Comparison of the Empirical Bayes and the Significance Analysis of Microarrays

Holger Schwender, Andreas Krause, and Katja Ickstadt*

Abstract

Microarrays enable to measure the expression levels of tens of thousands of genes simultaneously. One important statistical question in such experiments is which of the several thousand genes are differentially expressed. Answering this question requires methods that can deal with multiple testing problems. One such approach is the control of the False Discovery Rate (FDR). Two recently developed methods for the identification of differentially expressed genes and the estimation of the FDR are the SAM (Significance Analysis of Microarrays) procedure and an empirical Bayes approach.

In the two group case, both methods are based on a modified version of the standard t-statistic. However, it is also possible to use the Wilcoxon rank sum statistic. While there already exists a version of the empirical

*Holger Schwender is a Ph.D. Student, Collaborative Research Center SFB 475, University of Dortmund, 44221 Dortmund, Germany (E-mail: holgers@statistik.uni-dortmund.de). Andreas Krause is Senior Statistician, Modeling and Simulation Group, Novartis Pharma AG, Basel, Switzerland (E-mail: andreas.krause@pharma.novartis.com). Katja Ickstadt is Professor, Department of Statistics, University of Dortmund, Dortmund, Germany (E-mail: ickstadt@statistik.uni-dortmund.de).

Bayes approach based on this rank statistic, we introduce in this paper a new version of SAM based on Wilcoxon rank sums. We furthermore compare these four procedures by applying them to simulated and real gene expression data.

Key Words: Identification of differentially expressed genes; Gene expression; Multiple Testing; False Discovery Rate

Acknowledgement

The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures") is gratefully acknowledged.

1 Introduction

A recently developed biotechnology called microarray makes it possible to measure the expression levels of tens of thousands of genes simultaneously. Not only the vast amount of data produced in a microarray experiment but also the fact that the data are very noisy, and that there are usually only a few observations (less than 50) but many variables ($3,000 - 30,000+$ genes), has opened this field of molecular biology for statisticians. Interesting statistical questions reach from experimental design and normalization to multiple testing, clustering and classification.

In this paper, our interest is focused on multiple testing: Our goal is to identify differentially expressed genes, i.e. genes whose expression levels strongly

differ under several conditions (e.g., types of cancer, or treated vs. untreated).

In multiple testing problems, most commonly the family-wise error rate (FWER) is used as an error measure and controlled by some procedure (for a summary of such methods, see Shaffer 1995) like the Bonferroni correction or the adjusted p-values of Westfall and Young (1993). The latter are applied to gene expression data by Dudoit, Yang, Callow, and Speed (2002). It however turns out that it is more appropriate to use the False Discovery Rate (FDR) as an error measure in microarray experiments since the control of the FWER is usually much too conservative for the purpose of a microarray analysis.

Two of the methods for the identification of differentially expressed genes and the estimation of the FDR are the SAM (Significance Analysis of Microarrays) procedure introduced by Tusher, Tibshirani, and Chu (2001) and an empirical Bayes approach proposed by Efron, Tibshirani, Storey, and Tusher (2001). In both procedures, a modified version of the standard t-statistic is used to find genes whose expression levels strongly differ between two groups.

Instead of using a t-statistic, one can also compute a Wilcoxon rank sum for the identification of such genes. While there already exists a version of the empirical Bayes approach using Wilcoxon rank sums (see Efron and Tibshirani 2002), we here introduce a version of SAM based on Wilcoxon rank sums.

While Dudoit, Shaffer, and Boldrick (2003) compare SAM but not the empirical Bayes method with procedures that either control the FWER or the FDR, we here compare the empirical Bayes approaches and the SAM methods by applying them to simulated and real gene expression data. These applications are performed by using the functions contained in our `R` package (Ihaka and Gentleman 1996) called `siggenes`.

This paper is organized as follows. In Chapter 2, we give a description of our testing situation and show how the FDR can be estimated. In Chapter 3,

we present SAM and the two versions of the empirical Bayes approach. The new version of SAM based on Wilcoxon rank sums is introduced in Chapter 4. Chapter 5 contains the comparison of these four methods, and in Chapter 6, our results are summarized and discussed.

2 Multiple Testing and the FDR

An important and common task that arises in microarray experiments is the identification of differentially expressed genes. The goal in such an analysis is to find a fairly large number of genes, typically a few hundred, for further analyses. It will not even matter if a few of these findings are false positives, i.e. not differentially expressed genes that are declared to be differentially expressed, as long as the number of false positives is small in proportion to the number of identified genes.

More formally, denote the number of false positives by V and the number of rejected null hypotheses, i.e. the number of identified genes, by R. Our goal is to keep V/R very small. So a first idea for an error measure would be the expected value $\mathrm{E}(V/R)$. But this definition is useless, since $\mathrm{Prob}(R = 0) > 0$ in almost any case. Therefore Benjamini and Hochberg (1995) propose to use the False Discovery Rate

$$\mathrm{FDR} = \mathrm{E}\left(\frac{V}{R}\middle|\ R > 0\right)\mathrm{Prob}(R > 0),$$

where the FDR will be set to 0 if there is no significant finding, i.e. if $R = 0$.

