{measurement error})_{ij}) = X_{ij},$$
$$\varphi_2(B_i, (\text{measurement error})_{ij}) = Y_{ij},$$

where φ_1, φ_2 are some non-random functions. If the measurement error in the above equation does not depend on i, all genes on the same half-slide share the same slide-specific random effect. Since the number of genes is very large, an estimate of the slide-specific effect for a particular j is based on a very large sample, and therefore has negligible variability. For example, in the estimates of the slide-specific effect involving terms of the form $\mu_n = \frac{1}{n}\sum_{i=1}^{n} A_i$ we have $\mathrm{Var}(\mu_n) = \frac{1}{n^2}\left\{\mathrm{Var}(A_i) + 2\sum_{i>j}\mathrm{Cov}(A_i, A_j)\right\}$. Assuming that $\mathrm{Var}(A_i)$ and $\mathrm{Cov}(A_i, A_j)$ are uniformly bounded, and additionally $\#\{(i, j) : \mathrm{Cov}(A_i, A_j) > 0\} = o(n^2)$, we have $\mathrm{Var}(\mu_n) \to 0$, as $n \to \infty$. In other words we can assume that the variance $\mathrm{Var}(\mu_n)$ is negligibly small if the number of independent gene-pairs is much larger than the number of dependent gene-pairs.

Adjustment procedures commonly combine one or several of the following transformations: normalizing the mean intensity on a slide, using the ratio (or log-ratio) of the intensities in the two channels, normalizing the observations for each gene to have mean 0 and variance 1, using internal control genes, etc. These methods may have some intuitive sense. However, it is unclear whether and under what assumptions do they provide unbiased results. Not only can the results on differential expression be biased, but also measurement-induced correlation may seriously derange cluster analysis of between-gene relationships.

In this paper we investigate how specific assumptions about the error process (specifically about the functions φ_1 and φ_2) justify adjustment procedures and statistics for testing differential expression. The value of test statistics can be interpreted as a *measure of differential expression*, thus producing an ordering of the genes according to the extent of their differential expression. In a testing setup one would need to define a cutpoint for the statistic that has suitable type I error and power characteristics, including adjustment for multiple comparison with dependent samples. We do not consider this problem in our paper. Our primary focus is the comparison of the adjustment procedures based on their ability to restore the correct ordering of the genes. We will compare various procedures using both simulated and real data.

2 METHODS

2.1 Continuous models

It is difficult to specify the form of φ_k, $k = 1, 2$ on mechanistic grounds. A fairly general model combines additive (γ and δ) and multiplicative (α and β) measurement error:

$$X_{ij} = \alpha_{ij} A_i + \gamma_{ij} \quad Y_{ij} = \beta_{ij} B_i + \delta_{ij}. \qquad \text{(M0)}$$

The systematic part of α and β accounts for the difference in intensity associated with the type of fluorescent dye used with a specific channel. Experimental conditions such as uneven depth of hybridization gel within and between slides may be responsible for the dependence of

measurement errors on the gene number i (location of the corresponding spot on the slide), and on the slide number j. An additive error term comes from the background effect, which may be different for the two channels, and may vary from slide to slide and often within a slide as well. Without further restrictions on the measurement error structure, the model (M0) is not identifiable.

2.1.1 The random effect model. By far the most frequently used assumption is that the measurement error has a simple multiplicative structure. Under this model it is assumed that the multiplicative measurement error is slide-specific, and is shared by genes belonging to the same half-slide. In this model α, β depend on the half-slide number j, and do not depend on the gene number i: $\alpha_{ij} = \alpha_j$, $\beta_{ij} = \beta_j$. Based on the above assumptions, we have the model:

$$X_{ij} = \alpha_j A_i \quad Y_{ij} = \beta_j B_i. \tag{M1}$$

Denote by E_j the empirical expectation taken with respect to all the genes on the jth half-slide, that is $E_j(u) = \frac{1}{n}\sum_{i=1}^{n} u_{ij}$ for any u. Applying E_j to both sides of equation (M1) and dividing X_{ij} and Y_{ij} by their E_j-expectations, we have the following adjustment procedure:

$$\begin{aligned} X^*_{ij} &= X_{ij}/E_j(X_{ij}) = \frac{A_i}{E_j(A_i)} \\ Y^*_{ij} &= Y_{ij}/E_j(Y_{ij}) = \frac{B_i}{E_j(B_i)}. \end{aligned} \tag{REFF}$$

As mentioned in Section 1 the variability of E_j is ignored. Dividing the observations by the mean intensity on each slide is indeed a common practice. If model (M1) is assumed, then under the null hypothesis of no differential expression for a particular gene i, X^*_{ij} and Y^*_{ij} are i.i.d. random variables.

We, as well as other researchers (van der Laan and Bryan, 2000; Kerr *et al.*, 2000), have found that the marginal distribution of the raw data X_{ij}, Y_{ij} is approximately lognormal. To improve normality it makes sense to use log-transformed data. The corresponding adjustment has the form:

$$\begin{aligned} X^*_{ij} &= \frac{X_{ij}}{\exp\{E_j(\log X_{ij})\}} \\ Y^*_{ij} &= \frac{Y_{ij}}{\exp\{E_j(\log Y_{ij})\}}. \end{aligned} \tag{REFFLOG}$$

A two-sample test statistic applied to the transformed data can be used to quantify differential expression. In the subsequent analysis we use the t-test and the Kolmogorov–Smirnov test with the log-based adjustment (REFFLOG). For each gene i the hypothesis to be tested is that $\log(X^*_{ij})$ and $\log(Y^*_{ij})$ come from the same distribution.

