DSC 2003 Working Papers
(Draft Versions)
http://www.ci.tuwien.ac.at/Conferences/DSC-2003/

A False Discovery Rate approach to separate the score distributions of induced and non-induced genes

Stefanie Scheid and Rainer Spang

Max Planck Institute for Molecular Genetics
Ihnestraße 63-73, D-14195 Berlin, Germany
{stefanie.scheid,rainer.spang}@molgen.mpg.de

Abstract

The distribution of scores for differential gene expression observed in microarray experiments give rise to the assumption that the underlying score distributions of induced and non-induced genes share wide overlapping regions. Our aim is to reconstruct this mixture not only for extremal score regions but over the whole range of scores. We propose and evaluate a method based on the theory of False Discovery Rates.

1 Introduction

Microarray experiments allow an insight into a cell's current state by measuring simultaneously the abundance of mRNA transcripts of several thousands of genes. With samples from two different classes, e.g. tissue, disease status or treatment, we can search for genes that are differentially expressed among these classes. From a statistical point of view we are confronted with a multiple testing problem. We test for differences in mean gene expression of several thousands of genes all measured from the same set of samples. Traditional multiplicity corrections like Bonferroni or Bonferroni-Holm control the Family-wise Error Rate (FWER) which is the probability of making at least one false positive decision. Dudoit et al. [2] give an overview

of FWER controlling procedures including methods that correct for the dependence structure in gene expression experiments resulting from coregulation.

FWER controlling procedures are often too conservative unless we increase the desired probability of making at least one false positive decision. However, when we search for induced genes among thousands we can allow false positives if this leads to a higher number of true positives and a better understanding of biological processes. The proportion of false positives among all positives is called False Discovery Rate (FDR). A Bonferroni-like method to control this rate was first introduced by Benjamini and Hochberg [1]. Storey [5] points out that controlling the FDR is only interesting when positive decisions have occured. He introduces a procedure to control this conditional FDR which is called *positive* FDR. Storey and Tibshirani [6] suggest methods to estimate the positive FDR and conduct simulation studies under various stages of coregulation. There exist various additional procedures to control the FWER or the FDR. Keselman et al. [3] and Reiner et al. [4] provide comparison studies of selected methods where FDR controlling methods exhibit higher power than FWER controlling methods.

We assign a score that measures the difference in mean gene expression to each gene. A high score corresponds to overexpression in the first class and a low score to underexpression in the first class. A gene showing either overexpression or underexpression is called "induced". The overall score distribution in a microarray experiment is typically a mixture of score distributions resulting from induced and non-induced genes. Neither these distributions nor the fraction of induced genes are known. The distinction of induced and non-induced genes is easy if differential regulation leads to extreme scores such that the score distribution displays clearly separated modes for induced genes. However, in a typical setting the two distributions overlap and only the most highly induced genes can be selected with common methods. The majority of the induced genes merge with non-induced genes in extended "twilight zones". Classical significance testing aims for a cutoff score level, that assures that there is no more than a small fraction (e.g. 5%) of non-induced genes scoring higher than this level. This only describes the beginning of the twilight zone.

In this paper, we want to locate twilight zones. More precisely, for any set of genes with similar score, we want to estimate how many genes are induced. This is equivalent to reconstructing the mixture at this score level. Our approach is based on a FDR estimating method introduced by Tusher et al. [7]. The paper is organized as follows. We extend the common FDR in a bin-wise manner. In a simulation study we apply the new method on various stages of sparsity and compare the estimated to the true mixture.

2 Bin-wise False Discovery Rate

Given a microarray experiment with samples from two classes A and B, we assign a score measuring the difference in mean gene expression to each gene. A commonly used nonparametric score is Wilcoxon's ranksum score. For this score, we want to reconstruct the mixture of score distributions resulting from induced genes and

from non-induced genes. The mixture consists of the two score distributions and a mixing parameter, i.e. the fraction of induced genes among all genes.