Under the assumption that the test statistics are independent, Storey (in press) shows that the FDR can be estimated by

$$\widehat{\mathrm{FDR}}(\alpha) = \frac{\hat{\pi}_0 \alpha m}{\max\{\#\{p_i \leq \alpha\}, 1\}},$$

where m is the number of tests/genes, α is the acceptable error rate, p_i is the (uncorrected) p-value of the ith gene, $i = 1, \ldots, m$, and $\hat{\pi}_0$ is an estimate of the prior probability π_0 that a gene is not differentially expressed.

There are several ways how π_0 can be estimated. We use the following estimate proposed by Storey and Tibshirani (2003):

1. For $\lambda = 0, 0.01, \ldots, 0.95$, compute $\hat{\pi}_0(\lambda) = \#\{p_i > \lambda\}/\big((1-\lambda)m\big)$.
2. Fit a natural cubic spline h with 3 degrees of freedom through the data points $\big(\lambda, \hat{\pi}_0(\lambda)\big)$, where each data point is weighed by $1 - \lambda$.
3. Estimate π_0 by $\min\big\{h(1), 1\big\}$.

All these estimates are obtained by assuming that the m test statistics are independent. However, gene expression data can be highly correlated, and hence the test statistics are not all independent. Nevertheless, for large m, these estimates can also be used under dependence (Storey and Tibshirani 2001).

3 Identifying Differentially Expressed Genes

Suppose we have given a data matrix X containing the expression levels x_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n$, of m genes and n biological samples, and we have observed in addition a response y for each of these samples. Since we are interested in two class unpaired data (e.g., case/control) we call the response 1 for each of the n_1 samples in group 1, and 2 for the n_2 samples in group 2.

3.1 Test Statistics

We now would like to identify the genes whose expression levels strongly differ between the two groups. To test for this, one can compute the usual t-statistic for unpaired data for each gene. There is however one problem concerned with this t-statistic that is particularly encountered in microarray experiments: Genes with low expression levels. Since the variance of such genes is very small, their t-value can be very large. To avoid that these genes with low expression levels dominate the results of our analysis, a small, strictly positive constant s_0, the so called *fudge factor*, is added to the denominator of the usual t-statistic. This fudge factor is computed as the quantile of the standard deviations s_i, $i = 1, \ldots, m$, of the genes that fulfills an optimization criterion. For details on the computation of the fudge factor, see Appendix A.1, and for details on the effect of s_0, see Appendix A.2.

Instead of using the usual t-statistic $t_i = r_i/s_i$, where r_i is the difference in mean expression levels between group 2 and group 1, we thus compute for each gene i, $i = 1 \ldots, m$, the expression score

$$d_i = \frac{r_i}{s_i + s_0}. \tag{3.1}$$

Since the null distribution of the d_i-values is unknown, this distribution is estimated by taking B sets of permutations of the response variable, and computing the permuted expression scores d_i^b, $i = 1, \ldots, m$, for each permutation b, $b = 1, \ldots, B$.

An alternative to a modified t-statistic is a rank statistic like the Wilcoxon rank sum to test for differentially expressed genes. The advantages of the Wilcoxon rank sum are, on the one hand, that it is not necessary to adjust for genes with low expression levels, and on the other hand, that the exact null

distribution is known.

3.2 Empirical Bayes Analysis of Microarrays

Efron et al. (2001), and Efron and Tibshirani (2002) model the distribution of the expression scores d_i, $i = 1, \ldots, m$, as a mixture of two components, one component for the differentially expressed genes, and the other for the not differentially expressed genes. Denoting the density of the former by f_1, and the latter by f_0, the mixture density of the expression scores is given by

$$f(d) = \pi_0 f_0(d) + \pi_1 f_1(d), \tag{3.2}$$

where π_1 or $\pi_0 = 1 - \pi_1$, respectively, is the prior probability that a gene is differentially expressed respectively not. Applying Bayes' rule to (3.2) results in the posterior probability

$$p_1(d) = 1 - \pi_0 \frac{f_0(d)}{f(d)} \tag{3.3}$$

that a gene with expression score d is differentially expressed. Following Efron et al. (2001), and Efron and Tibshirani (2002), a gene will be called differentially expressed if its posterior probability (3.3) is larger than or equal to 0.9. The FDR for the resulting rejection region $\Gamma = \big\{d : p_1(d) \geq 0.9\big\}$ is then estimated by

$$\widehat{\mathrm{FDR}}(\Gamma) = \hat{\pi}_0 \frac{\#\{d_i^b \in \Gamma\}/B}{\max\big\{\#\{d_i \in \Gamma\}, 1\big\}}.$$

For the computation of (3.3), it is necessary to estimate both the prior probability π_0 and the densities f_0 and f.

In the empirical Bayes approach (EBAM in the following) based on the modified t-statistic (3.1), both f and f_0 have to be estimated. Instead of estimating these densities individually by a density estimation procedure, it is more convenient to estimate the ratio f/f_0 directly by a logistic regression with repeated observations, where a natural cubic spline with 5 degrees of freedom is used as the regression function. A detailed description of this logistic regression is given in Appendix B. In the empirical Bayes analysis using Wilcoxon rank sums (EBAM-Wilc for short), only f has to be estimated since the null density f_0 is known. Efron and Tibshirani (2002) estimate the density f of the observed expression scores by a Poisson regression with offset $\ln\{f_0\}$, where f is modeled by a natural cubic spline with 5 degrees of freedom.