Generally speaking, adjustments REFF and REFFLOG are not equivalent. The convexity of the log function implies that $E_j(\log U_{ij}) < \log E_j(U_{ij})$, by the Jensen inequality. Consequently (REFFLOG) will be shifted towards larger adjusted values than (REFF).

2.1.2 The ratio model. A number of factors are believed to affect both microarray channels simultaneously. For example, fluorescence intensity is proportional to the length of mRNA that is gene-specific and is shared by the expression measurements in both channels corresponding to a particular gene. For this reason it is argued that differential expression corresponding to a particular gene is best represented by the ratio of fluorescent intensities. A common adjustment procedure is to restore the ratio of ideal measurements rather than their absolute values. The ratio model is

$$X_{ij} = \alpha_{ij} A_i \quad Y_{ij} = \rho_j \alpha_{ij} B_i, \tag{M2}$$

where ρ_j represents the ratio of the reference intensities of the two channels. Denote by $R_{ij} = X_{ij}/Y_{ij}$ the observed ratio and by $R_i = A_i/B_i$ the ideal ratio. Then $R_{ij} = R_i/\rho_j$ and $E_j(R_{ij}) = E_j(R_i)/\rho_j$.

As discussed in Section 1, we ignore the variability of $E_j(R_i)$, since the empirical expectation E_j is taken over a very large number of genes. In some particular cases (for example if A_i and B_i were independent normal), the expectation ER_i does not exist, to say nothing about the variance. However, we will assume that both the expectation and the variance of the ratio are finite. This assumption holds in the special case of lognormally distributed gene expressions. Indeed, if A_i and B_i are independent lognormal with expectation μ and shape parameter σ then R_i is also lognormal with expectation 1 and shape parameter 2σ. Consequently, under the null hypothesis $E_j(R_i) = E(R_i) = 1$.

We have the adjustment

$$R^*_{ij} = R_{ij}/E_j(R_{ij}) = R_i. \tag{RATIO}$$

A one-sample t-test for $E(R^*_i) = 1$ will identify the differentially expressed genes.

As before, an adjustment after a log transformation is preferred to improve the properties of the t-test:

$$R^*_{ij} = R_{ij}/\exp[E_j \log R_{ij}]. \tag{RATIOLOG}$$

Testing for $E(\log(R)^*_i) = 0$ identifies differentially expressed genes.

2.2 Categorical adjustment: rank-based methods

We have used fairly strong assumptions to derive model-based adjustment procedures. They are likely to be

sensitive to outliers and to model misspecification. Note that adjustments (REFF) and (REFFLOG) preserve the ordering of observations: if $X_{ij_1} > X_{ij_2}$, then $X^*_{ij_1} > X^*_{ij_2}$. A more robust adjustment would be to replace each observation by its rank in the half-slide sample. To make the resulting values insensitive to the number of genes n, the ranks can be normalized by n. The corresponding adjustment procedure can be summarized as

$$X^*_{ij} = \text{rank}_j X_{ij}/n, \quad Y^*_{ij} = \text{rank}_j Y_{ij}/n, \quad \text{(RANKS)}$$

where $\text{rank}_j u_{kj}$ is the place (counted from the left) of u_{kj} in the series u_{ij}, $i = 1, \ldots, n$ arranged in a decreasing order for each j.

Adjustment (RANKS) restores the correct ordering of observations regardless of the error structure (multiplicative, additive, etc.) as long as it is order-preserving and error terms are shared by measurements on a half-slide. Also note that data represented by ranks are invariant to any monotone transformation. Again, the t-test for equality of the mean normalized ranks $E(X^*_i) = E(Y^*_i)$, or other pertinent tests can be used to identify differentially expressed genes.

A similar rank-based adjustment can be constructed for the expression ratios:

$$R^*_{ij} = \text{rank}_j R_{ij}/n. \quad \text{(RANKSRATIO)}$$

This procedure is less sensitive than (RATIO) and (RATIOLOG) to possible variations of ρ_j within a half-slide. To identify differentially expressed genes based on this adjustment, the hypothesis $E(R^*_i) = (n+1)/(2n)$ is tested, where the right side is the average of normalized ranks on a half-slide.

2.3 Categorical adjustment: scatter plot methods

2.3.1 Paired data. A scatter plot of expression measurements from a particular half-slide is a major tool of exploratory analysis of differential expression. Each spot on a half-slide corresponds to a particular gene. Measurements of fluorescent intensity in the two channels (x = green and y = red) gives a point (x, y) on the plane. A set of all such points for the genes associated with a given half-slide forms a scatter plot. Ideally, non-differentially expressed genes would preserve a constant green/red ratio of 1, the corresponding (x, y) points building a line on the plane. A differentially expressed gene would ideally show a different ratio, the corresponding points being away from the line.

Unfortunately, a number of factors complicate the observed picture:

- additive background effect provides for a non-zero intercept of the line;
- due to measurement errors and random character of gene expressions even the points corresponding to non-differentially expressed genes are scattered considerably around the line;
- a strong slide-specific effect makes the picture variable from slide to slide.

Statistical challenges of working with real expression data include putting all scatter plots on the same scale, restoring a line corresponding to ideal location of non-differentially expressed genes, quantification of differential expression as deviations from the line, and summarizing measurements from a series of scatter plots (half-slides) using an appropriate statistical test.