Classical approaches of separation, like estimating mixture models, require knowledge of underlying distribution functions. A method that does not need this information can be constructed from the FDR concept. In a multiple testing situation and a fixed rejection area, the FDR estimates the proportion of genes that are falsely called induced among all genes called induced (Benjamini and Hochberg [1], Storey [5]). If the random variable R denotes the number of positive outcomes (genes called "induced") and V denotes the number of false positive outcomes then the FDR is defined as their expected ratio:

$$FDR = \begin{cases} \mathrm{E}\left[\frac{V}{R}\right] & \text{if } R > 0, \\ 0 & \text{if } R = 0. \end{cases}$$

The proper definition of FDR gives rise to discussion (see Storey [5]). It is more intuitive not to consider the expected ratio of the two random variables but the ratio of the expectation of V and the (positive) observation r of R:

$$FDR = \frac{\mathrm{E}[V]}{r}.$$

The observed number r of positive outcomes depends on the choice of rejection rules. Typically, genes which score above or below given thresholds are called induced. This means rejection areas correspond to extremal scores. For example, if the choice of a threshold yields a FDR of 0.05, we expect 5% falsely called induced genes among the set of genes exceeding the threshold. In addition we expect 95% of the rejected genes to be truly induced. Thus we can reconstruct the mixture score distribution of induced and non-induced genes for an extremal rejection area. Note that the FDR is not a property of a single score level, but a property of the whole list of rejected genes.

Although extremal rejection areas are natural and intuitive in a multiple test setting, non extremal rejection areas are the basis of our method. We define rejection areas such that genes are called induced if they fall into a specific score interval (bin). Clearly the concept of FDRs is flexible enough to calculate a *bin-wise FDR*, which leads to an estimate of the fraction of induced genes in a certain bin. Hence for this bin the FDR reconstructs the mixture. Putting together results from many bin-wise FDRs we obtain the global reconstruction of the mixture score distribution.

The FDR estimator as given in Tusher et al. [7] is based upon class permutation: For each permutation the number of scores exceeding a given threshold is calculated. Their median number is divided by the total number of rejected genes. This ratio multiplied by an estimate of the overall fraction of non-induced genes is the estimated FDR. The procedure can be interpreted as an empirical Bayes approach, where the estimated fraction is the prior probability that a gene is non-induced. The commonly used estimate for this prior probability as given in Tusher et al. [7] is the number of observed scores contained in a quantile interval of all permutation scores, say between the 25% and 75% quantile, divided by its expected number, here 50% of all genes. To estimate a probability its upper bound is set to 1. In

Table 1: Bin-wise FDR algorithm.

1. For each gene g calculate the observed Wilcoxon ranksum score W_g and k permutation scores W_g^i $(i = 1, \ldots, k)$ using class permutation.
2. Divide the range of scores into b disjoint bins B_j $(j = 1, \ldots, b)$.
3. Calculate lower and upper quartiles $q_{.25}$ and $q_{.75}$ of *all* permutation scores. Estimate the prior probability π_0 that a gene is non-induced as:
$$\hat{\pi}_0 = \min\left(1, \frac{\#\{g : W_g \in [q_{.25}, q_{.75}]\}}{0.5 \cdot \text{number of genes}}\right).$$
4. For each bin B_j estimate the bin-wise false discovery rate FDR_j as:
$$\widehat{FDR}_j = \hat{\pi}_0 \cdot \frac{\underset{i=1,\ldots,k}{\text{median}} \#\{g : W_g^i \in B_j\}}{\#\{g : W_g \in B_j\}}.$$
$\widehat{FDR}_j$ estimates the percentage of non-induced genes and $1-\widehat{FDR}_j$ estimates the percentage of induced genes in bin B_j.

case of large twilight zones the interval above contains many induced genes and the method overestimates the overall amount of non-induced genes.

To obtain a bin-wise FDR we divide the range of scores into bins and define each bin separately as a rejection area, i.e. a gene is called induced if its score is contained within the bin. The estimator of the overall probability of non-induced genes is kept as described above. For each bin we obtain an estimated percentage of non-induced genes and by subtracting it from 1, we also get an estimated percentage of induced genes. Hence, we reconstruct the mixture for this bin. Putting these results together yields a separation of two previously mixed score distributions. Of course the two distributions are discrete due to the binning and hence only approximations of the true score distributions. Table 1 contains the estimation algorithm in detail.

3 Simulation study

We use simulated data to evaluate the performance of our method. The simulation is constructed such that we obtain log expression values for samples in two classes where a subset of genes is induced in one class. The advantage of simulated data is the knowledge of the true score distributions. We evaluate our method by comparing the estimated to the true binned mixture at several levels of sparsity.