Efron et al. (2001) recommend to use $\int_{\mathcal{A}} f(z)dz / \int_{\mathcal{A}} f_0(z)dz$ as an upper bound for π_0, where $\mathcal{A}$ is an interval near $z = 0$, and hence to estimate π_0 by

$$\hat{\pi}_0 = \frac{\int_{\mathcal{A}} \hat{f}(z)dz}{\int_{\mathcal{A}} \hat{f}_0(z)dz} = \frac{\#\{d_i \in \mathcal{A}\}}{\#\{d_i^b \in \mathcal{A}\}/B}. \tag{3.4}$$

If the lower and upper bound of $\mathcal{A}$ are now specified by $q_{\lambda/2}$ and $q_{1-\lambda/2}$, respectively, where q_λ is the λ quantile of the mB permuted d_i^b values, then (3.4) will become

$$\hat{\pi}_0(\lambda) = \frac{\#\Big\{d_i \in \big(q_{\lambda/2}, q_{1-\lambda/2}\big)\Big\}}{\#\Big\{d_i^b \in \big(q_{\lambda/2}, q_{1-\lambda/2}\big)\Big\}/B} = \frac{\#\{p_i > \lambda\}}{(1-\lambda)m}$$

which is exactly the same value that is computed in the first step of the π_0 estimation procedure described in Section 2. The most natural choice for $\mathcal{A}$ is $\mathcal{A} = \{0\}$. It would hence be reasonable to estimate π_0 by $\hat{f}(0)/\hat{f}_0(0)$ if this choice was not that instable. Since the π_0 estimation procedure described in Section 2 deals with this instability and computes $\hat{\pi}_0(1)$ which corresponds to

using $\mathcal{A} = \{0\}$, we use this algorithm to estimate π_0.

Instead of using $\lambda = 0, 0.01, \ldots, 0.95$ as in EBAM, we take

$$\lambda_p = 1 - \sum_{w=W_{\min}+p}^{W_{\max}-p} f_0(w), \quad p = 0, \ldots, \left[\frac{n_1 n_2}{2}\right] \tag{3.5}$$

in EBAM-Wilc, where $W_{\min}$ is the minimum and $W_{\max}$ is the maximum, respectively, of the possible values of the Wilcoxon rank sum. Otherwise $\hat{\pi}_0(\lambda)$ could not be determined unambiguously. Then $\hat{\pi}_0(\lambda)$ is computed by

$$\hat{\pi}_0(\lambda_p) = \frac{1}{1-\lambda_p} \sum_{w=W_{\min}+p}^{W_{\max}-p} \hat{f}(w), \tag{3.6}$$

where $\hat{f}$ is the Poisson regression estimate for f described above. Instead of using the observed numbers, the estimated numbers of observations with expression score w, $w \in \{W_{\min}, \ldots, W_{\max}\}$, are thus used, since it has been shown that this results in a better estimation of π_0 in our analyses. For example, in the analysis of the simulated data described in Section 5.1, where $\pi_0 = 0.9$, using $\hat{f}$ leads to a slightly conservative estimation of π_0 since $\hat{\pi}_0(\lambda) \in [0.9, 0.925]$ in almost any case, whereas $\hat{\pi}_0(\lambda) \in [0.68, 1]$, if f_{obs} is used in (3.6), where $f_{\text{obs}}(w)$ is the observed number of genes with expression score w divided by the number of genes m.

3.3 Significance Analysis of Microarrays

In the following, the SAM (Significance Analysis of Microarrays) procedure is described.

1. Compute the expression score d_i for each gene i, $i = 1, \ldots, m$, and order these values to obtain the observed order statistics $d_{(i)} \leq \ldots \leq d_{(m)}$.

2. Draw B random permutations of the group labels. For each permutation b, compute the permuted expression scores d_i^b, $i = 1 \ldots, m$, and order them. Estimate the expected order statistics by $\bar{d}_{(i)} = \sum_b d_{(i)}^b / B$, $i = 1, \ldots, m$.

3. Plot the observed order statistics $d_{(i)}$ against the expected order statistics $\bar{d}_{(i)}$ to obtain the SAM plot (see Figure 1).

4. For a fixed threshold $\Delta > 0$, find the first data point $\left(\bar{d}_{(i_1)}, d_{(i_1)}\right)$ to the right of the origin for which $d_{(i)} - \bar{d}_{(i)} \geq \Delta$, and set $d_{(i_1)} = \text{cut}_{\text{up}}(\Delta)$. Call any gene i with $d_i \geq \text{cut}_{\text{up}}(\Delta)$ positive significant. Similarly, find the first data point $\left(\bar{d}_{(i_2)}, d_{(i_2)}\right)$ to the left of the origin for which $d_{(i)} - \bar{d}_{(i)} \leq -\Delta$, set $d_{(i_2)} = \text{cut}_{\text{low}}(\Delta)$, and call any gene i with $d_i \leq \text{cut}_{\text{low}}(\Delta)$ negative significant.