2.3.2 Reference line. The use of the least squares line based on linear regression is not warranted in this case. The sample of x and y values corresponds to a system of dependent random variables with unknown dependency structure. The sample points $\{(x_i, y_i)\}_{i=1}^n$ of the scatter plot contain an unknown fraction of outliers that are not supposed to follow the line. Also, both x and y are subject to measurement error. Even with independent pairs, a linear structural relationship would be non-identifiable without additional constraints, if both x and y were measured with error. Fitting a least squares line under the simple model

$$X_i = U_i + \delta_i \qquad Y_i = V_i + \epsilon_i, \qquad (1)$$

where δ and ϵ are measurement errors, and $V = a + bU$, results in underestimation of the slope b of the latent structural relationship (Seber, 1977). The models considered so far for gene expression data were more complex than (1). Given all the complications mentioned above we resort to an *ad hoc* method to put a reference line through the scatter plot and will develop a method for statistical inference that is not based on the assumption of independent points in the scatter plot. Having explored a number of robust procedures for linear regression using real and simulated expression data we came up with a simple and computationally fast method based on the one studied by Bartlett (1949). A set of n plotted points is divided into three groups, the equal numbers k in the two extreme groups being chosen to be as close to $n/3$ as possible. This is accomplished by sorting and dividing the x-points. The line that joins points corresponding to the mean coordinates for the two extreme groups is used to determine the slope of the sought-for line. To determine the location of the line we use the point with coordinates $(\min_i(x_i), \min_i(y_i))$. Once the reference line is determined, all n points of the scatter plot are projected on the line. Let $v = a + bu$ be a line, and (y, x) be a point of the scatter plot. The projection

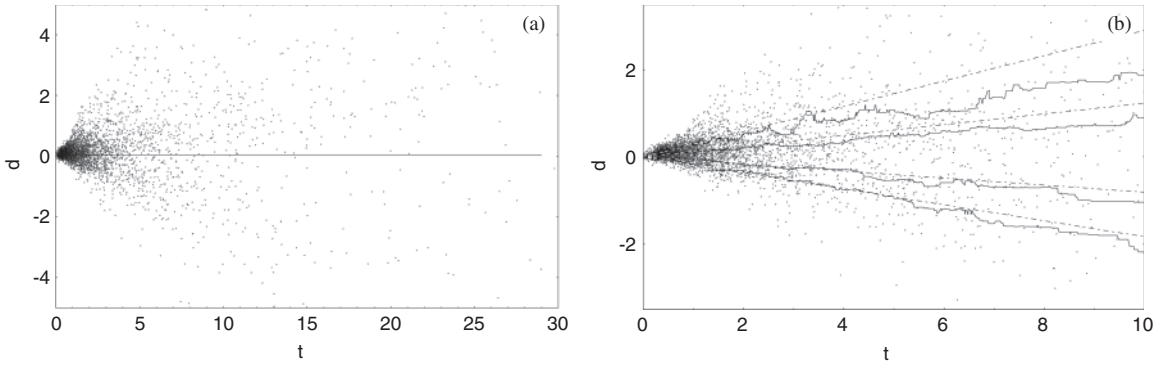

Fig. 1. (a) Scatterplot in transformed coordinates (t, d). The fitted line coincides with the t axis. (b) Estimated percentile bands. Real data, see Section 3.2.

of (x, y) on the line will have the coordinates $u' = (by + x - ab)/(b^2 + 1)$ and $v' = a + b(by + x - ab)/(b^2 + 1)$. Finally, the coordinate system is changed from (x, y) to (t, d), as shown in Figure 1a, where t is the distance from minimal projection on the reference line to (u', v'), and d is the signed distance from (x, y) to its projection (u', v'). The signed distance d quantifies an instance of differential expression for a particular half-slide. Positive d, corresponding to a point above the line is an indication of potential overexpression, while negative d is an indication of potential underexpression.

2.3.3 Percentile bands. In view of the considerable difficulties in formulating and testing an adequate statistical model for expression data, model-free statistical methods hold much promise. A summary measure of differential expression can be constructed by ranking genes with respect to the directional distance d to the reference line. Given the divergent structure of the scatter plot (Figure 1), d is an inadequate measure of differential expression when applied overall to all genes. Differential expression of a particular gene i with coordinates (t_i, d_i) should be measured against the scattering at the cross section of the scatter plot at t_i. To categorize differential expression define a cross section layer $W_t^+ = \{0 < d < \infty, t - \Delta(t) < t < t + \Delta(t)\}$, where $\Delta(t)$ is a bandwidth (see Figure 2). Similarly, $W_t^- = \{-\infty < d < 0, t - \Delta(t) < t < t + \Delta(t)\}$. Define a set of cutpoints α_j, $j = 0, \ldots, k + 1$ that break the interval of total probability [0, 1] down into $k + 1$ subintervals. By definition $\alpha_0 = 0$, $\alpha_{k+1} = 1$, $\alpha_{j-1} < \alpha_j$. A gene with coordinates (t_i, d_i) above the reference line is assigned a category of differential expression $\mathcal{C}_j^+$ if $C_{\alpha_j}^+ < d_i \leqslant C_{\alpha_{j+1}}^+$, where C_α^+ is the empirical α-percentile of the distribution of d for genes in the layer $W_{t_i}^+$. All genes in W_t^- under the line are categorized in a similar manner. In fact, as W_t depends on t, C_{α_j} is a function of t representing a moving-average estimator of