Simulated data should reflect two characteristic features of real data: 1) Each gene has a characteristic gene profile across samples and 2) groups of genes are correlated because they act in pathways, a property Storey and Tibshirani [6] called

Table 2: Simulation parameters.

No. of genes:	1 000
No. of samples per class:	30
No. of permutations:	5 000
No. of bins:	20
Prior percentage π:	5, 15, 25, 50%
Mean induction offset μ:	0.5, 0.7

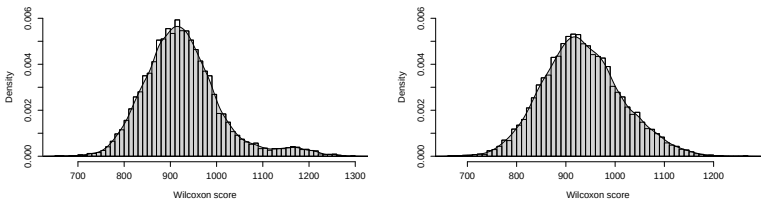

Figure 1: Densities of 10 000 Wilcoxon scores derived from simulation models with $\pi = 0.05$, $\mu = 1$ (left) and $\pi = 0.15$, $\mu = 0.5$ (right).

"clumpy dependence". The first criterion is obtained by randomly drawing a master expression value for each gene from a lognormal distribution with parameters $\mu_{logn} = 1.5$ and $\sigma_{logn} = 0.3$. For each gene in each sample we add an individual standard normal error term. The clumpy dependence is simulated as in Storey and Tibshirani [6]: We randomly divide the genes into blocks of 50 and add the same standard normal error term to all genes in a block.

No gene is induced so far. We simulate induced genes by randomly selecting a percentage π of genes and adding individual mean offset terms μ for all samples in the first class. The parameter π gives the proportion of induced genes and is connected to the prior probability π_0 in Table 1 via $\pi = 1 - \pi_0$. The mean offset terms μ are normally distributed according to $N(\mu, \sigma)$ with $\sigma = 0.2$. Table 2 contains all simulation parameters.

We choose the induction offset μ to be small ($\mu = 0.5$) and intermediate ($\mu = 0.7$). The higher it is, the less overlapping are the score distributions. With a high offset $\mu = 1$, the induced scores form a small "hill" easily separable by eye from the bulk of non-induced genes (cp Figure 1 left). With decreasing μ the problem becomes more interesting: The induced scores approach the mode of the overall score distribution. They are hidden in a twilight zone only recognizable as a subtle elevation in the otherwise symmetrical score distribution (cp Figure 1 right with $\mu = 0.5$).

The borders of the 20 bins are not chosen equidistantly but such that each bin

Table 3: Mean and standard deviation (in parentheses) of mean squared estimation errors.

π	$\mu = 0.5$	$\mu = 0.7$
5%	.0179 (.0046)	.0212 (.0065)
15%	.0220 (.0086)	.0257 (.0085)
25%	.0236 (.0062)	.0210 (.0071)
50%	.0446 (.0142)	.0305 (.0093)

contains approximately 5% of the genes: We compare the observed scores to each set of permutation scores and find a suitable division. The final borders are then given as the median borders of all comparisons.

Each parameter combination is repeated 20 times, resulting in 20 estimates of the proportion of non-induced genes for every bin. In simulations, we know which genes are induced and can therefore know the 20 true proportions for every bin. For each simulation we calculate the mean squared difference between the estimated and the true proportions over all bins and finally combine the 20 simulations by taking the average (and standard deviation). The result is one averaged error value for each parameter combination (π,μ). Table 3 shows the averaged mean squared estimation errors and their standard deviations. There is no observable dependence between estimation error and induction offset μ. That means, the estimator's performance does not degrade when twilight zones occur. The errors increase slightly with increasing proportion π but are not distinguishable with respect to their standard deviations.

An example of a reconstructed mixture is shown in Figure 2. Each step in the stairplot corresponds to a bin. The height of each step denotes the percentage of non-induced genes in that bin. The dashed black line gives the mean true percentage and the solid green line the mean estimated percentage, both averaged over simulations and truncated at 100%. In this case with rather wide twilight zone ($\mu = 0.5$) the bin-wise FDR performs well but overestimates the true percentage slightly in a score range between 950 and 1 000. The overestimation is due to the estimator's conservative character because it underestimates the percentage of induced genes. The underestimation increases with increasing twilight zones. Figure 3 shows reconstructed mixtures for a model with 50% induced genes for the hard problem $\mu = 0.5$ and the intermediate problem $\mu = 0.7$. For $\mu = 0.5$ the bin-wise FDR reconstructs the percentage curve in shape but underestimates the percentage of induced genes for every bin. This is due to the strong underestimation of π, here $\hat{\pi} \approx 41\%$. As mentioned above the problem arises in case of weak induction. With higher induction ($\mu = 0.7$) the estimator as given in Table 1 performs well, here $\hat{\pi} \approx 49\%$ (cp Figure 3).