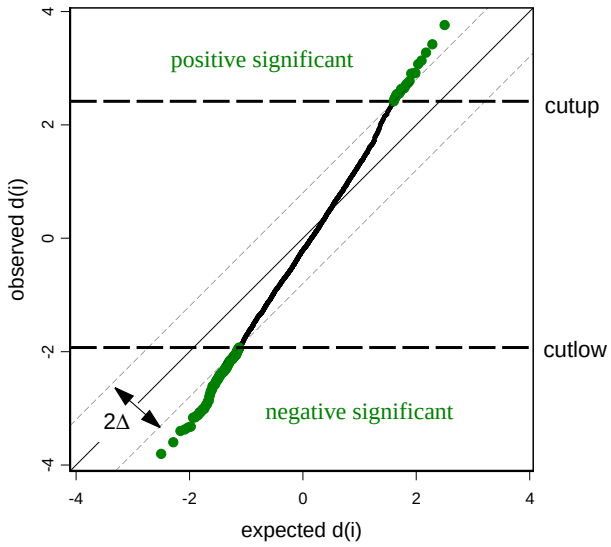

Figure 1: SAM Plot for $\Delta = 0.7$ using the Hedenfalk et al. (2001) data set (see Section 5). Plot of the ordered observed expression scores d_i against the ordered expected expression score $\bar{d}_i$. Each gene is represented by a dot. Differentially expressed genes, i.e. genes lying outside $\left(\text{cut}_{\text{low}}, \text{cut}_{\text{up}}\right)$, are marked by big dots.

5. Estimate the FDR by

$$\widehat{\mathrm{FDR}}(\Delta) = \hat{\pi}_0 \frac{(1/B)\sum_b \#\Big\{ d_i^b \notin \big(\mathrm{cut}_{\mathrm{low}}(\Delta), \mathrm{cut}_{\mathrm{up}}(\Delta)\big)\Big\}}{\max\Big\{\#\big\{\text{significant genes}\big\}, 1\Big\}},$$

where $\hat{\pi}_0$ is the natural cubic spline based estimate described in Section 2.

6. Repeat steps 4 and 5 for several values of the threshold Δ. Choose the value of Δ that provides the best balance between the number of identified genes and the estimated FDR.

4 SAM Using Wilcoxon Rank Sums

Because of the advantages of rank statistics (no need to adjust for genes with low expression levels, exact null distribution is known) it is not surprising that there already exists a SAM procedure based on rank sums called SAM-RS that is proposed by van de Wiel (2002). He suggests to estimate the null distribution of the rank sum by a permutation method. Thus, his approach exactly fits into the original SAM procedure, and one only has to replace the modified t-statistic by the rank statistic.

We argue that we know the null distribution, and hence should use this exact null distribution instead of van de Wiel's estimated one that is less exact and takes a lot of time to compute. Therefore, we here introduce a new Wilcoxon rank sum based SAM method called *SAM-Wilc* using the exact null distribution, and show in the following how the SAM procedure described in Section 3.3 has to be modified for SAM-Wilc:

1. Obtain the observed order statistics $W_{(1)} \leq \ldots \leq W_{(m)}$ by computing the Wilcoxon rank sum W_i for each gene i, $i = 1, \ldots, m$, and by ordering these rank sums.

2. Compute the ith expected order statistic $W^0_{(i)}$, $i = 1, \ldots, m$, by the $(i - 0.5)/m$ quantile of the exact null distribution of the Wilcoxon rank sum statistic.

3. For a positive integer Δ, find the first data point $\left(W^0_{(i_1)}, W_{(i_1)}\right)$ to the right of the mean of the null distribution given by $W_{\text{mean}} = n_1(n+1)/2$ for which $W_{(i)} - W^0_{(i)} \geq \Delta$. Set $W_{(i_1)} = \text{cut}_{\text{up}}(\Delta)$, and call any gene i with $W_i \geq \text{cut}_{\text{up}}(\Delta)$ positive significant. Similarly, find the first data point $\left(W^0_{(i_2)}, W_{(i_2)}\right)$ to the left of W_{mean} for which $W_{(i)} - W^0_{(i)} \leq -\Delta$, set $W_{(i_2)} = \text{cut}_{\text{low}}(\Delta)$, and call any gene i with $W_i \leq \text{cut}_{\text{low}}(\Delta)$ negative significant.

4. Estimate the FDR by

$$\widehat{\text{FDR}}(\Delta) = \hat{\pi}_0 \frac{m\left(1 - \sum_{w=\text{cut}_{\text{low}}(\Delta)+1}^{\text{cut}_{\text{up}}(\Delta)-1} f_0(w)\right)}{\max\Big\{\#\{\text{significant genes}\}, 1\Big\}},$$

   where f_0 is the null density of the Wilcoxon rank sum, and $\hat{\pi}_0$ is the natural cubic spline based estimate of π_0 computed by using (3.5) and (3.6).

5. Repeat steps 3 and 4 for a set of positive integers Δ. Choose the value of Δ that provides the best balance between the number of identified genes and the estimated FDR.