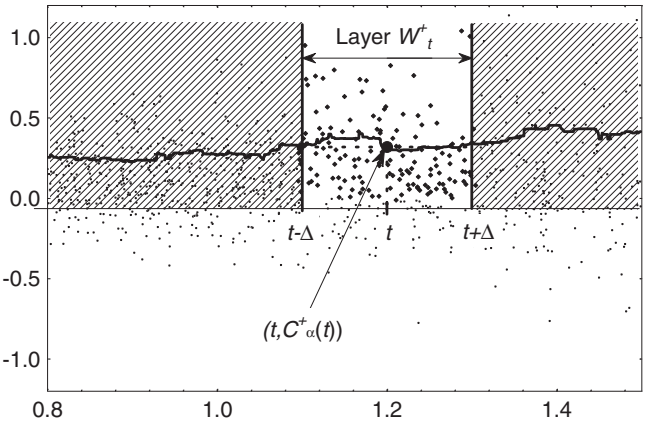

Fig. 2. Construction of a percentile band.

the α_j-percentile of the distribution of d given t (see Figure 2). The step-functions $C_{\alpha_j}(t)$ cut the plane into $2k + 1$ percentile bands $\mathcal{B}_j^+ = \{0 \leqslant t < \infty, C_{\alpha_j}^+ < d \leqslant C_{\alpha_{j+1}}^+\}$ and $\mathcal{B}_j^- = \{0 \leqslant t < \infty, C_{\alpha_{j+1}}^- < d \leqslant C_{\alpha_j}^-\}$ (the bands $\mathcal{B}_0^+$ and $\mathcal{B}_0^-$ are combined into a single one).

To keep the estimation accuracy constant, Δ is treated as data-adaptive and such that for any t the layer W_t contains approximately the same number of points. We also imposed a constraint on the maximal bandwidth.

With $k = 1$ a realization of gene expression associated with a particular half-slide can fall into one of the three categories: 'Overexpressed' (the point is in the upper band $\mathcal{B}_1^+$), 'Not differentially expressed' (the point is in the middle band $\mathcal{B}_0$) and 'Underexpressed' (the point is in the lower band $\mathcal{B}_1^-$). With $k > 1$ overexpression and underexpression are represented in more detail as a

number of categories:

$$(X_{ij}, Y_{ij}) \to \begin{cases} \text{Overexpr. } k & \text{if} \quad (t_{ij}, d_{ij}) \in \mathcal{B}^+_{kj} \\ \cdots & \cdots \quad \cdots \\ \text{Overexpr. 1} & \text{if} \quad (t_{ij}, d_{ij}) \in \mathcal{B}^+_{1j} \\ \text{Not diff. expr.} & \text{if} \quad (t_{ij}, d_{ij}) \in \mathcal{B}_{0j} \\ \text{Underexpr. 1} & \text{if} \quad (t_{ij}, d_{ij}) \in \mathcal{B}^-_{1j} \\ \cdots & \cdots \quad \cdots \\ \text{Underexpr. } k & \text{if} \quad (t_{ij}, d_{ij}) \in \mathcal{B}^-_{kj}. \end{cases} \tag{CAT}$$

An important feature of the proposed categorical summary measure of differential expression is that any rank preserving transformation (possibly dependent on the absolute expression level t) of ideal expression data will be adequately adjusted for.

2.3.4 Ratio. The non-parametric percentile curves C_α may show a parametric trend. In this case, to improve the sensitivity of estimation procedure, a parametric form of C_α can be assumed. The curves shown on Figure 1b, for example, indicate a linear relationship. If $C_\alpha(t)$ is linear in t, the estimation of percentile bands is a trivial issue of determining the percentiles $\tilde{C}_{\alpha_j}$, $j = 0, \ldots, k$ of the ratio d/t. We have the adjustment

$$(X_{ij}, Y_{ij}) \to \begin{cases} \text{Overexpr. } k & \text{if} \quad \tilde{C}^+_{\alpha_k} < \frac{d_{ij}}{t_{ij}} \leqslant \tilde{C}^+_{\alpha_{k+1}} \\ \cdots & \cdots \quad \cdots \\ \text{Overexpr. 1} & \text{if} \quad \tilde{C}^+_{\alpha_1} < \frac{d_{ij}}{t_{ij}} \leqslant \tilde{C}^+_{\alpha_2} \\ \text{Not diff. expr.} & \text{if} \quad \tilde{C}^-_{\alpha_1} \leqslant \frac{d_{ij}}{t_{ij}} \leqslant \tilde{C}^+_{\alpha_1} \\ \text{Underexpr. 1} & \text{if} \quad \tilde{C}^-_{\alpha_2} \leqslant \frac{d_{ij}}{t_{ij}} < \tilde{C}^-_{\alpha_1} \\ \cdots & \cdots \quad \cdots \\ \text{Underexpr. } k & \text{if} \quad \tilde{C}^-_{\alpha_{k+1}} \leqslant \frac{d_{ij}}{t_{ij}} < \tilde{C}^-_{\alpha_k}. \end{cases} \tag{CATRATIO}$$

It should be noted, however, that the ratio model (M2) described in Section 2.1 generally does not provide linear percentile bands. For this reason the ratio model should not be confused with the above method. Indeed, it follows from (M2) that $R_{ij} = R_i/\rho_j$. Given ρ_j and a sample from R_i, the points (x_{ij}, y_{ij}) follow the lines $y_{ij} = t_{ij} b_{ij}/\sqrt{1 + b^2_{ij}}$, where $b_{ij} = R_i/\rho_j$ and t_{ij} is arbitrary. Appropriately choosing t_{ij} we may reproduce an arbitrary shape of percentiles of the distribution of d given t, as a function of t.