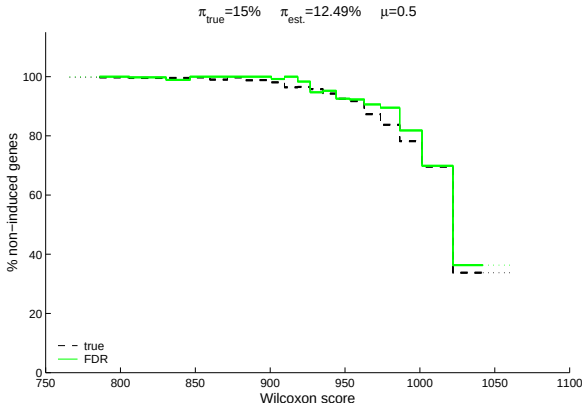

Figure 2: Stairplot of reconstructed mixture from 20 simulations with $\pi = 15\%$, $\mu = 0.5$. Dashed black line gives mean true percentage of non-induced genes, solid green line gives corresponding estimate. Each step denotes a bin.

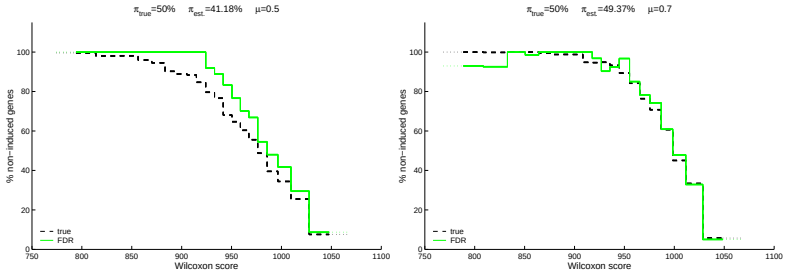

Figure 3: Stairplots of reconstructed mixtures from 20 simulations with $\pi = 50\%$, $\mu = 0.5$ (left) and $\mu = 0.7$ (right). For explanations see Figure 2.

4 Discussion

In case of weak induction the scores for differential gene expression of induced genes are hidden among scores of non-induced genes in rather wide twilight zones. The situation of overlapping score distributions can be observed in real microarray experiments. We cannot assign the labels "induced" or "non-induced" to genes in a twilight zone but can estimate the probability that a gene is induced given its score is contained within that zone.

We introduced the bin-wise FDR method to separate two overlapping score distributions by estimating their proportions after binning and evaluated its performance in a simulation study. The estimator performs well and does not degrade with respect to its mean estimation error when the two distributions share a wide overlapping region. Our estimator is based upon the FDR estimator given in Tusher et al. [7] which involves an estimator for the prior probability that a gene is non-induced. This estimator is sufficient in cases of high induction but overestimates the probability of non-induced genes when many genes are only slightly induced. The improvement of the prior estimation as well as the application to real microarray data are topics of future research.

References

[1] Benjamini, Y., Hochberg, Y. 1995. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society B*, 57(1):289–300.

[2] Dudoit, S., Yang, Y. H., Callow, M. J., Speed, T. P. 2002. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12:111–139.

[3] Keselman, H. J., Cribbie, R., Holland, B. 2002. Controlling the rate of Type I error over a large set of statistical tests. *British Journal of Mathematical and Statistical Psychology*, 55:27–39.

[4] Reiner, A., Yekutieli, D., Benjamini, Y. 2003. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–375.

[5] Storey, J. D. 2001. *The positive False Discovery Rate: A Bayesian Interpretation and the q-value*, Stanford University: Stanford Technical Report.

[6] Storey, J. D., Tibshirani, R. 2001. *Estimating False Discovery Rates Under Dependence, with Applications to DNA Microarrays*, Stanford University: Stanford Technical Report.

[7] Tusher, V. G., Tibshirani, R., Chu, G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121.