5 Comparison of SAM and EBAM

In this section, the performance of the two SAM procedures and the two empirical Bayes approaches is compared by applying these four methods to one simulated and two real gene expression data sets. In SAM and EBAM, B=1000 permutations are used to assess the null distribution.

5.1 Data Sets

Simulated data. The simulation is performed as follows:

1. Generate a 5,000 x 50 matrix Z containing random draws from the standard normal distribution. Compute the expression level x_{ij} of the ith gene, $i = 1, \ldots, 5000$, and the jth sample, $j = 1, \ldots, 50$, by

$$x_{ij} = z_{ij} + \begin{cases} \delta_{ij}, & \text{if } i \leq 250 \text{ and } j \leq 25 \\ \theta_{ij}, & \text{if } 251 \leq i \leq 500 \text{ and } j \leq 25 \\ 0 & \text{otherwise} \end{cases},$$

   where $\delta_{ij} \sim N(1.5, 1)$ and $\theta_{ij} \sim N(-1.5, 1)$, and suppose that the first 25 columns/samples belong to group 1, and the remaining samples belong to group 2. Thus, a data matrix is constructed that contains expression levels of 50 samples – 25 from each group – and 5000 genes from which 10% are differentially expressed.

2. Apply each of the four procedures to this data set, and record the numbers of differentially expressed genes and the FDRs obtained by these methods.

3. Repeat steps 1 and 2 k times (we have used $k = 100$). For each procedure, compute the mean number of differentially expressed genes and the mean

FDR by averaging over the iterations.

Hedenfalk data. An excerpt from the data set on hereditary breast cancer of Hedenfalk et al. (2001) is considered that contains the gene expression levels of 3,226 genes and 15 samples that were measured by using cDNA microarrays. 7 of the 15 samples come from patients who carry the BRCA1 mutation, and the remaining 8 samples correspond to carriers of the BRCA2 mutation, where mutations of the two genes BRCA1 and BRCA2 are known to lead to a greatly increased breast cancer risk.

Golub data. The Golub et al. (1999) data set consists of the expression levels of 3,051 genes from 38 patients with leucemia, where the expression values were measured by using Affymetrix high-density oligonucleotide chips. 27 of the 38 patients have acute lymphoblastic leucemia (ALL), and the remaining 11 patients have acute myeloid leucemia (AML).

5.2 Results

In the following, we take a closer look at these results of the SAM and the empirical Bayes analyses summarized in Table 1.

Simulated data. While controlling about the same FDR, SAM identifies more differentially expressed genes than EBAM and SAM-Wilc. In the analysis of this data set, SAM is hence more powerful than EBAM which in turn is more powerful than SAM-Wilc. To compare EBAM-Wilc with these three methods, the rejection region in EBAM-Wilc is chosen such that this approach controls about the same FDR as the other methods. This leads to calling a gene differentially expressed if its posterior probability $p_1(z)$ is larger than or equal to 0.933.

Table 1: Comparison of the SAM and the EBAM procedures applied to three data sets. For each method and data set, the number of identified genes, R, and the estimated FDR (in %) are listed.

	Simulation		Hedenfalk		Golub	
Method	R	FDR	R	FDR	R	FDR
SAM	386.5	0.84	158	5.93	707	2.72
SAM-Wilc	369.1	0.88	206	7.25	714	2.75
EBAM	380.9	0.86	162	5.52	714	2.76
EBAM-Wilc	395.8	1.25	178	6.04	711	2.68

In this case, the mean number of genes called differentially expressed is 367.08, and the mean FDR is 0.0087. The two methods based on Wilcoxon rank sums thus have about the same power.

Hedenfalk data. Here both EBAM procedures are more powerful than SAM since they find more differentially expressed genes than SAM, while all three methods control about the same FDR. If the rejection region in the EBAM analysis is chosen such that 178 genes are called differentially expressed, the FDR will be almost the same as in EBAM-Wilc, and hence both approaches have almost the same power. It is a bit harder to compare SAM-Wilc with the other approaches since the small number of samples results in only eight different values for the threshold Δ. For this comparison, the rejection regions of the other methods are computed such that 206 genes are identified. Using these rejection regions, the FDR in both EBAM procedures is 0.067, and in SAM 0.073. Since SAM-Wilc controls the FDR at a level of 0.072, the EBAM approaches are also more powerful than SAM-Wilc, whereas SAM-Wilc has about the same power as SAM.

A total of 248 genes are called differentially expressed by at least one procedure, where 108 genes are identified by all four methods, and 48 are called significant by only one approach.

Golub data. In the analysis of the Golub et al. (1999) data set, all four procedures have about the same power since they all identify about the same number of genes while the estimated FDRs differ only slightly.

There are 812 genes that are identified by at least one method. While 611 genes are called differentially expressed by all procedures, 38 are identified by only one method.

6 Discussion

In this paper, three procedures for the identification of differentially expressed genes and the estimation of the false discovery rate (FDR) have been presented. These are, on the one hand, the Significance Analysis of Microarrays (SAM) based on a modified t-statistic, and on the other hand, two empirical Bayes approaches – one based on the same modified t-statistic that is used by SAM, and the other one based on the Wilcoxon rank sum statistic.