2.3.5 Testing for symmetry. Under the null hypothesis of no differential expression, genes are expected to show overexpression approximately as often as they show underexpression. In other words, the distribution of a categorical measure of differential expression over a set of half-slides is symmetric under the null hypothesis.

For a given gene i, introduce the notation: n^+_i = the number of half-slides where the gene happened to be in category $\mathcal{C}^+_i$; n^-_i = the number of half-slides where the gene happened to be in category $\mathcal{C}^-_i$; n_0 = the number of half-slides where the gene happened to be in category $\mathcal{C}_0$; p^+_i = the probability for the gene of being in category $\mathcal{C}^+_i$; p^-_i = the probability for the gene of being in category $\mathcal{C}^-_i$; p^0_i = the probability for the gene of being in category $\mathcal{C}^0_i$. The total number of half-slides $m = \sum_{i=1}^k (n^+_i + n^-_i) + n_0$.

The null hypothesis that the gene is not differentially expressed can be formulated as $p^+_i = p^-_i = p_i$, $i = 1, \ldots, k$. Under the null hypothesis $\hat{p}_i = (n^+_i + n^-_i)/(2m)$, $\hat{p}_0 = n_0/m$. Under a saturated model, $\hat{p}^+_i = n^+_i/m$, $\hat{p}^-_i = n^-_i/m$, $\hat{p}_0 = n_0/m$.

The likelihood ratio statistics can be used to summarize and quantify differential expression over a series of experiments: $\mathrm{LR} = 2\sum_{i=1}^k (n^-_i \log(n^-_i) + n^+_i \log(n^+_i) - (n^-_i + n^+_i)\log(n^-_i + n^+_i))$. Under the null hypothesis LR is asymptotically χ^2-distributed with k degrees of freedom.

The LR statistic computed for each gene is used to order genes according to their differential expression.

3 RESULTS

3.1 Data analysis and simulation experiments

A method of scoring differential expression is based on a particular adjustment (data transformation) and subsequent application of an appropriate test statistic to the transformed sample. Below we compare the efficiency at restoring the correct order according to differential expression for the proposed adjustment procedures. As a reference we included the option of no adjustment (RAW) and random ordering.

We have simulated 'ideal' gene expression data and then incorporated error terms according to (M0). Differential expression was measured by the log-ratio $|\log(B_i/A_i)|$ that was set by the simulation program, so the correct ordering was known. Categorical adjustments were based on five categories. Constant probabilities $\alpha_j = \alpha = 0.33$ provided the best performance. With ratio-based adjustments RAWRATIO, RATIO, RATIOLOG and RANKSRATIO, a one-sample t-test statistic was used to quantify differential expression; with non-ratio based adjustments RAW, REFF, REFFLOG and RANKS the two-sample t-test was used. With categorical adjustments CAT and CATRATIO we used the Likelihood Ratio (LR) symmetry test described in Section 2.3.

3.1.1 Indicators of performance. The purpose of our analysis of differential expression is to rank genes according to their differential expression. Each method provides a score that is used for the ranking. The score is based on a test statistic appropriate for the type of transformed (adjusted) data. Using simulated data, the ranking suggested by the method can be verified against the 'true' ranking known as a result of simulation. Newton *et al.* (2000) sug-

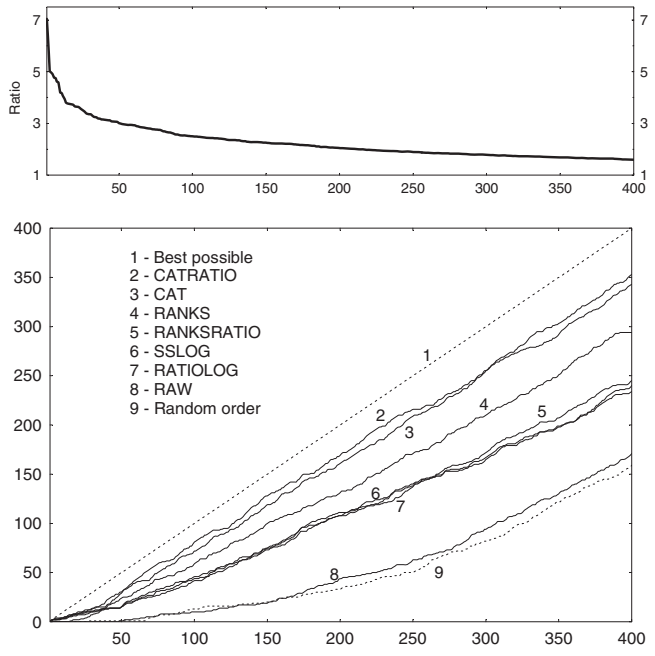

Fig. 3. Comparison of two-sample tests versus one-sample tests (based on ratio) used with different adjustment procedures to score differential expression. The top panel shows the ratio set by the simulation program; the bottom panel shows the quality function $M(N)$ (see text). Simulated data.

gested a simple plot to compare the performance of different methods in restoring the 'true' ranks. Consider the N 'truly' most differentially expressed genes as known from the simulations. Let $M(N)$ be the number of these genes ranked in the top N by a particular method. Ideally, if the method restores the true ranking error free, we will have $M(N) = N$ for any N. Generally, if errors are present, $M(N) \leqslant N$. Obviously, $M(N)$ is an increasing function such that $M(0) = 0$ and $M(n) = n$. The more the curve $M(N)$ arches towards the lower right corner $(0, n)$ on the plane $\{M\} \times \{N\}$, the poorer the method's performance. Figure 3 shows examples of such curves. Only the initial portion of the performance curve $M(N)$ is interesting. In Figure 3 this portion is limited to 400 (out of 1000) simulated genes. The extent of the simulated differential expression is illustrated by a plot of the ratio $\exp\{|\log(B_i/A_i)|\}$ by the 'true' gene rank i as used in the simulations. This value decreases from about 7 for the most differentially expressed genes to 1.5 for the gene ranked # 400.