We have furthermore introduced a new version of SAM that is based on Wilcoxon rank sums. Although it has only been shown how the SAM algorithm has to be modified when Wilcoxon rank sums are used, our approach called SAM-Wilc can easily be adjusted for other rank statistics by just exchanging the Wilcoxon rank sum statistic and its null distribution with the other rank statistic and its null distribution. The disadvantage of SAM-Wilc is that Wilcoxon rank sums can be too discrete, especially when the number of samples is small. In such a case, the use of normal rank scores might improve the analysis.

These four procedures have then been applied to one simulated data set and two real microarray data sets. While SAM is the most powerful method in the analysis of the simulated data set, it performs worse in the applications to the real data sets. There are however no big differences in the performance of the procedures. In particular, all four methods have almost the same power in the analysis of the Golub et al. (1999) data set. In the analysis of the simulated data, the approaches based on the modified t-statistic have shown a better performance than the procedures based on the Wilcoxon rank sum. A reason for this is that the data have been generated from normal distributions. In this case, the t-test is more powerful than the Wilcoxon test. In the analysis of the Hedenfalk et al. (2001) data, both EBAM approaches have shown a better performance than the SAM procedures.

There is one disadvantage in the way SAM and EBAM estimate the FDR: Both the computation of the rejection region and the estimation of the FDR are performed by using the same data set. This is comparable with using the same data set in a discrimination problem for both building a classifier and estimating the misclassification rate. The FDR is hence estimated anti-conservatively.

All the procedures presented here are long run methods in the sense that they only should be used if the number of genes is very large. If only 50 or 100 genes are to be analyzed, it will be likely that the results of these analyses are not meaningful, and thus other methods such as a Northern Blot analysis should be preferred.

The analyses described above were performed by using the functions contained in the `R` package `siggenes`. This package was programmed at the University of Dortmund and can be downloaded from `http://www.bioconductor.org`. There also exists an Microsoft Excel based SAM software programmed at Stanford (see `http://www-stat.stanford.edu/~tibs/SAM/index.html`).

Appendix

A Fudge Factor

In this appendix, we take a closer look on the fudge factor s_0. First, details on the computation of the fudge factor in both EBAM and SAM are given. Afterwards, it is shown how s_0 affects the expression score of a gene, in particular the expression score of a gene with low expression values.

A.1 Computation of the Fudge Factor

In both the EBAM and the SAM analysis, the fudge factor s_0 is specified by the quantile of the standard deviations s_i, $i = 1, \ldots, m$, of the genes that fulfills an specific optimization criterion.

Efron et al. (2001) argue that the information loss is reflected by the reduction of the number of genes with a convincingly large posterior probability. Thus, the larger the number of genes called differentially expressed is, the less information is lost. Efron et al. (2001) hence suggest to specify the optimal choice of the fudge factor in an EBAM analysis by running the EBAM procedure for several values of s_0, and by selecting the value of s_0 that leads to the most differentially expressed genes. When comparing the performance of the EBAM procedure for several values of s_0, one has to keep in mind that it is necessary to always have the same marginal distribution for the observed expression scores. Efron et al. (2001) therefore monotonically transform the observed expression scores to have a standard normal distribution. The permuted expression scores are then transformed accordingly.

In the SAM analysis, the fudge factor is computed by the following algorithm provided by Chu, Narasimhan, Tibshirani, and Tusher (2002):

1. Compute the 100 percentiles q_k, $k = 1, \ldots, 100$, of the s_i values.

2. For $\alpha \in \mathcal{R} = \{0, 0.05, 0.1, \ldots, 1\}$

   (a) compute $d_i^\alpha = r_i/(s_i + s^\alpha)$, where s^α denotes the α quantile of the s_i values, and $s^0 = q_0 = \min_{i=1,\ldots,m} \{s_i\}$,

   (b) calculate $v_k^\alpha = 1.4826 \cdot \text{MAD}\big\{d_i^\alpha | s_i \in [q_{k-1}, q_k)\big\}$, $k = 1, \ldots, 100$,

   (c) compute the coefficient of variation $\text{CV}(\alpha)$ of the v_k^α values.

3. Set $\hat{\alpha} = \arg\min_{\alpha \in \mathcal{R}} \big\{\text{CV}(\alpha)\big\}$, and $s_0 = s^{\hat{\alpha}}$.

A.2 Effects of the Fudge Factor

In this Section, we take a look on how the fudge factor affects the expression score of a gene (with low expression levels). As an example, the Hedenfalk et al. (2001) data set presented in Section 5.1 is used. The fudge factor for the

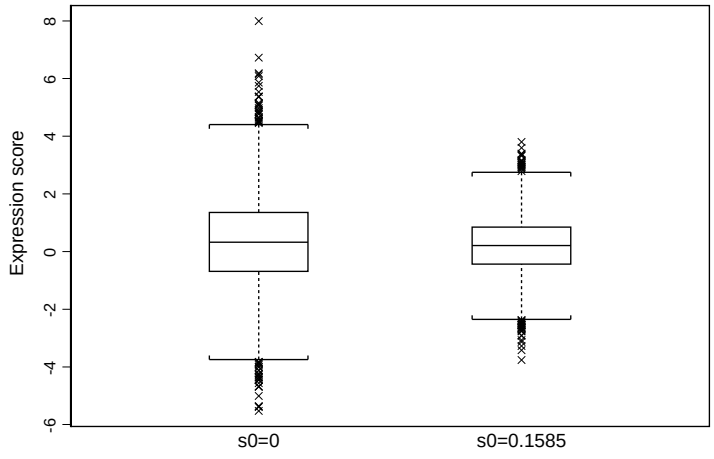

Figure 2: Boxplots of the t- and the d-statistics of the genes contained in the Hedenfalk et al. (2001) data set. Adding a small strictly positive constant s_0 to the denominator of a test statistic leads to a less dispersed distribution.