3.1.2 Simulation studies. Shown in Figure 3 is the comparison of various approaches. Without the adjustment component (curve 8), a two-sample test performs virtually like the worst possible procedure that ranks genes by random permutation without making any use of expression measurements (curve 9). Even the simplest adjustment provides a substantial improvement in performance. This is the commonly used procedure of normalizing expressions by means taken over genes on the same half-slide (REFFLOG, curve 6). Rank-based procedures (curves 4 and 5) are second best suggesting that robust approaches outperform procedures based on restrictive model assumptions. It is clear from Figure 3 that the categorical procedures (CATRATIO and CAT, curves 2 and 3) are by far the best in the comparison. With the exception of rank-based procedures, working with ratios allows us to restore the correct ordering almost as accurately as the corresponding two-sample approach (Figures 3, 6 versus 7, 2 versus 3). However, the two-sample test with RANKS adjustment outperforms the one-sample test applied after the ranking of the ratio (RANKSRATIO). Investing the information that percentile bands are approximately linear in the analysis slightly improves the performance of categorical procedures (curve 2 versus 3). Along with the performance improvement associated with robustness, categorical methods are subject to information loss due to creation of ties. It is the most differentially expressed genes reflective of the initial portion of the performance curve $M(N)$ that are affected by the loss of ordering information for tied observations. Indeed, the RANKS method outperforms the CATRATIO method for some 25 most differentially expressed genes. To combine robustness of the categorical procedures with more detailed information on ordering of tied observations, ties can be broken according to their order in the RANKS method. With this treatment of ties the CATRATIO curve becomes as good as the RANKS curve for the initial portion of $M(N)$ and maintains its advantage for the entire range.

All the above results were obtained with a fixed samples size $m = 22$ that matches the sample size available for the biological data described below. To investigate the effect of sample size we used the best ranking approach as determined in our simulation study: LR test score for the CATRATIO adjustment with ties broken according to the RANKS-based test score. Figure 4 shows the curves $M(N)$ estimated for various values of the sample size m. As expected, increasing the sample size improves the accuracy of the ordering of the genes, especially for moderately large 2–3-fold differences. This is consistent with the intuitive expectation that extremely large differences can be identified even with small sample sizes and identification of small differences (and especially the exact order of those values!) is difficult even with large sample sizes.

3.2 Analysis of biological data

We selected two commonly studied colon cancer cell lines for our analyses. HT29 cells represent advanced, highly aggressive colon tumors. They contain mutations in both

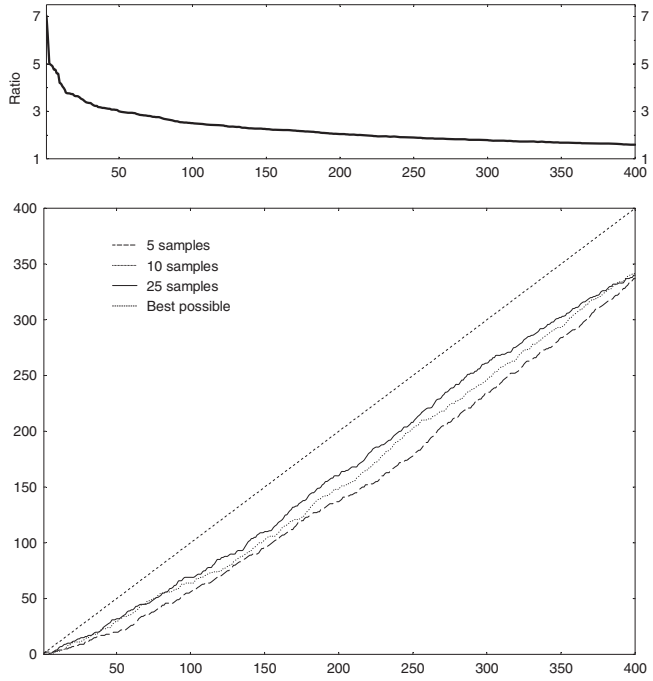

Fig. 4. The effect of sample size on restoring the correct ranking based on CATRATIO with ties broken according to RANKS. The top panel shows the ratio set by the simulation program; the bottom panel shows the quality function $M(N)$ (see text). Simulated data.

the APC gene and p53 gene, two tumor suppressor genes that frequently mutate during colon tumorigenesis. As another cell type, we selected HCT116 cells. This cell line models less aggressive colon tumors and harbors functional p53 and APC. However, they show a deficiency of those genetic systems that are responsible for the repair of mismatched regions of DNA. To generate the data, three samples of each mRNA (1 μg each) were labeled by production of first-strand cDNA in the presence of Cy3-dCTP (green) or Cy5-dCTP (red). Six identical comparison sets of samples were labeled in the reverse orientation. In the first three sets Cy-3 was used to label HCT116 cells while Cy-5 was used for HT29 cells. In the next three sets Cy-5 was used with HCT116 while HT29 was labeled using Cy-3. Each comparison set was hybridized against two microarray slides containing 4608 minimally redundant cDNAs spotted in duplicate. This experiment resulted in a total of 24 measurements for each gene on the microarrays (unpublished data). The top ten most differentially expressed genes based on CATRATIO with ties broken using the RANKS method are listed in Figure 5. It is interesting to note that both copies, not just one, of Cyclin-dependent kinase inhibitor 1 appear in the list.

The results of the simulation study described above ap-

```
1   Metallothionein 1L
2   Thymosin beta-10
3   Cyclin-dependent kinase inhibitor 1
4   P55-C-FOS proto-oncogene protein
5   Human effector cell protease receptor-1(EPR-1) gene, partial cds
6   Interferon regulatory factor 1
7   ESTs, Moderately similar to AGRIN PRECURSOR [Rattus norvegicus]
8   Human focal adhesion kinase(FAK)mRNA,complete cds
9   Cyclin-dependent kinase inhibitor 1
10  ESTs
```

Fig. 5. HT29 versus HCT119 cell lines: 10 most differentially expressed cDNAs.

pear to be corroborated by the analysis of real data. Unfortunately, the correct ordering of the genes is unknown in this (or any other real) dataset. Thus we cannot make a plot similar to Figure 3. The concept of a major proportion of housekeeping genes mixed with a minor proportion of differentially expressed genes would intuitively provide a score distribution approximately symmetric around zero with a small fraction of outliers. From Figures 6 and 7 it is evident that with no adjustments the center of the distribution of score is shifted away from zero, while adjustment procedures restore the proper shape for both the simulated and real data. It should be noted that categorical procedures that performed best in the simulation study provide the distribution of score that really looks like a mixture of a unimodal distribution associated with housekeeping genes and an outlier noise affecting the extremes (Figure 7, CAT).

4 DISCUSSION

'All models are wrong but some are useful' (Box, 1979). When building a statistical model of complex phenomena such as gene expression, model misspecification is inevitable. In the final analysis the statistical model is judged pragmatically by its ability to provide unbiased assessment of differential expression. When trying to make the model more realistic and more complex one has to reckon with two potential consequences. If added complexity captures the most essential features of the data, the corresponding statistical methods will be more powerful. On the other hand, the same complexity may be part of model misspecification contributing to worse performance of a statistical procedure. This tradeoff is resolved with a reasonably rough (robust) method. The results of our analysis of gene expression data suggest that robust categorical methods outperform methods associated with a precisely defined stochastic model, including ratio-based methods.

The analysis in the paper was focused on macro properties of the microarray methods as an exploratory tool to identify a reasonably sized pool of candidate genes for differential expression. Making statements about particular genes would be premature at this point of analysis. Be-

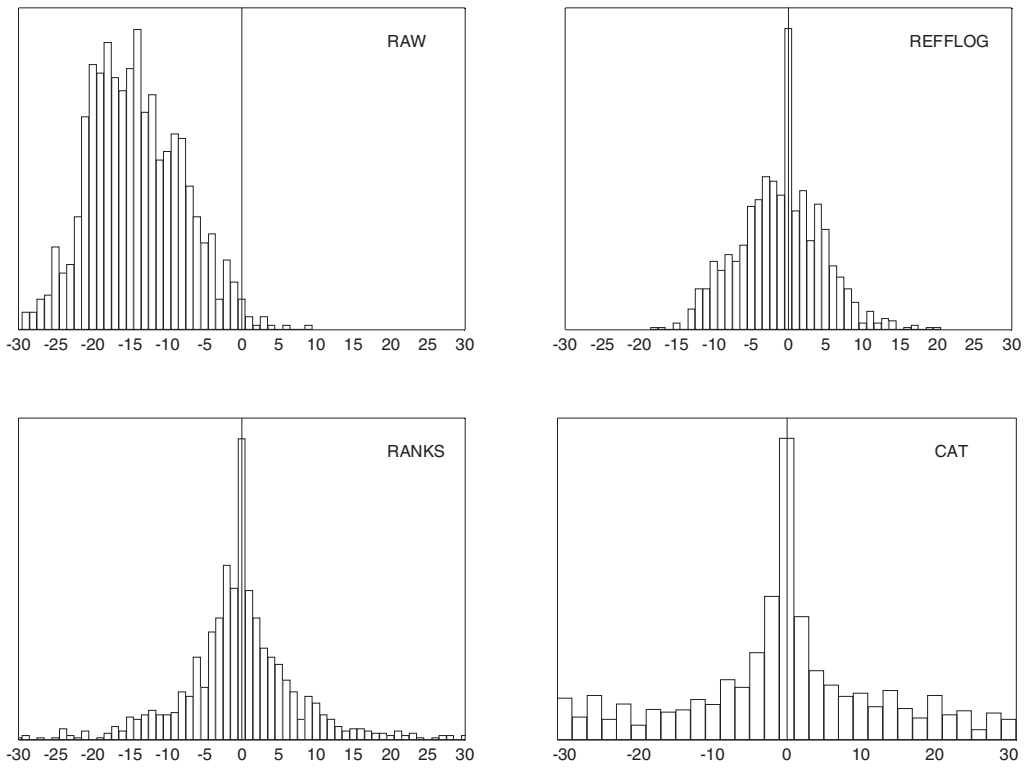

Fig. 6. Frequency histograms of differential expression scores associated with different scoring methods. Simulated data.