Hedenfalk data is calculated using the algorithm of Chu et al. (2002) described in Appendix A.1. The result of this computation is $s_0 = 0.1585$, the 5% quantile of the standard deviations of the genes. Furthermore, both the d-statistic (3.1) and the standard t-statistic for each of the 3226 genes are computed.

Figure 2 shows what generally happens when a small strictly positive constant is added to the denominator of a test statistic. The distribution of the d values ($s_0 = 0.1585$) is less dispersed than the distribution of the t values ($s_0 = 0$).

For the investigation of the influence of the fudge factor on genes with low expression levels, the rank of the d_i value of gene i, $i = 1, \ldots, m$, is plotted against the rank of its t-statistic (see Figure 3(a)). The bold black circles in Figure 3(a) symbolize genes with a standard deviation smaller than $s_0 = 0.1585$. This figure reveals that if s_0 is added to the denominator of the standard t-statistic the value of a gene with a small variance will much more shrink towards

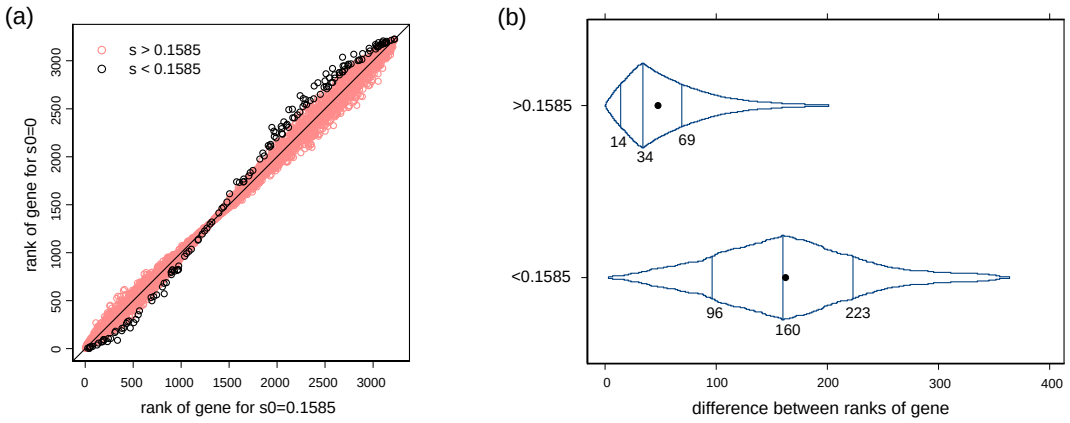

Figure 3: Influence of the fudge factor on genes with small variances contained in the Hedenfalk et al. (2001) data set: (a) Scatter plot of the ranks of the d values vs. the ranks of the corresponding t values, (b) box-percentile plots of the differences between these ranks of both genes with standard deviation smaller than $s_0 = 0.1585$ and larger than $s_0 = 0.1585$.

zero than the value of a gene with a higher variance, since in comparison to genes with a standard deviation larger than 0.1585, the rank of small variance genes with negative expression score increases very strongly, and the rank of small variance genes with positive expression score decreases very strongly. Figure 3(a) therefore indicates that the fudge factor has more influence on the small variance genes, and hence on genes with low expression values since most of the small variance genes have low expression levels.

To confirm this, two side-by-side box-percentile plots of the absolute differences between the ranks of the t values of the genes and the ranks of the corresponding d values are generated, one plot for genes with standard deviation smaller than or equal to $s_0 = 0.1585$, and the other for genes with a standard deviation larger than 0.1585.

Figure 3(b) shows that the differences between the ranks of small variance genes are much larger than the differences between the ranks of the genes with large variances, since about 75% of the former differences are larger than 100, whereas more than 80% of the latter differences are smaller than 100, and 50% are at most 34. The fudge factor has hence an increased influence on small variance genes which are mostly genes with low expression levels.