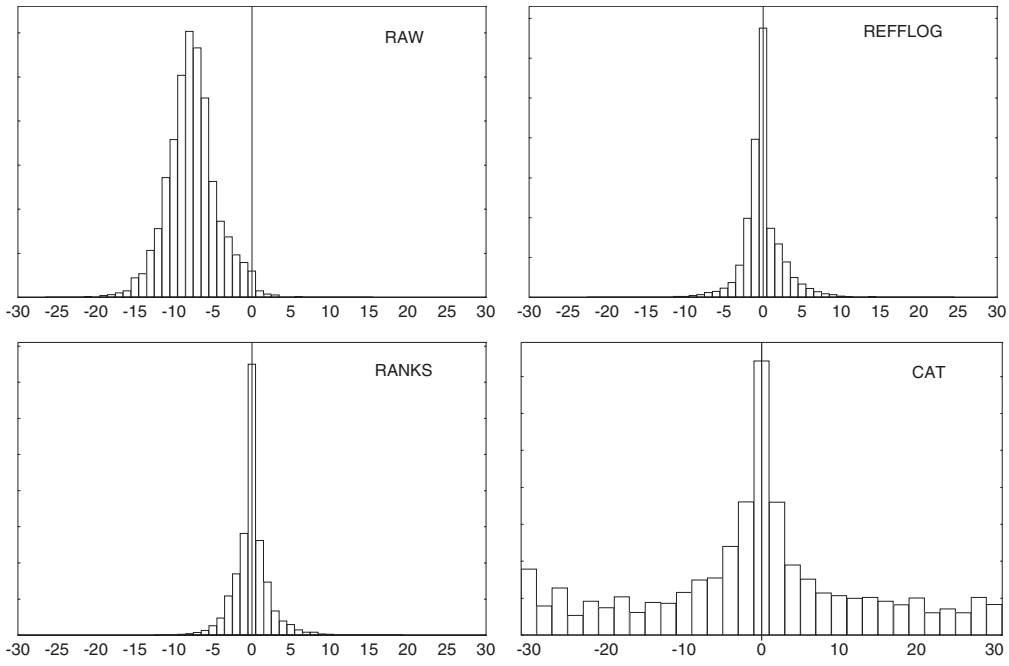

Fig. 7. Frequency histograms of differential expression scores associated with different scoring methods. HT29 versus HCT119 data.

fore we proceed with statements about particular genes, differential expression of candidate genes will be subject to second order validation using Northern Blot analysis focused on most promising candidates. Such analysis will be supplemented by clustering genes according to their function and dependency structure to obtain a clear picture. We plan to develop new biostatistical methods and refine existing ones on the way to better understanding the difference between colon cancer cell lines through analysis of gene expression.

ACKNOWLEDGEMENTS

The authors would like to thank Andrei Yakovlev for many helpful comments and support, Hans Albertsen for initiating this work and stimulating discussions, Peter Peterson, Brian Dalley and the MicroArray Core Facility at the Huntsman Cancer Institute for introduction to the microarray technology and data management. This research was supported, in part, by NCI Cancer Center Support Grant 2P30 CA 42014 and US Civilian Research & Development Foundation Grant AB2-2005.

REFERENCES

Alon,U., Barkai,N., Notterman,D.A., Gish,K., Ybarra,S., Mack,D. and Levine,A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.

Bartlett,M.S. (1949) Fitting a straight line when both variables are subject to error. *Biometrics*, **5**, 207–213.

Basett,G.E. Jr, Eisen,M.B. and Boguski,M.S. (1999) Gene expression informatics it's all in your mine. *Nature Genet.*, **21**, 51–55.

Ben-Dor,A., Shamir,R. and Yakhini,Z. (1999) Clustering gene expression patterns. *J. Comput. Biol.*, **6**, 281–97.

Box,G.E.P. (1979) Robustness in the strategy of scientific model building. In Launer,R.L. and Wilkinson,G.N. (eds), *Robustness in Statistics*. Academic, New York, pp. 201–236.

Carr,D.B., Somogyi,R. and Michaels,G. (1997) Templates for looking at gene expression clustering. *Stat. Comput. Stat. Graph. Newsl.*, 20–29.

Chen,Y., Dougherty,E. and Blittner,M. (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Opt.*, **2**, 364–374.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14 863–14 868.

Herwig,R., Poustka,A., Muller,C., Bull,C., Lehrach,H. and O'Brien,J. (1999) Large-scale clustering of cdna-fingerprinting data. *Genome Res.*, **9**, 1093–1105.

Heyer,L.J., Kruglyak,S. and Yooseph,S. (1999) Exploring expression data: indentification and analysis of co-expressed genes. *Genome Res.*, **9**, 1106–1115.

Kerr,M.K., Martin,M. and Churchill,G.A. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–837.

Khan,J., Simon,R., Bittner,M., Chen,Y., Leighton,S.B., Pohida,T., Smith,P.D., Jiang,Y., Gooden,G.C., Trent,J.M. and Meltzer,P.S. (1998) Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.*, **58**, 5009–5013.

van der Laan,M.J. and Bryan,J.F. (2000) *Gene Expression Analysis with the Parametric Bootstrap* (*Public Health Series* 86), University of California, Berkeley.

Michaels,G.S., Carr,D.B., Askenazi,M., Fuhrman,S., Wen,X. and Somogyi,R. (1998) Cluster analysis and data visualization of large-scale gene expression data. *Pac. Symp. Biocomput.*, **3**, 42–52.

Newton,M.A., Kendziorski,C.M., Richmond,C.S., Blattner,F.R. and Tsui,K.W. (2000) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.*, **8**, 37–52.

Seber,G.A.F. (1977) *Linear Regression Analysis*. Wiley, New York.