B Logistic Regression Estimate of f_0/f

Instead of estimating the two densities f_0 and f individually, it is more convenient to estimate the ratio f_0/f directly. For this, consider the observed expression scores d_i, $i = 1, \ldots, m$, as successes, and the permuted expression scores d_i^b, $i = 1, \ldots, m$, $b = 1, \ldots, B$, as failures. If these $m(B+1)$ scores are plotted on a line, then the probability $\varphi(d)$ of a success at point d can be computed by

$$\varphi(d) = \frac{f(d)}{f(d) + Bf_0(d)},$$

and the posterior probability (3.3) is given by

$$p_1(d) = 1 - \pi_0 \frac{f_0(z)}{f(d)} = 1 - \pi_0 \frac{1 - \varphi(d)}{B\varphi(d)}.$$

$\varphi(d)$ can now be estimated by a logistic regression. This is usually done by maximizing the log-likelihood function

$$\ell(\beta_1, \ldots, \beta_p) = \sum_{i=1}^{m(B+1)} y_i g(d_i) - \sum_{i=1}^{m(B+1)} \ln\Big(1 + \exp\big\{g(d_i)\big\}\Big), \qquad \text{(B.1)}$$

where $\beta_1, \ldots, \beta_p$ are the parameters of the regression function $g(d)$, and $y_i = 1$ if d_i is an observed expression score, and $y_i = 0$ if d_i is a permuted expression score. In the EBAM analysis, a natural cubic spline with five degrees of freedom is used as regression function $g(d)$. The probability $\varphi(d)$ of a success at point d is then estimated by

$$\hat{\varphi}(d) = \frac{\exp\big\{\hat{g}(d)\big\}}{1 + \exp\big\{\hat{g}(d)\big\}}. \qquad \text{(B.2)}$$

In an EBAM analysis, this means that we have to maximize over millions of components which is computationally not feasible. A solution to this problem is provided by the *logistic regression with repeated observations*. For such a logistic regression, the range of the observed expression scores is divided into K equally spaced intervals A_k, $k = 1, \ldots, K$. Efron et al. (2001), e.g., use $K = 139$ intervals.

We do not use the range of all expression scores, i.e. the observed and permuted d values, since it has turned out that the very few permuted expression

scores that lie outside the range of the observed d values can totally destabilize the logistic regression. Rather than excluding such permuted d values, they are set either to the minimum or the maximum of the observed expression scores so that we do not lose all the information in these expression scores.

For each interval A_k, $k = 1, \ldots, K$, the number R_k of the observed expression scores in A_k, the total number N_k of expression scores in A_k, and the center point $\tilde{d}_k$ of A_k is computed. If now each of the $m(B+1)$ expression scores d_i is replaced by the center point $\tilde{d}_k$ of the interval A_k into which d_i falls, then (B.1) becomes

$$\ell(\beta_1, \ldots, \beta_p) = \sum_{k=1}^{K} \left\{ \ln \binom{N_k}{R_k} + R_k g(\tilde{d}_k) - N_k \ln\Big(1 + \exp\big\{g(\tilde{d}_k)\big\}\Big) \right\}.$$

Instead of maximizing over $2m(B+1)$ components, i.e. over millions of components, we thus only have to maximize over $2K$ components, i.e. over a few hundred components. The probability $\varphi(d)$ of a success at point d can still be estimated by (B.2).

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,* Ser. B 57, 289–300.

Chu, G., Narasimhan, B., Tibshirani, R., and Tusher, V. (2002), "SAM "Significance Analysis of Microarrays" – Users guide and technical document," *Technical Report*, Stanford University. Available with their SAM software at `http://www-stat.stanford.edu/~tibs/SAM/index.html`.

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002), "Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments," *Statistica Sinica*, 12, 111–139.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003), "Multiple Hypothesis Testing in Microarray Experiments," *Statistical Science*, 18, 71–103.

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

Efron, B., and Tibshirani, R. (2002), "Empirical Bayes Methods and False Discovery Rates for Microarrays," *Genetic Epidemiology*, 23, 70–86.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caliguiri, M. A., Bloomfield, C. D., and Lander, E. S. (1999), "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, 286, 531–537.

Hedenfalk, I., Duggan, D., Chen, Y. D., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O. P., Wilfond, B., Borg, A., and Trent, J. (2001), "Gene-expression Profiles in Hereditary Breast Cancer," *New England Journal of Medicine*, 344, 539–544.

Ihaka, R., and Gentleman, R. (1996), "R: A Language for Data Analysis and Graphics," *Journal of Computational and Graphical Statistics,* 5, 299–314.

Shaffer, J. P. (1995), "Multiple Hypothesis Testing," *Annual Review of Psychology*, 46, 561–584.

Storey, J. D., and Tibshirani, R. (2001), "Estimating False Discovery Rates Under Dependence, with Applications to DNA Microarrays," Technical Report 2001-28, Stanford University, `http://faculty.washington.edu/~jstorey/papers/dep.pdf`.

Storey, J. D., and Tibshirani, R. (2003), "Statistical Significance for Genome-wide Studies," *Proceedings of the National Academy of Sciences,* 100, 9440-9445.

Storey, J. D. (in press), "The positive False Discovery Rate: A Bayesian Interpretation and the q-value," *Annals of Statistics.*

Tusher, V. G., Tibshirani, R., and Chu, G. (2001), "Significance Analysis of Microarrays Applied to the Ionizing Radiation Response," *Proceedings of the National Academy of Science,* 98, 5116–5121.

Westfall, P. H., and Young. S. S. (1993), *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustments*, Wiley, New York.

Wiel, M. A. van de (2002), "Significance Analysis of Microarrays Using Rank Scores," Technical Report, Department of Mathematics and Computer Science, Eindhoven University of Technology, `http://www.nr.no/documents/samba/research_areas/SMBI/seminar/npsam.ps`.